Voice Conversion Based on Mixtures of Factor Analyzers

Yosuke Uto[†], *Yoshihiko Nankaku*[†], *Tomoki Toda*[‡], *Akinobu Lee*[†], *and Keiichi Tokuda*[†]

[†] Nagoya Institute of Technology Gokiso-cho, Showa-ku, Nagoya, Aichi, 466-8555 Japan
[‡]Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara, 630-0101 Japan
{uto, nankaku, ri, tokuda}@ics.nitech.ac.jp
tomoki@is.naist.jp

Abstract

This paper describes the voice conversion based on the Mixtures of Factor Analyzers (MFA) which can provide an efficient modeling with a limited amount of training data. As a typical spectral conversion method, a mapping algorithm based on the Gaussian Mixture Model (GMM) has been proposed. In this method two kinds of covariance matrix structures are often used : the diagonal and full covariance matrices. GMM with diagonal covariance matrices requires a large number of mixture components for accurately estimating spectral features. On the other hand, GMM with full covariance matrices needs sufficient training data to estimate model parameters. In order to cope with these problems, we apply MFA to voice conversion. MFA can be regarded as intermediate model between GMM with diagonal covariance and with full covariance. Experimental results show that MFA can improve the conversion accuracy compared with the conventional GMM.

Index Terms: voice conversion, GMM (Gaussian Mixture Model), MFA (Mixtures of Factor Analyzers)

1. Introduction

Voice conversion is a potential technique for flexibly synthesizing various types of speech. This technique can modify speech characteristics using conversion rules statistically extracted from a small amount of training data. As a typical spectral conversion method, a mapping algorithm based on the Gaussian Mixture Model (GMM) has been proposed [1]. In this method, the mapping between spectral features of the source and target is determined based on GMM. In each mixture component, the conditional mean vector of target features given source features is calculated as a simple linear transformation using the covariance matrix of the concatenated feature vector. The converted vector is defined as the weighted sum of the conditional mean vectors, and the conditional occupancy probabilities of mixture components are used as weights. More accurate formularization of spectral conversion based on ML (Maximum Likelihood) criterion has been presented [2]. In the GMM-based techniques, it is important to determine the optimal number of mixtures and the structure of the covariance matrices. Typically two kinds of covariance matrices are used for training GMM: the diagonal and full covariance matrices. Although the diagonal covariance can reduce the number of parameters of each component , a large number of mixture components is required for accurate spectral estimation. On the other hand, full covariance matrices can represent the correlation between the source and target features with a few mixture components. However, it needs sufficient training data to estimate model parameters. In order to cope with these problems, we apply the Mixtures of Factor Analyzers (MFA) to the GMM-based voice conversion method. MFA can be regarded as intermediate model between GMM with diagonal and with full covariance matrices, and it provides an efficient modeling with a limited amount of training data.

The paper is organized as follows. Section 2 explains the voice conversion technique based on GMM. Section 3 describes the general formulation of MFA, and the voice conversion based on MFA. Experimental results are reported in Section 4. Finally, conclusions and future works are given in Section 5.

2. Voice Conversion Based on GMM

To convert spectral features of a source speaker X to a target speaker Y, the joint probability density of two speaker's features are modeled by GMM [3]. Let a vector $\boldsymbol{Z}_t = \left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top$ be a joint feature vector of the source one $\boldsymbol{X}_t$ and the target one $\boldsymbol{Y}_t$ at time t. In the GMM-based voice conversion, the vector sequence $\boldsymbol{Z} = \left[\boldsymbol{Z}_1^\top, \boldsymbol{Z}_2^\top, \ldots, \boldsymbol{Z}_T^\top\right]^\top$ is modeled by GMM $\boldsymbol{\lambda} = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \mid i = 1, 2, \ldots, M\}$. The output probability of $\boldsymbol{Z}$ given GMM $\boldsymbol{\lambda}$ can be written as follows:

$$p\left(\boldsymbol{Z}|\boldsymbol{\lambda}\right) = \prod_{t}^{T} \sum_{i}^{M} w_i \mathcal{N}(\boldsymbol{Z}_t|\boldsymbol{\mu}_i^{(Z)}, \boldsymbol{\Sigma}_i^{(Z)}) \tag{1}$$

$$\boldsymbol{\mu}_i^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_i^{(X)} \\ \boldsymbol{\mu}_i^{(Y)} \end{bmatrix}, \boldsymbol{\Sigma}_i^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(XX)} & \boldsymbol{\Sigma}_i^{(XY)} \\ \boldsymbol{\Sigma}_i^{(YX)} & \boldsymbol{\Sigma}_i^{(YY)} \end{bmatrix} \tag{2}$$

where M is the number of mixtures, w_i is the mixture weight of the i-th component, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is the mean vector and covariance matrix, respectively.

2.1. Maximum likelihood spectral conversion

In the maximum likelihood spectral conversion, the optimal sequence of the target feature vectors $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \ldots, \boldsymbol{Y}_T^\top\right]^\top$ given a source feature vector sequence $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \ldots, \boldsymbol{X}_T^\top\right]^\top$ is obtained by maximizing the following conditional distribution:

$$p\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}\right) = \sum_{all\,\boldsymbol{m}} p(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}) p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}) \tag{3}$$

$$= \prod_{t}^{T} \sum_{i}^{M} p(m_t = i|\boldsymbol{X}_t, \boldsymbol{\lambda}) p(\boldsymbol{Y}_t|\boldsymbol{X}_t, m_t = i, \boldsymbol{\lambda}) \tag{4}$$

where $\boldsymbol{m} = (m_1, m_2, \ldots, m_T)$ is a mixture number sequence. The conditional distribution can also be written as GMM, and its output probability distribution is presented as follows:

$$p\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, m_t = i, \boldsymbol{\lambda}\right) = \mathcal{N}(\boldsymbol{Y}_t; \boldsymbol{E}_i(t), \boldsymbol{D}_i) \tag{5}$$

This work was partly supported by MEXT e-Society project.

$$
\begin{aligned}
\boldsymbol{E}_i(t) &= \boldsymbol{\mu}_t^{(Y)} + \boldsymbol{\Sigma}_i^{(YX)} \boldsymbol{\Sigma}_i^{(XX)^{-1}} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_i^{(X)} \right) && (6) \\
\boldsymbol{D}_i &= \boldsymbol{\Sigma}_i^{(YY)} - \boldsymbol{\Sigma}_i^{(YX)} \boldsymbol{\Sigma}_i^{(XX)^{-1}} \boldsymbol{\Sigma}_i^{(XY)} && (7)
\end{aligned}
$$

Since the equation (3) includes latent variables, the optimal sequence of $\boldsymbol{Y}$ is estimated via the EM algorithm. The EM algorithm is an iterative method for approximating the maximum likelihood estimation. It maximizes the expectation of the complete data log-likelihood so called $\mathcal{Q}$-function (auxiliary function):

$$
\mathcal{Q}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{all\, \boldsymbol{m}} p(\boldsymbol{Y}, \boldsymbol{m} | \boldsymbol{X}, \boldsymbol{\lambda}) \log p \left(\hat{\boldsymbol{Y}}, \boldsymbol{m} | \boldsymbol{X}, \boldsymbol{\lambda} \right) \tag{8}
$$

Taking the derivative of the $\mathcal{Q}$-function, the spectral sequence $\hat{\boldsymbol{Y}}$ which maximizes the $\mathcal{Q}$-function is given by

$$
\hat{\boldsymbol{Y}} = \left(\overline{\boldsymbol{D}^{-1}} \right)^{-1} \overline{\boldsymbol{D}^{-1} \boldsymbol{E}} \tag{9}
$$

where

$$
\begin{aligned}
\overline{\boldsymbol{D}^{-1}} &= \mathrm{diag} \left[\overline{\boldsymbol{D}_1^{-1}}, \overline{\boldsymbol{D}_2^{-1}}, \cdots, \overline{\boldsymbol{D}_T^{-1}} \right] && (10) \\
\overline{\boldsymbol{D}_t^{-1}} &= \sum_{i=1}^{M} \gamma_i(t) \boldsymbol{D}_i^{-1} && (11) \\
\overline{\boldsymbol{D}^{-1} \boldsymbol{E}} &= \left[\overline{\boldsymbol{D}^{-1} \boldsymbol{E}_1}^{\top}, \overline{\boldsymbol{D}^{-1} \boldsymbol{E}_2}^{\top}, \cdots, \overline{\boldsymbol{D}^{-1} \boldsymbol{E}_T}^{\top} \right]^{\top} && (12) \\
\overline{\boldsymbol{D}^{-1} \boldsymbol{E}_t} &= \sum_{i=1}^{M} \gamma_i(t) \boldsymbol{D}_i^{-1} \boldsymbol{E}_i(t) && (13) \\
\gamma_i(t) &= p(m_t = i | \boldsymbol{X}_t, \boldsymbol{Y}_t, \boldsymbol{\lambda}) && (14)
\end{aligned}
$$

2.2. Maximum likelihood spectral estimation using dynamic features

In this paper, we use the spectral estimation technique using dynamic features as described in [2]. Let $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^{\top}, \Delta \boldsymbol{x}_t^{\top} \right]^{\top}$ and $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^{\top}, \Delta \boldsymbol{y}_t^{\top} \right]^{\top}$ be a source and a target feature vector with dynamic features, respectively. Where $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ denote static features, and the notation $\Delta\, \cdot$ represents the first order dynamic features calculated from the neighboring frames of time t. The relation between the static feature vector sequence $\boldsymbol{y} = [\boldsymbol{y}_1^{\top}, \boldsymbol{y}_2^{\top}, \ldots, \boldsymbol{y}_T^{\top}]^{\top}$ and the static and dynamic feature vector sequence $\boldsymbol{Y}$ can be represented as a linear transformation:

$$
\boldsymbol{Y} = \boldsymbol{W} \boldsymbol{y} \tag{15}
$$

where $\boldsymbol{W}$ is a matrix which concatenates dynamic features to the static feature sequence $\boldsymbol{y}$ [4]. Under this condition, the optimal static feature vector sequence $\hat{\boldsymbol{y}}$ which maximizes the $\mathcal{Q}$-function (equation (8)) is given by

$$
\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^{\top} \boldsymbol{D}^{-1} \boldsymbol{W} \right)^{-1} \boldsymbol{W}^{\top} \boldsymbol{D}^{-1} \boldsymbol{E} \tag{16}
$$

3. Voice Conversion Based on MFA

3.1. 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 [5]. In FA, a d-dimensional observation vector $\boldsymbol{o}$ is generated from a q-dimensional factor vector $\boldsymbol{a}(q < d)$ and an observation noise(d dimension), that is

$$
\boldsymbol{o} = \boldsymbol{C} \boldsymbol{a} + \boldsymbol{n} \tag{17}
$$

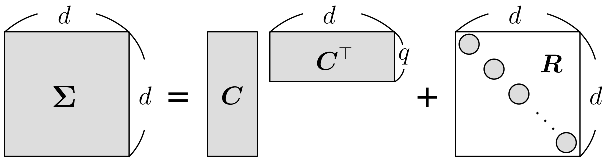

Figure 1: *Structure of covariance matrix in MFA*

where $\boldsymbol{C}$ is a $d \times q$ matrix called the factor loading matrix and q is the number of factors. It is assumed that the factor $\boldsymbol{a}$ and the noise $\boldsymbol{n}$ are distributed according to a Gaussian distribution:

$$
\begin{aligned}
p(\boldsymbol{a}) &= \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) && (18) \\
p(\boldsymbol{n}) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{R}) && (19)
\end{aligned}
$$

where $\boldsymbol{R}$ is a diagonal matrix. The conditional probability of $\boldsymbol{x}$ given $\boldsymbol{a}$ is written by

$$
p(\boldsymbol{o} | \boldsymbol{a}) = \mathcal{N}(\boldsymbol{C} \boldsymbol{a} + \boldsymbol{\mu}, \boldsymbol{R}) \tag{20}
$$

because if the variable $\boldsymbol{a}$ is fixed in equation (17), the product $\boldsymbol{C}\boldsymbol{a}$ becomes a constant vector which is added to the observation noise vector $\boldsymbol{n}$. Therefore, the marginal distribution of $\boldsymbol{x}$ is calculated by integrating out the latent variable $\boldsymbol{a}$

$$
\begin{aligned}
p(\boldsymbol{o}) &= \int p(\boldsymbol{o} | \boldsymbol{a}) p(\boldsymbol{a}) d\boldsymbol{a} \\
&= \mathcal{N} \left(\boldsymbol{\mu}, \boldsymbol{C} \boldsymbol{C}^{\top} + \boldsymbol{R} \right) && (21)
\end{aligned}
$$

From the equation, it can be seen that FA is the Gaussian distribution with the constraint covariance matrix composed of the factor loading matrix and the diagonal matrix $\boldsymbol{R}$. Figure 1 shows the constraint structure of the covariance matrix.

3.2. Extension of FA to MFA

FA is an effective model for correlated data with Gaussian distribution provided the number of factors is appropriately selected. However, the data is not usually distributed according to a Gaussian distribution. To deal with this problem, FA is often extended to the Mixtures of Factor Analyzers (MFA). MFA is defined as the mixtures of M factor analyzers. The likelihood of T independent feature vector $\boldsymbol{O} = \left[\boldsymbol{o}_1^{\top}, \boldsymbol{o}_2^{\top}, \ldots, \boldsymbol{o}_t^{\top} \right]^{\top}$ for a M-component MFA $\boldsymbol{\lambda} = \{ w_i, \boldsymbol{\mu}_i, \boldsymbol{C}_i, \boldsymbol{R}_i | i = 1, 2, \ldots, M \}$ is given by

$$
\begin{aligned}
p(\boldsymbol{X} | \boldsymbol{\lambda}) &= \prod_{t}^{T} \sum_{i}^{M} \int w_i p_i(\boldsymbol{o}_t | \boldsymbol{a}) p_i(\boldsymbol{a}) d\boldsymbol{a} \\
&= \prod_{t}^{T} \sum_{i}^{M} w_i \mathcal{N} \left(\boldsymbol{\mu}_i, \boldsymbol{C}_i \boldsymbol{C}_i^{\top} + \boldsymbol{R}_i \right) && (22)
\end{aligned}
$$

Similarly to equation (22), MFA can be regarded as GMM in which the covariance matrices are constrained by the factor loading matrices $\boldsymbol{C}_i$ and a diagonal matrix $\boldsymbol{R}_i$. MFA with zero factor is equivalent to the diagonal covariance GMM, and as increasing the number of factors, MFA becomes similar to the full covariance GMM. This means that MFA can be regarded as intermediate model between GMM with diagonal covariance and with full covariance. Since MFA only differs from GMM in the structure of covariance matrices, MFA can be converted to the general GMM, and it can be easily applied to the GMM-based voice conversion by replacing the training process $\left(\boldsymbol{Z} = \boldsymbol{O}, \boldsymbol{\Sigma}_i^{(Z)} = \boldsymbol{C}_i \boldsymbol{C}_i^{\top} + \boldsymbol{R}_i \right)$.

3.3. EM algorithm for MFA

The maximum likelihood (ML) solution of MFA can be obtained by the expectation maximization (EM) algorithm. The EM steps for MFA parameters $\boldsymbol{\lambda}$ are summarized as follows.

3.3.1. E-step

The E-step calculates the expectation of the latent vector $\boldsymbol{a}$:

$$\begin{aligned} \langle \boldsymbol{a}_{ti} \rangle &= E[\boldsymbol{a}_t | \boldsymbol{o}_t, i] = \boldsymbol{\beta}_i (\boldsymbol{o}_t - \boldsymbol{\mu}_i) \quad (23) \\ \langle \boldsymbol{a}\boldsymbol{a}_{ti} \rangle &= E\left[\boldsymbol{a}_t \boldsymbol{a}_t^\top | \boldsymbol{o}_t, i\right] \\ &= \boldsymbol{I} - \boldsymbol{\beta}_i \boldsymbol{C}_i \langle \boldsymbol{a}_{ti} \rangle \langle \boldsymbol{a}_{ti} \rangle^\top \quad (24) \end{aligned}$$

and the posterior of the i-th mixture component:

$$\gamma_i(t) \quad \propto \quad w_i \mathcal{N}\left(\boldsymbol{o}_t \middle| \boldsymbol{\mu}_i, \boldsymbol{C}_i \boldsymbol{C}_i^\top + \boldsymbol{R}_i\right) \quad (25)$$

where $\beta_i = \boldsymbol{C}_i^\top \left(\boldsymbol{R}_i + \boldsymbol{C}_i \boldsymbol{C}_i^\top\right)^{-1}$.

3.3.2. M-step

In the M-step, the new model parameters $\boldsymbol{\mu}_i'$, $\boldsymbol{C}_i'$, $\boldsymbol{R}_i'$ and w_i'can be obtained by re-estimation formulas. Using the following representations

$$\begin{aligned} \tilde{\boldsymbol{C}}_i &= (\boldsymbol{C}_i \ \boldsymbol{\mu}_i) \quad (26) \\ \tilde{\boldsymbol{a}}_{ti} &= \begin{pmatrix} \boldsymbol{a} \\ 1 \end{pmatrix} \quad (27) \end{aligned}$$

the re-estimation of $\tilde{\boldsymbol{C}}_i'$ and $\boldsymbol{R}_i'$ can be written by

$$\tilde{\boldsymbol{C}}_i' = \left(\sum_t^T \gamma_i(t) \boldsymbol{o}_t \langle \tilde{\boldsymbol{a}}_{ti} \rangle^\top\right) \cdot \left(\sum_l^T \gamma_i(l) \langle \tilde{\boldsymbol{a}}\tilde{\boldsymbol{a}}_{li} \rangle\right)^{-1} \quad (28)$$

$$\boldsymbol{R}_i' = \frac{1}{\sum_t^T \gamma_i(t)} \mathrm{diag}\left\{\sum_t^T \gamma_i(t) \left(\boldsymbol{o}_t - \tilde{\boldsymbol{C}}_i' \langle \tilde{\boldsymbol{a}}_{ti} \rangle\right) \boldsymbol{o}_t^\top\right\} \quad (29)$$

where

$$\begin{aligned} \langle \tilde{\boldsymbol{a}}_{ti} \rangle &= \begin{pmatrix} \langle \boldsymbol{a}_{ti} \rangle \\ 1 \end{pmatrix} \quad (30) \\ \langle \tilde{\boldsymbol{a}}\tilde{\boldsymbol{a}}_{ti} \rangle &= \begin{pmatrix} \langle \boldsymbol{a}\boldsymbol{a}_{ti} \rangle & \langle \boldsymbol{a}_{ti} \rangle \\ \langle \boldsymbol{a}_{ti} \rangle & 1 \end{pmatrix} \quad (31) \end{aligned}$$

and $\mathrm{diag}(\cdot)$ denotes the operator to set the off-diagonal elements to zeros. The mixture weight w_i' is re-estimated as

$$w_i' = \frac{1}{T} \sum_t^T \gamma_i(t) \quad (32)$$

4. Experimental Evaluation

4.1. Experimental conditions

Voice conversion experiments on the ATR Japanese speech database were performed. Two male speakers are selected as a source and a target speaker. One sentence uttered by the both speakers was used for training and 50 sentences were used for evaluation. The speech data were down-sampled from 20KHz to 16KHz, windowed at a 5-ms frame rate using a 25-ms Blackman window, and parameterized into 24 mel-cepstral coefficients excepting the zero-th coefficients and their first order derivative were used as the dynamic features. The parameters of the conventional GMM were initialized using an LBG codebook. From preliminary experiments, the initial value of MFA parameters were determined as follows: The factor loading matrices were initialized with random values. The covariance matrices of the noise factors are given by the global covariance. The number of factors and mixtures are varied among 2, 8, 32, and 1, 2, 4, 8, 16, respectively.

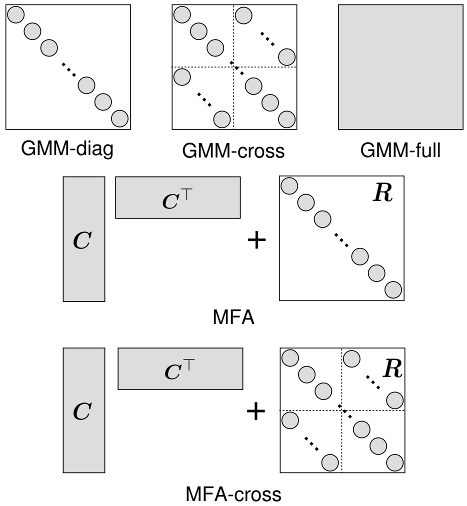

Figure 2: *Structure of covariance matrix for each model*

Figure 2 shows the structure of covariance matrices for each model. GMM-diag and GMM-full are GMM with diagonal and full covariance matrices, respectively. GMM-cross means GMM with covariance matrices which have the diagonal and cross-covariance elements. MFA is a generic MFA, and MFA-cross is MFA in which the covariance matrix of noise $\boldsymbol{R}$ is replaced with that of GMM-cross.

In this experiment, F0 sequences are converted by a simple linear conversion. Although the likelihood function in equation (3) is calculated by taking the sum of all possible mixture sequence $\boldsymbol{m}$, it was approximated by a single sequence which maximizes the output probability $p(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda})$.

4.2. Objective evaluation

The mel-cepstral distortion (Mel-CD) was used as an objective measure of the spectral conversion. The Mel-CD between the target mel-cepstram and estimated one is given by the following equation:

$$\text{Mel-CD} = \frac{10}{\log 10} \sqrt{2 \sum_{d=1}^{24} \left(mc_d^{(t)} - mc_d^{(e)}\right)^2} \quad (33)$$

where $mc_d^{(t)}$ and $mc_d^{(e)}$ denote the d-th coefficient of the target and the estimated mel-cepstra, respectively.

Figure 3 and 4 show the Mel-CD obtained by the spectral conversion with various structures of GMM and MFA. In the figures, each line indicates the change of the distortion with increasing the number of mixtures while keeping the number of factors fixed. In Figure 3, GMM-full with single mixture obtained a lower Mel-CD than GMM-diag due to the availability of the correlation. However, the performance of GMM-full was degraded as increasing the number of mixtures, because the lack of training data for each mixture components reduced the estimation accuracy. It can be

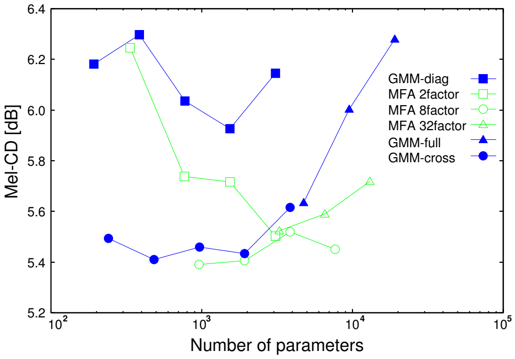

Figure 3: *Mel-CD of the conversion with GMM and MFA*

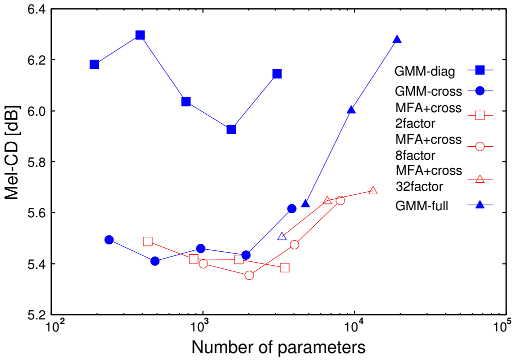

Figure 4: *Mel-CD of the conversion with GMM and MFA-cross*

also seen that the lines MFA is almost smoothly connected from GMM-diag to GMM-full as increasing the number of factors. This result confirms that MFA is intermediate model between GMM-diag and GMM-full, and by setting the appropriate number of factors and mixtures, MFA can improve the conversion accuracy.

Comparing GMM-cross with GMM-diag, the conversion accuracy of GMM-cross is better than GMM-diag. This result indicates that the diagonal elements of the cross-covariance are important for converting spectral features. Although both MFA and GMM-full can also represent the cross-covariance elements, MFA achieved an efficient modeling from a limited amount of data, due to the structure of the loading matrices.

From Figure 4, it seems that MFA-cross is better than MFA especially in the case that the number of factors is small. This is because MFA-cross can represent the diagonal elements of cross-covariance by using noise vector as well as GMM-cross. Furthermore, MFA-cross obtained a lower Mel-CD than GMM-cross in the same number of parameters, because MFA-cross can represent not only the diagonal elements of cross-covariance but also non-diagonal elements of covariance matrices.

4.3. Subjective evaluation

A DMOS (Differential Mean Opinion Score) test was performed for evaluating the similarity between the target and converted speech in speaker characteristics. In the test, the opinion score was set to a 5-point scale. Ten sentences were used for the evaluation set, and the number of listeners was 13.

Figure 5 shows the result of the DMOS test. For each model, the number of mixtures with the lowest Mel-CD in the objective test was selected. Although there is a large difference in the number of parameters, the score was not improved even if the number of mixtures was increased in preliminary experiments. It can be seen that GMM-full achieved a significantly higher score than GMM-diag. Although GMM-cross is better than GMM-full in the objective test, it could not improve the performance accuracy in the subjective test. In the results of MFA and MFA-cross, the score is improved as increasing the number of factors. Comparing MFA and MFA-cross, MFA-cross achieved a higher score than MFA in all number of factors. Although the score of MFA-cross with 32 factors is similar or slightly better than GMM-full, the number of parameters is significantly reduced. This means that the appropriate model structure was given dependently of the amount of training data.

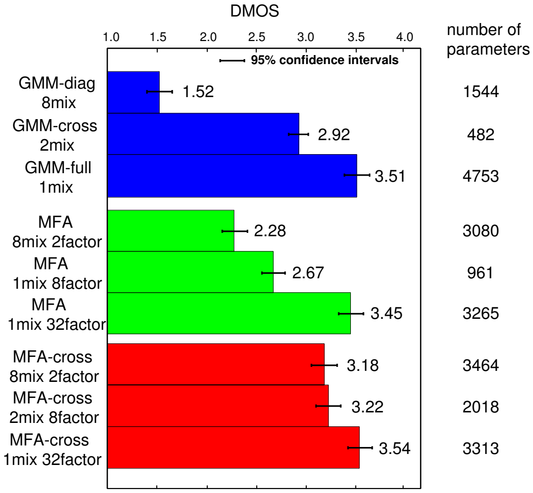

Figure 5: *Result of DMOS test*

5. Conclusion

In this paper, we proposed the voice conversion technique based on MFA. MFA can represent intermediate model between diagonal and full covariance GMM. In the objective test, the Mel-CD between the target and the estimated features was decreased by using MFA-cross. Although the score of MFA-cross in the DMOS test is similar or slightly better than full covariance GMM, MFA-cross can reduce the number of parameters and provide an appropriate model structure dependently of the amount of training data. Future works will focus on investigation parameter sharing of MFA.

6. References

[1] Yining Chen, Min Chu, Eric Chang, Jia Liu, and Runsheng Liu, “Voice conversion with smoothed GMM and MAP adaptation,” *Proc. of EUROSPEECH*, pp.2413–2416, Sep. 2003

[2] Tomoki Toda, Alan W Black, and Keiichi Tokuda, “Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis,” *Proc. of ISCA Speech Synthesis Workshop*, pp.31–36, Jun. 2004.

[3] Tomoki Toda, Alan W Black, and Keiichi Tokuda, “Spectral conversion based on maximum likelihood estimation considering global varinace of converted parameter,” *Proc. of ICASSP, vol.1*, pp.9–12, Mar. 2005.

[4] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” *Proc. of ICASSP, vol.3*, pp.1315–1318, Jun. 2000.

[5] Hiroyoshi Yamamoto, Yoshihiko Nankaku, Chiyomi Miyajima, Keiichi Tokuda, and Tadashi Kitamura, “Parameter Sharing and Minimum Classification Error Training of Mixtures of Factor Analyzers for Speaker Identification,” *Proc. of ICASSP, vol.1*, pp.29–32, Mar. 2004.