GAUSSIAN MIXTURE MODELS WITH COMMON PRINCIPAL AXES AND THEIR APPLICATION IN TEXT-INDEPENDENT SPEAKER IDENTIFICATION

Kuo-Hwei Yuo and Hsiao-Chuan Wang
Department of Electrical Engineering, National Tsing Hua University,
Hsinchu, Taiwan 30043
E-mail: hcwang@ee.nthu.edu.tw

ABSTRACT

Gaussian mixture models (GMM's) have been demonstrated as one of the powerful statistical methods for speaker identification. In GMM method, the covariance matrix is usually assumed to be diagonal. That means the feature components are relatively uncorrelated. This assumption may not be correct. This paper concentrates on finding an orthogonal speaker-dependent transformation to reduce the correlation between feature components. This transformation is based on the eigenvectors of the *within-class scatter matrix* which is attained in each stage of iterative training of GMM parameters. Hence the transformation matrix and GMM parameters are both updated in each iteration until the total log-likelihood converges. An experimental evaluation of the proposed method is conducted on a 100-person connected digit database for text independent speaker identification. The experimental result shows a reduction in the error rate by 42% when 7-digit utterances are used for testing.

I. INTRODUCTION

Gaussian Mixture Model (GMM) has been successfully applied to speaker identification [1]. In GMM method, diagonal covariance matrices are commonly used. However, the original feature components are not uncorrelated. This assumption of completely ignoring feature correlation may not be correct [2] [3] and will degrade the identification accuracy. Hence full convariance matrices have to be used in order to count the fact of the correlations between feature components. Unfortunately, such models are computationally expensive and require large amount of training data. An alternative way is to use orthogonal transformation to make the transformed feature components less correlated.

Here we propose a new transformation which is incorporated into model training. This means the feature transformation is jointly optimized with the diagonal GMM parameters.

II. GAUSSIAN MIXTURE SPEAKER MODELS

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

$$p(\bar{x}|\lambda)=\sum_{i=1}^{M}\omega_i b_i(\bar{x}), \tag{1}$$

where $\bar{x}$ is a D-dimensional feature vector, $b_i(\bar{x})$,*i =1,...,M,* are the component densities and ω_i, *i =1,...,M,* are the mixture weights. Each component density is a D-variate Gaussian function of the form

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

with mean vector $\bar{\mu}_i$ and covariance matrix Σ_i. The mixture weights satisfy the constraint that $\sum_{i=1}^{M}\omega_i=1$. The speaker model is then represented by the mean vectors, covariance matrices and mixture weights. These parameters are denoted by a set

$$\lambda=\{\omega_i,\bar{\mu}_i,\Sigma_i\}\ i=1,\dots,M. \tag{3}$$

Given a sequence of *T* training vectors $X=\{\bar{x}_1,\dots,\bar{x}_T\}$, the most popular method to derive the parameters λ is based on maximum likelihood (ML) estimation in which we begin with an initial model λ, and estimate a new model $\bar{\lambda}$, such that $L(X|\bar{\lambda})\geq L(X|\lambda)$ where

$$L(X|\lambda)=\sum_{t=1}^{T}\log p(\bar{x}_t|\lambda) \tag{4}$$

The reestimation formulas are as follows:

$$\overline{\omega}_i = \frac{1}{T}\sum_{t=1}^{T} p(i|\vec{x}_t,\lambda) \tag{5}$$

$$\overline{\vec{\mu}}_i = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t,\lambda)\vec{x}_t}{\sum_{t=1}^{T} p(i|\vec{x}_t,\lambda)} \tag{6}$$

$$\overline{\Sigma}_i = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t,\lambda)(\vec{x}_t-\vec{\mu}_i)(\vec{x}_t-\vec{\mu}_i)'}{\sum_{t=1}^{T} p(i|\vec{x}_t,\lambda)} \tag{7}$$

where the posteriori probability of the *i*th mixture is given by

$$p(i|\vec{x}_t,\lambda) = \frac{w_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} w_k b_k(\vec{x}_t)} \tag{8}$$

When the covariance matrix Σ_i is assumed to be diagonal, the equation (6) can be simplified to

$$\overline{\rho}_{i,j}^2 = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t,\lambda)x_{t,j}^2}{\sum_{t=1}^{T} p(i|\vec{x}_t,\lambda)} - \mu_{i,j}^2, \text{ for } 1 \le j \le D \tag{9}$$

where $x_{t,j}$ and $\mu_{i,j}$ are the *j*th element of the vector $\vec{x}_t$ and μ_i, respectively, and $\rho_{i,j}$ is the *j*th diagonal element of the diagonal matrix $\overline{\Sigma}_i$.

III. TRAINING OF TRANSFORMATIONS AND GMM PARAMETERS

In discriminant analysis of statistics [4], when samples are partitioned into more than one class, some measure criteria are used to formulate the separability among classes. *Within class scatter matrix*, which shows the scatter of samples around their respective expected vectors, is such a criterion and is expressed by

$$S_w = \sum_{i=1}^{M} \omega_i \Sigma_i \tag{14}$$

where M is the number of classes, ω_i and Σ_i are the prior probability and the covariance matrix of *i*th class, respectively. Instead of treating all training vectors as a single class, the transformation and the estimation will be performed alternatively for all classes.

When we estimate the GMM parameters with equations (5) through (8), we can define a within class scatter matrix according to equation (14) and express it as

$$\begin{aligned} S_w &= \sum_{i=1}^{M} \overline{\omega}_i \overline{\Sigma}_i \\ &= \sum_{i=1}^{M} \left(\frac{1}{T}\sum_{t=1}^{T} p(i|x_t,\lambda)\right) \bullet \\ &\quad \frac{\sum_{t=1}^{T} p(i|\vec{x}_t,\lambda)(\vec{x}_t-\vec{\mu}_i)(\vec{x}_t-\vec{\mu}_i)'}{\sum_{t=1}^{T} p(i|\vec{x}_t,\lambda)} \\ &= \frac{1}{T}\sum_{i=1}^{M}\sum_{t=1}^{T} p(i|\vec{x}_t,\lambda)(\vec{x}_t-\vec{\mu}_i)(\vec{x}_t-\vec{\mu}_i)' \end{aligned} \tag{15}$$

The transformation is an eigenvector matrix of this *within class scatter matrix* which is attained in each training iteration of GMM parameters. So, the transformation matrix and GMM parameters are both repeatly updated in each iteration until the total log-likelihood is convergent. The procedures are as follows:

1. Initialization

(a) Set initial model $\tilde{\lambda}^{(0)} = \{\tilde{\omega}_i^{(0)}, \tilde{\mu}_i^{(0)}, \tilde{\Delta}_i^{(0)}\}$ for $1 \le i \le M$. $\tilde{\Delta}_i^{(0)}$ is a diagonal covariance. This initial model is attained by training $X = \{\vec{x}_1, ..., \vec{x}_T\}$ with LBG algorithm.

(b) Define $x_t^{(0)} = x_t$ for $1 \le t \le T$. This vector set is denoted as $X^{(0)} = \{x_1^{(0)}, ..., x_T^{(0)}\}$

(c) Set the initial transformation ($\Omega^{(0)}$) be a $D \times D$ unit matrix where D is the dimension of feature vector

(d) Define $\Omega = (\Omega^{(0)})$ and $n = 1$. n indicates the iteration index.

2. Recursion

(a) Transform feature vector

$X^{(n-1)} = \{x_1^{(n-1)}, ..., x_T^{(n-1)}\}$ to $X^{(n)} = \{x_1^{(n)}, ..., x_T^{(n)}\}$ by the equation

$$x_t^{(n)} = (\Omega^{(n-1)})' x_t^{(n-1)} \text{ for } 1 \le t \le T . \tag{16}$$

$\tilde{\lambda}^{(n-1)} = \{\tilde{\omega}_i^{(n-1)}, \tilde{\mu}_i^{(n-1)}, \tilde{\Delta}_i^{(n-1)}\}$ is the GMM of $X^{(n-1)}$. Then we obtain a transformed model,

$\lambda^{(n)} = \{\omega_i^{(n)}, \mu_i^{(n)}, \Delta_i^{(n)}\}$.

$$\mu_i^{(n)} = (\Omega^{(n-1)})' \tilde{\mu}_i^{(n-1)} \text{ for } 1 \le i \le M \tag{17}$$

$$\Delta_i^{(n)} = (\Omega^{(n-1)})' \tilde{\Delta}_i^{(n-1)} (\Omega^{(n-1)}) \text{ for } 1 \le i \le M \qquad (18)$$

$$\omega_i^{(n)} = \tilde{\omega}_i^{(n-1)} \text{ for } 1 \le i \le M \qquad (19)$$

(b) Let $\lambda^{(n)}$ be the initial model and feature set $X^{(n)}$ be the training vectors. We can re-estimate the model parameters, $\tilde{\mu}_i^{(n)}, \tilde{\Delta}_i^{(n)}$, and $\tilde{\omega}_i^{(n)}$, by equations (4) to (6). In this iteration, we must compute the scatter matrix by

$$S_w^{(n)} = \frac{\sum_{i=1}^{M}\sum_{t=1}^{T} p\left(i \middle| x_t^{(n)}, \lambda^{(n)}\right)\left(x_t^{(n)} - \mu_i^{(n)}\right)\left(x_t^{(n)} - \mu_i^{(n)}\right)'}{T}, \qquad (20)$$

then we can decompose $S_u^{(n)}$ into the following equation

$$S_\omega^{(n)} = (\Omega^{(n)})\Lambda^{(n)}(\Omega^{(n)})' \qquad (21)$$

where $\Lambda^{(n)}$ is a diagonal matrix whose diagonal elements are eigenvalues of $S_u^{(n)}$, and the columns of the matrix $\Omega^{(n)}$ are eigenvectors.

(c) Compute the improvement ratio of the total log-likelihood by

$$\varepsilon = \frac{L(X^{(n)}|\tilde{\lambda}^{(n)}) - L(X^{(n-1)}|\tilde{\lambda}^{(n-1)})}{L(X^{(n-1)}|\tilde{\lambda}^{(n-1)})}, \qquad (22)$$

(d) Increment $n \rightarrow n+1$ and iterate (a) through (c), if the value ε is larger than a preset threshold.

3. Result

The final result of this training procedure is the transformation matrix

$$\Omega = \prod_{k=1}^{n-1} \Omega^{(k)}, \qquad (23)$$

and the GMM model is $\lambda^{(n)} = \{\omega_i^{(n)}, \mu_i^{(n)}, \Delta_i^{(n)}\}$

IV. SPEAKER IDENTIFICATION

For speaker identification, a group of S speakers is represented by a set of speaker models $\{\Gamma_1, \Gamma_2, ..\Gamma_s..\Gamma_S\}$. Each Γ_s is composed of a speaker dependent transformation matrix Ω_s and a GMM parameter set $\lambda_s = \{\omega_{s,i}, \mu_{s,i}, \Delta_{s,i}\}, i = 1,..., M,$ where $\mu_{s,i}$, $\Delta_{s,i}$, and $\omega_{s,i}$ represent mean, diagonal covariance, and weight of the *i*th mixture of the *s*th speaker, respectively. Note that each speaker has only one transformation matrix, but has a GMM codebook of size *M*. Given an unknown test utterance X=$\{x_1, x_2, ..x_t..x_L\}$, the most possible speaker identity $\hat{s}$ is found by the following equation.

$$\hat{s} = \arg\max_{1\le s\le S}\left(\sum_{t=1}^{L} \log p(\Omega'_s x_t | \lambda_s)\right) \qquad (24)$$

in which $p(\Omega'_s x_t|\lambda_s)$ is given in (1) with $\vec{x}$ replaced by $\Omega'_s x_t$.

Equation (24) indicates that before computing the log-likelihood of each speaker GMM parameters λ_s, we must transform the features by Ω'_s.

V.EXPERIMENTS

A Chinese connected-digit database collected from 100 speakers was used in this experiment. Each speaker was asked to provide 40 utterances in a recording session. Five sessions had been recorded. Each set of 40 utterances includes 10 single digits and 30 connected digit strings with 2 to 7 digits. The database was partitioned into two parts. The first three sessions were used for training while the other two sessions were used for testing. To generate features, the speech signal was sampled at a 10-kHz sampling rate, and weighted by 25.6-ms Hamming window shifted in every 12.8 ms. For each speech frame, a 20-channel filterbank spectrum with mel-scale frequency was obtained. Each speech spectral vector was then transformed to a cepstral vector. Each cepstral vector contained 12 cepstral coefficients.

There are several experiments conducted to demonstrate our proposed method. The first part of experiments is to examine the log-likelihood on the training data for different mixture numbers. The second part of experiments is to compare the performance of our feature transformed GMM with that of no transform GMM. Diagonal covariance matrix is used for all probabilistic models.

Table 1: The average log-likelihood of training data

model type	codebook size		
	8	16	32
Transform	-33.68	-32.44	-31.43
No Transform	-33.26	-32.09	-31.08

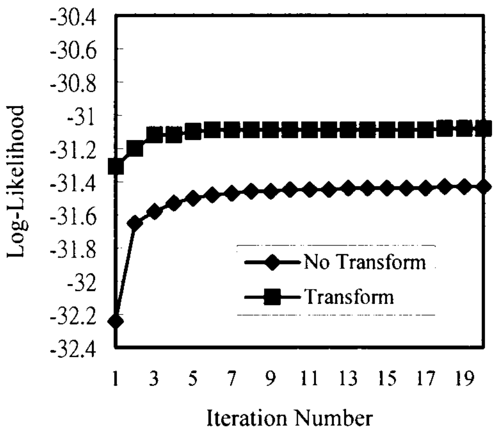

Figure 1. Comparison of log-likelihood on the training data during each iteration step

1. Comparing the log-likelihood of transformed GMM with that of no transform GMM.

Figure 1 shows the log-likelihood curves of training data for each iteration step. The codebook size is 32. Both methods become saturated very quickly, and the log-likelihood of transformed GMM is larger than that of no transformation. The reason is that the orthogonal transformation makes the feature vectors less correlated such that the diagonal GMM could be more fitted to the feature distribution. Table 1 shows the final log-likelihood of training data for different codebbook size. We see again a large difference between the models.

2. Comparing the performance of the transformed GMM with that of no transform GMM: .

Table 2 shows error rates of different digital lengths and codebook sizes for the two models. We can observe that the error rates of the two models reduce as the digital length or the codebook size increases. The mixture number in GMM represents the number of acoustic segments used for modeling speaker's characteristics. The larger the mixture number is, the higher the spectral resolution for speaker

The GMM parameters with no transformation of features is used as a baseline. We observe that the GMM with the feature transformation has a better performance than the baseline for different utterance length and mixture size. The experimental result shows a reduction in the error rate by 42% when employing 7-digit utterances are used for testing. This result is also consistent with the situation shown by the log-likelihood of training data.

VI.CONCLUSION

An orthogonal transformation of speech features are presented to reduce the correlation between feature components. This transformation is jointly optimized with GMM parameters in training phase. Experiments show that the performance of transformed GMM is better than that of the no transformation.

Table 2: The error rate (%) for the two models

model type		no transform	transform	no transform	transform	no transform	transform
size. of mixture		8	8	16	16	32	32
no. of digit length	1	42.4	34.7	29	23.2	21	19.4
	4	14.2	8.8	7	4.4	4.9	3.8
	7	7.6	4.1	4.2	2.3	2.6	1.5

REFERENCES

[1] D. A. Reynolds and R. C. Rose. “Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models”. IEEE Trans. Speech and Audio Processing, vol. 3, no. 1,Jan. 1995.

[2] A. Ljolje. “The importance of cepstral parameter correlations in speech recognition”. Computer Speech and Language (1994) 8, 223-232.

[3] B. L. Tseng, F. K. Soong, and A. E. Rosenberg “Continuous probabilistic acoustic map for speaker recognition”. In Proc. ICASSP-92, pages of IEEE, vol. 2, pages 161-164.

[4] K. Fukunaga. “Introduction to statistical pattern recognition” Academic Press, 1990