Information Theoretic Expectation Maximization based Gaussian Mixture Modeling for Speaker Verification

Sheeraz Memon, Margaret Lech, Namunu Maddage
School of Electrical & Computer Engineering, RMIT University, Melbourne, Australia
{sheeraz.memon,margaret.lech,namunu.maddage}@rmit.edu.au

Abstract

The expectation maximization (EM) algorithm is widely used in the Gaussian mixture model (GMM) as the state-of-art statistical modeling technique. Like the classical EM method, the proposed EM-Information Theoretic algorithm (EM-IT) adapts means, covariances and weights, however this process is not conducted directly on feature vectors but on a smaller set of centroids derived by the information theoretic procedure, which simultaneously minimizes the divergence between the Parzen estimates of the feature vector’s distribution within a given Gaussian component and the centroids distribution within the same Gaussian component. The EM-IT algorithm was applied to the speaker verification problem using NIST 2004 speech corpus and the MFCC with dynamic features. The results showed an improvement of the equal error rate (ERR) by 1.5% over the classical EM approach. The EM-IT also showed higher convergence rates compare to the EM method.

1. Introduction

In this paper the performance of the speaker recognition system using GMM based on information theoretic vector quantization (ITVQ) criteria [4] is evaluated. The GMM method [2] based on expectation maximization (EM) is commonly regarded as the state-of-the-art modeling and classification technique successfully applied in many pattern recognition problems including speech and speaker recognition, stress and emotion classification, face recognition, and many others [3,6].

In a variety of practical applications, the distribution of the parameters can be approximated by a family of finite mixture densities where, the density function is a weighted sum of component densities. The component densities are commonly modeled as Gaussians. It can be shown that any continuous probability density function can be approximated arbitrarily closely by a Gaussian mixture density [5]. In its classical form [1,2], the GMM applies the EM algorithm, which iteratively updates the means, covariances and weights for each class, and converges to a set of vectors, providing the maximum value of the likelihood function. Each set consisting of means, variances and weights constitutes a class model. The resulting models provide multivariate probability density functions for each class with the highest likelihood values for given training data.

Vector quantization and GMM methods were combined based on the fact that both methods represent the distribution of the data vectors in feature space in [7]. Study in [8] suggests that the learning rules used by a number of vector quantization methods are variants to EM algorithm. It is reported in the literature that EM appears to converge with local maxima [14]. The study in [10] suggests that it is sufficient to adapt the means only for the optimization of GMM parameters. The study in [9] proposes a vector quantization method based on the GMM, a method for optimizing GMM model parameters and an extension to the EM algorithm is suggested in this study. As we have evaluated in [17] that information theoretic (IT) based VQ has better performance than k-means and LBG. We then investigated the performance of speaker verification with different VQ methods and GMM described in [15] and [16]. In this paper we use the information theoretic metric nested within EM method and introduce EM-IT. Our proposed method shows better convergence rates and therefore the improved speaker recognition scores.
2. Gaussian mixture model

Given the original set of R-dimentional feature vectors \( X_i = \{x_{i0}, x_{i1}, \ldots, x_{iS}\} \) for \( N \) classes (\( k=1, \ldots, N \)), the Gaussian mixture model (GMM) iteratively develops different multivariate Gaussian probability density functions for each class. Given \( M \) mixtures within each class, the Gaussian pdf of a feature vector \( x \) for the \( i^{th} \) mixture within class \( k \), is given as:

\[
p_i^k(x) = \frac{1}{\sqrt{(2\pi)^R |\Sigma_i^k|}} \exp\left(-\frac{1}{2}(x - \mu_i^k)^T(\Sigma_i^k)^{-1}(x - \mu_i^k)\right)
\]

Where \( i=1,2,\ldots,M \), \( \mu_j \) is the mixture mean vector, \( \Sigma_i^k \) is the mixture covariance matrix and \( R \) is the dimension of the feature vectors. The set of pdfs, means and covariances for all mixtures within a given class constitutes a class model \( \lambda_k = \{w_i^k, \mu_i^k, \Sigma_i^k\} \). The probability that a feature vector \( x \) represented by a particular model \( \lambda_k \) belongs to any of the \( M \) mixtures representing \( k^{th} \) class is a weighted mixture of \( M \) Gaussian pdfs:

\[
p(x) = \sum_{i=1}^{M} w_i^k p_i^k(x)
\]

Where \( w_i^k \) are the mixture weights for class \( k \).

The EM algorithm [1] iteratively improves the estimates of \( \lambda_k \), and maximizes log-likelihood monotonically. The GMM/EM process can be viewed as a form of “soft” clustering (Fig.1 (a)), where, feature vectors representing each class \( k \) are divided into \( M \) mixtures (or clusters) with each mixture represented by a probabilistic model \( \lambda_k \).

3. Expectation maximization using information theoretic criteria

Aiming to improve the convergence rate of the classical EM algorithm, a new approach to the mean adaptation was introduced. It combines the classical EM technique with the vector quantization method based on the information theoretic learning and is referred to as the EM-IT method. In the EM-IT, the clustering process of the EM method (Fig.1(a)) is enhanced by data reduction achieved through the ITVQ clustering (Fig.1(b)). The algorithm convergence properties are reinforced by both maintaining the improvement of log-likelihood process using the EM algorithm, and iterative improvement of centroids calculation guided by the information theoretic criteria which simultaneously minimize the divergence measure between each vector within a given cluster and the centroid of this cluster, and maximize the divergence between centroids of neighboring clusters.

3.1. The EM-IT algorithm

The EM-IT algorithm proceeds in the following steps:

Step 1. Initialization. In this step an initial set of \( C \) centroids \( C_{init} = \{c_{0}, c_{1}, \ldots, c_{C}\} \) is generated for each class \( k \) using a relatively simple unsupervised clustering method such as the k-means algorithm. The centroids are then used to derive the initial models \( \lambda_{init}^k \).

Step 2. Checking the stopping criteria. The algorithm proceeds until an arbitrary number of iterations are reached or the increase of the expectation value over a number of consecutive iterations is less than an arbitrary threshold \( \zeta \).

Step 3. ITVQ. During this step, the ITVQ algorithm iteratively improves the centroids with respect to the information theoretic (IT) criteria producing a new set of centroids \( C_{iter} = \{c_{0}^{iter}, c_{1}^{iter}, \ldots, c_{C}^{iter}\} \) for each Gaussian component.

Step 4. Updating the model parameters. In this step the model’s parameters are updated using:

\[
\begin{align*}
\{w_i^k\}_{iter} &= \frac{1}{C} \sum_{n=1}^{C} p(i_n = k | c_n, \lambda_{iter}) \\
\{\mu_j\}_{iter} &= \frac{1}{C} \sum_{n=1}^{C} p(i_n = k | c_n, \lambda_{iter}) \mu_{i,iter}^k \\
\{\Sigma_i^k\}_{iter} &= \frac{1}{C} \sum_{n=1}^{C} p(i_n = k | c_n, \lambda_{iter}) \Sigma_{i,iter}^k
\end{align*}
\]

A replacement of the model components is then made to obtain the next set of estimates \( \lambda_{iter}^k \), and the procedure is repeated by returning to Step 2.

In general, the ITVQ process can be viewed as a “sharp” clustering (Fig.1(b)), where each Gaussian component is divided into a number of clusters and each cluster is represented by \( C \) centroid vectors. Thus, on each iteration of the new algorithm, the large number of \( S \) original feature vectors in each class is replaced by a smaller number of \( C \) (\( C < S \)) representative centroid vectors and the EM procedure is carried out on centroids instead of original feature vectors. The centroids are iteratively refined using a
number of ITVQ updates nested in EM. Thus, the EM-IT has a dynamic character as it applies the updating formulas of Eq.(3-5) not to a constant set of feature vectors but to a gradually more and more refined configurations of centroids which change at each iteration.

\[ \hat{g}(x) = (1/C) \sum_{j=1}^{C} \exp(-\frac{1}{2}(\|x - c_j\|^2 / \sigma_j^2)) \]  

(7)

Where \( \sigma_j^2 \) and \( \sigma^2 \) are the kernel variances. The cost function is the divergence \( J(c) \) between these two distributions given by the Cauchy-Schwarz formula:

\[ J(c) = \log \int f^2(x)dx - 2\log \int f(x)\hat{g}(x)dx + \log \int \hat{g}^2(x)dx \]  

(8)

The cost function is minimized by calculating the derivatives of \( J(c) \) with respect to the centroids \( c_j \), which leads to the following centroid updating formula:

\[ c_j (n+1) = c_j (n) - \eta (\Delta V / V) - 2(\Delta D / D) \]  

(9)

Where \( i=1,...,C \), \( n \) is the iteration index, \( \eta \) is a constant step size, and the vectors \( D \) and \( V \) are given as:

\[ D = \int J \frac{1}{2} (x) \hat{g}(x)dx, \quad V = \int \frac{1}{2} (x) \hat{g}(x)dx \]  

(10)

The terms \( \Delta D \) and \( \Delta V \) are the vectors of derivatives of \( D \) and \( V \) respectively, calculated with respect to the centroids \( c_i \). The \( \eta \) value of 0.03 provided satisfactory results when applied to the speaker verification problem.

4. Experiments

4.1. Proposed speaker recognition system

The speaker verification system that we used for the experiment to examine the effect on the performances with NIST 2004 SRE corpus, is shown in Fig. 2. The system operates in one of the three possible ways, such as universal background model (UBM) training, target speaker enrollment and testing. In each case identical speech detection and feature extraction method is used. In the speech detection block in Fig.2, an energy based silence detector which identifies the low energy portions of the signal [11] as silence regions is used. Previous research [13] have shown that MFCC based system is relatively robust to the changes in frame size (in the range 20-50 ms) and frame step (in the range 1/6 to 1/3 of the frame size). Thus we employed MFCC to characterize the speaker information. Twelve MFCC with delta, double delta, energy and zero crossings form a 38 dimensional feature vector. The sequences of feature vectors are then modeled with GMM; the most frequently and successfully employed density estimators in speaker recognition. We use 1024 Gaussians for every target speaker’s GMM; each model is parameterized by its \( a \) priori probability, mean vector, and diagonal covariance matrix. For training and testing the models, we used around 5 minute’s
utterances. After the enrollment, universal background model (UBM) parameter inference [12] is accomplished via the EM and EM-IT algorithms using a large corpus of speech from non-target speakers obtained from NIST 2001 and 2002 SRE. Target speaker model means are then adapted away from the UBM via maximum a posteriori (MAP) estimation, using only the target speaker’s speech. Testing proceeds by applying the same feature processing as for model training. The observed sequence of feature vectors is then scored by each speaker’s model. Performance is assessed using equal error rate (EER) measure and by plotting detection error trade-off (DET) curve.

Fig. 2. GMM based Speaker Verification System

4.2. Comparison of the training algorithms convergence rates

The EM-IT algorithm introduced additional complexity to the modeling process by including an ITVQ update procedure. As demonstrated in our experiments (Fig. 3) this additional complexity can be reduced to as little as 15 ITVQ updates while still maintaining significantly higher convergence rate of the EM-IT compared to the EM approach. Fig. 3 shows that while increasing the number of ITVQ updates from 3 to 15, the convergence rate of EM-ITVQ can be improved on average by 32% compared to the EM algorithm. An increase from 15 to 50 ITVQ updates provides further improvement of the average convergence rates by 13.5%. The % rate is computed by evaluating the change in likelihood with respect to iterative updates. In our approach the likelihood is evaluated based on the smaller set of centroids data which is derived from the entire cluster data used by EM updates. The optimization procedure of IT updates chooses the best representation of centroids data representing the

speaker model which ultimately leads to better likelihood.

4.3. Comparison of the speaker verification rates

Fig. 4 illustrates the percentage miss probability versus the percentage of false alarm probability for both the EM algorithm and the EM-IT algorithms. The miss probability was calculated as the probability that the system incorrectly declares a successful match between the input features and a non-matching model in the database. The false alarm probability was calculated as the probability that the system incorrectly declares failure of match between the input features and the matching model. The EER parameter represents the rate at which both the miss probability and the false alarm probability are equal. The lower the EER, the more accurate the system is considered to be. EM and EM-IT were tested using feature vectors of length R=38. EM-IT shows an improvement by 1.5% of the average EER value over the classical EM algorithm.

Fig. 3. Convergence rates for the EM and EM-ITVQ algorithms using the NIST 2004 data

5. Conclusion

A new approach to the GMM training of speaker models has been described. It has been empirically demonstrated that when applied to the speaker verification task, the EM-IT modeling algorithm achieves higher convergence rates and provides smaller EER values compared to the classical EM algorithm. Unlike the EM, which relies only on maximization of
the expectation function, the EM-ITVQ is guided by additional objectives helping to minimize a divergence measure between the distribution of the original feature vectors and the distribution of the centroids. In future we aim to investigate EM with kd-tree data structure and fast Gaussian Transform for speaker verification. The investigation of our proposed method with the recent NIST SRE data still needs to be investigated.

![Graph of Miss probability versus false alarm for EM and EM-ITVQ using NIST 2004.](image)

**Fig. 4.** Miss probability versus false alarm for EM and EM-ITVQ using NIST 2004.

### 6. References


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