A Flexible Method of Creating HMM Using Block-diagonalization of Covariance Matrices

R. Koshiba, M. Tachimori and H. Kanazawa
Kansai Research Laboratories, Toshiba Corporation
8-6-26 Motoyama-Minami-cho, Higashi-Nada-ku, Kobe 658-0015, Japan
e-mail: koshiba@krl.toshiba.co.jp

ABSTRACT

A new algorithm to reduce the amount of calculation in the likelihood computation of continuous mixture HMM(CMHMM) with block-diagonal covariance matrices while retaining high recognition rate is proposed. The block matrices are optimized by minimizing difference between the output probability calculated with full covariance matrices and that calculated with block-diagonal covariance matrices. The idea was implemented and tested on a continuous number recognition task.

1 Introduction

In speech recognition, two conventional methods have been used to calculate the likelihood computation of CMHMM. One is to use full covariance matrices and the other is to use diagonal covariance matrices[2]. The former realizes higher recognition rate. However, if the dimension of the covariance matrices is large, the amount of calculation is too large to realize a real-time system. The latter, in which only diagonal elements of the covariance matrices are used, reduces the amount of the calculation. However, using only diagonal elements results in a poorer recognition rate than using full covariance matrices, especially in a noisy environment. Although each method has its advantages, tradeoff between the recognition rate and the amount of calculation is not considered.

To solve this problem, a new algorithm using block-diagonal covariance matrices is considered. The crucial ideas are as follows:

1: Some non-diagonal elements of covariance matrices, which have a significant effect on the output probability, are preserved as block matrices.

2: Block-diagonal covariance matrices are created by flexibly determined structure of block matrices, considering tradeoff between the recognition rate and the amount of calculation.

3: The block matrices are optimized by minimizing difference between the output probability calculated with full covariance matrices and that calculated with block-diagonal covariance matrices.

The paper is organized as follows. Section 2 confirms the ability of the conventional method to calculate the likelihood computation of CMHMM. In Section 3 the method to calculate the likelihood with block-diagonal covariance matrices is shown. Section 4 reports the results of a recognition experiment and shows the effect of the proposed method. Finally, Section 5 concludes this paper and indicates some guidelines for future work.

2 Conventional Likelihood Computation of CMHMM

In CMHMM, the likelihood b_{ij} between an input feature vector and each Gaussian distribution is

$$b_{ij}(\boldsymbol{y}_t) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_{ij}|^{1/2}} \times \exp[-\frac{1}{2}(\boldsymbol{y}_t - \boldsymbol{\mu}_{ij})^T \boldsymbol{\Sigma}_{ij}^{-1}(\boldsymbol{y}_t - \boldsymbol{\mu}_{ij})], \quad (1)$$

where $\boldsymbol{y}_t$ denotes an n-dimensional input feature vector at time t, $\boldsymbol{\Sigma}_{ij}$ and $\boldsymbol{\mu}_{ij}$ are a covariance matrix and a mean vector of a Gaussian distribution respectively, and i and j denote state numbers. To make the expression short, the suffixes i and j are omitted in the following discussion.

The order of computation of the exponential member in the equation (1) is n^2. For a real-time recognition system, the reduction of the amount of the arithmetic operation is crucial.

One conventional method to reduce the calculation is to use diagonal covariance matrices. If a diagonal covariance matrix is

$$\boldsymbol{\Sigma}_{diag} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2), \tag{2}$$

where $\sigma_i^2 (i = 1, \ldots, n)$ are variances for each feature vector component, then the equation (1) becomes

$$b(\boldsymbol{y}_t) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_{diag}|^{1/2}} \times \exp[-\frac{1}{2}\sum_{i=0}^{n}\frac{(y_i - \mu_i)^2}{\sigma_i^2}]. \tag{3}$$

By using covariance matrices, the order of computation of the exponential member in the equation (3) is reduced to n.

In the equation (2), however, non-diagonal elements of a covariance matrix are replaced with zero. This replacement is equivalent to assuming that elements of a characteristic vector have no correlation to one another. If some feature vector components have significant correlation to one another, the difference between the output probability calculated with full covariance matrices and that calculated with diagonal covariance matrices becomes large and causes the degradation of recognition rate. By using block-diagonal covariance matrices, this problem can be solved.

3 Likelihood Computation of CMHMM using Block–diagonal Covariance Matrices

3.1 Continuous Mixture HMM using Block-diagonal Covariance Matrices

A block-diagonal covariance matrix can be denoted as

$$\boldsymbol{\Sigma}_{BD} = \begin{bmatrix} \boldsymbol{A}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{A}_2 & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{A}_D \end{bmatrix}, \tag{4}$$

where $\boldsymbol{A}_i$ is a d_i-dimensional square, symmetric and positive definite matrix and d_i satisfies

$$\sum_{i=1}^{D} d_i = n. \tag{5}$$

If $\boldsymbol{\Sigma}$ in the equation (1) is replaced with $\boldsymbol{\Sigma}_{BD}$, then the equation (1) becomes

$$b(\boldsymbol{y}_t) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_{BD}|^{1/2}} \times \exp[-\frac{1}{2}\sum_{k=1}^{D}\boldsymbol{x}_k^t \boldsymbol{A}_k^{-1}\boldsymbol{x}_k], \tag{6}$$

where $\boldsymbol{x}_k$ is a d_k-dimensional vector and satisfies

$$\boldsymbol{y}_t - \boldsymbol{\mu} = (\boldsymbol{x}_1 \boldsymbol{x}_2 \cdots \boldsymbol{x}_D)^T. \tag{7}$$

The order of computation of the exponential member in the equation (6) is $\sum_i^D (d_i \times d_i)$, whereas that using full covariance matrices is n^2.

The structure of block matrices, d_i and D, is determined according to specifications of a supposed system. If a faster system is needed, the number and dimension of block matrices should be small, namely $\sum_i^D (d_i \times d_i)$ should be small. On the other hand, if a higher recognition rate is required, they should be large. In this case the amount of calculation becomes large.

After the determination of the structure, the next problem is how to choose block-matrices.

3.2 How to Obtain "Optimal" Block-diagonal Covariance Matrices

If sufficient training data are prepared and full covariance matrices are well trained, the performance using the full covariance matrices is higher than using any block-diagonal covariance matrices. Considering this fact, block-diagonal covariance matrices have been chosen to minimize the difference between the equation (1) and (6). $|\boldsymbol{\Sigma}|$ in (1) and $|\boldsymbol{\Sigma}_{BD}|$ in (6) are constant. So block-diagonal matrices are determined by minimizing

$$\begin{aligned} &|\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{\Sigma}_{BD}^{-1}\boldsymbol{x}| \\ &= |\boldsymbol{x}^T(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_{BD}^{-1})\boldsymbol{x}|, \end{aligned} \tag{8}$$

where $\boldsymbol{x} = \boldsymbol{y}_l - \boldsymbol{\mu}$.

To obtain $\boldsymbol{\Sigma}_{BD}$ which minimizes the equation (8), the following lemma is available:

Lemma 1 *Let $\boldsymbol{A}, \boldsymbol{B} \in R^{n\times n}$ * be symmetric matrices and $\boldsymbol{A}$ be positive definite. Then, under the condition $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \leq 1$, the maximum value of $|\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}|$ is $\max_\lambda |\lambda(\boldsymbol{B}\boldsymbol{A}^{-1})|$ †.*

A full covariance matrix $\boldsymbol{\Sigma}$, a block-diagonal covariance matrix $\boldsymbol{\Sigma}_{BD}$ and their inverse matrices $\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Sigma}_{BD}^{-1}$ are square, symmetric and positive definite. So $\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_{BD}^{-1}$ is a square and symmetric matrix (not necessarily positive definite). If $\boldsymbol{A}$ and $\boldsymbol{B}$ in the lemma 1 are replaced with $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_{BD}^{-1}$ respectively, then, under the condition $\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} \leq 1$, the maximum value of the equation (8) is

$$\max_\lambda |\lambda((\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_{BD}^{-1})\boldsymbol{\Sigma})| = \max_\lambda |\lambda(I - \boldsymbol{\Sigma}_{BD}^{-1}\boldsymbol{\Sigma})|. \tag{9}$$

A vector $\boldsymbol{y}_l$ which satisfies the condition $\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} = 1$ is a vector one standard deviation away from the mean $\boldsymbol{\mu}$. Then a block-diagonal covariance matrix which minimizes the equation (9) under $\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} \leq 1$ has been chosen. Namely, an *optimal* block-diagonal covariance matrix is determined by

$$\boldsymbol{\Sigma}_{BD} = \arg\min_{\boldsymbol{\Sigma}_{BD}} \max_\lambda |\lambda(I - \boldsymbol{\Sigma}_{BD}^{-1}\boldsymbol{\Sigma})|. \tag{10}$$

* $R^{n\times n}$ means a class of $n \times n$-dimensional real matrices.

† $\lambda(A)$ is an eigen value of a matrix A.

The $\boldsymbol{\Sigma}_{BD}$ cannot be solved analytically. So this problem is solved by choosing a combination of feature vector components which minimizes the equation (9). If elements of a full covariance matrix are σ_{ij}, then the i-th block matrix in the equation (4) is expressed as

$$\boldsymbol{A}_i = \begin{bmatrix} \sigma_{pp} & \cdots & \sigma_{p,p+d_i-1} \\ \vdots & \ddots & \vdots \\ \sigma_{p+d_i-1,p} & \cdots & \sigma_{p+d_i-1,p+d_i-1} \end{bmatrix}, \tag{11}$$

where $p = \sum_{j=1}^{d_{i-1}}$.

The procedure of the algorithm to choose the combination is as follows:

S1: Compute full covariance matrices($n \times n$ dimension) for each HMM model.

S2: Determine the number(D) and dimension($d \times d$) of block matrices according to specifications of a required system.

S3: For each possible combination of d input feature vector components, make a block-diagonal covariance matrix $\boldsymbol{\Sigma}_{BD}$ with their covariance elements and all diagonal elements. Then calculate the equation (9) using the $\boldsymbol{\Sigma}_{BD}$ and obtain the combination which minimizes (9).

S4: If the number of block matrices reaches the determined number(D), use the obtained block-diagonal covariance matrices $\boldsymbol{\Sigma}_{BD}$ for recognition. If not, remove rows and columns that include the obtained block matrices from full covariance matrices $\boldsymbol{\Sigma}$ and reiterate from S3.

In S2 the number and dimension of block matrices do not need to be the same values for all covariance matrices.

4 Simulations

The algorithm proposed in the previous section was implemented and tested on a continuous number recognition task.

The training and evaluation conditions were as follows:

- A feature vector was 30-dimension.

- Training data were 35 balanced four-digit continuous numbers uttered by 222 different speakers.

- Evaluation data were the same numbers uttered by 10 different speakers.

- Each evaluation datum was recognized under city car noise of 20, 10, 0 dB.

Table 1 shows the average recognition rate under each condition. Each full covariance matrix was block-diagonalized to fifteen 2×2-block matrices, ten 3×3-block matrices and six 5×5-block matrices.

Table 1: Speech recognition performance under city car noise(%)

	20dB	10dB	0dB
diagonal	96.51	97.09	96.51
$2 \times 2 \times 15$	97.09	97.09	96.22
$3 \times 3 \times 10$	97.09	97.09	97.38
$5 \times 5 \times 6$	97.97	98.26	98.26
full	99.42	100.00	99.13

The reduction rates of the calculation of diagonal, $2 \times 2 \times 15$, $3 \times 3 \times 10$ and $5 \times 5 \times 6$ block-diagonal covariance matrices as compared with full covariance matrices were 96.7%, 94.7%, 90.0% and 83.3% respectively. Table 1 also shows that the tradeoff between the recognition rate and the amount of calculation can be determined flexibly through the number and dimension of block matrix.

5 Conclusion and Further Work

A new algorithm to reduce the amount of calculation in the likelihood computation of CMHMM with flexibly created block-diagonal covariance matrices has been proposed. The idea was implemented and tested on a continuous number recognition task. The result of the simulations showed that tradeoff between the amount of calculation and the recognition rate can be determined flexibly with the proposed algorithm.

Further work will include:

- Vector quantization of the combined feature vectors and table look-up.

- Block-diagonal training when training-data are poor.

Vector quantization and table look-up is a conventional method to reduce the amount of calculation in the likelihood computation of CMHMM. But quantization error becomes large in proportion to the dimension of a feature vector. In the proposed method each dimension of a block matrix can be small. This leads to small quantization error.

This paper assumed that full covariance matrices were well-trained. If training-data are limited and full covariance matrices are poorly estimated, the proposed algorithm doesn't provide reliable block matrices[1]. In this case, the following method is possible:

- First, create full covariance matrices with small number of mixtures.

- Secondly, obtain an *optimal* combination of feature vector components with the proposed method.

- Finally, retrain block-diagonal covariance matrices with more mixtures using the combination obtained above.

Verification of these methods is a subject for future work.

6 References

1. Rabiner, L., and Juang, B.-H., *Fundamentals of Speech Recognition*, Prentice Hall International Editions, 1993

2. Sagayama, S., and Takahashi, S., "On the use of scalar quantization for fast HMM computation," Proc. of ICASSP 95, pp. 213-216, 1995