ON THE USE OF ORTHOGONAL GMM IN SPEAKER RECOGNITION

Li Liu and Jialong He

Dept. of Speech and Hearing Science
Arizona State University
Tempe, AZ 85287-1908, USA

ABSTRACT

The Gaussian mixture modeling (GMM) techniques are increasingly being used for both speaker identification and verification. Most of these models assume diagonal covariance matrices. Although empirically any distribution can be approximated with a diagonal GMM, a large number of mixture components are usually needed to obtain a good approximation. A consequence of using a large GMM is that its training is time consuming and its response speed is very slow. This paper proposes a modification to the standard diagonal GMM approach. The proposed scheme includes an orthogonal transformation: feature vectors are first transformed to the space spanned by the eigenvectors of the covariance matrix before applying to the diagonal GMM. Only a small computational load is introduced by this transformation, but results from both speaker identification and verification experiments indicated that the orthogonal transformation considerably improves the recognition performance. For a specific performance level, the GMM with orthogonal transform needs only one-fourth the number of Gaussian functions required by the standard GMM.

1. INTRODUCTION

The Gaussian mixture speaker model has been successfully used for both speaker identification and verification [1][2]. The basis of this approach is to represent the distribution of training vectors from each speaker with a weighted sum of several multivariate Gaussian functions. The parameters in the model can be estimated using the iterative Expectation-Maximization (EM) algorithm.

In theory each Gaussian function may have a full covariance matrix. However, the diagonal covariance matrix has been used almost exclusively in the practice. Since the inverse of covariance matrices have to be calculated repeatedly during the EM iteration, using the diagonal matrices are clearly more advantageous from the calculation perspective. In addition, if available training data are very limited, a full covariance matrix is more likely to be ill-conditioned. In many practical situations, even the diagonal elements (i.e., variances) could become too small in magnitude. This is particularly true when a mixture model with a large number of component densities is used. To solve this problem, a floor value (minimum limit) is usually set to the variances [1].

Generally, the elements of feature vectors extracted from a speech signal are correlated. Even though a linear combination of diagonal covariance Gaussian functions is capable of modeling the correlation [1], a large number of mixtures has to be used in order to provide a good approximation. A large GMM takes a long time to train, and is very slow in response. This paper proposes a modification to the commonly used diagonal GMM (called standard GMM). The basic idea is that before applying to the diagonal GMM the feature vectors are first transformed to the space spanned by the eigenvectors of the covariance matrix so that the correlation among the elements is reduced.

2. ORTHOGONAL GAUSSIAN MIXTURE MODEL

2.1 Description

A Gaussian mixture model is a weighted sum of *M* multivariate Gaussian functions

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} c_i b_i(\vec{x}) \qquad (1)$$

where $b_i(\vec{x})$ has the form

$$b_i(\vec{x}) = \frac{\exp\left\{-(\vec{x}-\vec{\mu}_{xi})^T \Sigma_{xi}^{-1}(\vec{x}-\vec{\mu}_{xi})/2\right\}}{(2\pi)^{D/2}\left|\Sigma_{xi}\right|^{1/2}} \qquad (2)$$

In most situations, all covariance matrices Σ_{xi} are assumed to be diagonal. From the linear algebra theory, we know that a covariance matrix can be diagonalized if the vectors are linearly transformed to the space spanned by the eigenvectors of the original covariance matrix.

Suppose the covariance matrix of the current speaker is Σ_x and the transform matrix Ω is composed of the eigenvectors of Σ_x, then after the linear transformation,

$y = \Omega^T x$, the covariance matrix in the $\boldsymbol{y}$ space, Σ_y, is diagonal. Σ_y is related to Σ_x according to the following equation

$$\Sigma_y = \Omega^T \Sigma_x \Omega \tag{3}$$

Since Ω^T is composed of the eigenvectors of Σ_x, it has the property that $\Omega\Omega^T = I$. Replacing x in Eq. (2) with Ωy, we have,

$$b_i(\vec{y}) = \frac{\exp\left\{-\left(\vec{y}-\vec{\mu}_{yi}\right)^T \left(\Sigma_{yi}\right)^{-1}\left(\vec{y}-\vec{\mu}_{yi}\right)/2\right\}}{(2\pi)^{D/2}\left|\Sigma_{yi}\right|^{1/2}} \tag{4}$$

where Σ_{yi} and μ_{yi} are defined as,

$$\begin{aligned} \Sigma_{yi} &= \Omega^T \Sigma_{xi} \Omega \\ \mu_{yi} &= \Omega^T \mu_{xi} \end{aligned} \tag{5}$$

Comparing Eq. (3) and (5), it is easy to see that if there is only one Gaussian function (uni-modal Gaussian model), the diagonal Gaussian function in $\boldsymbol{y}$ space is equivalence to the Gaussian function with full covariance in the $\boldsymbol{x}$ space. The uni-modal Gaussian function with a full covariance has also been shown to be an effective speaker model [3][4]. Generally, the GMM has more than one Gaussian components, therefore, the covariance matrices Σ_{yi} are not truly diagonal. However, it is more reasonable to assume that Σ_{yi} are diagonal dominated than Σ_{xi}. In other words, a diagonal GMM in $\boldsymbol{y}$ space would provide a better approximation to the distribution of feature vectors. We name the new GMM with orthogonal transform as orthogonal GMM (OGMM).

Figure 1 shows a block diagram of OGMM. The model is composed of a linear transform matrix and a normal diagonal GMM. Please note, the transform matrix is speaker dependent, each model has its own transform matrix. In fact, this model has exactly the same structure as that proposed by Yuo *et al.* [5], but the method for generating the transform matrix is difference. Yuo *et al.* estimated the transform matrix and the diagonal GMM parameters jointly in the EM iteration process. In contrast, the training in our method is done in two steps. The first step is the calculation of the orthogonal transform matrix. Since the covariance matrix is a real symmetric matrix, there are well developed and efficient methods to find its eigenvectors.

After obtaining the transform matrix, the second step is to multiply all training vectors belonging to this model with the transform matrix. This transformation is done only once during the training phase. Then the parameters in the diagonal GMM are estimated with the commonly used EM algorithm. In the test phase, the test vectors are also transformed to the new space before they are applied to the diagonal GMM.

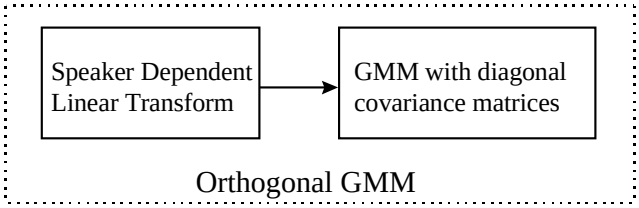

Figure 1 A block diagram of the OGMM.

2.2 Illustration in Two-dimensional Space

In this section, we give a conceptual interpretation to the orthogonal transform in the two dimensional space. Figure 2(a) shows the case of representing the data with a diagonal GMM. Each ellipse stands for a Gaussian component. Since diagonal covariance matrices are used, the axes of all ellipses are parallel to the x- and y-axis. However, the principle components of the feature vectors may not be along these directions. If we first turn the coordinate system according to the distribution direction of the principle components, it is more likely that a better approximation to the data distribution can be obtained with the same number of Gaussian functions.

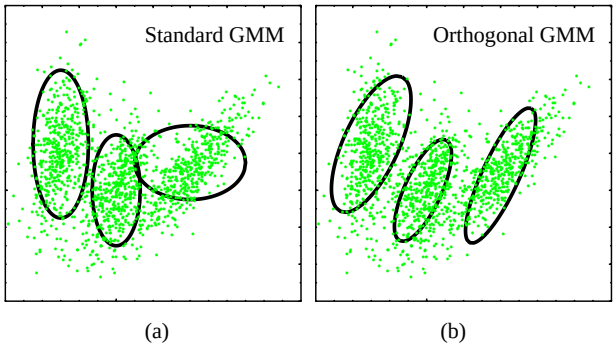

Figure 2 A conceptual illustration of (a) the standard GMM, and (b) the OGMM.

3. EXPERIMENTS

3.1 Evaluation Speech

We did both speaker identification and verification experiments to show the effectiveness of the orthogonal transform. The speech data were taken from the YOHO database [6]. A subset of the database including 40 speakers (20 males, 20 females) were used in both

experiments. For the verification experiment, another 20 speakers (10 males, 10 females) from the same database were selected as impostors. No model was created for the impostors, only their test data were used.

Feature vectors were extracted with the popular short-time analysis techniques. The analysis window size was 32 ms (256 samples) and the advancing step was 16 ms. Sentences from all enrollment sessions were used for training. However, silence and unvoiced segments were discarded based on an energy threshold. Discarding unvoiced segments will inevitably degrade the overall performance, but this is not a serious problem because we only want to compare the relative performance of different models. There were about 4500 training vectors from each speaker. A feature vector was composed of 16 MFCC coefficients. In the identification experiment, individual phrases from the verification sessions were used as test sentences. The average length of the test sentences was about 48 frames (48×16 ms, after removing silence and unvoiced frames). In the verification experiment, we tested both cases with a single string as a test sentence and with four strings as a test sentence.

3.2 Computational Efficiency

Suppose the feature vector dimension is D, and the number of mixtures in the GMM is M. Then each standard diagonal GMM will have $2D \times M + M$ parameters. An OGMM needs extra $D \times D$ storage for the transform matrix.

Although the OGMM needs extra steps of obtaining the transform matrix, and then multiplying the training vectors with the transform matrix, the computational costs of all these steps are negligible with a modern computer. In fact, the most time-consuming part during the training phase is the EM iteration process. Since the OGMM usually converges more quickly than the standard GMM, we found that the OGMM does not need more training time than the GMM.

In the test phase, the OGMM involves a linear transform for all test vectors. However, this small calculation overhead has no noticeable influence on the response speed. Totally speaking, for the same number of mixtures, there is no clear difference between OGMM and standard GMM in terms of both computational efficiency and response speed.

Throughout the discussions above, we have assumed that the same number of Gaussian components are used in the standard GMM and in the OGMM. As we will see in the next two subsections, the OGMM can give the same performance level with much less mixture components, usually only one-fourth of that needed by the standard GMM. For example, if we use 64 Gaussian functions in the standard GMM, only 16 Gaussian components is needed for the OGMM to obtain the same performance. Suppose the vector dimension is 16, then the standard GMM will have 2112 parameters, while the OGMM only has 784. It is known that the training and test time increases rapidly with the number of mixtures. Therefore, to achieve a specific performance level, OGMM is much faster than the standard GMM.

3.3 Identification Performance

Figure 3 shows the relationship between the speaker identification rate and the number of mixtures. The dotted line is obtained from the standard GMM and the solid line is from the OGMM. In this experiment, single phrase was used as a test sentence, thus each point in the figure was obtained from 1600 trials (40 speakers × 40 test sentences). From the figure we see that if the same number of mixtures is used, the OGMM always gives a higher identification rate. A more careful examination here also reveals that to reach the same performance level, standard GMM needs about four times the number of mixtures used by the OGMM. As we said above, a OGMM with 16 components needs less storage and is faster than the standard diagonal GMM with 64 mixtures.

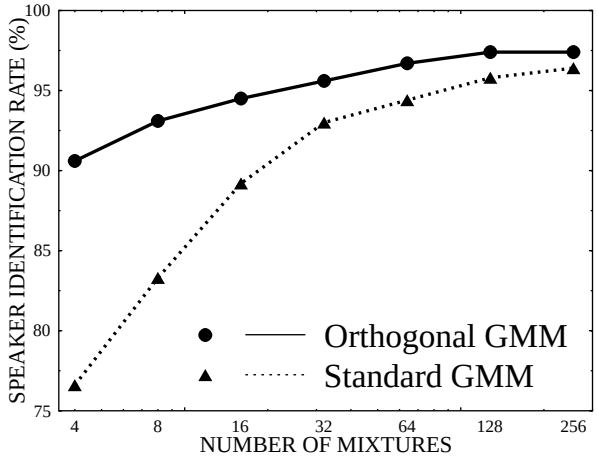

Figure 3 Speaker identification performance with the standard GMM and the OGMM.

3.4 Verification Performance

In the verification experiment, we evaluated the error rates with two different test lengths. The reason we did this is because we have found in another study that in order to characterize the relationship between the verification performance and the length of test sentences, evaluation results at two different lengths are necessary. Besides, we want to compare our results with that obtained by Renolds [2]. In his paper, he had concatenated four strings together as one test sentence.

By using one string as a test sentence, there were 40 sentences from each customer and 800 sentences from 20 impostors. Since the total number of speaker models was 40 and each impostor's sentence was applied to all 40 models, there were 40×40 customer scores and 40×800 impostor scores. In the case of using 4 strings as a test sentence, each verification performance was obtained from 400 customer scores and 8000 impostor scores. The decision threshold value was post-determined after obtaining all customer and impostor scores. We tested two threshold values. One was selected so that the false rejection (*FR*) rate equals to the false acceptance (*FA*) rate. Another case is that *FA* rate was pre-selected as 0.1%, and the *FR* had to be determined.

	Standard GMM		Orthogonal GMM	
Mixtures / String length	EER (%)	FR (%) @ FA=0.1 %	EER (%)	FR (%) @ FA=0.1%
16 / 1	6.2	43.1	4.1	21.4
64 / 1	4.5	22.9	2.8	14.2
16 / 4	5.0	15.0	2.5	8.0
64 / 4	2.6	7.5	1.1	3.7

Table 1 Verification error rates with the standard GMM and the OGMM.

The verification results are summarized in Table 1. As expected, the verification performance improves with the number of mixtures and the length of test sentences. Let's first look at the results with short test sentences (String length = 1, 1st & 2nd rows in Table 1). As in the speaker identification experiment, for the same number of mixtures, OGMM always performs better than the standard GMM. We see that the performance of the OGMM with 16 mixtures is comparable with that of the GMM with 64 mixtures. Similar conclusions can be made with longer test sentences (String length = 4, 3rd & 4th rows in Table 1). Remember that the OGMM with 16 mixtures needs less memory space and is more faster than the standard GMM with 64 mixtures.

Now we want to compare our results with that of Renolds [2], who used the standard diagonal GMM in his experiment. Even though both studies are based on the GMM and the same YOHO database is used, the verification error obtained in our experiment is slightly higher. We think the main cause of this discrepancy is the different feature vectors used in the two systems. In our experiment, each feature vector was composed of 16 MFCC coefficients, while Renolds used a higher dimension of feature vectors. In general, performance improves with the dimension of feature vectors. Another reason might be that we simply took all speakers of the same gender as the background speakers, while in Renolds' experiment, he made a careful selection for the reference speakers.

4. SUMMARY

In most systems based on the Gaussian mixture speaker models, diagonal covariance matrices are used. To provide a better approximation to the distribution, a large number mixture components has to be used. Here we have proposed a modification to the standard diagonal GMM. In the new model (named as OGMM), there is an orthogonal transform matrix. Feature vectors are first transformed to the space spanned by the eigenvectors of the covariance matrix before applying to the diagonal GMM. It is shown that with the same number of mixtures, the OGMM always gives a better performance. To reach a specific performance level, the OGMM needs only one-fourth the number of mixtures used by the standard GMM, therefore, the OGMM is more faster and needs less storage.

5. REFERENCES

[1] Reynolds D., Rose R. C., "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, Vol. 3, pp. 72-83, Jan. 1995.

[2] Reynolds D., "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, Vol. 17, pp. 91-108, 1995.

[3] Bimbot F., Chagnolleau I., Mathan L., "Second-order statistical measures for text-independent speaker identification," *Speech Communication*, Vol. 17, pp. 177-192, 1995.

[4] Gish H. Schmidt M. "Text-independent speaker identification," *IEEE Signal Processing Magazine*, pp. 18-32, Oct. 1994.

[5] Yuo K. And Wang H., "Gaussian mixture models with common principal axes and their application in text-independent speaker identification," *Proceedings of Eurospeech Conference*, pp. 2279-2282, Rhodes, Greece, 1997.

[6] Godfrey J., Graff D. and Martin A., "Public databases for speaker recognition and verification," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 39-42. Martigny, Switzerland, 1994.