Using Reversed MFCC and IT-EM for Automatic Speaker Verification

SHEERAZ MEMON*, SANIA BHATTI**, AND TARIQ JAMIL SAIFULLAH KHANZADA*

RECEIVED ON 21.09.2011 ACCEPTED ON 01.12.2011

ABSTRACT

This paper proposes text independent automatic speaker verification system using IMFCC (Inverse/ Reverse Mel Frequency Coefficients) and IT-EM (Information Theoretic Expectation Maximization). To perform speaker verification, feature extraction using Mel scale has been widely applied and has established better results. The IMFCC is based on inverse Mel-scale. The IMFCC effectively captures information available at the high frequency formants which is ignored by the MFCC. In this paper the fusion of MFCC and IMFCC at input level is proposed. GMMs (Gaussian Mixture Models) based on EM (Expectation Maximization) have been widely used for classification of text independent verification. However EM comes across the convergence issue. In this paper we use our proposed IT-EM which has faster convergence, to train speaker models. IT-EM uses information theory principles such as PDE (Parzen Density Estimation) and KL (Kullback-Leibler) divergence measure. IT-EM acclimatizes the weights, means and covariances, like EM. However, IT-EM process is not performed on feature vector sets but on a set of centroids obtained using IT (Information Theoretic) metric. The IT-EM process at once diminishes divergence measure between PDE estimates of features distribution within a given class and the centroids distribution within the same class. The feature level fusion and IT-EM is tested for the task of speaker verification using NIST2001 and NIST2004. The experimental evaluation validates that MFCC/IMFCC has better results than the conventional delta/MFCC feature set. The MFCC/IMFCC feature vector size is also much smaller than the delta MFCC thus reducing the computational burden as well. IT-EM method also showed faster convergence, than the conventional EM method, and thus it leads to higher speaker recognition scores.

Key Words: Information Theory, Expectation Maximization, MFCC, Gaussian Mixture Model, Speaker Verification.

1. INTRODUCTION

For the past decade, MFCCs [1] and GMM based on EM have been widely applied to text-independent speaker verification. For feature extraction the performance improvements are achieved when dynamic features are fused with MFCC [2] at the input level. For modeling GMM/EM [3] has remained successful for speech and speaker recognition [4,5]. In this paper the task of text-independent speaker verification is evaluated using MFCC/IMFCC input level fusion strategy for feature extraction. The performance of MFCC/

* Assistant Professor, Department of Computer Systems Engineering, Mehran University of Engineering & Technology, Jamshoro.
** Assistant Professor, Department of Software Engineering, Mehran University of Engineering & Technology, Jamshoro.

IMFCC features is compared with delta-MFCC, which is widely used feature extraction method. After the feature extraction, for modeling, GMM based on IT-EM instead of EM is proposed and evaluated. IT-EM method is devised using an ITVQ (Information Theoretic Vector Quantization) criterion [6]. The convergence rates of IT-EM based GMM are also compared with EM based GMM. The experiments are performed on speaker verification corpora available from NIST. The system evaluation is performed using EER (Equal Error Rate) measure.

This paper has further following sections, Section 2 demonstrates the proposed IMFCC and the fusion strategy for feature extraction, and the proposed modeling method is discussed in Section 3. Experiments based on the proposed techniques are summarized in Section 4, followed by the conclusion in Section 5.

2. INVERSE MFCC AND FEATURE FUSION AT INPUT LEVEL

In this section we discuss the delta MFCC, IMFCC and the feature fusion performed.

2.1 MFCC and Delta-MFCC

The psychophysical studies have discovered that the human perception of sound and its frequency content follow a subjectively defined nonlinear scale which is known as Mel scale. The Mel (derived from the word melody) scale, is a heuristically determined perceptual scale and provides the relation between subjectively perceived frequency (or pitch) of a pure tone as a function of its objective acoustic frequency [6]. Studies of speaker, stress and emotion recognition in speech clearly indicate that characteristic features based on human auditory characteristics provide better performance than features that do not take these characteristics into account [7]. The widely used MFCC [8] provide an example of feature parameters based on the human auditory perception. It was demonstrated in [9] that in noisy conditions MFCC show higher robustness than features such as Linear Prediction Cepstral Coefficients (LPCC), Perceptual Linear Prediction (PLP), which do not incorporate human auditory characteristics.

The Mel scale is defined as "A logarithmic scale to map frequency that is based upon human pitch perception. Equal intervals in Mel units correspond to equal pitch intervals. The following mapping formula between frequency in Hz and the corresponding subjective pitch in Mels is the building block of MFCC:

$$f_{mel} = 2591 \log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

In Equation (1) f_{mel} represents subjective pitch, which is recorded in Mels. f_{mel} corresponds to f, that is actual frequency of sound in Hz. The calculation of the MFCC parameters takes place in four stages, including calculation of mel-spectrum for speech frames, evaluation of sub-band energies, and sub-bands. The steps are outlined in Fig. 1.

The Mel spectrum generated is shown in Fig. 2. The time derivatives of spectrum based features such as MFCC is called delta MFCC. Delta cepstral features and double delta cepstral features have played an important function in capturing transitional characteristics of sound [2] and thus delta-MFCCs have established better results for speaker verification [2]. In this paper a feature extraction strategy based on delta MFCC with frame energy and zero crossings is used as the baseline feature extraction method. It is further detailed in experimental section.

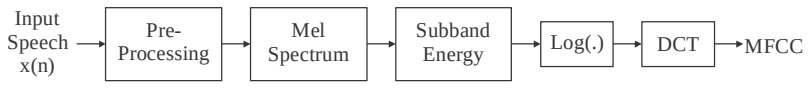

FIG. 1. CALCULATION OF THE MFCC PARAMETERS

2.2 IMFCC

The filter bank used with the MFCC is good at capturing efficiently the vocal tract characteristics at lower frequencies. In this paper we have evaluated a feature set which uses a complementary filter-bank called IMFCC. The IMFCC [10] captures speaker specific features which are present at the high frequency regions. The high level features [11-13] are often difficult to extract, however the IMFCC offer computational simplicity during the extraction process. The calculation steps for the IMFCCs are almost identical to the steps involved in the calculation of MFCCs, however the only difference lies with the filter bank structure. In [10] a parallel implementation of the MFCC and IMFCC was tested. However, the integration of MFCC and IMFCC features was performed at the classifier level.

The IMFCC inverts the filter bank structure used in the MFCC method such that the lower frequencies are averaged by using small number of widely spaced filters and the higher frequencies are averaged by using narrower spacing of filters as shown in Fig. 3. Thus, the IMFCC effectively capture information available at the high frequency formants which is ignored by the MFCC. The frequency range considered for the speaker recognition is between 100-3900Hz, thus the reversed Mel scale can be obtained when the filter bank is flipped at the point f=2kHz. The reverse/inverse Mel scale can be calculated as [10]:

$$\hat{f}_{mel}(f) = 2195.2860 - 2595 \log_{10}\left(1 + \frac{4031.25 - f}{700}\right) \quad (2)$$

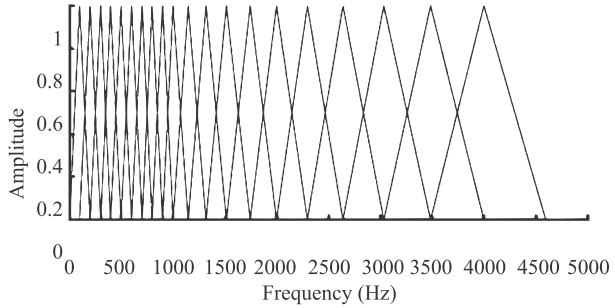

FIG. 2. A MEL SPACED FILTER BANK

Where $\hat{f}_{mel}$ is the inverted mel scale pitch value in mels. For details on the mathematical background involved in obtaining the reversed mel scale refer [10].

The MFCC method transforms speech spectrum to perceptually useful subjective spectrum using human auditory standards [14], with low resolution at high frequency ranges. However the reversed Mel scale provides a complimentary structure capturing high frequency formants with higher accuracy than the Mel scale.

2.3 Fusion of MFCC with delta MFCC and IMFCC

In the past few years research interest has grown a lot on applying information or results fusion strategies at feature and/or classifier level [15]. The fusion taking place at feature level is called input level and the fusion taking place at classification level is called output level. The input level fusion can be seen in one of two possible forms, known as multi-feature and multi-sample. In multi-sample fusion approach the instance or speaker may be required to utter single phrase, multiple times and therefore the result is based on combining the match scores. However with multi-feature, similar utterance yields different features. The example is the use of MFCC cepstra with its delta cepstra. In this paper the input level fusion is performed using multi-feature approach to obtain delta-MFCC and MFCC/IMFCC feature set. The fusion at input level is performed by concatenating the feature arrays horizontally. The idea behind the fusion of MFCC and IMFCC was to capture formant characteristics at both low and high frequency ranges.

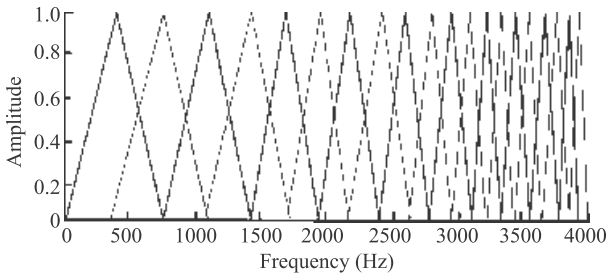

FIG. 3. STRUCTURE OF THE FILTERS FOR THE INVERSED MEL SCALE

3. INFORMATION THEORETIC EXPECTATION MAXIMIZATION BASED GAUSSIAN MIXTURE MODELLING

GMM uses EM algorithm; EM iteratively updates means, covariance matrices and weights for each speaker model and converges to a set of vectors providing the maximum value of the likelihood function [16-18]. A class model is obtained for each set consisting of means, covariances and weights.

GMM and the VQ (Vector Quantization) are combined because both methods represent the distribution of the data vectors in feature space [19]. In a number of studies [20-23] VQ is used with GMM in order to improve the results and avoid the drawbacks caused by EM algorithm. We have evaluated in [24] that IT based VQ has better performance than k-means and LBG cluster techniques. We also investigated the performance of speaker verification with different VQ methods and GMM in [25,26]. We proposed and validated IT-EM method with delta-MFCC in [27]. This paper evaluates the performance of our proposed IT-EM method with the MFCC/IMFCC.

ITVQ uses the IT principles such as PDE estimation and KL divergence measure. PDE and KL are applied to enhance the convergence rate of EM procedure. Our novel approach used for parameter optimization of GMM is elaborated in Fig. 4. The proposed procedure is referred as IT-EM, as it mingles the EM algorithm with IT metric. Using IT-EM, the clustering process of EM algorithm is improved by selection of centroids achieved by IT metric, as shown in Fig. 5(a-b) respectively. In IT-EM algorithm, the convergence is maintained by both, that is by preserving the maximization properties of EM as well as iterative upgrading of centroids calculation. The iterative upgrading of centroid calculation is guided by the information theoretic criteria. The IT criterion simultaneously minimizes divergence measure between each vector within a given cluster and centroids of this cluster, and maximizes the divergence between centroids of neighboring clusters. For further details on IT-EM algorithm, the sequence of steps used by IT-EM algorithm, and the computations used to evaluate centroids [24,27].

The clustering can be classified to sharp and hard clustering approaches. IT-EM approach can be regarded as sharp clustering (Fig.5(b)), because with every update of EM, the number of feature vectors for every speaker class is replaced with a small number of centroids. However IT-EM is performed on centroid vectors instead of original feature vectors. The centroid vectors are updates by applying a number of IT updates nested within EM procedure. Therefore IT-EM has bi-optimization character, since it uses updating not only for input set of features but also to more refined arrangement of centroid vectors.

4. EXPERIMENTAL EVALUATION OF SPEAKER VERIFICATION BASED ON IMFCC AND IT-EM

4.1 Speaker Verification System

The arrangement of the speaker verification system, used in the experimental evaluation of the proposed MFCC/ IMFCC and IT-EM methods is discussed in this section. The verification system works in three possible fashions, UBM (Universal Background Model) training mode using MA (Maximum a Posteriori) estimation, Target speaker enrollment and Testing/recognition.

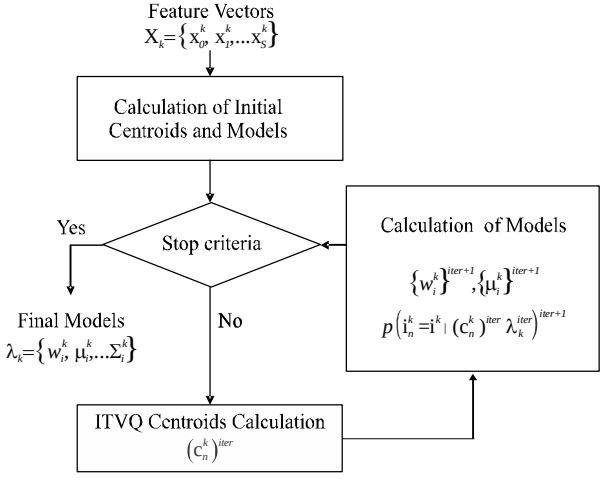

FIG. 4. BLOCK DIAGRAM OF IT-EM PROCEDURE

For UBM training, speaker enrolment and verification stages, similar speech detection and speech feature extraction techniques are followed. For speech detection an energy based silence detector described in [28] is used. It has been concluded in several research papers that MFCC performs better and is relatively robust when frame size is in the range 20-50ms and frame step is in the range 5-15ms of the frame size. Therefore, we have also used MFCC to characterize the speaker information using 30ms frame size and 10ms frame step. For each frame 12 MFCCs, 12 delta-MFCCs, 12 double-delta-MFCCs, 12 IMFCCs, 1 averaged spectral energy coefficient and 1 zero-crossing coefficient is calculated.

For each frame the following feature vectors are generated: 12-dimensional MFCC feature vector, 12-dimensional IMFCC feature vector, 24-dimensional MFCC/IMFCC fused feature vector and 38-dimensional delta-MFCC feature vector. The delta-MFCC is used as the baseline feature extractor.

MAP-UBM based GMM is then used to model the sequence of feature vectors. The GMM based modeling is tuned by both EM and IT-EM algorithms and the trained models are stored separately. For each speaker, the Gaussian components used are 1024. Approximately 5 minutes training utterances are taken from NIST2004 speech corpus and approximately same length of test utterances are used to evaluate the system performance. Once the enrollment is complete, the UBM [3] parameter conjecture is achieved using both EM and IT-EM non-target speakers, this is obtained using NIST2001. The target speaker means are then adjusted away from the UBM using MAP estimation. Testing/verification is performed by using same feature extraction as for speaker model training. In verification mode the tested set of feature vectors is scored by speaker's training model.

With NIST2004 protocol the set of verification tests is defined, the defined set is used to evaluate the proposed MFCC/IMFCC fusion and IT-EM for speaker verification task. The system performance is evaluated using EER measure, as defined above and by plotting the two probabilities to constitute DET (Detection Error Trade-off) curve.

4.2 Comparison of the Training Algorithms Convergence Rates

Fig. 6 shows the relationship amongst ITVQ updates and the log likelihood calculated at each updated stage. It can be observed that IT-EM preserves the monotonic

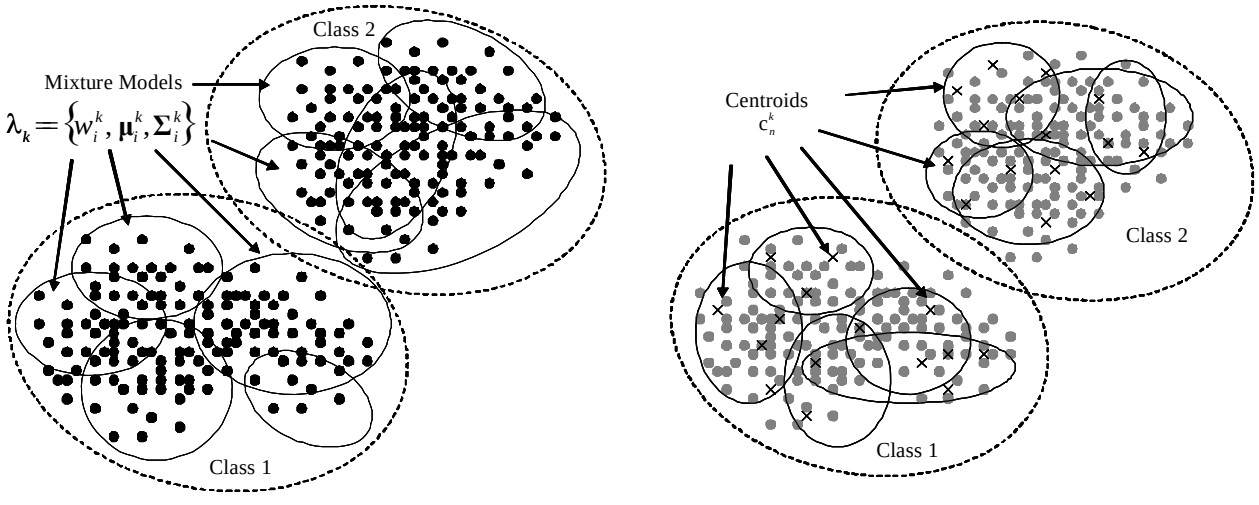

FIG. 5(a). EM CLUSTERING (b) IT-EM CLUSTERING, THE BLACK CROSSES REPRESENT IT-EM CENTROIDS

character of EM. It also improves (as 1/log-likelihood is descending) log-likelihood at each update stage. It is obvious that IT-EM has added to complexity by using ITVQ metric. In [27] we have given a detailed description about convergence rates of IT-EM.

4.3 Speaker Verification Results

The speaker verification results are shown in Figs. 7-8. Using different feature extractors the percentage miss probability and the percentage false alarm probability using GMM based on EM and IT-EM is obtained; EER is the value at which the two probability measures are equal. Greater the EER value means poor the system performance. A GMM based system summarized above is established, it involves training under EM and IT-EM algorithms. The IT-EM procedure is examined using MFCC/IMFCC, delta/MFCC, MFCC and IMFCC feature vectors. The most important finding is that MFCC/IMFCC has the better performance compared to delta-MFCC, since MFCC/IMFCC is low dimensional feature vector compared to delta-MFCC, therefore MFCC/IMFCC can be regarded as the efficient algorithm. The average improvement of MFCC/IMFCC over delta-MFCC is 0.75% for EM based modeling and 0.25% for IT-EM based modeling.

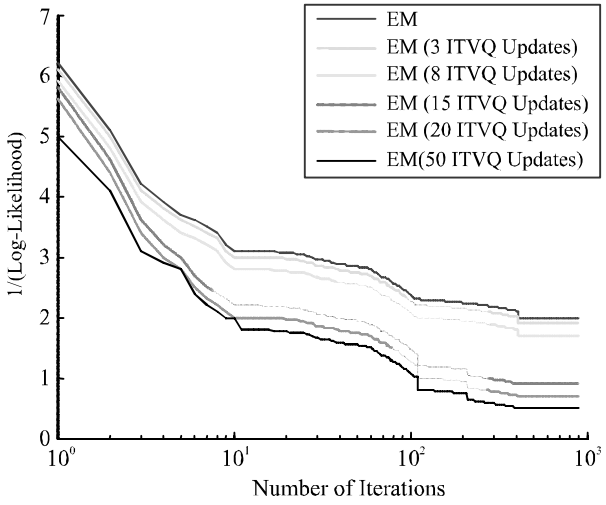

FIG. 6. EM, IT-EM CONVERGENCE RATES

It is also evident from the details shown in Figs. 7-8 that the IT-EM based modeling shows an improvement of the average EER's values over the classical EM algorithm. The average improvement of EER is about 0.65% (for MFCC/IMFCC) and 1.15% (for delta MFCC) for NIST2004.

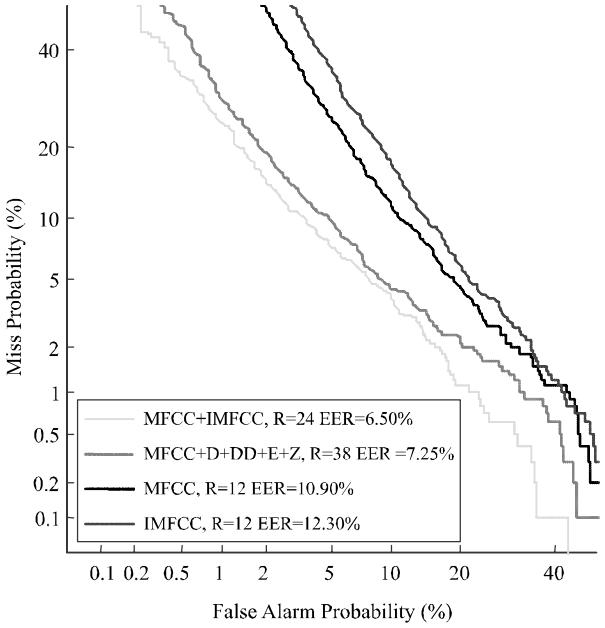

FIG. 7. EER BASED ON EM MODELLING.

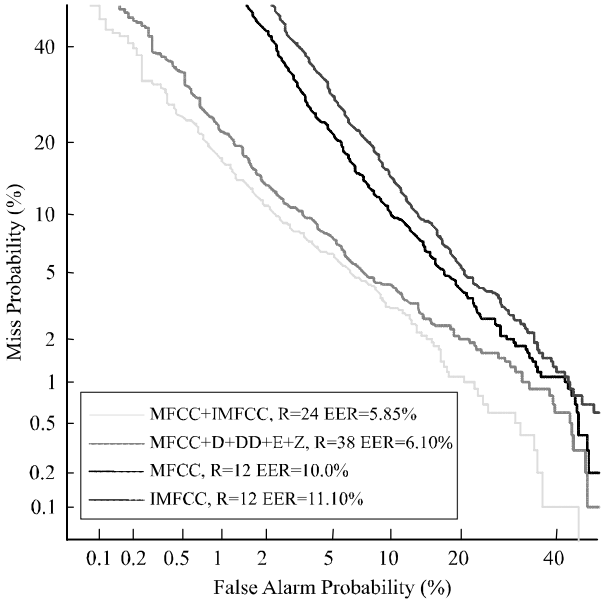

FIG. 8. EER BASED ON IT-EM MODELLING

5. CONCLUSION

A feature extraction strategy with input level fusion is tested. A novel modeling approach is described and validated. The results indicate that the speaker characteristic information is present in both low and high frequency ranges. Although the MFCC with their high resolution at the low frequency range provide relatively good speaker verification rates, the addition of IMFCC with high resolution at the high frequencies improves the verification results. The proposed IT-EM achieves improved convergence rate and thus leads to smaller EER values for speaker verification task. IT-EM is applied on averaged feature vectors called ITVQ centroids. IT-EM is directed by objective of minimizing divergence between original feature vectors and centroids. IT-EM is actually the sequential implementation of ITVQ, which derives cluster centroids and it is followed by EM on ITVQ centroids to approximate the Gaussian mixture parameters.

ACKNOWLEDGEMENTS

The authors would like to thank Prof. Dr. Abdul Qadeer Khan Rajput, Vice-Chancellor, and Prof. Dr Mukhtiar Ali Unar, Director, Institute of Information & Communication Technologies, Mehran University of Engineering & Technology, Jamshoro, Pakistan, for their support and guidance in order to complete this research.

REFERENCES

[1] Quatieri, T.F., "Discrete-Time Speech Signal Processing Principles and Practice", Prentice Hall, 2002.

[2] Liu, Y., Russell, M., and Carey, M.," The Role of Dynamic Features in Text-Dependent and Independent Speaker Verification", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Volume 1, May, 2006.

[3] Reynolds, D.A., Quatieri, T., and Dunn, R., "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, Volume 10, No. 1, pp. 19-41, 2000.

[4] Bozkurt, E., Erzin, E., Erdem, C.E., and Erdem, A.T., "Automatic Emotion Recognition for Facial Expression Animation from Speech", IEEE Conference on Signal Processing and Communications, pp. 989-992, 2009.

[5] Lehn-Schiøler, T., Hegde, A., Erdogmuz, D., and Principe, J., "Vector-Quantization Using Information Theoretic Concepts", Natural Computing, Voume 4, No. 1, pp. 39-51, January, 2005.

[6] Zwicker, E., and Terhardt, E., "Analytical Expressions for Critical Band Rate and Critical Bandwidth as a Function of Frequency", Journal of Acoustical Society of America, Volume 68, No. 5, pp. 1523-1525, 1980.

[7] Scherer, K.R., Johnstone, T., Klasmeyer, G., and Bänziger, T., "Can Automatic Speaker Verification be Improved by Training the Algorithms on Emotional Speech", University of Geneva, Switzerland, 2000.

[8] Davis, S.B., and Mermelstein, P., "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentence", IEEE Transactions on Acoustic Speech and Signal Processing, Volume 28, No. 4, pp. 357-366, 1980.

[9] Reynolds, D.A., "Speaker Identification and Verification Using Gaussian Mixture Speaker Models", Speech Communication, Volume 17, pp. 91-108, 1995.

[10] Chakroborty, S., Roy, A., Majumdar, S., and Saha, G., "Capturing Complementary Information via Reversed Filter Bank and Parallel Implementation with MFCC for Improved Text-Independent Speaker Identification", International Conference on Computing Theory and Applications, pp. 463-467, March, 2007.

[11] Yegnanarayana, B., Prasanna, S.R.M., Zachariah, J.M., and Gupta, C.S., "Combining Evidence from Source, Suprasegmental and Spectral Features for a Fixed-Text Speaker Verification System", IEEE Transactions on Speech and Audio Processing, Volume 13, No. 4, pp. 575-582, July, 2005.

[12] Murty K.S.R., and Yegnanarayana, B., "Combining Evidence from Residual Phase and MFCC Features for Speaker Recognition", IEEE Signal Processing Letters, Volume 13, No. 1, pp. 52-55, January, 2006.

[13] Prasanna, S.R.M., Cheedella, S.G., and Yegnanarayana, B., "Extraction of Speaker-Specific Excitation Information from Linear Prediction Residual of Speech", Speech Communication, Volume 48, No. 10, pp. 1243-1261, October, 2006.

[14] Gold, B., and Morgan, N., "Speech and Audio Signal Processing", Part-IV, Chapter-14, pp. 189-203, John Willy & Sons, 2002.

[15] Damper, R., and Higgins, J., "Improving Speaker Identification in Noise by Sub Band Processing and Decision Fusion", Pattern Recognition Letters, Volume 24, pp. 2167-2173, 2003.

[16] Sorenson, H.W., and Aspach, D.L., "Recursive Bayesian Estimation Using Gaussian Sums", Automatica, Volume 7, pp. 465-479, 1971.

[17 Duda, R.O., and Hart, P.E., "Pattern Classification and Scene Analysis", Wiley, New York, 1973.

[18] Reynolds, D.A., and Rose, R.C., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Transactions on Speech and Audio Processing, Volume 3, No.1, pp. 72-83, 1995.

[19] Pelecanos, J., Myers, S., Sridharan, S., and Chandran, V., "Vector Quantization Based Gaussian Modeling for Speaker Verification", International Conference on Pattern Recognition, Volume 3, pp. 294-297, Spain, 2000

[20] Alpaydm, E., "Soft Vector Quantization and the EM Algorithm", Neural Networks, Volume 11, No. 3, pp. 467-477, April, 1998.

[22] Ueda, N., and Nakano, R., "Deterministic Annealing EM Algorithm", Neural Networks, Volume 11, pp. 271-282, 1998.

[23] Ververidis, D., and Kotropoulos, C., "Gaussian Mixture Modeling by Exploiting the Mahalanobis Distance", IEEE Transactions on Signal Processing, Volume 56, No. 7, pp. 2797-2811, July, 2008.

[24] Hedelin, P., and Skoglund, J., "Vector Quantization Based on Gaussian Mixture Models", IEEE Transactions on Speech and Audio Processing, Volume 8, No. 4, pp. 385-401, 2000.

[25] Memon, S., and Lech, M., "Speaker Verification Based on Information Theoretic Vector Quantization", Communications in Computer and Information Science, Wireless Networks, Information Processing and Systems, Springer Berlin Heidelberg, 2009.

[26] Memon, S., and Lech, M., "Using Information Theoretic Vector Quantization for GMM Based Speaker Verification", 16th European Signal Processing Conference, Lausanne, Switzerland, August 25-29, 2008.

[27] Memon, S., Lech, M., and Maddage, N., "Speaker Verification Based on Different Vector Quantization Techniques with Gaussian Mixture Models", Third International Conference on Network and System Security, 2009.

[28] Memon, S., Lech, M., and Maddage, N., "Information Theoretic Expectation Maximization Based Gaussian Mixture Modeling for Speaker Verification", 20th International Conference on Pattern Recognition, pp. 4536-4540, [ISBN: 978-0-7695-4109-9], 2010.

[29] Reynolds, D.A, Rose, R.C., and Smith, M.J.T., "The PC-Based TMS 320C30 Implementation of the Gaussian Mixture Model Text-Independent Speaker Recognition System", Proceedings of ICSPAT, pp. 967-973, November, 1992.