2012 IACSIT Hong Kong Conferences
IPCSIT vol. 29 (2012) © (2012) IACSIT Press, Singapore

A New Method for Computing EM Algorithm Parameters in Speaker Identification Using Gaussian Mixture Models

Mohsen Bazyar [1+], Ahmad Keshavarz [2], and Khatoon Bazyar [3]

[1]Department of Communication, Bushehr Branch, Islamic Azad University Bushehr, Iran
[2]Persian Gulf University of Bushehr, Iran
[3]Department of Communication, Bushehr Branch, Islamic Azad University Bushehr, Iran

Abstract. In speaker identification, most of the computational processing time is required to calculate the likelihood of the test utterance of the unknown speaker with respect to the speaker models in the database. The time required for identifying a speaker is a function of feature vectors and their dimensionality and the number of speakers in the database. In this paper, we focus on optimizing the performance of Gaussian mixture (GMM) based speaker identification system. An improved approach for model parameter calculation is presented. The advantage of proposed approach lies in the reduction in computational time by a significant amount over an approach which uses expectation maximization (EM) algorithm to calculate the model parameter values. This approach is based on forming clusters and assigning weights to them depending upon the number of mixtures used for modeling the speaker. The reduction in computation time depends upon how many mixtures are used for training the speaker model.

Keywords: Speaker Recognition, Gaussian mixture model, Feature extraction, Vector quantization

1. Introduction

Over the past several years, there has been a significant amount of research in the field of speaker recognition. Various algorithms have been developed to model the speakers; these include HMM (Hidden Markov Models), NN (Neural Networks), SVM (Support Vector Machines) and GMM (Gaussian Mixture Models).A speaker recognition system typically consists of three stages: feature extraction, speaker modeling, and decision making. In this paper, we focus on the text-independent identification task using GMM. In this paper, we focus on an approach to reduce the computational time in speaker modeling. Model parameter calculation is an important step in speaker modeling. In this paper, we propose speaker identification using the approach described in [1]. In [1] the model parameters are calculated using the EM algorithm. We have investigated another approach for calculating the model parameters. This approach is based on the VQ technique. The identification rates and computational time for training the model are compared for both approaches, i.e., GMM based on EM and GMM based on VQ. It has been shown that the identification accuracy for both approaches is almost equal but the computational time has been greatly reduced in the new approach. The rest of the paper is organized as follows: Section 2 introduces the basic idea of GMM based speaker identification. In this section, the front-end processing technique, MFCC, used for feature extraction of speech is also discussed in short. Section 3 introduces the approach of model parameter calculation using Vector Quantization. In Section 4 experimental results are presented, and conclusions are drawn in Section5.

2. GMM Based Identification System

2.1. Speech Feature Extraction

+ Corresponding author. Tel.: + 07714542211;
E-mail address: mohsenbazyar114@yahoo.com.

Speech spectrum has been proven effective for speaker identification. Speech spectrum reflects user's vocal tract structure which distinguishes him/her from others. Studies done in the past have proven the filterbank technique to be very effective for speech recognition. In this paper, we use MFCCs which take human ear frequency response into consideration Fig. 1 illustrates the process of feature extraction [2] and the detailed procedure is explained in [3].

2.2. Gaussian Mixture Model

In GMM, we model the speaker data (feature vectors obtained from the above step) using statistical variations of the features. Hence, it provides us a statistical representation of how speaker produces sounds. Gaussian mixture density is shown to provide a smooth approximation to the underlying long-term sample distribution of observations obtained from utterances by a given speaker. These are the important motivations for using GMM as a modeling technique [1].

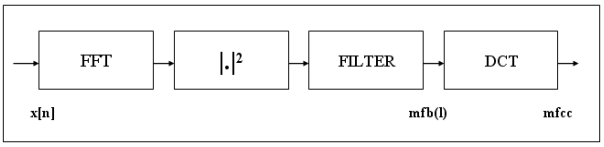

Fig.1: Mel-frequency Cepstral Coefficients feature extraction process

A Gaussian mixture density is a weighted sum of M component densities and is given by the equation

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

Where $\vec{x}$ is a D-dimensional random vector, $b_i(\vec{x}), i=1,\ldots,M$, are the component densities and $p_i, i=1,\ldots,M$, are the mixture weights. Each component density is a D-variate Gaussian function of the form

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

The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by $\lambda=\{p_i,\vec{\mu}_i,\Sigma_i\}, i=1,..,M$

2.3. Maximum Likelihood Parameter Estimation

Given a training data, the goal of model training is to calculate model parameters, l , which best matches the distribution of training vectors. The goal of the technique is to maximize

$$p(X|\lambda)=\prod_{t=1}^{T} p(\vec{x}_t|\lambda) \tag{3}$$

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

The following formulae are used to guarantee the above condition:

Mixture weights:
$$\bar{p}_t=\frac{1}{T}\sum_{t=1}^{T} p(i_t=i|\vec{x}_t,\lambda) \tag{4}$$

Means:
$$\bar{\vec{\mu}}_i=\frac{\sum_{t=1}^{T} p(i_t=i|\vec{x}_t,\lambda)\vec{x}_t}{\sum_{t=1}^{T} p(i_t=i|\vec{x}_t,\lambda)} \tag{5}$$

Variances:
$$\bar{\vec{\sigma}}_i^2=\frac{\sum_{t=1}^{T} p(i_t=i|\vec{x}_t,\lambda)\vec{x}_t^2}{\sum_{t=1}^{T} p(i_t=i|\vec{x}_t,\lambda)} \tag{6}$$

where the a posteriori probability for acoustic class *i* is given by

$$p\left(i|\vec{x}_t,\lambda\right)=\frac{p_i b_i\left(\vec{x}_t\right)}{\sum_{k=1}^{M} p_k b_k\left(\vec{x}_t\right)} \tag{7}$$

2.4. Identification

For speaker identification a group of S speakers S={1,2,…,S} is represented by GMM's $\{\lambda_1,...,\lambda_s\}$. The goal in identification is to find the speaker model which has the maximum a posteriori probability for a given observation sequence.

$$\hat{S}=\arg\max_{1\le k\le S} p\left(X|\lambda_k\right) \tag{8}$$

3. Vector Quantization Approach

In this section, we will introduce the vector quantization (VQ) approach first and its use for calculating model parameters. VQ was used as a modeling technique for speaker recognition [5]. In this technique, the entire speech data (feature vectors obtained from front-end processing) are divided into certain number of clusters, M, also known as codebook size using the approach in [4]. Each cluster has one centroid associated with it representing the mean of all the feature vectors belonging t that cluster. The size of the codebook has direct effect on the identification error percentage as mentioned in [5].We have used this technique to find the M clusters such that each cluster has a weight of at least 1/M. The centroids of these clusters are used as means in equation (2). The covariance matrix is calculated using feature vectors belonging to each of the M clusters. This way all the parameters required for equation (2) are obtained. In the previous approach, using EM based model parameter calculations, the means for M mixtures are randomly initialized. Random initialization could allocate feature vectors, which are at higher distance from other feature vectors, as means.This minimizes the value of $(\vec{x}-\vec{\mu}_i)$ in equation (2). This approach is fast. This results in an efficient approach for calculating model parameter keeping all the advantages Gaussian Mixture Modeling technique has to offer.

4. Experiments

The experiments were carried out on a speaker database containing 100 speakers. The database has almost equal distribution of male and female speakers. Two sessions were carried out to get data for training and testing sessions. Feature extraction process is performed as follows: Every 12ms speech signal is multiplied by 24ms Hamming window. 12 Mel-frequency cepstral coefficients are calculated using a bank of 13 filters as mentioned in [3].This may hamper identification accuracy to some extent. Thus we have obtained 12-dimensional feature vectors. For training phase, system was trained for different durations: 30s and 60s. Testing was done using 10s test frames. For first set of experiments, EM algorithm is used for training the model. In the next set of experiments, we have used Vector Quantization for calculating the model parameters. After final centroids splitting, it is checked that the cluster for every centroid has weight of 1/M where M represents number of mixtures. If the cluster has weight less than 1/M, then the centroids are splitted again. For shorter training data, selection of M is important. The covariance is calculated based on the data for each cluster. Identification accuracy is calculated for 10s testing data using training model parameters obtained from above steps. The following table shows comparison for both approaches considering the accuracy and time required for model parameter calculations.

Table.1: Identification Accuracy and running time for both approaches (Training with 30s and testing with 10s)

M	EM-GMM		VQ-GMM	
	Accuracy	time	Accuracy	time
8	80	3.9122	84	1.7246
16	85	6.4272	88	2.1770
32	86	11.8212	88	2.6905
64	88	22.6620	89	3.9328
128	88	42.2443	88	6.7476

The DET curve is also plotted for the above set of experiments as shown in Figure 2. The number of mixtures used is 128. Training duration is of 60sec and test duration isof 10sec. The curves show that, a slightly better performance can is obtained using VQ-GMM approach. The next set of experiments consisted of plotting the effect of model size on speaker identification performance using VQ-GMM approach. The training duration used was 60sec and testing duration was 10sec. The curves in Figure 3 show that as we increase the number of mixtures, the performance of the system increases.

Table.2: Identification Accuracy and running time for both approaches (Training with 60s and testing with 10s)

M	EM-GMM		VQ-GMM	
	Accuracy	time	Accuracy	time
8	83	7.3245	86	3.2199
16	92	13.2304	91	4.2370
32	91	24.7670	94	6.3149
64	92	42.6920	94	8.9794
128	92	82.1192	95	18.6309
256	92	157.7441	92	31.5249

5. Conclusion

This paper has addressed the implementation of GMM based speaker identification. We have implemented two approaches for training the speaker model. It has been shown that there is a slight improvement in identification accuracy using VQ based model parameter calculation. Considerable improvement is observed in computational time. A speed-up factor of 7 was achieved in first set of experiments: 128 mixtures and training duration of 30 sec, as shown in table.1, while in the second set of experiments a speed-up factor 5 was achieved which contained 256 mixtures and training data of 60sec as shown in table.2.

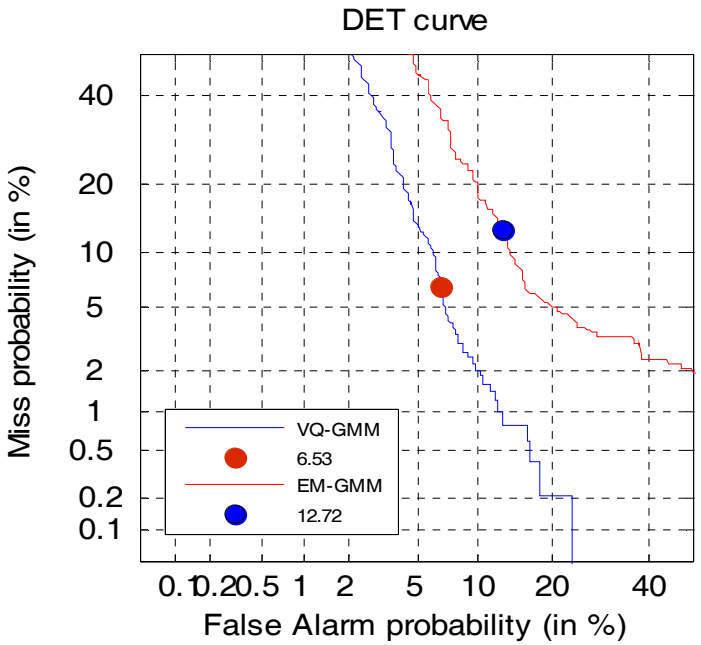

Fig.2:comparison of VQ-GMM and EM-GMM approach for 128 mixtures

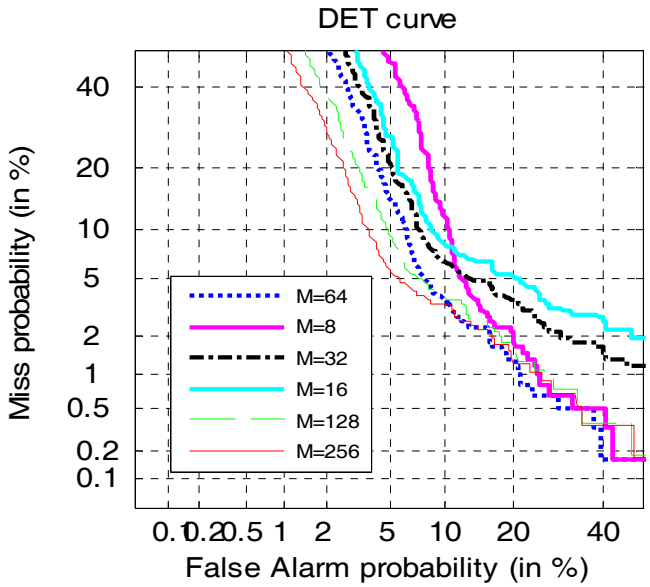

Fig. 3: Effect of model size on speaker identification using VQ-GMM approach

6. References

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

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

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

[5] F. Soong et al., "A vector quantization approach to speaker recognition," in *Proc. IEEE ICASSP*, 1985, pp. 387-390.

[6] D. A . Reynolds, T . F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing*, vol. 10, p19-41, Jan. 2000.