Calculating Model Parameters Using Gaussian Mixture Models; Based on Vector Quantization in Speaker Identification

Hamideh Rezaei-Nezhad
h.rezaei@iauk.ac.ir
Department of Computer Engineering, Qeshm Branch, Islamic Azad University, Qeshm, Iran.

Summary
The use of Gaussian Mixture Model (GMM) is most common in speaker identification. The most of the computational processing time in GMM is required to compute the likelihood of the test speech of the unknown speaker with consider to the speaker models in the database. The time required for speaker identification is depending to the feature vectors, their dimensionality and the number of speakers in the database. In this paper, we focused on optimizing the performance of Gaussian mixture (GMM) and adapted Gaussian mixture model (GMM-UBM) based speaker identification system and proposed a new approach for calculation of model parameters by using vector quantization (VQ) techniques to increase recognition accuracy and reduce the processing time. Our proposed modeling is based on forming clusters and assigning weights to them according to upon the number of mixtures used for modeling the speaker. The advantage of this method is in the reduction in computation time which depends upon how many mixtures are used for training the speaker model by a substantial value compared with approaches which use expectation maximization (EM) algorithm for computing the model parameters.

Key words:
Speaker identification, Gaussian mixture model, EM algorithm, Vector quantization, Feature extraction.

1. Introduction

Statistical models such as Hidden Markov Models (HMM), Neural Networks (NN), Support Vector Machines (SVM) and Gaussian Mixture Models (GMM) have been used in speaker recognition in the past several years. Using a Gaussian mixture model due to having a number of advantages has become a classic and successful method in speaker recognition and makes it suitable for modeling the probability distributions over vectors of input features [1]. Generally, speaker identification system consists of three phases [2-3]. Feature extraction is the first step and followed by feature selection where the speech signal is extracted to feature vectors. Speech can be characterized in terms of the signal carrying message

Information and also this kind of signal has been very useful in some applications. Feature extraction could get three main types of information: Speech Text, Language and Speaker Identity. The second step is speaker modeling and developing a speaker model database. By using feature vectors extracted from a given speaker's training utterance(s), a speaker model is trained and stored into the system database. In text-dependent mode, the model is utterance-specific and it includes the temporal dependencies between the feature vectors. The last step is decision making.

This paper focused on the calculation of model parameter in text-independent speaker identification systems using Gaussian mixture model and adapted Gaussian mixture model. We are optimist in searching an approach to reduce the computational time in speaker modeling [4-6]. The most important step of speaker modeling is the calculation of model parameters [7]. In this paper EM algorithm is used for the calculation of model parameters for both GMM and GMM-UBM approaches.

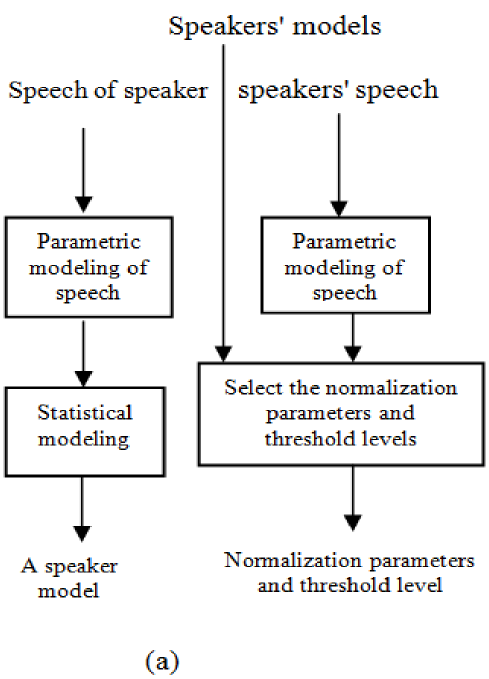
We consider a speaker identification system based on the EM algorithm for calculating the model parameters and we have investigated another method by using the VQ technique to calculate the model parameters. For these approaches, i.e., GMM based on EM and GMM based on VQ (GMM-UBM based on EM and GMM-UBM based on VQ), the recognition rates and computation time are compared. It has been shown out that even though the recognition accuracy of two methods is nearly equal but the computation time is considerably reduced in the new method. In Section II, we discuss how to use Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction of speech and explain the front-end processing technique in short [8-10].

Section III and IV explain the Gaussian mixture model based speaker identification and maximum likelihood parameter estimation to compute model parameters respectively [11-14]. Adapted Gaussian mixture model is explained in section V. The sixth section is dedicated to introducing the new method for computing model parameters using Vector Quantization. Experimental results and its discussion are presented In Section VII, and conclusions are shown in last Section.

Manuscript received February 5, 2017
Manuscript revised February 20, 2017

2. Speech Feature Extraction

For all the recognition systems, there are two main phases. The first phase is called training phase and the next phase is called identification or (testing) phase. Training and Testing are two important steps of an identification system. Training phase is to get the speaker models or voiceprints for speaker database. In this phase, the most useful features are extracted from speech signal for speaker identification or verification, and train models to get optimal system parameters. Feature extraction is the heart of the speaker identification system. In testing phase, the same method for extracting features as in the first phase is used for the incoming speech signal, and then the speaker models getting from enrollment phase are used to calculate the similarity between the new speech signal model and all the speaker models in the database. Fig. 1 shows the training and testing phases for speaker identification. The human speech signal conveys many levels of information ranging from phonetic content to speaker identity and even emotional status. Human voice before its final form passes through two different systems. The first system is vocal folds and the second system is vocal tract [15]. The aim of feature extraction stage is to extract the speaker particular information as feature vectors.



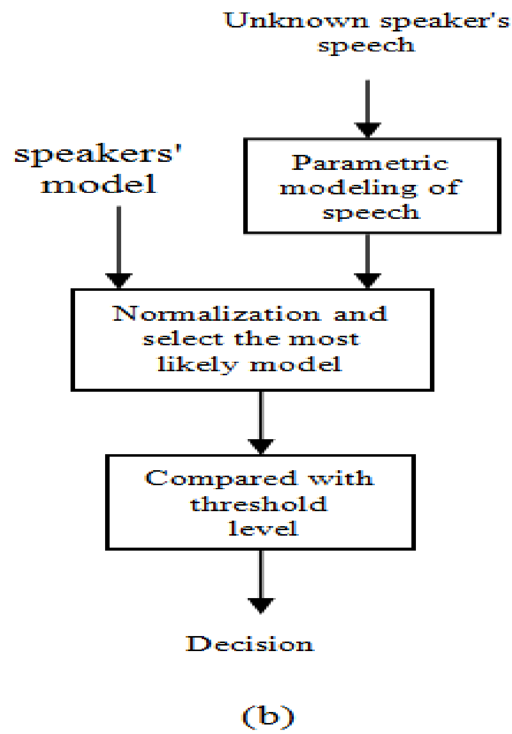


Fig. 1 (a) Training phase, (b) Testing phase.

The MFCC is a method that analyzes how the Fourier transform extracts frequency components of a signal in the time-domain. Popular features for describing speech signal are MFCC. In this paper, MFCCs is used for feature extraction which take human ear's frequency response into consideration. Process of feature extraction is shown in Fig. 2 and the comprehensive method is explained in [16-17]. In the frame blocking the input speech waveform is divided into frames of approximately 30 milliseconds. The windowing block minimizes the discontinuities of the signal by tapering the beginning and end of each frame to zero.

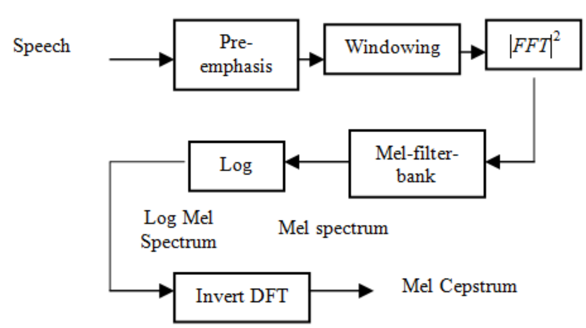


Fig. 2 Mel-frequency Cepstral Coefficients feature extraction process.

Each frame converts from the time domain to the frequency domain by the FFT block then the magnitude spectrum of the utterance is passed through a bank of triangular-shaped filters. The energy output of each filter is compressed and transformed to the Cepstral domain via the DCT.

Cepstrum, $c(n)$ in its simplest form is the discrete cosine transformation of the Mel-spectrum of a signal, $s(n)$ in logarithmic amplitudes and can be mathematically defined as

$$c(n) = ifft(lof \,|\, fft(s(n))|) \tag{1}$$

Speech signal consists of many features that all of them are not important for speaker discrimination [22], [24]. An ideal feature would:

- have large between-speaker variability and small within-speaker variability
- be robust against noise and distortion
- occur frequently and naturally in speech
- be easy to measure from speech signal
- be difficult to impersonate/mimic
- Not be affected by the speaker's health or long-term variations in voice.

3. Gaussian Mixture Model

GMM is a classic parametric method best used for speaker modeling due to the fact that Gaussian components have the capability of representing some general speaker dependent spectral shapes. Modeling techniques like GMM are to generate speaker models from Feature vectors that obtained from the above step by using statistical variations of the features. A GMM is a parametric probability density function represented as a weighted sum of Gaussian component densities that is shown to provide a smooth approximation to the underlying long-term sample distribution of observations obtained from utterances by a given speaker [18]. An important step in the implementation of the likelihood ratio detector is selection of the actual likelihood function. The choice of this function is largely dependent on the features being used as well as specifics of the application. For text-independent speaker recognition, where there is no prior knowledge of what the speaker will say, the most successful likelihood function has been Gaussian mixture models. A Gaussian mixture model is generated from a mixture of a finite number of Gaussian distributions that each distribution is defined by a mean vector $\vec{\mu}_i$, a covariance matrix $\sum_{i=1}^{M}$ and a mixture weight $\vec{\rho}_i$. In a GMM model, the probability distribution density given by the following equation:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{2}$$

Where $\vec{x}$ is a D-dimensional random vector, $b_i(\vec{x}), i =$ 1,…, M , are the component densities and $p_i, i =$ 1,…, M, are the mixture weights.

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)' \Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)\right\} \tag{3}$$

Mean vectors, covariance matrices and mixture weights parameterize the complete Gaussian mixture density. These parameters are represented by $\lambda = \{\rho_i, \mu_i, \Sigma_i\}$, i=1,.., M.

Each speaker in a speaker identification system can be represented by a GMM and is referred to by the speaker's respective model λ. The parameters of a GMM model can be estimated using maximum likelihood (ML) estimation. The main objective of the ML estimation is to derive the optimum model parameters that can maximize the likelihood of GMM.

4. Maximum Likelihood Parameter Estimation

Unfortunately direct maximization using ML estimation is not possible and therefore a special case of ML estimation known as Expectation-Maximization (EM) algorithm is used to extract the model parameters. The goal of this technique is to maximize the following formula with best matches in distribution of training vectors.

$$p(X|\lambda) = \prod_{t=1}^{T} p(\vec{x}_t|\lambda) \tag{4}$$

Where $X = \{\vec{x}_1, ..., \vec{x}_T\}$ is a set of T training vectors. The EM algorithm is used to estimate parameters. The goal of the EM algorithm is to compute the model parameters iteratively till $p(X|\lambda^{k+1}) \geq p(X|\lambda^k)$.

To guarantee the above condition, the following formulae are used:

Mixture weights: $$\bar{p}_t = \frac{1}{T}\sum_{t=1}^{T} p\left(i_t = i \middle| \vec{x}_t, \lambda\right) \quad (5)$$

MEANS: $$\vec{\bar{\mu}}_i = \frac{\sum_{t=1}^{T} p\left(i_t = i \middle| \vec{x}_t, \lambda\right)\vec{x}_t}{\sum_{t=1}^{T} p\left(i_t = i \middle| \vec{x}_t, \lambda\right)} \quad (6)$$

VARIANCES: $$\vec{\bar{\sigma}}_i^2 = \frac{\sum_{t=1}^{T} p\left(i_t = i \middle| \vec{x}_t, \lambda\right)\vec{x}_t^2}{\sum_{t=1}^{T} p\left(i_t = i \middle| \vec{x}_t, \lambda\right)} \quad (7)$$

Where for acoustic class *i* is given by

$$p\left(i \middle| \vec{x}_t, \lambda\right) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)} \quad (8)$$

A group of S speakers S={1,2,…,S} is represented by GMM's $\{\lambda_1,\ldots,\lambda_s\}$ For speaker identification . Finding the speaker model which has the maximum a posteriori probability of a given observation sequence is the identification's aim.

$$\hat{S} = \arg\max_{1 \le k \le S} p\left(X \middle| \lambda_k\right) \quad (9)$$

5. Adapted Gaussian mixture model

In the GMM-UBM system we use a single, speaker independent background model to represent $p(X|\lambda)$. The UBM is a large GMM trained to represent the speaker-independent distribution of features. Specifically, we want to select speech that is reflective of the expected alternative speech to be encountered during recognition. This applies to both the type and the quality of speech, as well as the composition of speakers. There are many approaches that can be used to obtain the final model to train the UBM [19].

The simplest is to merely pool all the data to train the UBM via the EM algorithm. One should be careful that the pooled data are balanced over the subpopulations within the data. In this system, we derive the hypothesized speaker model by adapting the parameters of the UBM using the speaker's training speech and a form of Bayesian adaptation Unlike the standard approach of maximum likelihood training of a model for the speaker independently of the UBM, the basic idea in the adaptation approach is to derive the speaker's model by updating the well-trained parameters in the UBM via adaptation. Like the EM algorithm, the adaptation is a two steps estimation process. The first step is identical to the expectation step of the EM algorithm. Unlike the second step of the EM algorithm, for adaptation these new sufficient statistic estimates are then combined with the old sufficient statistics from the UBM mixture parameters using a data-dependent mixing coefficient. For mixture *i* in the UBM, we compute equation (3) that is the same as the expectation step in the EM algorithm. Finally, these new sufficient statistics from the training data are used to update the old UBM sufficient statistics for mixture *i* to create the adapted parameters for mixture *i* with the equations:

$$\hat{w}_i = \left[\alpha_i^w n_i / T + \left(1 - \alpha_i^w\right) w_i\right]\gamma \quad (10)$$

$$\hat{\vec{\mu}}_i = \alpha_i^m E_i(\vec{x}) + \left(1 - \alpha_i^m\right)\vec{\mu}_i \quad (11)$$

$$\hat{\vec{\sigma}}_i^2 = \alpha_i^v E_i\left(\vec{x}\vec{x}^T\right) + \left(1 - \alpha_I^V\right)\left(\vec{\sigma}_i^2 + \vec{\mu}_i^2\right) - \hat{\vec{\mu}}_i^2 \quad (12)$$

The adaptation coefficients controlling the balance between old and new estimates are $\{\alpha_i^w, \alpha_i^m, \alpha_i^v\}$ for the weights, means and variances, respectively. The scale factor γ is computed over all adapted mixture weights to ensure they sum to unity. For each mixture and each parameter, a data-dependent adaptation coefficient $\alpha_i^\rho, \rho \in \{w, m, v\}$, is used in the above equations. This is defined as

$$\alpha_i^\rho = \frac{n_i}{n_i + r^\rho} \quad (13)$$

Where r^ρ is a fixed relevance factor for parameter ρ .

6. Vector Quantization Approach

For calculating model parameters, the vector quantization (VQ) method and its uses are introduced in this section [20]. Vector quantization is an approach to mapping vectors to a finite number of regions in the space as it shows in Fig. 3. Regions are called clusters and can be represented by their centers. Each center called a code word and the collection of all code words is a code book. The VQ codebook has a small number of highly representative vectors that efficiently represent the speaker specific characteristics.

This is a method used for reducing or compressing the number of training vectors required in a recognition system. In this new approach, VQ as a modeling technique was used in speaker identification [21]. After obtaining the feature vectors of the input speech segment, feature vectors are divided into a certain number of clusters, M that introduced as a codebook size using the approach in [22]. Each cluster has one centroid which representing the mean of all the feature vectors related to that cluster. The identification error percentage is directly depending on the size of the codebook [23]. This technique was used to find the M clusters such that each cluster has a weight of at least 1/M. In equation (3), we have used the centroids of these clusters as a mean. By using feature vectors belonging to each of the M clusters the covariance matrix is obtained. By this way all the parameters are obtained in equation (3).

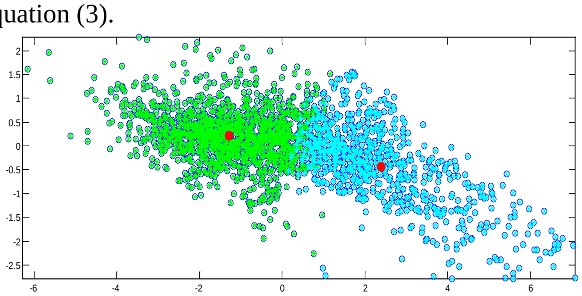


Fig. 3 Mapping vectors in the space using VQ approach.

The means for M mixtures are randomly initialized in the previous approach (using EM for calculating model parameters). In a 2-dimensional feature vector space, it may this space be dense in certain regions where most of the feature vectors are located and some feature vectors are at larger distances from feature vectors in the dense area, so it could devote feature vectors which are at higher distances from other feature vectors, as a means but VQ technique takes into consideration feature vectors belonging to that cluster only by using clustering approach. This minimizes the value of $(\vec{x} - \vec{\mu}_i)$ in equation (3). This approach despite keeping all the advantages Gaussian Mixture Modeling technique has efficient result for calculating model parameter.

7. Experiments setup, result and discussion

As a test material for our experiments we used the Farsdat database. The experiments were made using a speaker database containing speech data from 100 speakers. Distribution of male and female speakers on the speaker database is almost equal. Two sessions for training and testing sessions were used. Our experiments operate on Cepstral features, extracted using a 24-ms Hamming window with 10 milliseconds overlapping. The signal was pre emphasized by the filter H (z) =1-0.97·z-1 and silence frame was removed before the feature extraction. 12 MFCCs together with log energy were calculated using a bank of 13 filters as mentioned in [24]. Thus we have obtained 12-dimensional feature vectors. For training phase, training was done in different durations: 30sec and 60sec. System was tested using 10sec test frames.

Two sets of experience were done; EM algorithm is used for training the model in the first set of experiments for GMM and GMM-UBM approaches. In the second set, Vector Quantization have used for computing the model parameters.

Table1: Identification accuracy and running time for EM-GMM and VQ-GMM approach (training with 30sec and testing with 10sec)

M	EM-GMM		VQ-GMM	
	Accurac	Time	Accurac	Time
8	82	9.6321	85	5.8866
16	89	14.6712	91	7.7798
32	89	28.0978	91	12.3149
64	90	45.8909	92	13.9794
128	90	78.1192	94	22.4089

Table2: Identification accuracy and running time for EM-GMM and VQ-GMM approach (training with 60sec and testing with 10sec)

M	EM-GMM		VQ-GMM	
	Accurac	Time	Accurac	Time
8	78	4.2298	82	2.6370
16	83	6.4272	85	3.3267
32	83	11.8467	85	3.6709
64	85	21.6620	86	5.9567
128	85	40.2443	86	8.9870

If the cluster for every centroid has weight of 1/M (M is the number of mixtures), stops splitting but if the cluster has a weight less than 1/M, the centroids are split again. Selection of M is important for shorter training data. The covariance is computed based on the data for each cluster. Identification accuracy is calculated for 10sec testing data using training model parameters obtained from the above steps.

Tables 1 and Table 2 show a comparison of considering the accuracy and timelines required between GMM based on EM and GMM based on VQ for different training and testing times. Furthermore, Table 3 shows GMM-UBM based on EM and GMM-UBM based on VQ for training and testing times as for Table 1 and 2.

Table3: Identification accuracy and running time for EM-GMM-UBM and VQ-GMM-UBM approach (training with 30sec and testing with 10sec)

M	EM-GMM-UBM		VQ-GMM-UBM	
	Accuracy	**Time**	**Accuracy**	**Time**
8	84	7.5466	86	3.7655
16	91	12.6672	93	6.1312
32	91	19.9879	93	9.8087
64	93	36.1243	95	12.5998
128	94	58.2556	95	19.4497

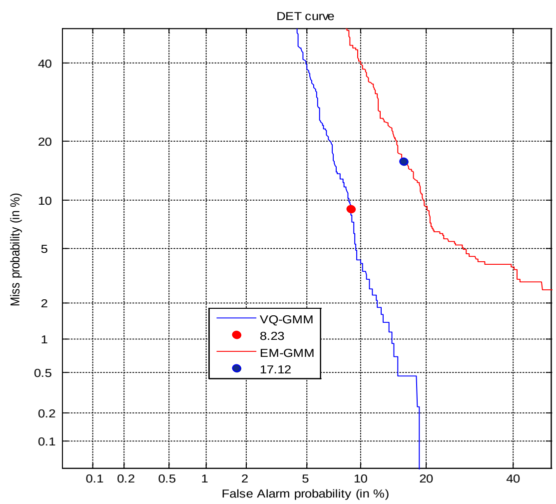


Fig. 4 Comparison of VQ-GMM and EM-GMM approach for 128 mixtures.

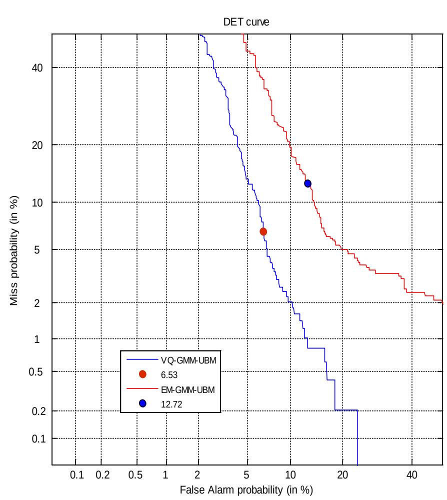


Fig. 5: Comparison of VQ-GMM-UBM and EM-GMM-UBM approach for 128 mixtures.

Fig. 4 and Fig. 5 show the above set of experiments which are plotted on the DET curve. 128 mixtures are used in this experiment. Training duration is of 60sec and test duration is of 10sec. We can see that, VQ-GMM and VQ-GMM-UBM approaches have a slightly better performance rather than EM-GMM and EM-GMM-UBM respectively. We compare the effect of model size using VQ-GMM and VQ-GMM-UBM approaches on speaker identification performance in the next experiment. The training is 60sec and testing is 10sec. Referring to Fig. 6 and Fig. 7, the curves showed that the increase in the number of mixtures increases the performance of the system.

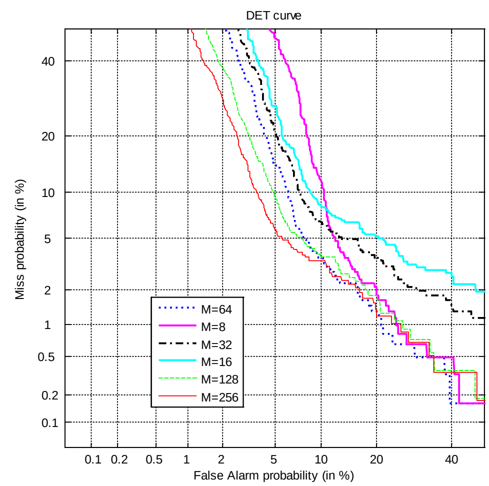


Fig. 6 Effect of model size on speaker identification using VQ-GMM approach.

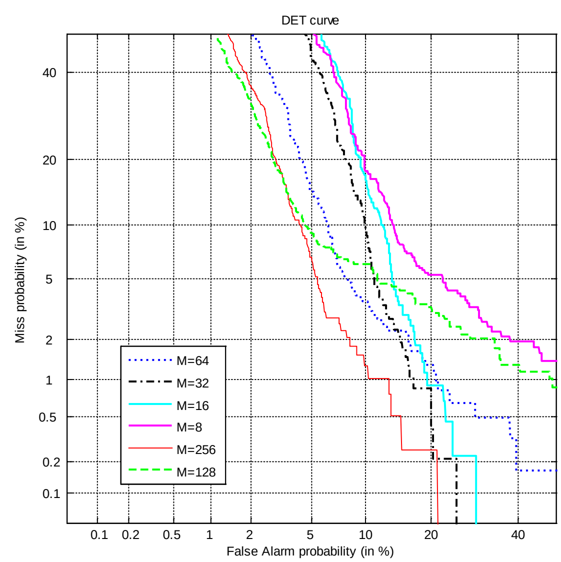


Fig. 7 Effect of model size on speaker identification using VQ-GMM-UBM approach.

8. Conclusion

The implementation of GMM based speaker identification has been addressed in this paper. Four approaches were used for training the speaker model. It has been shown that using VQ-GMM and VQ-GMM-UBM in the model parameters calculation have a slight improvement in identification accuracy. Considerable improvement is observed in computational time. Table I and Table III showed a speedup factor of 7 was achieved in the first set of experiments with 128 mixtures and training duration of 30 Sec, while in the second set of experiments a speedup factor 5 was achieved with 256 mixtures and training data of 60sec as shown in the Table II.

References

[1] D. A. Reynolds and R. C. Rose , "Robust text-independent speaker identification using Gaussian speaker models", *IEEE Trans. Speech Audio Process*. pp. 72-83, 1995.

[2] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "A robust speaker-adaptive HMM based text-to-speech synthesis," *IEEE Trans. Speech, Audio & Language Process.*, Vol. 17, No. 6, pp. 1208–1230, Aug. 2009.

[3] Tomi Kinnuen et al., "Real-time speaker identification and verification", *IEEE Transactions on audio, speech and language processing*, Vol. 14, No. 1, Jan. 2006.

[4] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, Vol. 10, pp. 19-41, 2000.

[5] Xiong, Z., Zheng, T.F., Song, Z., Soong, F., Wu, W., "A tree-based kernel selection approach to efficient Gaussian mixture model universal background model based speaker identification," *Speech Communication* , Vol. 48, pp. 1273–1282, 2006.

[6] Saeidi, R., Sadegh Mohammadi, H.R., Rodman, R.D., Kinnunen, T., "A new segmentation algorithm combined with transient frames power for text independent speaker verification," In: *Proc. ICASSP* 2007, Vol. 1, pp. 305–308, April 2007.

[7] Arthur Chan et al., "Four-layer categorization scheme of fast GMM computation in large vocabulary continuous speech recognition systems", in *Proc. Of Interspeech*, pp. 689-692, 2004.

[8] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Lang. Process*., Vol. 19, No. 4, pp. 788–798, 2011

[9] Sandipan Chakroborty, Anindya Roy and Goutam Saha "Improved Closed Set Text- Independent Speaker Identification by Combining MFCC with Evidence from Flipped Filter Banks" *International Journal of Signal Processing*, Vol. 4, pp.114-121, Nov. 2006.

[10] M. Sahidullah and G. Saha, "Design, Analysis and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition," *Speech Communication*, Vol. 54, No. 4, 2012, pp. 543-565.

[11] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, Vol. 10, No. 2, pp. 19-41, 2000

[12] D. A. Reynolds, "Gaussian Mixture Models," *Technical Report, MIT Lincoln Laboratory, Cincinnati*, 2001.

[13] Reynolds D. A, Rose R. C. , " Robust Text Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. Speech Audio Processing*, Vol. 3, pp. 72—83, 1995.

[14] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via EM Algorithm," *J. Royal Statist. Soc*., Vol. 39, No. 1, pp. 1-38, 1997.

[15] O. Schleusing, T. Kinnunen, B. Story, J.-M. Vesin, "Joint Source-Filter Optimization for Accurate Vocal Tract Estimation Using Differential Evolution", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 21, No. 8, pp. 1560--1572, August 2013.

[16] D. A. Reynolds, "A Gaussian mixture modeling approach to text-independent speaker identification", Ph.D. thesis, Georgia Institute of Technology, September 1992.

[17] T. Kinnunen, R. Saeidi, F. Sedlak, K.A. Lee, J. Sandberg, M. Hansson-Sandsten, H. Li, "Low-Variance Multitaper MFCC Features: a Case Study in Robust Speaker Verification", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 20, No. 7, pp. 1990-2001, September 2012.

[18] F. Bimbot et al., "A Tutorial on Text-Independent Speaker Verification," *EURASIP J. Appl. Signal Process.*, No. 4, pp. 430-451, 2004.

[19] T. Kinnunen, E. Karpov, and P. Franti, "Real-time speaker identification and verification," *IEEE Trans. Audio, Speech, and Language Process.*, Vol. 14, No. 1, pp. 277–288, Jan. 2006.

[20] A.Srinivasan, "Speaker Identification and verification using Vector Quantization and Mel frequency Cepstral Coefficients", *Research Journal of Applied Sciences, Engineering and Technology*, Vol. 4(I), pp. 33-40, 2012.

[21] 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.

[22] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for Vector Quantization," *IEEE Trans. on Communications*, Vol. COM48, No. 1, pp. 84-95, January 1980.

[23] A. Revathi, R. Ganapathy and Y. Venkataramani, "Text Independent Speaker Recognition and Speaker Independent Speech Recognition Using Iterative Clustering Approach," *International Journal of Computer Science & Information Technology*, Vol. 1, No. 2, pp. 30-42, 2009.

[24] Premakanthan and W.B. Mikhael, Speaker verification/ recognition and the importance of selective feature extraction: Review, *Proceedings of the 44th IEEE 2001, Midwest Symposium*, Vol. 1, pp. 14-17, 2001.