PAPER *Special Section on Corpus-Based Speech Technologies*

Parameter Sharing in Mixture of Factor Analyzers for Speaker Identification

Hiroyoshi YAMAMOTO[†a)], ***Nonmember***, **Yoshihiko NANKAKU**[†], **Chiyomi MIYAJIMA**[††], **Keiichi TOKUDA**[†], ***and*** **Tadashi KITAMURA**[†], ***Members***

SUMMARY This paper investigates the parameter tying structures of a mixture of factor analyzers (MFA) and discriminative training of MFA for speaker identification. The parameters of factor loading matrices or diagonal matrices are shared in different mixtures of MFA. Then, minimum classification error (MCE) training is applied to the MFA parameters to enhance the discrimination ability. The result of a text-independent speaker identification experiment shows that MFA outperforms the conventional Gaussian mixture model (GMM) with diagonal or full covariance matrices and achieves the best performance when sharing the diagonal matrices, resulting in a relative gain of 26% over the GMM with diagonal covariance matrices. The improvement is more significant especially in sparse training data condition. The recognition performance is further improved by MCE training with an additional gain of 3% error reduction.

key words: *speaker identification, GMM, mixture of factor analyzers, parameter sharing, minimum classification error training*

1. Introduction

Gaussian mixture model (GMM) is widely used for text-independent speaker identification [1]. It is well known that GMM with full covariance matrices needs sufficient training data to guarantee the reliability of the estimated model parameters. Furthermore, GMM with diagonal covariance matrices requires a relatively large number of Gaussians to provide high recognition performance. To cope with this problem, a mixture of factor analyzers (MFA) [2] has been proposed as a mixture model (or linear combination) of factor analysis (FA) models.

FA is a statistical method in which high dimensional observation vectors are assumed to be generated from several latent *factors* which can capture the correlation between the feature vectors. In FA, a feature vector is defined as the sum of a lower dimensional factor vector weighted by a factor loading matrix and a noise vector. A covariance matrix of FA is modeled by the factor loading matrix and a diagonal matrix which represents the covariance of the noise vectors. To deal with data distributed intricately, FA model is often extended to a *mixture* of factor analyzers (MFA). MFA allows us to reduce the degree of freedom of the covariance matrices maintaining the recognition performance. MFA has been successfully applied to feature modeling in several areas such as speech and speaker recognition [3], [4].

Moreover, the reliability of the estimated parameters of MFA can be improved by sharing parameters in different mixture components. Some researchers used MFA with several covariance structures. For example, *Ghahramani, et al.* derived an expectation maximization (EM) algorithm of MFA with shared diagonal covariance matrices [2]. On the other hand, *Ding, et al.* adopted MFA sharing factor loading matrices among different mixture components, or among different models [4]. *Saul and Rahim* used FA-based hidden Markov model (HMM) without parameter tying in a speech recognition task [3]. They also evaluated minimum classification error (MCE) training of FA-HMM in the same task. However, there are no reports comparing these possible covariance structures of MFA in previous papers.

In this paper, parameter sharing structures of MFA are investigated for speaker identification [5]. Factor loading matrices or diagonal matrices of MFA-based speaker models are shared in different mixture components assuming that all the mixture components have the same number of factors. We compare the following three kinds of MFAs with different parameter sharing structures.

1) MFA without parameter sharing
2) MFA with shared diagonal matrices
3) MFA with shared factor loading matrices

In addition, MCE training is applied to MFA to improve the speaker recognition performance. The effectiveness of the MCE training for the parameter shared MFA is evaluated in a text-independent speaker identification task.

This paper is organized as follows. Sections 2 and 3 describe the general formulation of MFA and parameter sharing structures, respectively. Section 4 presents the MCE training of MFA, and the experimental results are reported in Sect. 5. Finally, conclusions and future works are given in Sect. 6.

2. Mixture of Factor Analyzers

2.1 Factor Analysis

Factor analysis (FA) is a statistical method for modeling the covariance structure of high dimensional data using a small number of latent variables [6]. In FA, a d-dimensional

Manuscript received July 8, 2004.
Manuscript revised September 24, 2004.

[†] The authors are with the Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya-shi, 466–8555 Japan.

[††] The author is with the Department of Media Science, Nagoya University, Nagoya-shi, 466–8603 Japan.

a) E-mail: legacy@ics.nitech.ac.jp

DOI: 10.1093/ietisy/e88–d.3.418

speech feature vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)^T$ is modeled using a q-dimensional vector $\boldsymbol{z} = (z_1, z_2, \ldots, z_q)^T$ and a d-dimensional observation noise $\boldsymbol{n} = (n_1, n_2, \ldots, n_d)^T$:

$$\begin{aligned}\boldsymbol{x} &= \boldsymbol{\mu} + \sum_{i=1}^{q} z_i \boldsymbol{w}_i + \boldsymbol{n} \\ &= \boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{z} + \boldsymbol{n}, \end{aligned} \tag{1}$$

where $\boldsymbol{\mu}$ denotes a mean vector, and $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_q)$, $\boldsymbol{w}_i = (w_{i1}, w_{i2}, \ldots, w_{id})^T$ is a $d \times q$ matrix known as a factor loading matrix. Vector $\boldsymbol{z}$ is a latent variable assumed to be distributed according to a Gaussian density $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, i.e., zero-mean independent normals with unit variance. Each element of $\boldsymbol{z}$ is referred to as "factor". The noise vector $\boldsymbol{n}$ is distributed according to $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is a diagonal matrix.

The likelihood of an observation $\boldsymbol{x}$ is given by

$$p(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{z}, \boldsymbol{\Psi}) \tag{2}$$

because when $\boldsymbol{z}$ is given, the product $\boldsymbol{W}\boldsymbol{z}$ is a constant vector added to the observation noise vector $\boldsymbol{n}$. Therefore, distribution for $\boldsymbol{x}$ is obtained by integrating out the latent variable $\boldsymbol{z}$:

$$\begin{aligned} p(\boldsymbol{x}) &= \int p(\boldsymbol{x} \mid \boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} \\ &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{\Psi}). \end{aligned} \tag{3}$$

2.2 Extension of FA to MFA

The FA model works well for correlated data with Gaussian distribution provided the number of factors is appropriately selected. In reality, the data, such as speech feature vectors, are not always Gaussian distributed. To deal with such data, FA model is often extended to a *mixture* of factor analyzers (MFA). MFA is defined as a mixture of M factor analyzers (Fig. 1). The likelihood of T independent feature vectors $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ for the M-component MFA $\boldsymbol{\theta} = \{c_m, \boldsymbol{\mu}_m, \boldsymbol{W}_m, \boldsymbol{\Psi}_m \mid m = 1, \ldots, M\}$ is given by

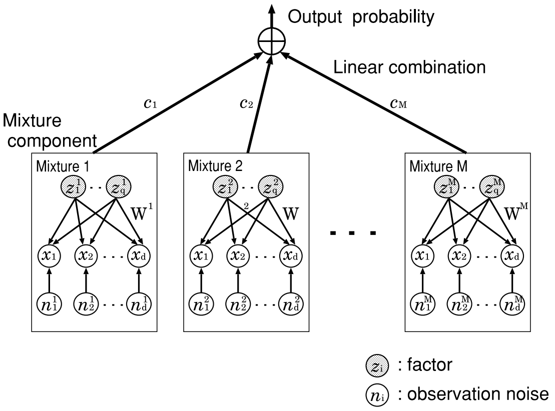


Fig. 1 Mixture of factor analyzers.

$$\begin{aligned} p(\boldsymbol{X} \mid \boldsymbol{\theta}) &= \prod_{t=1}^{T} \sum_{m=1}^{M} \int p_m(\boldsymbol{x}_t \mid \boldsymbol{z}) p_m(\boldsymbol{z}) c_m d\boldsymbol{z} && (4) \\ &= \prod_{t=1}^{T} \sum_{m=1}^{M} c_m \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_m), && (5) \end{aligned}$$

where c_m denotes the weight of the m-th mixture component and $\boldsymbol{\Sigma}_m = \boldsymbol{W}_m \boldsymbol{W}_m^T + \boldsymbol{\Psi}_m$.

3. Parameter Sharing

Covariance matrices $\boldsymbol{\Sigma}_m$ of MFA consist of $\boldsymbol{W}_m$ and $\boldsymbol{\Psi}_m$. The reliability of the MFA can be improved by sharing these parameters of the covariance matrices. In this section, some variations of parameter sharing structures are presented.

Some researches have been conducted to evaluate the effectiveness of MFA with several covariance structures [2]–[4]. However, detailed comparison of these possible parameter sharing structures has not given yet in previous papers.

In this paper, we compare the following three kinds of MFAs with different parameter sharing structures shown in Fig. 2, assuming that all the mixture components have the same number of factors:

1) **Non-shared MFA:** MFA without parameter sharing [3].
2) **$\boldsymbol{\Psi}$-shared MFA:** MFA with shared diagonal covariance matrices, where $\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2 = \cdots = \boldsymbol{\Psi}$ and $\boldsymbol{n}$ is assumed to be a sensor noise [2].
3) **$\boldsymbol{W}$-shared MFA:** MFA with shared factor loading matrices, where $\boldsymbol{W}_1 = \boldsymbol{W}_2 = \cdots = \boldsymbol{W}$, i.e., the weights of each factor in different mixtures are the same [4].

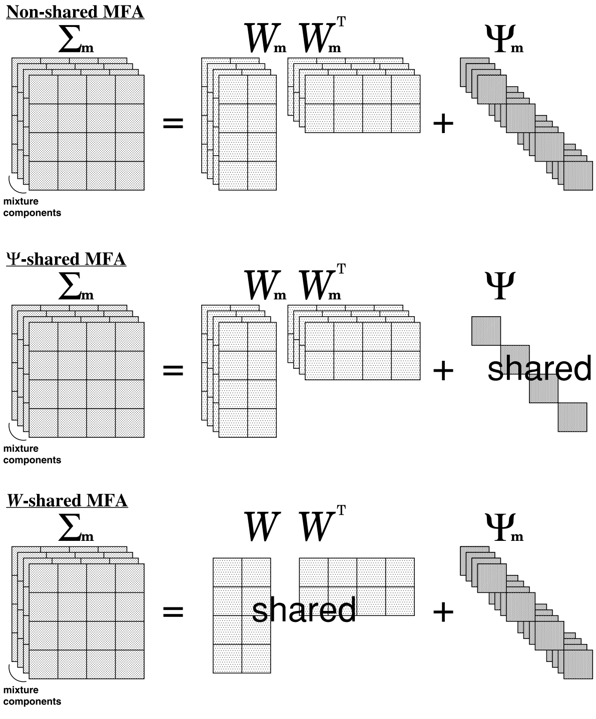


Fig. 2 The structures of covariance matrices.

The maximum likelihood (ML) solution better suits the linear Gaussian model framework since the expectation maximization (EM) algorithm can be used. The EM steps for the MFA parameters $\boldsymbol{\theta}$ are summarized as follows.

3.1 E-Step

The E-step calculates the expectation of latent vector z and the posterior of the m-th mixture component:

$$\begin{aligned} \langle z_{tm} \rangle &= E[z|\boldsymbol{x}_t, m] = \boldsymbol{\beta}_m(\boldsymbol{x}_t - \boldsymbol{\mu}_m), & (6)\\ \langle zz_{tm} \rangle &= E[zz^T|\boldsymbol{x}_t, m] \\ &= I - \boldsymbol{\beta}_m \boldsymbol{W}_m + \langle z_{tm} \rangle \langle z_{tm} \rangle^T, & (7)\\ h_{tm} &= \frac{c_m \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_m c_m \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}, & (8) \end{aligned}$$

where $\boldsymbol{\beta}_m = \boldsymbol{W}_m^T \boldsymbol{\Sigma}_m^{-1}$ and $\boldsymbol{\Sigma}_m = \boldsymbol{W}_m \boldsymbol{W}_m^T + \boldsymbol{\Psi}_m$.

3.2 M-Step

The M-step is also very straightforward. The new model parameters $\boldsymbol{\mu}'$, $\boldsymbol{W}'$, $\boldsymbol{\Psi}'$, and c'_m for the three kinds of MFA mentioned above can be obtained by the following re-estimation formulae.

1) Non-shared MFA

The re-estimation formulae require some manipulation to obtain the new MFA parameters using the following convenient matrix operations.

$$\tilde{\boldsymbol{W}}_m = (\boldsymbol{W}_m \; \boldsymbol{\mu}_m) \qquad (9)$$

$$\tilde{z}_{tm} = \begin{pmatrix} z \\ 1 \end{pmatrix} \qquad (10)$$

The re-estimates of $\tilde{\boldsymbol{W}}'_m$ and $\boldsymbol{\Psi}_m$ are obtained by

$$\tilde{\boldsymbol{W}}'_m = \left(\sum_t h_{tm} \boldsymbol{x}_t \langle \tilde{z}_{tm} \rangle^T \right) \cdot \left(\sum_l h_{lm} \langle \tilde{z}\tilde{z}_{lm} \rangle \right)^{-1}, \qquad (11)$$

$$\boldsymbol{\Psi}'_m = \frac{1}{\sum_t h_{tm}} \mathrm{diag} \left\{ \sum_t h_{tm} \left(\boldsymbol{x}_t - \tilde{\boldsymbol{W}}'_m \langle \tilde{z}_{tm} \rangle \right) \boldsymbol{x}_t^T \right\}, \qquad (12)$$

where

$$\langle \tilde{z}_{tm} \rangle = \begin{pmatrix} \langle z_{tm} \rangle \\ 1 \end{pmatrix}, \qquad (13)$$

$$\langle \tilde{z}\tilde{z}_{tm} \rangle = \begin{pmatrix} \langle zz_{tm} \rangle & \langle z_{tm} \rangle \\ \langle z_{tm} \rangle & 1 \end{pmatrix}, \qquad (14)$$

and diag(·) denotes setting the elements outside the main diagonal to zeros. The mixture weight c_m is re-estimated as follows.

$$c'_m = \frac{1}{T} \sum_{t=1}^{T} h_{tm} \qquad (15)$$

2) $\boldsymbol{\Psi}$-shared MFA

The re-estimation formulae for $\boldsymbol{\Psi}$-shared MFA are the same as those for Non-shared MFA except for the diagonal covariance matrix:

$$\boldsymbol{\Psi}' = \frac{1}{T} \mathrm{diag} \left\{ \sum_{t,m} h_{tm} \left(\boldsymbol{x}_t - \tilde{\boldsymbol{W}}'_m \langle \tilde{z}_{tm} \rangle \right) \boldsymbol{x}_t^T \right\}. \qquad (16)$$

3) $\boldsymbol{W}$-shared MFA

The new model parameters of $\boldsymbol{W}$-shared MFA is re-estimated as follows. The new factor loading matrix $\boldsymbol{W}'$ is given by

$$\begin{aligned} \boldsymbol{W}'_{(k)} = &\left(\sum_{t,m} h_{tm} \boldsymbol{\Psi}^{-1}_{m(k)} (\boldsymbol{x}_t - \boldsymbol{\mu}_m)_{(k)} \langle z_{tm} \rangle^T \right) \\ &\cdot \left(\sum_{t,m} h_{tm} \boldsymbol{\Psi}^{-1}_{m(k)} \langle zz_{tm} \rangle \right)^{-1} \qquad (17) \end{aligned}$$

where $\boldsymbol{W}_{(k)}$ is k-th row vector in the factor loading matrix $\boldsymbol{W}$. In the followings, the individual component parameters $\boldsymbol{\mu}'_m$ and $\boldsymbol{\Psi}'_m$ can be re-estimated:

$$\begin{aligned} \boldsymbol{\mu}'_m &= \frac{\sum_t h_{tm}(\boldsymbol{x}_t - \boldsymbol{W}' \langle z_{tm} \rangle)}{\sum_t h_{tm}}, & (18)\\ \boldsymbol{\Psi}'_m &= \frac{1}{\sum_t h_{tm}} \mathrm{diag} \sum_t \Big\{ h_{tm}(\boldsymbol{x}_t - \boldsymbol{\mu}'_m)(\boldsymbol{x}_t - \boldsymbol{\mu}'_m)^T \\ &\quad - h_{tm} \boldsymbol{W}' \left(2 \langle z_{tm} \rangle (\boldsymbol{x}_t - \boldsymbol{\mu}'_m)^T - \langle zz_{tm} \rangle \boldsymbol{W}'^T \right) \Big\}. & (19) \end{aligned}$$

4. MCE Training for MFA Speaker Model

To enhance the discrimination abilities of MFA-based speaker models, MCE training based on the generalized probabilistic descent (GPD) method [7] is applied to the parameters of MFA [3].

4.1 Definition of Loss Function

For the MCE training, the misclassification measure of training data $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ for speaker s is defined as

$$\begin{aligned} d_s(\boldsymbol{X}; \boldsymbol{\Theta}) = &-g_s(\boldsymbol{X}; \boldsymbol{\Theta}) \\ &+ \log \left[\frac{1}{S-1} \sum_{y \neq s} \exp \left\{ g_y(\boldsymbol{X}; \boldsymbol{\Theta}) \eta \right\} \right]^{\frac{1}{\eta}} \qquad (20) \end{aligned}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_S\}$ denotes the speaker model parameter set of MFA, and $g_s(\,\cdot\,;\cdot\,)$ is defined by the log likelihood of $\boldsymbol{X}$ for speaker model $\boldsymbol{\theta}_s$. When assuming $\eta = \infty$, we have

$$d_s(\boldsymbol{X}; \boldsymbol{\Theta}) = -g_s(\boldsymbol{X}; \boldsymbol{\Theta}) + \max_{y \neq s} g_y(\boldsymbol{X}; \boldsymbol{\Theta}). \qquad (21)$$

Equation (21) is the approximation of the log likelihood ratio between the competing models and the correct one. The loss function is defined as a differentiable sigmoid function approximating the 0-1 step loss function:

$$l_s(\boldsymbol{X};\,\boldsymbol{\theta}) = \Big(1 + \exp(-\gamma \cdot d_s + \alpha)\Big)^{-1}, \tag{22}$$

where γ denotes the gradient of the sigmoid function and α represents the offset of the sigmoid function. In this experiment, α is set to zero. The goal of the discriminative training is to minimize the loss function based on the probabilistic descent method.

4.2 Parameter Adjustment of MFA

During the parameter adaptation in the MCE training, the constraints of the MFA parameters, e.g., $c_m > 0$, should be satisfied. Hence, the MFA parameter set $\boldsymbol{\Theta}$ is transformed into a new model parameter set $\tilde{\boldsymbol{\Theta}}$.

$$\tilde{\boldsymbol{\Theta}} = \{\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \ldots, \tilde{\boldsymbol{\theta}}_S\}, \tag{23}$$

$$\tilde{\boldsymbol{\theta}} = \{\tilde{c}_m, \tilde{\boldsymbol{\mu}}_m, \boldsymbol{W}_m, \tilde{\boldsymbol{\Psi}}_m \mid m = 1, 2, \ldots, M\}, \tag{24}$$

where $\tilde{c}_m = \log c_m$, $\tilde{\mu}_{mi} = \frac{\mu_{mi}}{\Sigma_{mii}}$, $\tilde{\Psi}_{mii} = \log \Psi_{mii}$. $\tilde{\boldsymbol{\Theta}}$ is updated at each iteration r as

$$\tilde{\boldsymbol{\Theta}}(r+1) = \tilde{\boldsymbol{\Theta}}(r) - \varepsilon_r \nabla l_s(\boldsymbol{X};\,\tilde{\boldsymbol{\theta}}), \tag{25}$$

where ε_r is a monotonically decreasing learning step size at the r-th iteration. In this paper, $\tilde{\boldsymbol{\Theta}}$ is sequentially adjusted every time a training sample $\boldsymbol{X}$ is given (i.e., sample-by-sample mode).

The gradient of (25) is obtained as follows.

$$\nabla_{\tilde{\boldsymbol{\theta}}_y} l_s(\boldsymbol{X};\tilde{\boldsymbol{\theta}}) = \frac{\partial l_s}{\partial d_s}\frac{\partial d_s}{\partial g_y} \cdot \nabla_{\tilde{\boldsymbol{\theta}}_y} g_y(\boldsymbol{X};\tilde{\boldsymbol{\theta}}), \tag{26}$$

where $\frac{\partial l_s}{\partial d_s}$, $\frac{\partial d_s}{\partial g_y}$, $\nabla_{\tilde{\boldsymbol{\theta}}_y} g_y(\boldsymbol{X};\tilde{\boldsymbol{\theta}})$ are given by

$$\frac{\partial l_s}{\partial d_s} = \gamma l_s(1 - l_s), \quad \frac{\partial d_s}{\partial g_y} = \begin{cases} -1, & y = s \\ 1, & y \neq s \end{cases}, \tag{27}$$

$$\nabla_{\tilde{\boldsymbol{\theta}}_y} g_y(\boldsymbol{X};\tilde{\boldsymbol{\theta}}) = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{b_y(\boldsymbol{x}_t)} \nabla_{\tilde{\boldsymbol{\theta}}_y} b_y(\boldsymbol{x}_t). \tag{28}$$

For the three kinds of MFA, the gradient of $b_y(\boldsymbol{x}_t)$ with respect to each element in $\tilde{\theta}_y$ is obtained by the following formulae, where the subscript y is dropped for the simplicity of notation.

1) Non-shared MFA

For the Non-shared MFA, the gradients are obtained as follows.

$$\frac{\partial b(\boldsymbol{x}_t)}{\partial \tilde{c}_m} = f_m, \quad \frac{\partial b(\boldsymbol{x}_t)}{\partial \tilde{\mu}_{mi}} = f_m \delta_{mi} \Sigma_{mii}, \tag{29}$$

$$\frac{\partial b(\boldsymbol{x}_t)}{\partial W_{mij}} = -f_m \left\{ (\boldsymbol{\Sigma}_m^{-1} \boldsymbol{W}_m)_{ij} - \delta_{mi} [\boldsymbol{\delta}_m^T \boldsymbol{W}_m]_j \right\}, \tag{30}$$

$$\frac{\partial b(\boldsymbol{x}_t)}{\partial \tilde{\Psi}_{mii}} = -\frac{1}{2} f_m \left\{ \Sigma_{mii}^{-1} - \delta_{mi}^2 \right\} \Psi_{mii}, \tag{31}$$

where $f_m = c_m \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$, $\boldsymbol{\delta}_m = \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_m)$, and $[\cdot]_i$ denotes the i-th vector element.

2) $\boldsymbol{\Psi}$-shared MFA

In the case of $\boldsymbol{\Psi}$-shared MFA, the gradients with respect to mixture weights, mean vectors and factor loading matrices are also obtained as in (29) and (30), respectively. Only (31) for diagonal covariance matrices is changed as follows:

$$\frac{\partial b(\boldsymbol{x}_t)}{\partial \tilde{\Psi}_{ii}} = \sum_{m=1}^{M} \frac{\partial b(\boldsymbol{x}_t)}{\partial \tilde{\Psi}_{mii}}. \tag{32}$$

3) $\boldsymbol{W}$-shared MFA

The gradients in (29) and (31) apply to the $\boldsymbol{W}$-shared MFA case, and (30) is changed as follows.

$$\frac{\partial b(\boldsymbol{x}_t)}{\partial W_{ij}} = \sum_{m=1}^{M} \frac{\partial b(\boldsymbol{x}_t)}{\partial W_{mij}} \tag{33}$$

5. Experimental Evaluation

5.1 Database and Experimental Conditions

Text-independent speaker identification experiments were conducted using the ATR Japanese speech database and the NTT database [8].

We used word data spoken by 80 speakers (40 males and 40 females) in "c-set" of the ATR database. Phonetically-balanced 216 words are used for training each speaker model, and 520 common words are used for testing. The number of tests was 41600 in total. The NTT database consists of sentence data uttered at three speeds (normal, fast and slow) by 35 Japanese speakers (22 males and 13 females) on five sessions over ten months (Aug., Sept., Dec. 1990, Mar., June 1991), among which, the normal-speed data set was used. In each session, 15 sentences were recorded for each speaker. Ten sentences are common to all speakers and all sessions (A-set), and five sentences are different for each speaker and each session (B-set). The duration of each sentence is approximately four second. We used 15 sentences (A-set + B-set from the first session) per speaker for training, and 20 sentences (B-set from the other four sessions) per speaker for testing. The number of tests was 700 in total.

The speech data was down-sampled from 20 kHz to 10 kHz, windowed at a 10-ms frame rate using a 25.6-ms Blackman window, and parameterized into 12 mel-cepstral coefficients excluding zero-th coefficients with a mel-cepstral analysis technique. Session-dependent utterance variation was normalized for the multi-session NTT database using cepstrum mean subtraction (CMS) method, which is a well-known technique for canceling the effect of channels and utterance variation in speaker recognition [9].

GMM parameters were initialized using an LBG codebook. Mixture weights and mean vectors of MFA were also initialized using the LBG codebook, and factor loading matrices were initialized with random values considering full covariance. Diagonal covariance matrices were initialized using diagonal elements of full covariance matrices $\mathbf{\Sigma}$ [2]. The number of mixture components was changed from 4 to 64 for MFA and GMM with full covariance matrices, and from 4 to 256 for GMM with diagonal covariance matrices. The number of factors was changed from 2 to 10.

5.2 Results

Figures 3–5 compare the identification error rates among the three kinds of MFAs and the conventional GMMs with full or diagonal covariance matrices (full-GMM and diag-GMM). All speaker models in Figs. 3–5 were trained with 216 words based on ML-estimation and the number of factors q is changed as 2, 4, 6, 8 and 10. The horizontal axis corresponds to the number of model parameters in a logarithmic scale. MFA model parameters include c_m, $\boldsymbol{\mu}_m$, $\boldsymbol{W}_m$ and $\boldsymbol{\Psi}_m$, where c_m is a scalar, $\boldsymbol{\mu}_m$ is a D-dimensional vector, $\boldsymbol{W}_m$ is a $D \times q$ matrix, and $\boldsymbol{\Psi}_m$ is a D-dimensional diagonal vector. The total numbers of model parameters (N) of GMM and MFA are calculated as follows.

- Diag-GMM

$$N = (2D + 1)M, \tag{34}$$

- Full-GMM

$$N = \frac{M}{2}(D + 1)(D + 2), \tag{35}$$

- Non-shared MFA

$$N = \{(q + 2)D + 1\} M, \tag{36}$$

- $\boldsymbol{\Psi}$-shared MFA

$$N = \{(q + 1)D + 1\} M + D, \tag{37}$$

- $\boldsymbol{W}$-shared MFA

$$N = (2D + 1)M + Dq. \tag{38}$$

Figure 3 compares the results of Non-shared MFA with the conventional GMMs. We can see that the 6-factor Non-shared MFA, which had almost the same number of parameters as the full-GMM with the same number of Gaussians, gave the same or better performance than the full-GMM. Non-shared MFA gave better results than the diag-GMM with smaller numbers of factors, while the performance got worse than the full-GMM as the number of factors increased. The 64-mixture MFA with 2 factors achieved an error reduction rate of 6% over the diag-GMM. Figure 4 shows the results of $\boldsymbol{\Psi}$-shared MFA. $\boldsymbol{\Psi}$-shared MFA achieved a significant improvement over the conventional GMMs with larger number of mixtures. In the case of 64-mixture models, error reductions of 19% (q = 2) and 26% (q = 6) over diag-GMM were obtained. Figure 5 shows the results of $\boldsymbol{W}$-shared MFA. The performance of $\boldsymbol{W}$-shared MFA is almost equivalent to that of diag-GMM, because the model structure of $\boldsymbol{W}$-shared MFA is similar to that of diag-GMM, and has the lowest flexibility among the three kinds of sharing structures.

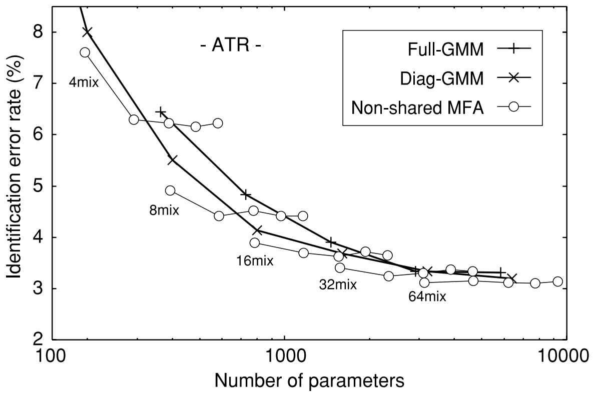


Fig. 3 Comparison among non-shared MFA and conventional GMMs with diagonal or full covariance matrices ($q = 2, \ldots, 10$).

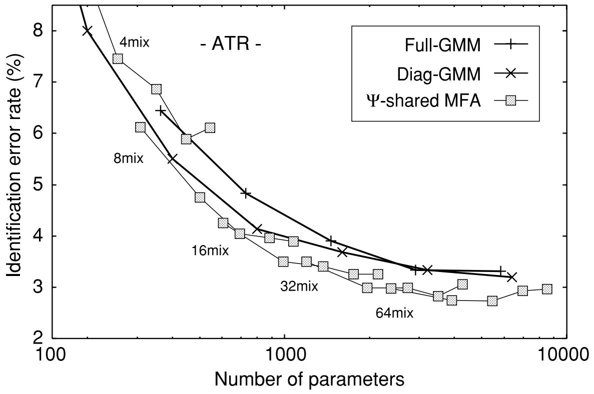


Fig. 4 Comparison among $\boldsymbol{\Psi}$-shared MFA and conventional GMMs with diagonal or full covariance matrices ($q = 2, \ldots, 10$).

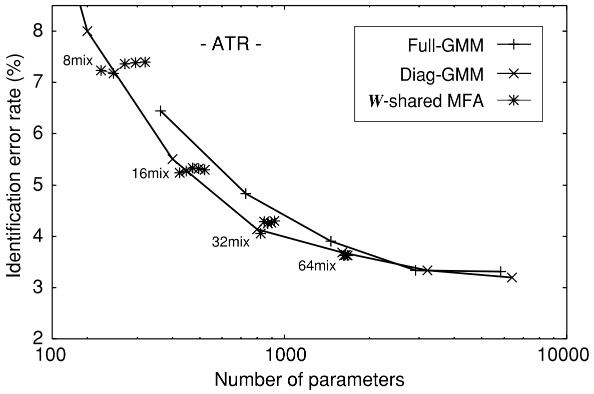


Fig. 5 Comparison among $\boldsymbol{W}$-shared MFA and conventional GMMs with diagonal or full covariance matrices ($q = 2, \ldots, 10$).

Figure 6 shows the results of the three kinds of MFAs with the number of factors q = 2 and diag-GMM, where the amount of training data was changed as 27, 54, and 216 words, i.e., one eighth, a quarter and all of the 216 words, respectively. In the first two cases, one of every eight and four words were selected for training. We can see that the MFA-based speaker models achieved relatively high performance

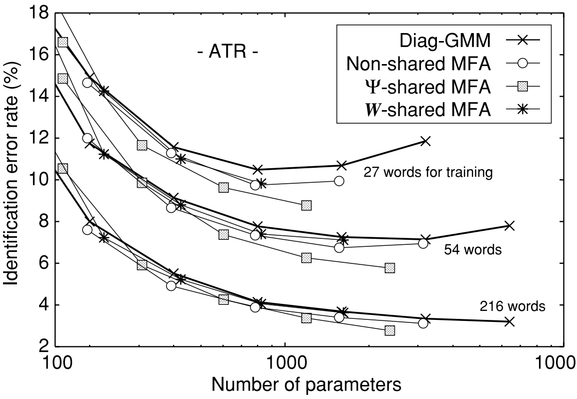


Fig. 6 Comparison among diag-GMM and three kinds MFAs (q = 2) with increasing the number of mixtures, using 27 words (upper), 54 words (middle), 216 words (lower) for training. Error rates of 4–64 mixture models are connected with a line.

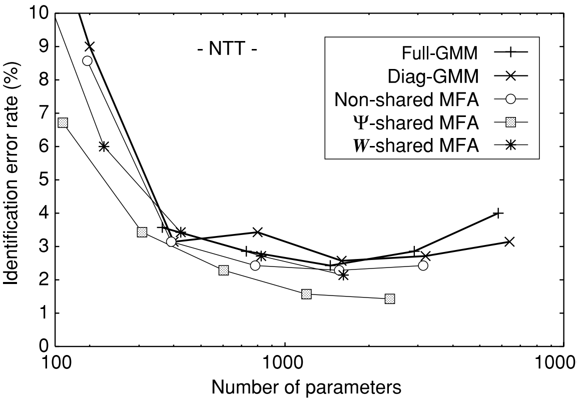


Fig. 7 Comparison among three kinds of MFAs and conventional GMMs with diagonal or full covariance matrices using the NTT database (q = 2). Error rates of 4–64 mixture models are connected with a line.

with such a small number of factors, and all the MFA-based speaker models outperformed the conventional GMM with any amount of training data, and improvement was more significant especially under sparse training data conditions. In the case of 32-mixture **Ψ**-shared MFA, error reduction rates compared to diag-GMM were 18%, 13%, and 8% for 27 words, 54 words, and 216 words, respectively. **Ψ**-shared MFA achieved the best performance among the three kinds of MFA and a significant difference is found with smaller amounts of training data.

Figure 7 shows the results of the three kinds of MFAs with the conventional GMMs using the NTT database. We can see that MFA-based speaker models also gave better results than the conventional GMMs in the multi-session speaker identification task and **Ψ**-shared MFA achieved a significant improvement over the conventional GMMs. In the case of 64-mixture models with 2-factors, relative gains of 10%, 47% and 17% over diag-GMM were obtained by Non-shared MFA, **Ψ**-shared MFA and ***W***-shared MFA, respectively.

Finally, MCE training was applied to the MFA-based speaker models. Figure 8 compares the performance of **Ψ**-shared MFA with two factors before and after the MCE training for the ATR database. We can see that the performance was further improved by the MCE training and the 2.98% error rate was reduced to 2.85% with a 3% error reduction. The performances of the MFAs with other structures of covariance matrices as well as conventional GMMs were also improved after MCE training. However, MFAs still outperformed the conventional GMMs, and **Ψ**-shared MFA gave the best result.

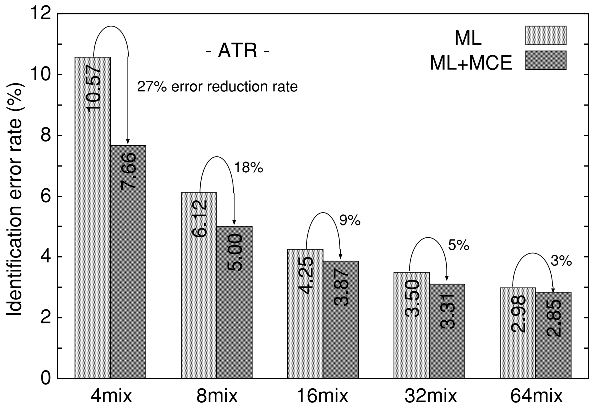


Fig. 8 Comparison of **Ψ**-shared MFA before and after MCE training (q = 2).

In this paper, factor loading matrices or diagonal covariance matrices were shared in all the mixture components. We believe that more effective parameter tying strategies will be available with clustering techniques. In [4], two clustering methods for factor loading matrices were evaluated: a target driven method and a hierarchical cluster tree method. In the hierarchical cluster tree method, the tying structure is determined based on Gaussian distributions composed of FA models of mixture components. However, the structure obtained from the Gaussians is not always appropriate for sharing MFA model parameters. On the other hand, although the target driven method takes account of the model structure of MFA, it is difficult to apply it to **Ψ**-shared MFAs. Accordingly a further investigation is needed to find an appropriate clustering criterion for the different types sharing structures, though it is beyond the scope of this paper.

6. Conclusions

This paper has investigated the parameter tying structures of MFA for speaker identification and MCE training has been applied to the parameter shared MFA. Sharing diagonal covariance matrices provided the best performance leading to a relative gain of 26% over the GMM with diagonal covariance matrices. The MCE training has further improved the recognition performance.

Our future works include the application of other variations of MFA to speaker identification [10], and automatic determination of the optimal number of mixture components and factors using the variational Bayesian approach [11].

More flexible parameter tying structure will be obtained by clustering MFA model parameters.

Acknowledgment

The authors would also like to thank Professor Sadaoki Furui of Tokyo Institute of Technology, and Professor Tomoko Matsui of the Institute of Statistical Mathematics, for providing the NTT database.

References

[1] D.A. Reynolds and R.C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” IEEE Trans. Speech Audio Process., vol.3, no.1, pp.72–83, Jan. 1995.

[2] Z. Ghahramani and G.E. Hinton, “The EM algorithm for mixtures of factor analyzers,” Technical Report Univ. of Toronto, CRGTR-96-1, May 1996.

[3] L.K. Saul and M.G. Rahim, “Maximum likelihood and minimum classification error factor analysis for automatic speech recognition,” IEEE Trans. Speech Audio Process., vol.8, no.2, pp.115–125, March 2000.

[4] P. Ding, Y. Liu, and B. Xu, “Factor analyzed Gaussian mixture models for speaker identification,” Proc. ICSLP-2002, vol.2, pp.1341–1344, Sept. 2002.

[5] H. Yamamoto, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura, “Parameter sharing and minimum classification error training of mixtures of factor analyzers for speaker identification,” Proc. ICASSP-2004, vol.1, pp.29–32, May 2004.

[6] M.E. Tipping and C.M. Bishop, “Mixtures of probabilistic principal component analyzers,” Neural Comput., vol.11, no.2, pp.443–482, 1999.

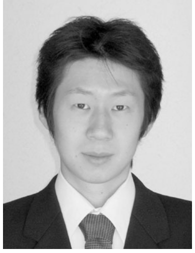
[7] B.-H. Juang and S. Katagiri, “Discriminative learning for minimum error classification,” IEEE Trans. Signal Process., vol.40, no.12, pp.3043–3054, Dec. 1992.

[8] T. Matsui and S. Furui, “Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs,” Proc. ICASSP ’92, vol.2, pp.157–160, March 1992.

[9] S. Furui, “Cepstral analysis technique for automatic speaker verification,” IEEE Trans. Acoust. Speech Signal Process., vol.29, no.2, pp.254–272, April 1981.

[10] A-V.I. Rosti and M.J.F. Gales, “Generalised linear Gaussian models,” Technical Report Cambridge Univ., CUED/F-INFENG/TR.420, Nov. 2001.

[11] Z. Ghahramani and M.J. Beal, “Variational inference for Bayesian mixtures of factor analysers,” Advances in Neural Inf. Process. Syst., vol.12, pp.449–455, 2000.



Hiroyoshi Yamamoto was born in 1980. He recieved the B.E. degree in Computer Science from the Nagoya Institute of technology, Nagoya, Japan in 2003. He is currently a Master’s candidate of the Nagoya Institute of technology. His research interests include speech and speaker recognition. He is a student member of the Acoustical Society of Japan.

Yoshihiko Nankaku received the B.E. degree in Computer Science, and the M.E. and Dr.Eng. degrees in the Department of Electrical and Electronic Engineering from the Nagoya Institute of Technology, Nagoya Japan, in 1999, 2001, and 2004 respectively. He is currently a postdoctoral fellow at the Nagoya Institute of Technology. His research interests include statistical machine learning, image recognition, speech recognition and synthesis and multimodal interface. He is a member of ASJ.

Chiyomi Miyajima received the B.E. degree in computer science and M.E. and Dr.Eng. degrees in electrical and computer engineering from Nagoya Institute of Technology, Nagoya, Japan, in 1996, 1998, and 2001, respectively. From 2001 to 2003, she was a Research Associate of the Department of Computer Science, Nagoya Institute of Technology. Currently she is a Research Associate of the Graduate School of Information Science, Nagoya University, Nagoya, Japan. Her research interests include automatic speaker recognition and multi-modal speech processing. She received the Awaya Award in 2000 from the Acoustical Society of Japan. She is a member of ASJ and JASL.

Keiichi Tokuda received the B.E. degree in electrical and electronic engineering from the Nagoya Institute of Technology, Nagoya, Japan, the M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986, and 1989, respectively. From 1989 to 1996 he was a Research Associate at the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. From 1996 to 2004 he was a Associate Professor at the Department of Computer Science, Nagoya Institute of Technology as Associate Professor, and now he is a Professor at the same institute. He is also an Invited Researcher at ATR Spoken Language Translation Research Laboratories, Japan and was a Visiting Researcher at Carnegie Mellon University from 2001 to 2002. He is a co-recipient of the Paper Award and the Inose Award both from the IEICE in 2001, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2001. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society. His research interests include speech speech coding, speech synthesis and recognition, and statistical machine learning.

Tadashi Kitamura received the B.E. degree in electronics engineering from Nagoya Institute of Technology, Nagoya, in 1973, and M.E. and Dr.Eng. degrees from Tokyo Institute of Technology, Tokyo, in 1975 and 1978, respectively. In 1978 He joined the Research Laboratory of Precision Machinery and Electronics, Tokyo Institute of Technology, as a Research Associate. In 1983 he joined Nagoya Institute of Technology, as a Assistant Professor. He is currently a Professor of Graduate School of Engineering of Nagoya Institute of Technology. His current interests include speech processing, image processing and multi-modal biometrics. He is a member of IEEE, ISCA, ASJ, IPSJ and ITE.