INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Voice conversion based on matrix variate Gaussian mixture model using multiple frame features

Yi Yang, Hidetsugu Uchida, Daisuke Saito, Nobuaki Minematsu

The University of Tokyo, Japan
{yang,uchida,dsk_saito,mine}@gavo.t.u-tokyo.ac.jp

Abstract

This paper presents a novel voice conversion method based on matrix variate Gaussian mixture model (MV-GMM) using features of multiple frames. In voice conversion studies, approaches based on Gaussian mixture models (GMM) are still widely utilized because of their flexibility and easiness in handling. They treat the joint probability density function (PDF) of feature vectors from source and target speakers as that of joint vectors of the two vectors. Addition of dynamic features to the feature vectors in GMM-based approaches achieves certain performance improvements because the correlation between multiple frames is taken into account. Recently, a voice conversion framework based on MV-GMM, in which the joint PDF is modeled in a matrix variate space, has been proposed and it is able to precisely model both the characteristics of the feature spaces and the relation between the source and target speakers. In this paper, in order to additionally model the correlation between multiple frames in the framework more consistently, MV-GMM is constructed in a matrix variate space containing the features of neighboring frames. Experimental results show that an certain performance improvement in both objective and subjective evaluations is observed.

Index Terms: voice conversion, Gaussian mixture model, matrix variate Gaussian mixture model, multiple frame features

1. Introduction

Voice conversion (VC), specifically speaker conversion discussed in this paper, is a technique to modify non-linguistic information — in this case speaker characteristics — while maintaining linguistic information unchanged. In speaker conversion, a statistical mapping function is constructed using pairs of features extracted from utterances of source and target speakers. Gaussian mixture model (GMM) [1] is widely used to construct mapping functions, as well as Neural Network (NN) [2] and Nonnegative Matrix Factorization (NMF) [3].

Founded on flexibilities of GMM-based approaches, several expanded models are proposed, such as a method in which dynamic features are taken into account [4] . In this method, relationship between features of adjacent frames are considered, and a conversion function is designed to maximize the likelihood of a time sequence of the target features given that of the source ones. Since dynamics of speech are captured, conversion performance of the method is improved well.

Generally in GMM-based VC, both the features extracted from the source and target speakers are concatenated and they are represented as a joint vector. In addition, when dynamic features are taken into account, they are also included in the joint vector. Finally, the characteristics of the static feature spaces, those of the dynamic ones, and the correlation of the source and target speakers are mixed together in the joint vector space. In order to model VC functions precisely, these information should be properly treated.

Voice conversion based on matrix variate Gaussian mixture models (MV-GMM), in which joint features from the source and target are represented as matrices, has been proposed [5]. In this model, a separable structure is derived to covariances of matrix features. Two types of covariance matrices, i.e. row and column matrices capture the characteristics of the feature spaces and the correlation between the source and target, respectively. Since they induce an effective training algorithm for the model, MV-GMM constructs a precise conversion function.

Basically both GMM and MV-GMM are mixture models of normal distributions. Hence, similar extensions to GMM-based VC can be applied to the MV-GMM approach. In this paper, in order to additionally model the correlation between adjacent frames in MV-GMM, voice conversion based on MV-GMM using multiple frame features is proposed. In the proposed method, it is expected that the separable structure in MV-GMM enables the model to capture the correlation between adjacent frames more precisely than the joint vector approach.

2. GMM-based VC with joint vectors

In this section, GMM-based voice conversion is briefly described [1]. A feature vector of time index t from an utterance of a source speaker is defined as $\boldsymbol{x}_t = [x_1, x_2, \ldots, x_n]^\top$, while $\boldsymbol{y}_t = [y_1, y_2, \ldots, y_n]^\top$ represents that from an utterance of a target speaker, where n denotes the dimension of the feature vectors and $(\cdot)^\top$ notifies the transposition of vector or matrix. Note that these utterances include the same linguistic content. The probability density of the joint vector $\boldsymbol{z}_t = [\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]^\top$ is modeled by GMM as follows:

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

where w_m, $\boldsymbol{\mu}_m^{(z)}$ and $\boldsymbol{\Sigma}_m^{(z)}$ denote the weight, the mean vector, and the covariance matrix of the m-the Gaussian component, respectively. $\boldsymbol{\mu}_m^{(z)}$, $\boldsymbol{\Sigma}_m^{(z)}$ is indicated by mean vectors and covariance matrices of the source and target speakers separately, which shown as follows:

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

$\boldsymbol{\Sigma}_m^{(\cdot)}$ shows covariance matrices of the source and target speakers, or cross-covariance matrices between them, which are usually restricted to diagonal matrices in order to mitigate the influence of overfitting.

http://dx.doi.org/10.21437/Interspeech.2016-705

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

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

where

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

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

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

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

The converted feature $\hat{\boldsymbol{y}}_t$ when $\boldsymbol{x}_t$ is given can be generated by the following equation using maximum likelihood criterion:

$$\hat{\boldsymbol{y}}_t=\left(\sum_{m=1}^{M}\gamma_{m,t}\boldsymbol{D}_m^{(y)-1}\right)^{-1}\left(\sum_{m=1}^{M}\gamma_{m,t}\boldsymbol{D}_m^{(y)-1}\boldsymbol{E}_{m,t}^{(y)}\right),$$
$$\gamma_{m,t}=P\left(m|\boldsymbol{x}_t,\boldsymbol{y}_t,\boldsymbol{\lambda}^{(z)}\right). \quad (8)$$

Dynamic features may also be contained in a GMM-based model [4]. They are usually defined as difference between the previous and next frames, and they are denoted by $\Delta\boldsymbol{x}_t$ and $\Delta\boldsymbol{y}_t$. $\boldsymbol{X}_t=[\boldsymbol{x}_t^\top,\Delta\boldsymbol{x}_t^\top]^\top$, $\boldsymbol{Y}_t=[\boldsymbol{y}_t^\top,\Delta\boldsymbol{y}_t^\top]^\top$ denote $2D$-dimensional feature vectors which consist of both static and dynamic features. Time sequences of the source and target feature vectors are vectorized and described as $\boldsymbol{x}=[\boldsymbol{x}_1^\top,\boldsymbol{x}_2^\top,\ldots,\boldsymbol{x}_T^\top]^\top$, $\boldsymbol{y}=[\boldsymbol{y}_1^\top,\boldsymbol{y}_2^\top,\ldots,\boldsymbol{y}_T^\top]^\top$, $\boldsymbol{X}=[\boldsymbol{X}_1^\top,\boldsymbol{X}_2^\top,\ldots,\boldsymbol{X}_T^\top]^\top$, and $\boldsymbol{Y}=[\boldsymbol{Y}_1^\top,\boldsymbol{Y}_2^\top,\ldots,\boldsymbol{Y}_T^\top]^\top$. For each frame, a joint vector is defined as $\boldsymbol{Z}_t=[\boldsymbol{X}_t^\top,\boldsymbol{Y}_t^\top]^\top$, and the joint probability density is modeled by GMM. When a time sequence $\boldsymbol{X}$ is given, the conditional probability density of $\boldsymbol{Y}$ can be shown as:

$$P\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)}\right)=\sum_{m=1}^{M}P\left(m|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)}\right)P\left(\boldsymbol{Y}|\boldsymbol{X},m,\boldsymbol{\lambda}^{(Z)}\right)$$
$$=\prod_{t=1}^{T}\sum_{m=1}^{M}P\left(m|\boldsymbol{X}_t,\boldsymbol{\lambda}^{(Z)}\right)P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t,m,\boldsymbol{\lambda}^{(Z)}\right). \quad (9)$$

The conditional probability density at each frame can also be modeled as GMM. At frame t , the posterior of the m-th mixture component and its corresponding conditional probability density of $\boldsymbol{Y}_t$ in Equation 9 are described as

$$P\left(m|\boldsymbol{X}_t,\boldsymbol{\lambda}^{(Z)}\right)=\frac{w_m\mathcal{N}\left(\boldsymbol{X}_t;\boldsymbol{\mu}_m^{(X)},\boldsymbol{\Sigma}_m^{(XX)}\right)}{\sum_{m=1}^{M}w_m\mathcal{N}\left(\boldsymbol{X}_t;\boldsymbol{\mu}_m^{(X)},\boldsymbol{\Sigma}_m^{(XX)}\right)}, \quad (10)$$

$$P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t,m,\boldsymbol{\lambda}^{(Z)}\right)=\mathcal{N}\left(\boldsymbol{Y}_t;\boldsymbol{E}_{m,t}^{(Y)},\boldsymbol{D}_m^{(Y)}\right), \quad (11)$$

$$\boldsymbol{E}_{m,t}^{(Y)}=\boldsymbol{\mu}_m^{(Y)}+\boldsymbol{\Sigma}_m^{(YX)}\boldsymbol{\Sigma}_m^{(XX)-1}(\boldsymbol{X}_t-\boldsymbol{\mu}_m^{(X)}) \quad (12)$$

$$\boldsymbol{D}_m^{(Y)}=\boldsymbol{\Sigma}_m^{(YY)}-\boldsymbol{\Sigma}_m^{(YX)}\boldsymbol{\Sigma}_m^{(XX)-1}\boldsymbol{\Sigma}_m^{(XY)} \quad (13)$$

Finally, a target sequence $\hat{\boldsymbol{y}}$ can be derived as a solution of the following optimization problem:

$$\hat{\boldsymbol{y}}=\arg\max P\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)}\right)\ \text{s.t.}\ \boldsymbol{Y}=\boldsymbol{W}\boldsymbol{y}. \quad (14)$$

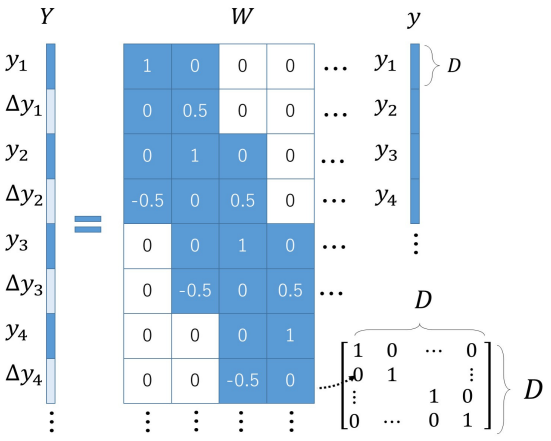

Figure 1: Relation between $\boldsymbol{y}_t$ and $\boldsymbol{Y}_t$. $\Delta\boldsymbol{y}_t$ is defined as $0.5(\boldsymbol{y}_{t+1}-\boldsymbol{y}_{t-1})$.

Figure 1 is an example of $\boldsymbol{W}$. In this case, the dynamic feature $\Delta\boldsymbol{y}_t$ is defined as $0.5(\boldsymbol{y}_{t+1}-\boldsymbol{y}_{t-1})$, and the corresponding window matrix $\boldsymbol{W}$ is derived. The conversion function is written as the equation bellow based on maximum likelihood estimation:

$$\hat{\boldsymbol{y}}=\left(\boldsymbol{W}^\top\overline{\boldsymbol{D}^{(Y)-1}}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top\overline{\boldsymbol{D}^{(Y)-1}\boldsymbol{E}^{(Y)}}. \quad (15)$$

$\overline{\boldsymbol{D}^{(Y)-1}}$, $\overline{\boldsymbol{D}^{(Y)-1}\boldsymbol{E}^{(Y)}}$ are parameters in time sequence, which written as:

$$\overline{\boldsymbol{D}^{(Y)-1}} = \mathrm{diag}[\overline{\boldsymbol{D}_1^{(Y)-1}},\overline{\boldsymbol{D}_2^{(Y)-1}},\ldots,\overline{\boldsymbol{D}_t^{(Y)-1}},\ldots,\overline{\boldsymbol{D}_T^{(Y)-1}}] \quad (16)$$

$$\overline{\boldsymbol{D}^{(Y)-1}\boldsymbol{E}^{(Y)}} = [\overline{\boldsymbol{D}_1^{(Y)-1}\boldsymbol{E}^{(Y)}}^\top,\overline{\boldsymbol{D}_2^{(Y)-1}\boldsymbol{E}^{(Y)}}^\top,\ldots,\overline{\boldsymbol{D}_t^{(Y)-1}\boldsymbol{E}^{(Y)}}^\top,\ldots,\overline{\boldsymbol{D}_T^{(Y)-1}\boldsymbol{E}^{(Y)}}^\top]^\top \quad (17)$$

$$\overline{\boldsymbol{D}_t^{(Y)-1}} = \sum_{m=1}^{M}\gamma_{m,t}\boldsymbol{D}_m^{(Y)-1} \quad (18)$$

$$\overline{\boldsymbol{D}_t^{(Y)-1}\boldsymbol{E}^{(Y)}} = \sum_{m=1}^{M}\gamma_{m,t}\boldsymbol{D}_m^{(Y)-1}\boldsymbol{E}_{m,t}^{(Y)} \quad (19)$$

$$\gamma_{m,t} = P\left(m|\boldsymbol{X}_t,\boldsymbol{Y}_t,\boldsymbol{\lambda}^{(z)}\right)$$

Maximum likelihood parameter generation described in Equation 15 achieves a certain improvement of conversion performance since time correlations between adjacent frames are taken into account. However, dynamic features has different properties from those of static features. The joint vector contains both static and dynamic features, which tends to present characteristics of feature space with different properties together. In this concern, this model may not be so reasonable in considering correlations between adjacent frames.

3. MV-GMM for VC

In this section, voice conversion based on matrix variate Gaussian model is briefly explained [5]. Here, $\boldsymbol{X}$ is a random matrix whose size is $n\times p$. When $\boldsymbol{X}$ follows a normal distribution of matrix variate, it can be written as follow:

$$\boldsymbol{X}\sim\mathcal{N}_{mv}(\boldsymbol{X};\boldsymbol{M},\boldsymbol{U},\boldsymbol{V}), \quad (20)$$

where $\boldsymbol{M}$ is an matrix whose size is $n \times p$, representing the mean information of the normal distribution. $\boldsymbol{U}$ and $\boldsymbol{V}$ are matrices whose sizes are $n \times n$ and $p \times p$, which represent the covariance structure of row and column spaces, respectively. A normal distribution of matrix variate corresponds to that of vector variate which is represented as follows [9]:

$$P\left(\mathrm{vec}(\boldsymbol{X})|\boldsymbol{\lambda}\right) = \mathcal{N}\left(\mathrm{vec}(\boldsymbol{X}); \mathrm{vec}(\boldsymbol{M}), \boldsymbol{V} \otimes \boldsymbol{U}\right), \quad (21)$$

where vec() is the operator that change a matrix into a vector. $\boldsymbol{V} \otimes \boldsymbol{U}$ means Kronecker product of the two matrices. According to Equation 21, the matrix variate normal distribution differs from the vector variate normal distribution. Compared with the vector variate normal distribution, it has a restricted covariance structure based on Kronecker product. Separating the covariance structure into two matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ makes it possible to represent properties of the row and column spaces, respectively.

Normal distribution of matrix can be applied to voice conversion, in a similar way to that of normal distribution of vector. Let $\boldsymbol{x}_t$, $\boldsymbol{y}_t$ be the feature vectors of the source and target speakers. $\boldsymbol{x}_t$, $\boldsymbol{y}_t$ are combined into one joint matrix $\boldsymbol{Z}_t = [\boldsymbol{x}_t, \boldsymbol{y}_t] \in \mathcal{R}^{D \times S}$, where D means the dimension of feature space, and S the dimension of speaker space. In the case of one source speaker and one target speaker, $S = 2$. The probability density of $\boldsymbol{Z}_t$ is represented by a mixture model shown as follows:

$$P\left(\boldsymbol{Z}_t|\boldsymbol{\lambda}^{(Z)}\right) = \sum_{m=1}^{M} w_m \mathcal{N}_{mv}(\boldsymbol{Z}_t; \boldsymbol{M}_m, \boldsymbol{U}_m, \boldsymbol{V}_m) \quad (22)$$

According to Equation 22, the probability density is defined as the weighted sum of normal distribution of each component.In each component, the matrix variate normal distribution is indicated by three matrix parameters $\boldsymbol{M}_m$, $\boldsymbol{U}_m$, and $\boldsymbol{V}_m$; $\boldsymbol{M}_m$ is the mean matrix, $\boldsymbol{U}_m$ the covariance matrix of the feature space, and $\boldsymbol{V}_m$ the covariance matrix capturing the correlation between the source and target. These parameters can be estimated by EM algorithm shown as follows:

$$\gamma_{m,t} = \frac{w_m \mathcal{N}_{mv}(\boldsymbol{Z}_t; \boldsymbol{M}_m, \boldsymbol{U}_m, \boldsymbol{V}_m)}{\sum_{m=1}^{M} w_m \mathcal{N}_{mv}(\boldsymbol{Z}_t; \boldsymbol{M}_m, \boldsymbol{U}_m, \boldsymbol{V}_m)} \quad (23)$$

$$\hat{\boldsymbol{M}}_m = \frac{1}{T_m} \sum_{t=1}^{T} \gamma_{m,t} \boldsymbol{Z}_t \quad (24)$$

$$\hat{\boldsymbol{U}}_m = \frac{1}{S T_m} \sum_{t=1}^{T} \gamma_{m,t} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m) \hat{\boldsymbol{V}}_m^{-1} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m)^\top \quad (25)$$

$$\hat{\boldsymbol{V}}_m = \frac{1}{D T_m} \sum_{t=1}^{T} \gamma_{m,t} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m)^\top \hat{\boldsymbol{U}}_m^{-1} (\boldsymbol{Z}_t - \hat{\boldsymbol{M}}_m) \quad (26)$$

$$T_m = \sum_{t=1}^{T} \gamma_{m,t} \quad (27)$$

A conversion function is derived based on the conditional probability density $P\left(\boldsymbol{y}_t|\boldsymbol{x}_t\right)$, referred to Equation 3. The conditional probability density of the m-th component is shown as follows:

$$P\left(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(Z)}\right) = \mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}; \boldsymbol{D}_m\right), \quad (28)$$

$$\boldsymbol{E}_{m.t} = \boldsymbol{\mu}_m^{(y)} + \frac{v_m^{(yx)}}{v_m^{(xx)}} (\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}), \quad (29)$$

$$\boldsymbol{D}_m = \left(v_m^{(yy)} - \frac{v_m^{(yx)} v_m^{(xy)}}{v_m^{(xx)}} \right) \boldsymbol{U}_m, \quad (30)$$

where $v_m^{(\cdot)}$ shows elements of $\boldsymbol{V}_m$. Separating the covariance structure into two directions induces an effective estimation as Equations 25 and 26. Hence MVGMM-based VC achieves more concise and proper modeling.

4. MVGMM VC with multiple frame features

In this section, we introduce features from multiple frames into MVGMM-based VC. Let $\boldsymbol{x}_t$, $\boldsymbol{y}_t$ be feature vectors from utterances of the source and target speakers at frame t. In order to capture time correlations, features of the source ranging from $\boldsymbol{x}_{t-N_x}$ to $\boldsymbol{x}_{t+N_x}$ and those of the target ranging from $\boldsymbol{y}_{t-N_y}$ to $\boldsymbol{y}_{t+N_y}$ are focused on. The joint matrix $\boldsymbol{Z}_t$ is defined as $\boldsymbol{Z}_t = [\boldsymbol{x}_{t-N_x}, \ldots, \boldsymbol{x}_t \ldots \boldsymbol{x}_{t+N_x}, \boldsymbol{y}_{t-N_y} \cdots \boldsymbol{y}_t \cdots \boldsymbol{y}_{t+N_y}] \in \mathcal{R}^{D \times S}$, where D denotes the dimension of the feature space, and where S denotes the total number of dimension considering multiple speakers and frames. The probability density of $\boldsymbol{Z}_t$ modeled by MVGMM is described as the same equation of Equation 22. In this case, $\boldsymbol{U}_m$, i.e. the covariance structure for the feature space has the same dimension with that of when multiple frames are not taken into account. Only the dimension of $\boldsymbol{V}_m$ increases as Equation 31;

$$\boldsymbol{V}_m = \begin{bmatrix} \boldsymbol{V}_m^{(xx)} & \boldsymbol{V}_m^{(xy)} \\ \boldsymbol{V}_m^{(yx)} & \boldsymbol{V}_m^{(yy)} \end{bmatrix}. \quad (31)$$

Two types of the conversion functions are derived according to whether multiple frames of the target are modeled. When a single frame feature of the target speaker is used, the conversion function is derived on each frame. The conditional probability density of $\boldsymbol{y}_t$ given $\boldsymbol{X}_t = [\boldsymbol{x}_{t-N_x}, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_{t+N_x}]$ is

$$P\left(\boldsymbol{y}_t|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(Z)}\right) = \sum_{m=1}^{M} P\left(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(Z)}\right) \times P\left(\boldsymbol{y}_t|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(Z)}\right), \quad (32)$$

where

$$P\left(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(Z)}\right) = \frac{w_m \mathcal{N}\left(\mathrm{vec}(\boldsymbol{X}_t); \mathrm{vec}(\boldsymbol{M}_m^{(x)}), \boldsymbol{V}_m^{(xx)} \otimes \boldsymbol{U}_m\right)}{\sum_{m=1}^{M} w_m \mathcal{N}\left(\mathrm{vec}(\boldsymbol{X}_t); \mathrm{vec}(\boldsymbol{M}_m^{(x)}), \boldsymbol{V}_m^{(xx)} \otimes \boldsymbol{U}_m\right)}, \quad (33)$$

$$P\left(\boldsymbol{y}_t|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(Z)}\right) = \mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(y)}, \boldsymbol{D}_m^{(y)}\right), \quad (34)$$

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{V}_m^{(yx)} \boldsymbol{V}_m^{(xx)-1} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}\right), \quad (35)$$

$$\boldsymbol{D}_m^{(y)} = \left(V_m^{(yy)} - \boldsymbol{V}_m^{(yx)} \boldsymbol{V}_m^{(xx)-1} \boldsymbol{V}_m^{(xy)}\right) \boldsymbol{U}_m. \quad (36)$$

The target $\hat{\boldsymbol{y}}_t$ can be generated by a simillar way to Equation 8.

When multiple frames of the target speaker are taken into account, maximum likelihood sequence estimation [4] is adopted according to Equation 15, where the relation between $\mathrm{vec}(\boldsymbol{Y})$ and $\boldsymbol{y}$ is shown in Figure 2 [6].

Comparing the proposed method with GMM-based model containing dynamic features, both of them are models considering correlations between adjacent frames. Instead of containing dynamic features which has different properties from static features, the proposed method contains also the static features of adjacent frames which shows similar properties to the current frame. The proposed method is regarded as a more reasonable model because all elements of the joint variate are static features instead of mixing static and dynamic features together.

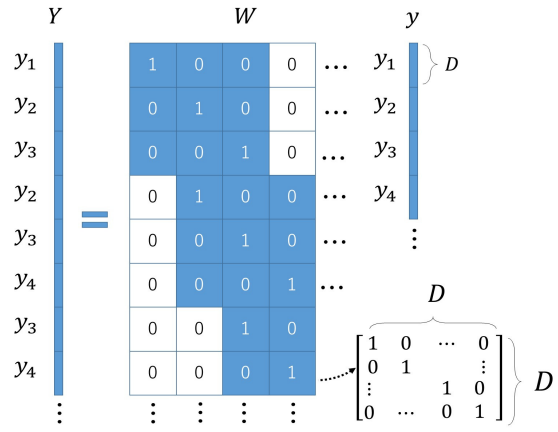

Figure 2: Relation between vec($\boldsymbol{Y}$) and $\boldsymbol{y}$

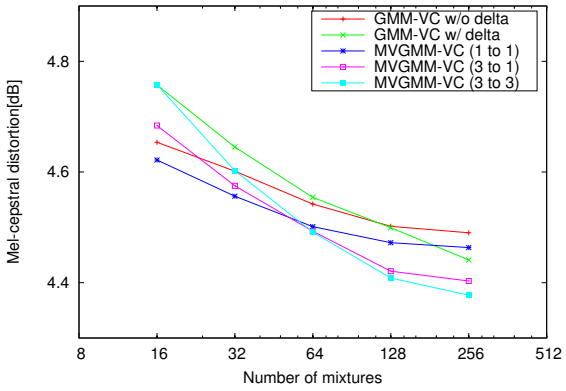

Figure 3: Results of objective evaluations

5. Experimental evaluation

5.1. Experimental setups

To evaluate the performance of the proposed method, objective and subjective evaluations were carried out. We used speech data of two male speakers (MMY as the source speaker, MHT as the target speaker) from ATR Japanese speech database [7]. 250 pairs of utterances were selected for training, and 50 pairs not included in the training data were selected for testing. As the proposed methods, we considered the method in which only multiple frame features from the source are modeled and the method in which multiple frame features both from the source and target are modeled. GMM-based approaches with/without dynamic features, and MVGMM-based one without multiple frames were compared with the proposed approaches. Diagonal structures were adopted for cross-covariance matrices in the conventional GMM-based methods, and full covariance matrices are used in the MV-GMM methods. 24-dimensional mel-cepstrum derived from STRAIGHT analysis [8] was used as feature vectors.

5.2. Objective evaluation

For the objective evaluations, Mel-cepstral distortion was used. From Figure 3, when the number of mixture components is larger than 64, the proposed methods outperformed the conventional ones. The performance was much better when multiple frames features is contained for both the source and target speakers (MVGMM-VC (3 to 3)). Besides, the proposed method (MVGMM-VC (3 to 3)) achieved better performance than the joint vector approach with dynamic features (GMM w/ delta). Note that both the methods take information of neighboring frames into account. This means that the separable structure in the proposed method has a positive effect in voice conversion.

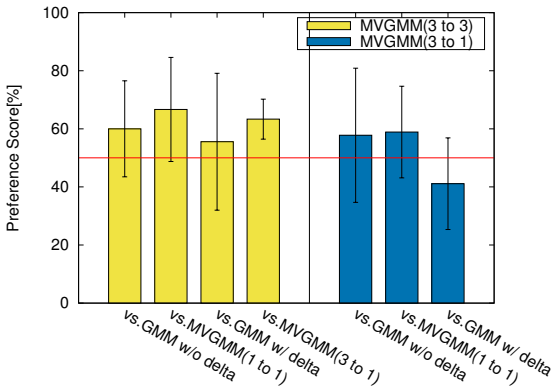

Figure 4: Results of subjective evaluations

5.3. Subjective evaluation

A subjective evaluation was processed in XAB method, where each listener was provided with utterances converted from two different kinds of conversion methods and he/she was asked to select the more similar one to the original speech of the target speaker. Ten Japanese native speakers aging from 20 to 35 , each of whom were provided with 70 pairs of sentences in 7 groups of comparison, including comparison between the proposed and conventional methods, and among the proposed ones. Mixture of 256 was used for all conversion methods. For waveform generation, $\log F_0$ values were linearly converted based on mean and variance values, and then STRAIGHT vocoder was adopted.

Figure 4 shows the result of the subjective evaluation. The error bars in the figure represent 95% confidence intervals. From Figure 4, the proposed method with multiple frame features of both the source and target (MVGMM-VC (3 to 3)) is more likely to be selected comparing to all other methods, while the proposed method with multiple frames feature only on the source side (MVGMM-VC (3 to 1)) performs better those not considering relationship between frames, however, it is not as preferable when compared with the joint vector approach with dynamic features (GMM w/ delta). This result suggests that modeling the multiple frames of both the source and target has a considerable influence on perception.

6. Conclusion

This paper presents a novel voice conversion method based on matrix variate Gaussian mixture model (MV-GMM) using features of multiple frames. The proposed method can effectively model features of utterances from the source and target speakers. Both objective and subjective evaluations show that the proposed method containing multiple frame features of both the source and target achieved a certain improvement of conversion performance, compared with the conventional methods.

7. Acknowledgment

This work was supported by MEXT KAKENHI Grant Number JP26118002 and JSPS KAKENHI Grant Number JP25730105.

8. References

[1] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," IEEE International Conference on, 1998, pp. 285-288

[2] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black and K. Prahallad, "Voice conversion using artificial neural networks," ICASSP, pp. 3893-3896 (2009).

[3] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Exemplar-based voice conversion in noisy environment," in SLT, pp. 313-317, (2012).

[4] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajctory, " IEEE TRANSACTIONS 15 (8) pp. 2222-2235 (2007).

[5] Daisuke Saito, Hidenobu Doi, Nobuaki Minematsu, and Keikichi Hirose, "Application of Matrix Variate Gaussian Mixture Model to Statistical Voice Conversion, " INTERSPEECH 2014

[6] Ling-Hui Chen, Zhen-Hua Ling, and Li-Rong Dai, "Voice Conversion Using Generative Trained Deep Neural Networks with Multiple Frame Spectral Envelopes, " Interspeech, pp. 2313-2317, (2014).

[7] Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, vol.9, pp.357-363, (1990).

[8] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigne, "Re-structuring 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.

[9] A. K. Gupta, D. K. Nagar, Matrix Variate Distributions, 2000