Application of Matrix Variate Gaussian Mixture Model to Statistical Voice Conversion

Daisuke Saito[1], *Hidenobu Doi*[1], *Nobuaki Minematsu*[2], *Keikichi Hirose*[1]

[1]Graduate School of Information Science and Technology, The University of Tokyo, Japan
[2]Graduate School of Engineering, The University of Tokyo, Japan
{dsk_saito,hdoi,mine,hirose}@gavo.t.u-tokyo.ac.jp

10.21437/Interspeech.2014-537

Abstract

This paper describes a novel approach to construct a mapping function between a given speaker pair using probability density functions (PDF) of matrix variate. In voice conversion studies, two important functions should be realized: 1) precise modeling of both the source and target feature spaces, and 2) construction of a proper transform function between these spaces. Voice conversion based on Gaussian mixture model (GMM) is the de facto standard because of their flexibility and easiness in handling. In GMM-based approaches, a joint vector space of the source and target is first constructed, and the joint PDF of the two vectors is modeled as GMM in the joint vector space. The joint vector approach mainly focuses on precise modeling of the 'joint' feature space, and does not always construct a proper transform between two feature spaces. In contrast, the proposed method constructs the joint PDF as GMM in a matrix variate space whose row and column respectively correspond to the two functions, and it has potential to precisely model both the characteristics of the feature spaces and the relation between the source and target spaces.

Index Terms: voice conversion, Gaussian mixture model, matrix variate distribution, matrix variate normal, matrix variate Gaussian mixture model

1. Introduction

Voice conversion (VC), or speaker conversion is a technique to transform an input utterance of a speaker to another utterance that sounds like another speaker with its linguistic content preserved [1]. Besides speech synthesis, VC techniques can be applied to various applications such as feature enhancement in ASR [2, 3]. Among several statistical approaches to construct the conversion model, approaches using Gaussian mixture model (GMM) are widely used because of their flexibility and easiness in handling [2, 4].

GMM-based techniques for statistical mapping use a mixture of Gaussians to model the probabilistic density function (PDF) of source feature vectors [2] or those of joint vectors of the source and the target vectors [4]. Both approaches derive a transformation function as a weighted sum of linear transformations from the constructed PDF. Each linear transformation corresponds to each Gaussian component, while the weights are calculated as posterior probabilities of source vectors. Since these approaches utilize Gaussian modeling as probabilistic densities, training algorithms for their models can be easily derived. In addition, several adaptation techniques based on maximum likelihood linear regression (MLLR) or maximum a posterior (MAP) adaptation [5, 6], and a target speaker model as prior knowledge [7] can be flexibly introduced.

In statistical voice conversion, two important functions should be achieved: to model the source and target feature spaces precisely, and to construct a proper transformation between these spaces. In the joint vector approach in GMM-based voice conversion [4], a joint vector space of the source and target is first constructed, and the joint PDF of the two vectors is modeled in a single vector space for the joint vector. That is to say, realization of the two functions in VC is implicitly founded on the concatenation of the source and target vectors. Once the source and the target feature vectors are concatenated, characteristics of the source space and those of the target space are not explicitly modeled in the training phase of the joint PDF. The approach can be regarded as precise modeling on the 'joint' feature space. However, since the dimensionality of the vector space is doubled, the approach easily suffers from overtraining effects when the complexity of the model is not properly configured. Although constraints such as assumption of diagonality in cross covariance matrices mitigate the problems, they are not always suitable for modeling of voice conversion. To realize the two functions in VC, multiple factors in the joint modeling should be explicitly modeled, particularly correlation in a feature space and relation between the source and target spaces.

In arbitrary speaker conversion, introduction of matrix representation has succeeded in dealing with multiple factors of acoustic variations. We have recently proposed a new representation of speaker space based on tensor analysis for arbitrary speaker conversion [8]. In our previous approach, an arbitrary speaker is not represented as a supervector, but as a matrix whose row and column dimensions respectively correspond to the component of GMM and the dimension of the mean vector. Inspired by this approach, the current paper introduces matrix representation to joint modeling of the joint PDF. The proposed method constructs the joint PDF as GMM in a matrix variate space whose row and column dimensions respectively correspond to the feature dimensionality and the speaker indices. The proposed approach has potential to precisely model both the characteristics of the two feature spaces and the relation between them.

2. Joint modeling of GMM for VC

In this section, the joint density GMM method is briefly described [4]. Let $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_x}]$ be a D-dimensional vector sequence characterizing an utterance from the source speaker, and $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n_y}]$ be that of the target speaker. Note that the two utterances contain the same linguistic content. The dynamic time warping algorithm (DTW) is applied to align the source vectors to their corresponding vectors in the target sequence. Then, a new sequence of $2D$-

dimensional joint vectors $\boldsymbol{z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n]$ where $\boldsymbol{z}_t = [\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]^\top$ is created. The notation $^\top$ denotes transposition of the vector, and t denotes a new time index after DTW is applied. The joint probability density of the source and the target vectors is modeled by a GMM for the joint vector $\boldsymbol{z}_t$ as follows:

$$P(\boldsymbol{z}_t|\boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}). \quad (1)$$

In Equation 1, $\mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)})$ denotes the normal distribution with mean vector $\boldsymbol{\mu}_m^{(z)}$ and covariance matrix $\boldsymbol{\Sigma}_m^{(z)}$, m is the mixture component index, and the total number of mixture components is M. The weight of the m-th component is w_m and $\sum_{m=1}^{M} w_m = 1$. $\boldsymbol{\lambda}^{(z)}$ denotes a parameter set of the GMM, which consists of weights, mean vectors, and covariance matrices for individual mixture components. Since the feature space of the joint vector $\boldsymbol{z}$ includes the feature spaces for the source and the target speakers as its subspaces, $\boldsymbol{\mu}_m^{(z)}$ and $\boldsymbol{\Sigma}_m^{(z)}$ are written as

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}, \quad (2)$$

where $\boldsymbol{\mu}_m^{(x)}$ and $\boldsymbol{\mu}_m^{(y)}$ are the mean vector of the m-th component for the source and that for the target, respectively. Similarly, $\boldsymbol{\Sigma}_m^{(xx)}$ and $\boldsymbol{\Sigma}_m^{(yy)}$ are the covariance matrices of the m-th component for the source and that for the target, respectively. The matrices $\boldsymbol{\Sigma}_m^{(xy)}$ and $\boldsymbol{\Sigma}_m^{(yx)}$ are the cross-covariance matrices of the m-th component for the source and the target. To mitigate the overfitting problems, the variance-covariance structures are sometimes constrained. For example, the covariance and cross-covariance matrices are restricted to diagonal matrices [9]. These parameters in the GMM are estimated by the EM algorithm using the sequence of the joint vectors ($\boldsymbol{z}$).

A mapping function $\mathcal{F}(\cdot)$ to convert the source vector $\boldsymbol{x}_t$ to the target vector $\boldsymbol{y}_t$ is derived based on the conditional probability density of $\boldsymbol{y}_t$, given $\boldsymbol{x}_t$. This probability density can be represented by the parameters of the joint density model as follows:

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}), \quad (3)$$

where

$$P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) = \frac{w_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}, \quad (4)$$

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(y)}, \boldsymbol{D}_m^{(y)}), \quad (5)$$

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}), \quad (6)$$

$$\boldsymbol{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)-1} \boldsymbol{\Sigma}_m^{(xy)}. \quad (7)$$

By minimizing the mean square error, the mapping function $\mathcal{F}$ is derived as

$$\mathcal{F}(\boldsymbol{x}_t) = \sum_{m=1}^{M} P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) \boldsymbol{E}_{m,t}^{(y)}. \quad (8)$$

When maximum likelihood estimation is adopted for parameter generation instead [9], the covariance matrix of the conditional probability density in Equation 7 is also taken into account and the target parameters are generated by the following updating equations:

$$\hat{\boldsymbol{y}}_t = \left(\sum_{m=1}^{M} \beta_{m,t} \boldsymbol{D}_m^{(y)-1} \right)^{-1} \left(\sum_{m=1}^{M} \beta_{m,t} \boldsymbol{D}_m^{(y)-1} \boldsymbol{E}_{m,t}^{(y)} \right),$$
$$\beta_{m,t} = P(m|\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{\lambda}^{(z)}). \quad (9)$$

Compared with Equation 8, in parameter generation by Equation 9, the inverse covariance matrices of each Gaussian component are contained, and they play a role of a kind of confidence measures for the conditional mean vectors in Equation 6.

3. Matrix variate GMM for VC

3.1. Matrix variate normal distribution

In this section, statistical modeling based on matrix variate is introduced to voice conversion. First, we introduce some fundamentals of matrix variate probability density functions [10]. Let $\boldsymbol{X}$ be a random matrix, whose row and column sizes are n and p respectively. In addition, let $\boldsymbol{M}, \boldsymbol{U}, \boldsymbol{V}$ be an $n \times p$, $n \times n$, $p \times p$ matrices, respectively, with $\boldsymbol{U}$ and $\boldsymbol{V}$ are positive definite. $\boldsymbol{X}$ has a matrix normal distribution with parameters $\boldsymbol{M}, \boldsymbol{U}$, and $\boldsymbol{V}$, if $\boldsymbol{X}$ has a moment generating function as follows:

$$M_X(\boldsymbol{T}) = \exp\left\{ \mathrm{tr}\left(\boldsymbol{M}^\top \boldsymbol{T}\right) + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{T}^\top \boldsymbol{U} \boldsymbol{T} \boldsymbol{V}\right) \right\}, \quad (10)$$

where $\boldsymbol{T}$ is an $n \times p$ matrix. The following notation

$$\boldsymbol{X} \sim \mathcal{N}_{\mathrm{mv}}(\boldsymbol{X}; \boldsymbol{M}, \boldsymbol{U}, \boldsymbol{V}) \quad (11)$$

is used hereafter. A property of the matrix variate normal distribution corresponding to that of vector variate normal distribution, i.e. our familiar Gaussian distribution, is derived from the vec operator and the Kronecker product. Equation 11 is equivalent to the following probability density for the vector $\mathrm{vec}(\boldsymbol{X})$;

$$P(\mathrm{vec}(\boldsymbol{X})|\boldsymbol{\lambda}) = \mathcal{N}(\mathrm{vec}(\boldsymbol{X}); \mathrm{vec}(\boldsymbol{M}), \boldsymbol{V} \otimes \boldsymbol{U}). \quad (12)$$

where $\mathrm{vec}()$ is the vec-operator that stacks the columns of a matrix into a vector, and $\boldsymbol{\lambda}$ denotes a parameter set. Finally, the probability density function for matrix $\boldsymbol{X}$ is written as

$$P(\boldsymbol{X}|\boldsymbol{\lambda}) = c^{-1} \exp\left[-\frac{1}{2} \mathrm{tr}\left\{ \boldsymbol{U}^{-1} (\boldsymbol{X} - \boldsymbol{M}) \boldsymbol{V}^{-1} (\boldsymbol{X} - \boldsymbol{M}) \right\} \right]$$
where $c = (2\pi)^{(1/2)np} |\boldsymbol{U}|^{(1/2)p} |\boldsymbol{V}|^{(1/2)n}$. (13)

When matrix samples $\mathcal{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_T]$ are observed as generated results of Equation 13, maximum likelihood estimator for $\boldsymbol{M}, \boldsymbol{U}$, and $\boldsymbol{V}$ are derived as follows:

$$\hat{\boldsymbol{M}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{X}_t, \quad (14)$$

$$\hat{\boldsymbol{U}} = \frac{1}{pT} \sum_{t=1}^{T} (\boldsymbol{X}_t - \hat{\boldsymbol{M}}) \hat{\boldsymbol{V}}^{-1} (\boldsymbol{X}_t - \hat{\boldsymbol{M}})^\top, \quad (15)$$

$$\hat{\boldsymbol{V}} = \frac{1}{nT} \sum_{t=1}^{T} (\boldsymbol{X}_t - \hat{\boldsymbol{M}})^\top \hat{\boldsymbol{U}}^{-1} (\boldsymbol{X}_t - \hat{\boldsymbol{M}}). \quad (16)$$

According to Equation 12, matrix variate normal distribution is not identical to Gaussian distribution with free covariance structure, but it supplies the PDF with a regular structure based on the Kronecker product. This 'separatable' structure for

the variance-covariance matrix can give the explicitly different characteristics to rows and columns. Namely, $\boldsymbol{U}$ and $\boldsymbol{V}$ capture variance-covariance structures in row and column directions, respectively. In addition, from Equations 15 and 16, an advantage for parameter estimation is revealed. Although the number of matrix samples is T in the above case, the effective numbers of samples to estimate $\boldsymbol{U}$ and $\boldsymbol{V}$ are pT and nT, respectively. That is to say, proper introduction of matrix variate modeling is expected to realize more efficient and precise inference.

3.2. Matrix variate GMM

In a similar manner of expanding the single Gaussian to GMM, a mixture model of matrix variate normal distribution can be derived [11]. This paper calls the model as matrix variate Gaussian mixture model (MV-GMM). Here, the joint density modeling based on MV-GMM for VC is described. Similarly to Section 2, let $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_x}]$ and $\boldsymbol{y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n_y}]$ be the feature vectors of the source and target speakers, respectively. In the proposed method, after DTW alignment between the source and target sequences, a new sequence is constructed as a sequence of joint matrices $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n]$ where $\boldsymbol{Z}_t = [\boldsymbol{x}_t, \boldsymbol{y}_t] \in \mathcal{R}^{D \times S}$. The notation S denotes the number of speakers and $S = 2$ in the above case. Note that the proposed approach can be easily expanded into model training using multiple speakers unlike the joint vector approach. The joint density is modeled by an MV-GMM for the joint matrix $\boldsymbol{Z}_t$ as follows:

$$P(\boldsymbol{Z}_t|\boldsymbol{\lambda}^{(Z)}) = \sum_{m=1}^{M} w_m \mathcal{N}_{\mathrm{mv}}(\boldsymbol{Z}_t; \boldsymbol{M}_m, \boldsymbol{U}_m, \boldsymbol{V}_m). \quad (17)$$

From Equation 17, the joint density model is represented as weighted sum of the matrix variate normal distribution of each component. In general, the notations in Equation 17 are the same as those in Equation 1 except for $\boldsymbol{U}_m$ and $\boldsymbol{V}_m$. The matrix $\boldsymbol{U}_m \in \mathcal{R}^{D \times D}$ is the covariance matrix of the m-th component representing variance-covariance structures of the feature space. The matrix $\boldsymbol{V}_m \in \mathcal{R}^{S \times S}$ is the covariance matrix representing the correlation between the source and the target speakers. The EM algorithm is used to estimate these parameters, and the update equations have a similar form to Equations 14–16 as follows:

$$\gamma_{m,t} = \frac{w_m \mathcal{N}_{\mathrm{mv}}(\boldsymbol{Z}_t; \boldsymbol{M}_m, \boldsymbol{U}_m, \boldsymbol{V}_m)}{\sum_{m=1}^{M} w_m \mathcal{N}_{\mathrm{mv}}(\boldsymbol{Z}_t; \boldsymbol{M}_m, \boldsymbol{U}_m, \boldsymbol{V}_m)}, \quad (18)$$

$$\hat{\boldsymbol{M}}_m = \frac{1}{T_m} \sum_{t=1}^{T} \gamma_{m,t} \boldsymbol{Z}_t, \quad (19)$$

$$\hat{\boldsymbol{U}}_m = \frac{1}{S T_m} \sum_{t=1}^{T} \gamma_{m,t} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m) \hat{\boldsymbol{V}}_m^{-1} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m)^{\top}, \quad (20)$$

$$\hat{\boldsymbol{V}}_m = \frac{1}{D T_m} \sum_{t=1}^{T} \gamma_{m,t} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m)^{\top} \hat{\boldsymbol{U}}_m^{-1} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m), \quad (21)$$

where $T_m = \sum_{t=1}^{T} \gamma_{m,t}$ means the effective number of samples corresponding to the m-th component. Like maximum likelihood estimation for the single matrix normal distribution, the efficient inference can be realized in Equations 20 and 21.

In a similar way, the mapping function is also derived based on the conditional probability of $\boldsymbol{y}_t$, given $\boldsymbol{x}_t$. Since MV-GMM has an explicitly separatable variance-covariance structure, the conditional probability density of the m-th component is simply represented as follows:

$$P(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(Z)}) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(y)}, \boldsymbol{D}_m^{(y)}), \quad (22)$$

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \frac{v_m^{(yx)}}{v_m^{(xx)}} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}), \quad (23)$$

$$\boldsymbol{D}_m^{(y)} = \left(v_m^{(yy)} - \frac{v_m^{(yx)} v_m^{(xy)}}{v_m^{(xx)}} \right) \boldsymbol{U}_m, \quad (24)$$

where $\boldsymbol{M}_m = [\boldsymbol{\mu}_m^{(x)}, \boldsymbol{\mu}_m^{(y)}]$ and $v_m^{(\cdot)}$ denotes the corresponding element in $\boldsymbol{V}_m$. Equations 23 and 24 mean that the conversion function corresponding to the m-th component is derived from parameters in the covariance matrix $\boldsymbol{V}_m$ corresponding to correlation of speakers. Compared with Equations 6 and 7, parameter generation based on MV-GMM is realized at low computational costs since it does not require calculating inverse matrices.

3.3. Model training using multiple speakers

Since the proposed modeling based on MV-GMM separates variance-covariance structures in the joint feature space explicitly, the joint feature space is easily expanded. Adding the feature column vectors from additional speakers to the joint matrix, the proposed method realizes model training using three or more speakers. In the conventional approach where the multiple feature vectors are concatenated into a high dimensional feature vector, model training using multiple speakers easily suffers from the overtraining problem. In contrast, the proposed method could utilize the additional column vectors for performance improvement, because a kind of adaptive training is realized in the training phase of MV-GMM (See Equation 20) [12, 13].

4. Experimental evaluation

4.1. Experimental conditions

To evaluate the performance of our proposed method and the effects of model training using multiple speakers, two kinds of voice conversion experiments were carried out. There are two objectives of the experiments. The first objective is to verify that the proposed approach based on MV-GMM effectively models both the characteristics of the feature spaces and the relation between the source and target spaces compared with the conventional joint modeling. The second objective is to verify the effectiveness of model training when the additional column vectors is added.

For the first objective, we used speech data of two male speakers from the CMU ARCTIC database [14] (bdl and rms). Voice conversion was performed using bdl as the source speaker and rms as the target one. We selected 256 sentences (from a0001 to a0256) for training. The evaluation set consisting of 50 sentences (from a0544 to a0593) were selected.

For the second objective, we used speech samples from three male speakers (MHT as the source speaker, MMY as the target speaker and MSH as the additional speaker) in the ATR Japanese speech database B-set [15]. For training, the first 250 sentences from the database were selected. The last 53 sentences were selected for test data. In this experiment, the DTW was first carried out between a feature sequence from MHT and that from MMY. Then, a sequence of averaged feature vectors of them was calculated based on the alignment information. Finally the DTW between the sequence of the averaged feature

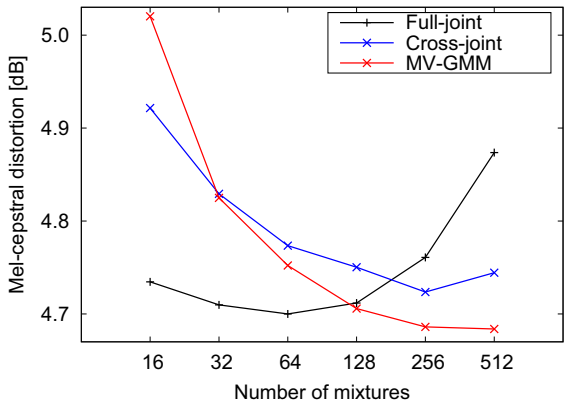

Figure 1: Results of objective evaluations by mel-cepstral distortion (MCD).

vectors and a feature sequence from MSH was carried out to construct the joint matrix in the proposed method.

We used 24-dimensional mel-cepstrum vectors for spectrum representation ($D=24$) in both the above setups. These were derived by STRAIGHT analysis [16]. Conditional maximum likelihood criterion is used for parameter generation (Equation 9). Note that parameter generation considering dynamic features is not adopted in the experiment [9].

In the first experiment using CMU ARCTIC database, we evaluated three methods; the joint vector approach where variance-covariance structure is not restricted (Full-joint), the joint vector approach where the covariance matrices $\mathbf{\Sigma}_m^{(xx)}$ and $\mathbf{\Sigma}_m^{(yy)}$, the cross covariance matrices $\mathbf{\Sigma}_m^{(xy)}$ and $\mathbf{\Sigma}_m^{(yx)}$ are diagonal (Cross-joint), and the proposed method (MV-GMM). $\boldsymbol{U}_m$ and $\boldsymbol{V}_m$ are the full covariance matrices. The number of mixture components (M) was varied from 16 to 512.

In the second experiment using the ATR Japanese speech database, we evaluated two methods based MV-GMM; one trained by the joint matrices from the source and target speakers ($S = 2$) and that trained by the joint matrices from the source, target and additional speakers ($S = 3$). The number of mixture components (M) was fixed to 256.

4.2. Objective evaluations

We evaluated the conversion performance using mel-cepstral distortion between the converted vectors and the vectors of the targets. Figure 1 shows the result of average mel-cepstral distortion for the test data as a function of the number of mixture components in GMMs. When the number of mixture components is small, the performance of the joint vector approach without restriction of variance-covariance structures is better than the performances of the other methods. However when the number of mixture components is bigger than 64, the performance of "Full-joint" is drastically degraded. This means that no restriction for variance-covariance structures in the joint modeling suffers from the overfitting effects when the model is complex. Compared with the joint vector approach with diagonal constraints, the proposed method based on MV-GMM shows the similar trends as a function of the number of mixture components. When the number of mixtures is bigger than 32, the proposed method outperforms the "Cross-joint" approach. This means that the proposed method models the detailed characteristics of the feature spaces both using the source and target features effectively in Equation 20. Compared with the "Full-joint" approach, the performance of the proposed method is slightly better in the optimal condition of the model complexity. This means that constraints derived from matrix variate modeling effectively work for the modeling of the two functions in voice conversion; to model the source and target feature spaces precisely, and to construct a proper transformation between these spaces.

Table 1: Results of objective evaluations by MCD in the optimal conditions. The optimal numbers of mixture components were selected. # of parameters means the number of variance-covariance parameters which should be estimated in the model.

	MCD [dB]	M	# of parameters
Full-joint	4.70	64	75264
Cross-joint	4.72	256	18432
MV-GMM	4.68	512	155136

Table 2: Results of objective evaluations by MCD when model training using multiple speaker is applied.

	MCD [dB]
$S = 2$	4.643
$S = 3$	4.635

Table 1 shows the results of objective evaluations in the optimal conditions. From Table 1, the proposed method has the most parameters for variance-covariance structure which should be estimated. Nevertheless, the proposed method effectively works and does not suffer from the overfitting effects. This means that the algorithm for parameter estimation in MV-GMM realizes efficient and precise inference.

4.3. Effects of model training using multiple speakers

Table 2 shows the results of objective evaluations when the number of training speakers for the joint model in MV-GMM is varied. When the additional speaker is included in the training phase of the proposed model, the performance of conversion is slightly improved. Note that the additional vectors are neither from the source nor target speakers. That is to say, the proposed framework effectively utilizes the additional column vectors for the improvement of the performance. The proposed method can be regarded as a kind of adaptive training which realizes an efficient parameter tying.

5. Conclusions

This paper proposed a novel approach to construct a mapping function between a given speaker pair using matrix variate Gaussian mixture model. The proposed method can effectively model both the characteristics of the feature spaces and the relation between the source and target spaces. In addition, the proposed framework has possibility to use the additional data from non-target speakers for improvement of the conversion performance. For further works, the effectiveness of the proposed method should be investigated in large-scale subjective evaluations. In addition, the proposed modeling with dynamic features, or long-span features is interesting. Integration with other parameter tying approaches focusing on precise modeling of variance-covariance structure is another further direction.

6. Acknowledgment

This work was supported by KAKENHI Grant-in-Aid for Young Scientists (B) (Number 25730105).

7. References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," Proc. ICASSP, pp. 655–658, 1988.

[2] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," IEEE Trans. on Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.

[3] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, "High-performance robust speech recognition using stereo training data," Proc. ICASSP, pp. 301–304, 2001.

[4] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," Proc. ICASSP, vol. 1, pp. 285–288, 1998.

[5] A. Mouchtaris, J. V. der Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," IEEE Trans. on Audio, Speech, and Language Processing, vol. 14, no. 3, pp. 952–963, 2006.

[6] C. H. Lee and C. H. Wu, "Map-based adaptation for speech conversion using adaptation data selection and non-parallel training," Proc. INTERSPEECH, pp. 2254–2257, 2006.

[7] D. Saito, S. Watanabe, A. Nakamura and N. Minematsu, "Statistical voice conversion based on noisy channel model," IEEE Trans. on Audio, Speech, and Language Processing, vol. 20, no. 6, pp. 1784–1794, 2012.

[8] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," Proc. INTERSPEECH, pp. 653–656, 2011.

[9] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Trans. on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222–2235, 2007.

[10] P. Dutilleul, "The MLE algorithm for the matrix normal distribution," Journal of Statistical Computation and Simulation, vol. 64, pp. 105–123, 1999.

[11] C. Viroli, "Finite mixture of matrix normal distributions for classifying three-wya data," Journal of Statistics and Computing, vol. 21, Issue 4, pp. 511–522, 2011.

[12] T. Anastasakos, J. McDonough, R. Schwarts and J. Makhoul, "A compact model for speaker adaptive training," Proc. ICSLP, vol. 2, pp. 1137–1140, 1996.

[13] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Speaker adaptive training for one-to-many eigenvoice conversion based on Gaussian mixture model," Proc. INTERSPEECH, pp. 1981–1984, 2007.

[14] J. Kominek and A. W. Black, "CMU ARCTIC Databases for Speech Synthesis," Lang. Technol. Inst., Carnegie Mellon Univ., Pittsburgh, PA. 2003 [Online].
Available: http://festvox.org/cmu_arctic/index.html

[15] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K.Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, vol.9, pp.357–363, 1990.

[16] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, vol.27, pp.187–207, 1999.