PARAMETER SHARING AND MINIMUM CLASSIFICATION ERROR TRAINING OF MIXTURES OF FACTOR ANALYZERS FOR SPEAKER IDENTIFICATION

Hiroyoshi Yamamoto†, *Yoshihoko Nankaku*†, *Chiyomi Miyajima*‡, *Keiichi Tokuda*†, *and Tadashi Kitamura*†

† Department of Computer Science and Engineering, Graduate School of Engineering
Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan

‡ Department of Media Science, Graduate School of Information Science
Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

ABSTRACT

This paper investigates the parameter tying strategies of mixtures 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. The minimum classification error (MCE) training is applied to the MFA parameters to enhance the discrimination abilities. The results of text-independent speaker identification experiments show that MFA outperforms the conventional Gaussian mixture models (GMMs) with diagonal or full covariance matrices and achieve the best performance when sharing the diagonal matrices, resulting in a relative gain of 26% over the GMM with diagonal covariance matrices. The recognition performance is further improved by the MCE training with an additional 3% error reduction.

1. INTRODUCTION

Gaussian mixture models (GMMs) are 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. In order to cope with the problem, mixtures of factor analyzers (MFA) [2] have been applied to speech as well as speaker recognition [3], [4]. MFA allows us to reduce the degree of freedom of the covariance matrices maintaining the recognition performance. Moreover, the reliability of the estimated parameters can be improved by sharing parameters in different mixture components of MFA.

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

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

In addition, minimum classification error (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 tying strategies, respectively. Section 4 presents the MCE training of MFA, and the experimental results are reported in Section 5. Finally, conclusions and future works are given in Section 6.

This work was partially supported by a Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science, Encouragement of Young Scientists.

2. MIXTURES 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. In FA, a d-dimensional 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 $\boldsymbol{n} = (n_1, n_2, \ldots, n_d)^T$:

$$\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{n}, \tag{1}$$

where $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_q)$, $\boldsymbol{w}_i = (w_{i1}, w_{i2}, \ldots, w_{id})^T$ is a $p \times d$ matrix known as a factor loading matrix, and $\boldsymbol{z}$ 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 $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Psi})$, where $\boldsymbol{\mu}$ denotes a mean vector and $\boldsymbol{\Psi}$ 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

MFA is defined as mixtures of M factor analyzers. The likelihood of T independent feature vectors $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T)$ 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

$$p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \prod_{t=1}^{T} \sum_{m=1}^{M} \int p_m(\boldsymbol{x}_t \mid \boldsymbol{z}) p_m(\boldsymbol{z}) c_m d\boldsymbol{z}, \tag{4}$$

where c_m denotes the weight of the m-th mixture component.

3. TYING STRATEGIES

Covariance matrices $\boldsymbol{\Sigma}_m$ of MFA consist of $\boldsymbol{W}_m$ and $\boldsymbol{\Psi}_m$. In this section, we present the various tying strategies of these parameters in MFA. We assume that all the mixture components have the same number of factors, and compare the following three kinds of MFA with different parameter sharing structures.

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

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 $\boldsymbol{z}$ and the posterior of the m-th mixture component:

$$\langle \boldsymbol{z}_{tm} \rangle = E[\boldsymbol{z}|\boldsymbol{x}_t, m] = \boldsymbol{\beta}_m(\boldsymbol{x}_t - \boldsymbol{\mu}_m), \tag{5}$$

$$\begin{aligned}\langle \boldsymbol{z}\boldsymbol{z}_{tm} \rangle &= E[\boldsymbol{z}\boldsymbol{z}^T|\boldsymbol{x}_t, m] \\ &= I - \boldsymbol{\beta}_m \boldsymbol{W}_m + \langle \boldsymbol{z}_{tm} \rangle \langle \boldsymbol{z}_{tm} \rangle^T,\end{aligned} \tag{6}$$

$$h_{tm} = \frac{c_m \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_m c_m \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}, \tag{7}$$

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) Generic MFA

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

$$\tilde{\boldsymbol{W}}_m = (\boldsymbol{W}_m \ \boldsymbol{\mu}_m) \tag{8}$$

$$\tilde{\boldsymbol{z}}_{tm} = \begin{pmatrix} \boldsymbol{z} \\ 1 \end{pmatrix} \tag{9}$$

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

$$\tilde{\boldsymbol{W}}'_m = \left(\sum_t h_{tm} \boldsymbol{x}_t \langle \tilde{\boldsymbol{z}}_{tm} \rangle^T \right) \cdot \left(\sum_l h_{lm} \langle \tilde{\boldsymbol{z}}\tilde{\boldsymbol{z}}_{lm} \rangle \right)^{-1}, \tag{10}$$

$$\boldsymbol{\Psi}'_m = \frac{1}{\sum_t h_{tm}} \mathrm{diag} \left\{ \sum_t h_{tm} \left(\boldsymbol{x}_t - \tilde{\boldsymbol{W}}'_m \langle \tilde{\boldsymbol{z}}_{tm} \rangle \right) \boldsymbol{x}_t^T \right\}, \tag{11}$$

where

$$\langle \tilde{\boldsymbol{z}}_{tm} \rangle = \begin{pmatrix} \langle \boldsymbol{z}_{tm} \rangle \\ 1 \end{pmatrix}, \tag{12}$$

$$\langle \tilde{\boldsymbol{z}}\tilde{\boldsymbol{z}}_{tm} \rangle = \begin{pmatrix} \langle \boldsymbol{z}\boldsymbol{z}_{tm} \rangle & \langle \boldsymbol{z}_{tm} \rangle \\ \langle \boldsymbol{z}_{tm} \rangle & 1 \end{pmatrix}, \tag{13}$$

and $\mathrm{diag}(\cdot)$ denotes setting the elements outside the main diagonal to zeros. The mixture weight c_m is re-estimated as follows.

$$c'_m = \frac{1}{T} \sum_{t=1}^{T} h_{tm} \tag{14}$$

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

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

$$\boldsymbol{\Psi}' = \frac{1}{T} \mathrm{diag} \left\{ \sum_{t,m} h_{tm} \left(\boldsymbol{x}_t - \tilde{\boldsymbol{W}}'_m \langle \tilde{\boldsymbol{z}}_{tm} \rangle \right) \boldsymbol{x}_t^T \right\}. \tag{15}$$

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

$$\begin{aligned}\boldsymbol{W}'_{(k)} = &\left(\sum_{t,m} h_{tm} \boldsymbol{\Psi}^{-1}_{m(k)} (\boldsymbol{x}_t - \boldsymbol{\mu}_m)_{(k)} \langle \boldsymbol{z}_{tm} \rangle^T \right) \\ &\cdot \left(\sum_{t,m} h_{tm} \boldsymbol{\Psi}^{-1}_{m(k)} \langle \boldsymbol{z}\boldsymbol{z}_{tm} \rangle \right)^{-1}\end{aligned} \tag{16}$$

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:

$$\boldsymbol{\mu}'_m = \frac{\sum_t h_{tm} (\boldsymbol{x}_t - \boldsymbol{W}' \langle \boldsymbol{z}_{tm} \rangle)}{\sum_t h_{tm}}, \tag{17}$$

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

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 [5] is applied to the parameters of MFA [3].

4.1. Definition of Loss Function

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

$$d_s(\boldsymbol{X}; \boldsymbol{\Theta}) = -g_s(\boldsymbol{X}; \boldsymbol{\Theta}) + \max_{y \neq s} g_y(\boldsymbol{X}; \boldsymbol{\Theta}), \tag{19}$$

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 $\boldsymbol{\theta}_s$. Equation (19) 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:

$$l_s(\boldsymbol{X}; \boldsymbol{\theta}) = \big(1 + \exp(-\gamma \cdot d_s)\big)^{-1}, \tag{20}$$

where γ denotes the gradient of the sigmoid function. 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 $\tilde{\boldsymbol{\Theta}}$.

$$\tilde{\boldsymbol{\Theta}} = \{\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \ldots, \tilde{\boldsymbol{\theta}}_S\}, \tag{21}$$

$$\tilde{\boldsymbol{\theta}} = \{\tilde{c}_m, \tilde{\boldsymbol{\mu}}_m, \boldsymbol{W}_m, \tilde{\boldsymbol{\Psi}}_m \mid m = 1, 2, \ldots, M\}, \tag{22}$$

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

$$\tilde{\boldsymbol{\Theta}}(r+1) = \tilde{\boldsymbol{\Theta}}(r) - \varepsilon_r \nabla l_s(\boldsymbol{X}; \tilde{\boldsymbol{\theta}}), \tag{23}$$

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 (23) is obtained as follows.

$$\nabla_{\tilde{\boldsymbol{\theta}}_y} l_s(\boldsymbol{X}; \tilde{\boldsymbol{\theta}}) = \frac{\partial l_s}{\partial d_s} \frac{\partial d_s}{\partial g_y} \cdot \nabla_{\tilde{\boldsymbol{\theta}}_y} g_y(\boldsymbol{X}; \tilde{\boldsymbol{\theta}}), \tag{24}$$

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

$$\frac{\partial l_s}{\partial d_s} = \gamma l_s (1 - l_s), \quad \frac{\partial d_s}{\partial g_y} = \begin{cases} -1, & y = s \\ 1, & y \neq s \end{cases}, \tag{25}$$

$$\nabla_{\tilde{\boldsymbol{\theta}}_y} g_y(\boldsymbol{X}; \tilde{\boldsymbol{\theta}}) = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{b_y(\boldsymbol{x}_t)} \nabla_{\tilde{\boldsymbol{\theta}}_y} b_y(\boldsymbol{x}_t). \tag{26}$$

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) generic MFA

For the generic MFA, the gradients are obtained as follows.

$$\frac{\partial b(\boldsymbol{x}_t)}{\partial \tilde{c}_m} = f_m, \quad \frac{\partial b(\boldsymbol{x}_t)}{\partial \tilde{\mu}_{mi}} = f_m \delta_{mi} \Sigma_{mii}, \tag{27}$$

$$\frac{\partial b(\boldsymbol{x}_t)}{\partial W_{mij}} = -f_m \left\{ (\boldsymbol{\Sigma}_m^{-1} \boldsymbol{W}_m)_{ij} - \delta_{mi} [\boldsymbol{\delta}_m^T \boldsymbol{W}_m]_j \right\}, \tag{28}$$

$$\frac{\partial b(\boldsymbol{x}_t)}{\partial \tilde{\Psi}_{mii}} = -\frac{1}{2} f_m \left\{ \Sigma_{mii}^{-1} - \delta_{mi}^2 \right\} \Psi_{mii}, \tag{29}$$

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 obtained by follows: (27) and (28), respectively, and only (29) is changed as

$$\frac{\partial b(\boldsymbol{x}_t)}{\partial \tilde{\Psi}_{ii}} = \sum_{m=1}^{M} \frac{\partial b(\boldsymbol{x}_t)}{\partial \tilde{\Psi}_{mii}}. \tag{30}$$

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

The gradients in (27) and (29) apply to the $\boldsymbol{W}$-shared MFA case, and (28) is changed as follows.

$$\frac{\partial b(\boldsymbol{x}_t)}{\partial W_{ij}} = \sum_{m=1}^{M} \frac{\partial b(\boldsymbol{x}_t)}{\partial W_{mij}} \tag{31}$$

5. EXPERIMENTAL EVALUATION

5.1. Database and Experimental Conditions

A text-independent speaker identification experiment was conducted for 80 speakers (40 males and 40 females) in the ATR Japanese speech database. 216 words were used for training each speaker model, and 520 words were used for testing. The number of tests was 41600 in total. The speech data was down-sampled from 20kHz to 10kHz, windowed at a 10-ms frame rate using a 25-ms Blackman window, and parameterized into 12 mel-cepstral coefficients excluding zero-th coefficients with a mel-cepstral analysis technique.

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 $\boldsymbol{\Sigma}$ [2]. The number of mixture components was changed from 4 to 64, and the number of factors was changed from 2 to 10.

5.2. Results

Figures 1–3 compare the identification error rates between the three kinds of MFA and the conventional GMMs with full or diagonal covariance matrices (full-GMM and diag-GMM). All speaker models in Figs.1–3 were trained with 216 words based on ML-estimation. Figure 1 compares the results of generic MFA with the conventional GMMs, where 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. Generic MFA show better performance with a smaller number of factors. However, the error rate of generic MFA with a larger number of factors is close to that of full-GMM, because the model structure of generic MFA with $q = 12$ is almost the same as the full-GMM. Figure 2 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 3 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.

Figure 4 shows the results of the three kinds of MFA with the number of factors $q = 2$ and diag-GMM, where the amount of training data was changed as 27, 54, and 216 words. We can see that the MFA-based speaker models show relatively high performance with such a small number of factors, and all the MFA-based speaker models outperform the conventional GMM with any amount of training data. Among the three kinds of MFA, $\boldsymbol{\Psi}$-shared MFA achieves the best performance and a significant difference is found with smaller amounts of training data.

Finally, MCE training is applied to the MFA-based speaker models. Figure 5 compares the performance of $\boldsymbol{\Psi}$-shared MFA with six factors before and after the MCE training. We can see that the performance is further improved by the MCE training and the 2.73% error rate was reduced to 2.65% with a 3% error reduction.

6. CONCLUSIONS

This paper has investigated the parameter tying strategies 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 [6] and automatic determination of the optimal number of mixture components and factors using the variational Bayesian approach [7].

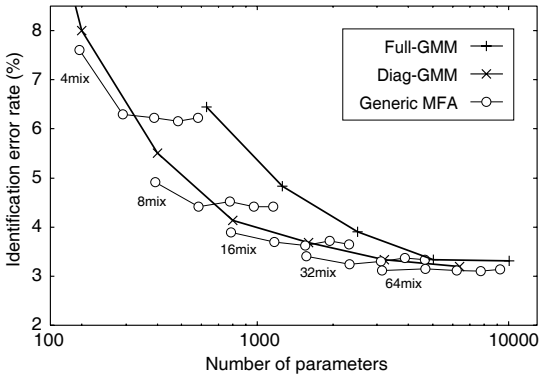

Fig. 1. Comparison between generic MFA and conventional GMMs with diagonal or full covariance matrices.

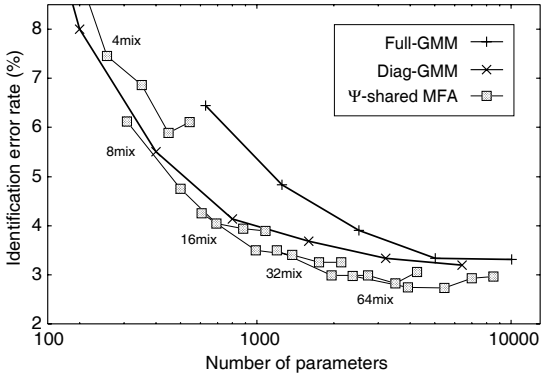

Fig. 2. Comparison between $\mathbf{\Psi}$-shared MFA and conventional GMMs with diagonal or full covariance matrices.

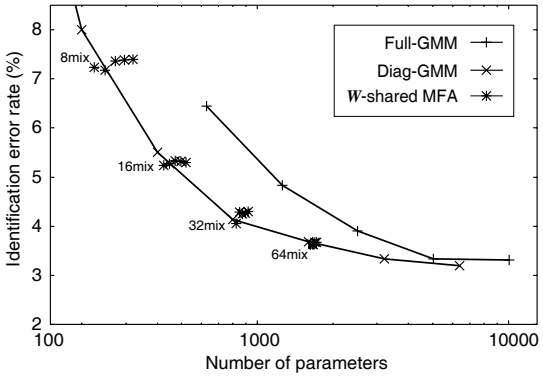

Fig. 3. Comparison between $\boldsymbol{W}$-shared MFA and conventional GMMs with diagonal or full covariance matrices.

7. REFERENCES

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

[2] Z. Ghahramani and G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," *Tech. Rep. Univ. of Toronto*, CRGTR-96-1, May 1996.

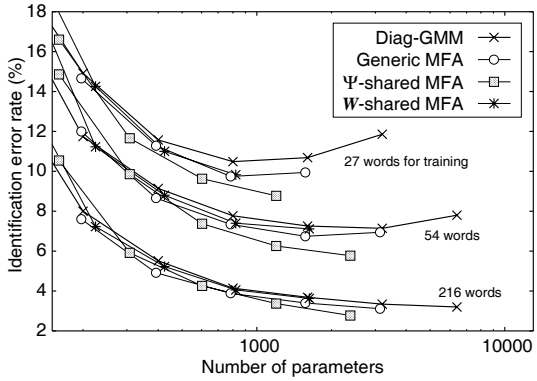

Fig. 4. Comparison between diag-GMM and three kinds MFA ($q = 2$) with increasing the number of mixtures, using 27 words (upper), 54 words (middle), 216 words (lower) for training.

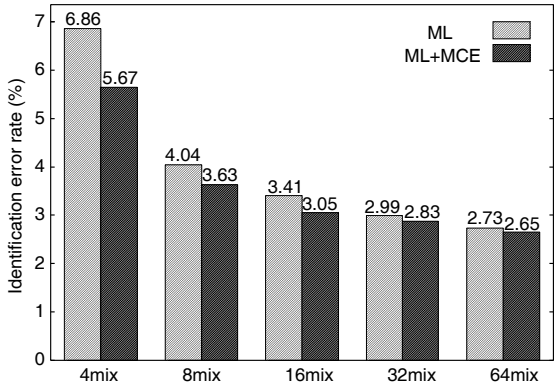

Fig. 5. Comparison of $\mathbf{\Psi}$-shared MFA before and after MCE training ($q = 6$).

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

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

[5] 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.

[6] A-V. I. Ṙosti and M.J.F. Gales, "Generalised linear Gaussian models," *Tech. Rep. Cambridge Univ.*, CUED/F-INFENG/TR.420, Nov. 2001.

[7] Z. Ghahramani and M.J. Beal, "Variational inference for Bayesian mixtures of factor analysers," *Neural Information Processing Systems 12*, 1999.