MODULATION SPECTRUM-CONSTRAINED TRAJECTORY TRAINING ALGORITHM FOR GMM-BASED VOICE CONVERSION

Shinnosuke Takamichi[†,‡]*, Tomoki Toda*[†]*, Alan W. Black*[‡]*, and Satoshi Nakamura*[†]

† Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan
‡ Language Technologies Institute, Carnegie Mellon University (CMU), U.S.A
Email: shinnosuke-t@is.naist.jp

ABSTRACT

This paper presents a novel training algorithm for Gaussian Mixture Model (GMM)-based Voice Conversion (VC). One of the advantages of GMM-based VC is computationally efficient conversion processing enabling to achieve real-time VC applications. On the other hand, the quality of the converted speech is still significantly worse than that of natural speech. In order to address this problem while preserving the computationally efficient conversion processing, the proposed training method enables 1) to use a consistent optimization criterion between training and conversion and 2) to compensate a Modulation Spectrum (MS) of the converted parameter trajectory as a feature sensitively correlated with over-smoothing effects causing quality degradation of the converted speech. The experimental results demonstrate that the proposed algorithm yields significant improvements in term of both the converted speech quality and the conversion accuracy for speaker individuality compared to the basic training algorithm.

Index Terms— GMM-based voice conversion, over-smoothing, modulation spectrum, training algorithm

1. INTRODUCTION

Statistical Voice Conversion (VC) is an effective technique for modifying speech parameters to convert non-linguistic information while keeping linguistic information unchanged, and making it possible to enhance various speech-based systems [1, 2, 3, 4]. Recently, several state-of-the-art methods have been applied to VC [5, 6, 7], but Gaussian Mixture Model (GMM)-based VC [8, 9] has still gained popularity thanks to its computationally efficient conversion processing. This framework models the joint probability density of the static and dynamic features of speech parameters from both source and target voices using a GMM. The conversion stage performs the Maximum Likelihood (ML)-based trajectory conversion [9] using the conditional probability density analytically derived from the GMM and an additional constraint between the static and dynamic features. In this method, the converted parameter trajectories can be determined analytically, also enabling the computationally efficient real-time conversion processing [10, 11]. However, they suffer from the over-smoothing effect, which makes the converted speech sound muffled.

There are various attempts at addressing the over-smoothing issue in statistical parametric speech synthesis. Zen *et al.* [12] proposed the trajectory training method for Hidden Markov Model (HMM)-based speech synthesis [13] by imposing the constraint between the static and dynamic features in synthesis on the training criterion. The use of a consistent optimization criterion between training and synthesis enables to effectively optimize the trained model for synthesis. Toda and Young [14] has further incorporated an additional criterion on Global Variance (GV) into the optimization criterion, where the GV is well known as a feature sensitively correlated to the over-smoothing effect [9]. This method enables the use of the computationally efficient parameter generation algorithm [15] for generating the parameter trajectory while keeping its GV close to natural one. Hwang *et al.* [16] proposed the training algorithm for GMM-based VC considering the GV. Although it significantly improves the quality of the converted speech while keeping the computationally efficient conversion, there still remain some problems to be solved: 1) it suffers from the inconsistent optimization criteria between training and conversion, and 2) the GV is still insufficient to detect the over-smoothing effect.

Recently, Takamichi *et al.* [17] have found that a Modulation Spectrum (MS) of the generated parameter trajectory is more sensitively correlated to the over-smoothing effect than the GV. They have also reported that the MS can be regarded as an extension of the GV and the synthetic speech quality is significantly improved by considering the MS rather than the GV in HMM-based speech synthesis. It is expected that a better consistent optimization criterion will be designed by incorporating a criterion on the MS.

This paper proposes an MS-constrained trajectory training algorithm as a novel training algorithm for GMM-based VC. After implementing the trajectory training, we further integrate the MS into the trajectory training. The proposed training algorithm gives a unified framework for both training and conversion which provides both a consistent optimization criterion and a closed form solution for parameter conversion considering the MS. The experimental results demonstrate that the proposed algorithm yields significant improvements in term of both speech quality and speaker individuality.

2. BASIC FRAMEWORK

2.1. GMM Training [8]

A joint probability density function of speech parameters of the source and target speakers is modeled with a GMM using a parallel data as follows:

$$P\left(\left[\begin{array}{c}\boldsymbol{X}_t\\ \boldsymbol{Y}_t\end{array}\right]\middle|\boldsymbol{\lambda}\right)=\sum_{m=1}^{M}\alpha_m\mathcal{N}\left(\left[\begin{array}{c}\boldsymbol{X}_t\\ \boldsymbol{Y}_t\end{array}\right];\boldsymbol{\mu}_m^{(X,Y)},\boldsymbol{\Sigma}_m^{(X,Y)}\right), \quad (1)$$

$$\boldsymbol{\mu}_m^{(X,Y)}=\left[\begin{array}{c}\boldsymbol{\mu}_m^{(X)}\\ \boldsymbol{\mu}_m^{(Y)}\end{array}\right], \boldsymbol{\Sigma}_m^{(X,Y)}=\left[\begin{array}{cc}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)}\\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{array}\right], \quad (2)$$

where $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ are the source and target feature vectors at frame t, respectively. $\boldsymbol{Y}_t$ is given by $2D$-dimensional joint static and dynamic feature vectors, $\left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$, where $\boldsymbol{y}_t$ is represented as a D-dimensional vector, $\left[y_t\left(1\right),\cdots,y_t\left(D\right)\right]^\top$. The source feature vector is also given by the same form in this paper. $\mathcal{N}\left(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$ denotes the Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The total number of mixture components is M. $\boldsymbol{\lambda}$ is a GMM parameter set consisting of the mixture-component weight

α_m, the mean vector $\boldsymbol{\mu}_m^{(X,Y)}$ and the covariance matrix $\boldsymbol{\Sigma}_m^{(X,Y)}$ of the m-th mixture component. $\boldsymbol{\mu}_m^{(X,Y)}$ consists of the source and target mean vectors, $\boldsymbol{\mu}_m^{(X)}$ and $\boldsymbol{\mu}_m^{(Y)}$. $\boldsymbol{\Sigma}_m^{(X,Y)}$ consists of the source and target covariance matrices, $\boldsymbol{\Sigma}_m^{(XX)}$ and $\boldsymbol{\Sigma}_m^{(YY)}$ and cross-covariance matrices, $\boldsymbol{\Sigma}_m^{(YX)}$ and $\boldsymbol{\Sigma}_m^{(XY)}$, where they are diagonal matrices in this paper. The GMM parameters are estimated by maximizing the joint probability density given by Eq. (1).

2.2. Parameter Conversion [9]

Given the T-frame feature sequence $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$ of the source speaker's voice, the converted parameter sequence $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}} = \left[\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top\right]^\top$ is analytically determined by maximizing the conditional probability of the target feature vector $\boldsymbol{Y}$ given $\boldsymbol{X}$ under a constraint $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$ as follows:

$$\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P\left(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}\right) \quad (3)$$

$$= \boldsymbol{R}_{\hat{\boldsymbol{m}}}^{-1}\boldsymbol{r}_{\hat{\boldsymbol{m}}} = \left(\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{-1}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{-1}\boldsymbol{E}_{\hat{\boldsymbol{m}}}, \quad (4)$$

where $\boldsymbol{W}$ is a $2DT$-by-DT weight matrix to calculate the dynamic features [15]. $\hat{\boldsymbol{m}} = [\hat{m}_1, \cdots, \hat{m}_T]^\top$ is the sub-optimum mixture component sequence determined by maximizing $P\left(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}\right)$, and the other parameters are given by

$$\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{-1} = \operatorname{diag}_{2D}\left[\boldsymbol{\Sigma}_{\hat{m}_1}^{(Y|X)^{-1}}, \cdots, \boldsymbol{\Sigma}_{\hat{m}_T}^{(Y|X)^{-1}}\right]^\top, \quad (5)$$

$$\boldsymbol{E}_{\hat{\boldsymbol{m}}} = \left[\boldsymbol{\mu}_{\hat{m}_1,1}^{(Y|X)^\top}, \cdots, \boldsymbol{\mu}_{\hat{m}_T,T}^{(Y|X)^\top}\right]^\top, \quad (6)$$

$$\boldsymbol{\mu}_{\hat{m}_t,t}^{(Y|X)} = \boldsymbol{A}_{\hat{m}}\boldsymbol{X}_t + \boldsymbol{b}_{\hat{m}}, \quad (7)$$

$$\boldsymbol{A}_{\hat{m}} = \boldsymbol{\Sigma}_{\hat{m}}^{(YX)}\boldsymbol{\Sigma}_{\hat{m}}^{(XX)^{-1}}, \quad (8)$$

$$\boldsymbol{b}_{\hat{m}} = \boldsymbol{\mu}_{\hat{m}}^{(Y)} - \boldsymbol{\Sigma}_{\hat{m}}^{(YX)}\boldsymbol{\Sigma}_{\hat{m}}^{(XX)^{-1}}\boldsymbol{\mu}_{\hat{m}}^{(X)}, \quad (9)$$

$$\boldsymbol{\Sigma}_{\hat{m}}^{(Y|X)} = \boldsymbol{\Sigma}_{\hat{m}}^{(YY)} - \boldsymbol{\Sigma}_{\hat{m}}^{(YX)}\boldsymbol{\Sigma}_{\hat{m}}^{(XX)^{-1}}\boldsymbol{\Sigma}_{\hat{m}}^{(XY)}. \quad (10)$$

The notation diag_{2D} denotes the construction of a block diagonal matrix that has the $2D$-by-$2D$ diagonal elements. The converted parameter sequence $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}$ is efficiently calculated sequence by sequence using the Cholesky decomposition [15], or recursively calculated frame by frame using the low-delay conversion algorithm [18].

3. CONVENTIONAL GV-CONSTRAINED TRAINING

3.1. Global Variance (GV) [9]

The GV $\boldsymbol{v}\left(\boldsymbol{y}\right) = \left[v\left(1\right), \cdots, v\left(D\right)\right]^\top$ is defined as the second order moment of the trajectory $\boldsymbol{y}$, and its d-th component is given as

$$v\left(d\right) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t\left(d\right) - \frac{1}{T}\sum_{\tau=1}^{T} y_\tau\left(d\right)\right)^2. \quad (11)$$

3.2. Objective Function for GV-Constrained Training [16]

A part of the GMM parameter set, $\{\boldsymbol{A}_m, \boldsymbol{b}_m, \boldsymbol{\Sigma}_m^{(Y|X)}\}$ is updated by maximizing the following objective function L_{gv} consisting of the GMM and GV likelihoods:

$$L_{\mathrm{gv}} = P\left(\boldsymbol{W}\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}\right) P\left(\boldsymbol{v}\left(\boldsymbol{y}\right)|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_{\mathrm{v}}\right)^{2\omega_{\mathrm{v}}T}, \quad (12)$$

$$P\left(\boldsymbol{v}\left(\boldsymbol{y}\right)|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_{\mathrm{v}}\right) = \mathcal{N}\left(\boldsymbol{v}\left(\boldsymbol{y}\right); \boldsymbol{v}\left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}\right), \boldsymbol{\Sigma}_{\mathrm{v}}\right), \quad (13)$$

where ω_{v} is a weight of the GV likelihood, $\boldsymbol{\Sigma}_{\mathrm{v}}$ is a covariance matrix of the GV, and $\boldsymbol{\lambda}_{\mathrm{v}}$ is a model parameter set of the GV. The GMM and GV likelihoods are normalized by the ratio of the number of feature dimensions when $\omega_{\mathrm{v}} = 1.0$. This training algorithm updates the model parameters to make the GV of the converted parameter sequence close to natural one.

3.3. Parameter Conversion

The basic parameter conversion process is performed as described in Section 2.2. The converted parameter sequence $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}$ is equivalent to that determined by maximizing L_{gv} under the constraint $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$. Therefore, L_{gv} is also regarded as the objective function for conversion although the constraint needs to be additionally used.

3.4. Problems

There is inconsistency of the optimization criterion between the training and the conversion; i.e., the explicit relationship between the static and dynamic features given by $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$ is ignored in training while it is considered in conversion. Therefore, the trained model parameters are not optimum for conversion.

The GV likelihood is used as a penalty term to alleviate the over-smoothness. Although it can improve the quality of the converted speech, the improved quality is still far from the natural one.

4. IMPLEMENTATION OF TRAJECTORY TRAINING

In [19], the trajectory training has been implemented for the joint probability density modeling in GMM-based voice conversion. In this section, we present yet another implementation by reformulating the conditional probability density function in Eq. (3) by imposing the explicit relationship between the static and dynamic features.

4.1. Objective Function for Training and Conversion

The objective function for the trajectory training is written as:

$$L_{\mathrm{trj}} = P\left(\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}\right) = \mathcal{N}\left(\boldsymbol{y}; \hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}, \boldsymbol{R}_{\hat{\boldsymbol{m}}}^{-1}\right) \quad (14)$$

The mean vector $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}$ is given by Eq. (4) and the inter-frame correlation is effectively modeled by the temporal covariance matrix $\boldsymbol{R}_{\hat{\boldsymbol{m}}}^{-1}$. In training, the GMM parameters are updated by maximizing L_{trj}. In conversion, the basic conversion process described in Section 2.2 is performed. Note that the mean vector $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}$ is equivalent to the generated parameter sequence in the basic conversion process. Therefore, L_{trj} can be regarded as the objective function not only for training but also for conversion.

4.2. Estimation of Model Parameters

Here, let $\boldsymbol{\xi}_{\mathrm{A}} = \left[\boldsymbol{A}_1^\top, \cdots, \boldsymbol{A}_M^\top\right]^\top$, $\boldsymbol{\xi}_{\mathrm{b}} = \left[\boldsymbol{b}_1^\top, \cdots, \boldsymbol{b}_M^\top\right]^\top$, and $\boldsymbol{\Sigma}^{-1} = \left[\boldsymbol{\Sigma}_1^{(Y|X)^{-1}}, \cdots, \boldsymbol{\Sigma}_M^{(Y|X)^{-1}}\right]^\top$ be the joint parameters of $\boldsymbol{A}_m$, $\boldsymbol{b}_m$, and $\boldsymbol{\Sigma}_m^{(Y|X)^{-1}}$ over all mixture components, respectively. To optimize these model parameters for the objective function, we employ the steepest descent algorithm[1] as follows:

$$\boldsymbol{\xi}_{\mathrm{A}}^{(i+1)} = \boldsymbol{\xi}_{\mathrm{A}}^{(i)} + \alpha\left.\frac{\partial \log L_{\mathrm{trj}}}{\partial \boldsymbol{\xi}_{\mathrm{A}}}\right|_{\boldsymbol{\xi}_{\mathrm{A}}=\boldsymbol{\xi}_{\mathrm{A}}^{(i)}}, \quad (15)$$

where α is a learning rate, and i is an iteration index. $\boldsymbol{\xi}_{\mathrm{b}}$ and $\boldsymbol{\Sigma}^{-1}$ are also optimized in the same manner. The gradients are given by:

$$\frac{\partial \log L_{\mathrm{trj}}}{\partial \boldsymbol{\xi}_{\mathrm{A}}} = \boldsymbol{S}_{\hat{\boldsymbol{m}}}^\top \operatorname{diag}_{2D}^{-1}\left[\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{-1}\boldsymbol{W}\left(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}\right)\boldsymbol{X}^\top\right], \quad (16)$$

$$\frac{\partial \log L_{\mathrm{trj}}}{\partial \boldsymbol{\xi}_{\mathrm{b}}} = \boldsymbol{S}_{\hat{\boldsymbol{m}}}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{-1}\boldsymbol{W}\left(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}\right), \quad (17)$$

$$\frac{\partial \log L_{\mathrm{trj}}}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{1}{2}\boldsymbol{S}_{\hat{\boldsymbol{m}}}^\top \operatorname{diag}_{2D}^{-1}\left[\boldsymbol{W}\left(\boldsymbol{R}_{\hat{\boldsymbol{m}}}^{-1} + \hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}^\top - \boldsymbol{y}\boldsymbol{y}^\top\right) - \boldsymbol{E}_{\hat{\boldsymbol{m}}}\left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}} - \boldsymbol{y}\right)^\top \boldsymbol{W}^\top - \boldsymbol{W}\left(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}} - \boldsymbol{y}\right)\boldsymbol{E}_{\hat{\boldsymbol{m}}}^\top\right], \quad (18)$$

[1] Closed form solutions also exist for $\boldsymbol{\xi}_{\mathrm{A}}$ and $\boldsymbol{\xi}_{\mathrm{b}}$.

where $\boldsymbol{S}_{\hat{\boldsymbol{m}}} = [\boldsymbol{S}_{\hat{m}_1}, \cdots, \boldsymbol{S}_{\hat{m}_T}]^\top \otimes \boldsymbol{I}_{2D}$ is a $2DT$-by-$2DM$ matrix, $\boldsymbol{S}_{\hat{m}_t}$ is an M-dimensional vector of which the m-th component is 1 when $m = \hat{m}_t$ and otherwise are 0, and $\boldsymbol{I}_{2D}$ indicates the $2D$-by-$2D$ identity matrix.

In this paper, the traditional joint density training is firstly performed to estimate $\boldsymbol{\lambda}$. Then, the proposed algorithms updates $\{\boldsymbol{\xi}_{\mathrm{A}}, \boldsymbol{\xi}_{\mathrm{b}}, \boldsymbol{\Sigma}^{-1}\}$ while keeping $\{\omega_m, \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_m^{(XX)}\}$ constant. Note that the sub-optimum mixture component sequence $\hat{\boldsymbol{m}}$ never changes in this setting.

5. PROPOSED MODULATION SPECTRUM-CONSTRAINED TRAJECTORY TRAINING

5.1. Modulation Spectrum

Whereas the GV represents the temporal scaling of the parameter trajectory as a scalar value in each feature dimension, the MS explicitly represents the temporal fluctuation as a vector [17]. In this paper, the MS $\boldsymbol{s}(\boldsymbol{y})$ of the parameter sequence $\boldsymbol{y}$ is defined as

$$\boldsymbol{s}(\boldsymbol{y}) = \left[\boldsymbol{s}(1)^\top, \cdots, \boldsymbol{s}(d)^\top, \cdots, \boldsymbol{s}(D)^\top\right]^\top, \quad (19)$$

$$\boldsymbol{s}(d) = \left[s_d(0), \cdots, s_d(f), \cdots, s_d\left(D'_{\mathrm{s}}-1\right)\right]^\top, \quad (20)$$

$$s_d(f) = R_{d,f}^2 + I_{d,f}^2 \quad (21)$$

$$= \left(\sum_{t=1}^{T} y_t(d)\cos kt\right)^2 + \left(\sum_{t=1}^{T} y_t(d)\sin kt\right)^2, \quad (22)$$

where $2D_{\mathrm{s}}$ is a length of Discrete Fourier Transform (DFT), $k = -\pi f / D_{\mathrm{s}}$ is a modulation frequency, and D'_{s} is the number of MS dimension in each feature dimension, where $D'_{\mathrm{s}} < D_{\mathrm{s}}$.[2] In this paper, the MS is calculated utterance by utterance.

5.2. Proposed Objective Function

We integrate the MS compensation into the trajectory training. The objective function consists of both the trajectory likelihood and the MS likelihood as follows:

$$L_{\mathrm{mstrj}} = P(\boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda})\, P(\boldsymbol{s}(\boldsymbol{y})|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_{\mathrm{s}})^{\omega_{\mathrm{s}} T / D'_{\mathrm{s}}}, \quad (23)$$

$$P(\boldsymbol{s}(\boldsymbol{y})|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}, \boldsymbol{\lambda}_{\mathrm{s}}) = \mathcal{N}(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{s}(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}), \boldsymbol{\Sigma}_{\mathrm{s}}), \quad (24)$$

where $\boldsymbol{\lambda}_{\mathrm{s}}$ is a model parameter set of the MS, and $\boldsymbol{\Sigma}_{\mathrm{s}}$ is a $D'_{\mathrm{s}}D$-by-$D'_{\mathrm{s}}D$ covariance matrix, and ω_{s} is a weight of the MS likelihood. The trajectory likelihood and the MS likelihood are normalized by the ratio of the number of feature dimensions when $\omega_{\mathrm{s}} = 1$. $\boldsymbol{\Sigma}_{\mathrm{s}}^{-1}$ is represented as $\left[\boldsymbol{p}_{\mathrm{s}}^{(1)}, \cdots, \boldsymbol{p}_{\mathrm{s}}^{(d)}, \cdots, \boldsymbol{p}_{\mathrm{s}}^{(D)}\right]$, where $\boldsymbol{p}_{\mathrm{s}}^{(d)}$ is $D'_{\mathrm{s}}D$-by-D matrix of which columns correspond to $\boldsymbol{s}(d)$. The MS likelihood works as a penalty term to the reduction of the temporal fluctuation of the converted parameter sequence.

5.3. Estimation of Model Parameters

The model parameters are estimated in the same way as the trajectory training. Let L_{ms} be the MS likelihood $\mathcal{N}(\boldsymbol{s}(\boldsymbol{y}); \boldsymbol{s}(\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}), \boldsymbol{\Sigma}_{\mathrm{s}})$. The logarithm function of L_{mstrj} is given by

$$\log L_{\mathrm{mstrj}} = \log L_{\mathrm{trj}} + \omega_{\mathrm{s}} \frac{T}{D'_{\mathrm{s}}} \log L_{\mathrm{ms}}, \quad (25)$$

[2] Because lower modulation frequency components mainly affect speech perception [20, 21], only these components are considered in this paper.

and the gradients of $\log L_{\mathrm{ms}}$ are given as

$$\frac{\partial \log L_{\mathrm{ms}}}{\partial \boldsymbol{\xi}_{\mathrm{A}}} = \boldsymbol{S}_{\hat{\boldsymbol{m}}}^\top \mathrm{diag}_{2D}^{-1}\left[\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{-1}\boldsymbol{W}\boldsymbol{R}_{\hat{\boldsymbol{m}}}^{-1}\boldsymbol{s}_{\hat{\boldsymbol{m}}}\boldsymbol{X}^\top\right], \quad (26)$$

$$\frac{\partial \log L_{\mathrm{ms}}}{\partial \boldsymbol{\xi}_{\mathrm{b}}} = \boldsymbol{S}_{\hat{\boldsymbol{m}}}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{-1}\boldsymbol{W}\boldsymbol{R}_{\hat{\boldsymbol{m}}}^{-1}\boldsymbol{s}_{\hat{\boldsymbol{m}}}, \quad (27)$$

$$\frac{\partial \log L_{\mathrm{ms}}}{\partial \boldsymbol{\Sigma}^{-1}} = \boldsymbol{S}_{\hat{\boldsymbol{m}}}^\top \mathrm{diag}_{2D}^{-1}\left[\boldsymbol{W}\boldsymbol{R}_{\hat{\boldsymbol{m}}}^{-1}\boldsymbol{s}_{\hat{\boldsymbol{m}}}\left(\boldsymbol{E}_{\hat{\boldsymbol{m}}} - \boldsymbol{W}\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}\right)\right], \quad (28)$$

where

$$\boldsymbol{s}_{\hat{\boldsymbol{m}}} = \left[\boldsymbol{s}_1'^\top, \cdots, \boldsymbol{s}_t'^\top, \cdots, \boldsymbol{s}_T'^\top\right]^\top, \quad (29)$$

$$\boldsymbol{s}'_t = \left[s_t(1), \cdots, s_t(d), \cdots, s_t(D)\right]^\top, \quad (30)$$

$$s_t(d) = 2\boldsymbol{f}_t(d)\,\boldsymbol{p}_{\mathrm{s}}^{(d)\top}\left(\boldsymbol{s}(\boldsymbol{y}) - \boldsymbol{s}(\hat{\boldsymbol{y}})\right), \quad (31)$$

$$\boldsymbol{f}_t(d) = \left[f_{t,d}(0), \cdots, f_{t,d}(f), \cdots, f_{t,d}\left(D'_{\mathrm{s}}-1\right)\right]^\top, \quad (32)$$

$$f_{t,d}(f) = \hat{R}_{d,f}\cos kt + \hat{I}_{d,f}\sin kt, \quad (33)$$

where $\hat{R}_{d,f}$ and $\hat{I}_{d,f}$ are calculated using the d-th dimensional components of $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}$.

5.4. Parameter Conversion

It is unnecessary to consider the MS in parameter conversion because the GMM parameters are optimized to make the MS of the converted parameter sequence close to the natural one. Consequently, the basic parameter conversion algorithm is straightforwardly employed. If the proposed objective function L_{mstrj} is used in the parameter conversion, the converted parameter sequence to maximize it is equivalent to $\hat{\boldsymbol{y}}_{\hat{\boldsymbol{m}}}$ which is analytically solved. Therefore, the proposed framework can also be regarded as a unified framework between the training and conversion process.

Because the MS involves the GV [17], the proposed MS-constrained algorithm well recovers not only the MS but also the GV. The GV is effectively recovered by the MS-constrained trajectory training ("MSTRJ") as observed in Fig. 1.

6. EXPERIMENTAL EVALUATION

6.1. Experimental Conditions

In our experiments, we prepared 2 English speakers (rms and slt) in the CMU ARCTIC database [22]. We used 50 sentences were selected for training and remaining 50 sentences for evaluation. We trained the slt-to-rms GMM. Speech signals were sampled at 16 kHz. The shift length was set to 5 ms. The 1st-through-24th mel-cepstral coefficients were used as a spectral parameter and log-scaled F_0 and 5 band-aperiodicity [23, 24] were used as excitation parameters. The STRAIGHT analysis-synthesis system [25] was employed for pa-

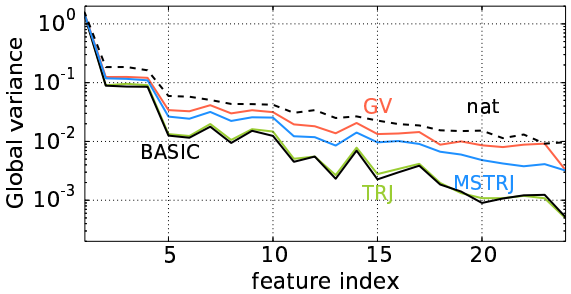

Fig. 1. An example of the GVs of the converted mel-cepstral coefficients ("nat" indicates natural speech parameter trajectories).

rameter extraction and waveform generation. The spectral parameters and aperiodic components were converted through a 64-mixture GMM and a 16-mixture GMM, respectively. The log-scaled F_0 was linearly converted. The DFT length to calculate MS was set to 2048 that covers the length of all training utterances. The conventional GV-constrained training algorithm and the proposed algorithms were applied only to spectral component, and the likelihood weight ω_{v} and ω_{s} were set to 1.0. According to the results of our preliminary test[3], D'_{s} was set to $D_{\mathrm{s}}/2$ ($= 50$ Hz).

We compared the following training algorithms:

BASIC : Basic training [8]
GV : Conventional training considering the GV [16]
TRJ : Proposed trajectory training
MSTRJ : Proposed trajectory training considering the MS

The trajectory likelihood and the MS likelihood for the natural parameter trajectories of the evaluation data were firstly calculated to analyze the effect of the proposed algorithms. Then, the speech quality and the speaker individuality of the converted speech are evaluated in the perceptual evaluation.

6.2. Objective Evaluation

Fig. 2 and Fig. 3 illustrate the trajectory likelihood L_{trj} and the MS likelihood L_{ms} for the natural parameter trajectories of the evaluation data, respectively. The trajectory likelihood is normalized by the total number of frames T. The trajectory training ("TRJ") slightly improves the MS likelihood as well as the trajectory likelihood compared to the basic training ("BASIC"). This result shows that the proposed trajectory training models the parameter trajectories more accurately than the basic training.

The MS-constrained trajectory training ("MSTRJ") well improves the MS likelihood than the other algorithms. This result demonstrates that the MS of the converted parameter trajectory is well recovered by "MSTRJ." It is observed that the trajectory likelihood is significantly degraded by the conventional GV-constrained training ("GV"). This is because "GV" uses the inconsistent criteria between training and conversion. This likelihood degradation is reduced by "MSTRJ."

6.3. Subjective Evaluation

In the evaluation of the speech quality, a preference test (AB test) was conducted. We presented every pair of converted speech of 4 algorithms in a random order, and we forced listeners to select speech sample that sounds better quality. Similarity, XAB test on speaker individuality was conducted using the analysis-synthesized speech as a reference "X." 6 listeners participated in each assessment.

The results are illustrated in Fig. 4 and Fig. 5. It is observed that "TRJ" has higher scores than "BASIC" in term of both the speech quality and the speaker individuality. On the other hand, the scores of "TRJ" are lower than "GV." Therefore, the effect of the GV compensation on the converted speech is larger than that of the trajectory training. We can see that "MSTRJ" achieves the best scores than others in term of the speech quality. This result demonstrates that the proposed MS-constrained trajectory training yields the best performance among these training methods.

[3]We conducted the preliminary subjective test to investigate the quality-wise effect of higher modulation frequency component of MS. As a result, there is no significant difference in quality between analysis-synthesized speech and the speech that MS over 50 Hz was cut.

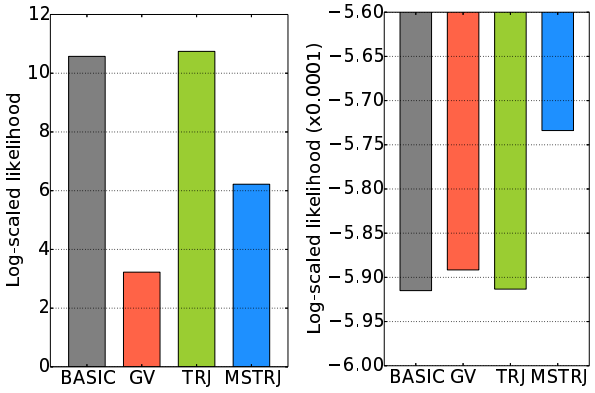

Fig. 2. Trajectory likelihood for the natural parameter trajectories.

Fig. 3. MS likelihood for the natural parameter trajectories.

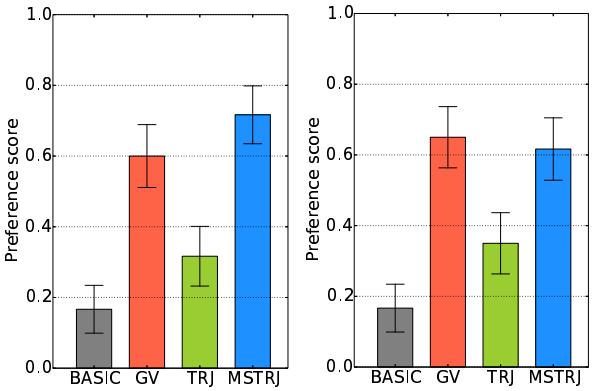

Fig. 4. Preference scores on speech quality with 95 % confidence intervals.

Fig. 5. Preference scores on speaker individuality with 95 % confidence intervals.

7. CONCLUSION

This paper have proposed novel training algorithms for GMM-based voice conversion in order to produce the high-quality speech while preserving the computationally-efficient conversion algorithm. Traditional GMM have been firstly reformulated as the trajectory model called "trajectory GMM," then, the Modulation Spectrum (MS) have been integrated into the trajectory training. The experimental results yielded the significant improvements in term of both the speech quality and the speaker individuality of the converted speech. As future work, we will implement the proposed algorithm for HMM-based speech synthesis, and apply the proposed algorithm to the voice conversion for arbitrary speakers [26, 27].

Acknowledgements: Part of this work was supported by JSPS KAKENHI Grant Number 26280060 and Grant-in-Aid for JSPS Fellows Grant Number $26 \cdot 10354$, and part of this work was executed under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation."

8. REFERENCES

[1] K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. An evaluation of excitation feature prediction in a hybrid approach to electrolaryngeal speech enhancement. In *Proc. ICASSP*, pp. 4521–4525, Florence, Italy, May 2014.

[2] K. Kobayashi, T. Toda, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura. Regression approaches to perceptual age control in singing voice conversion. In *Proc. ICASSP*, pp. 7954–7958, Florence, Italy, May 2014.

[3] N. Hattori, T. Toda, H. Kawai, H. Saruwatari, and K. Shikano. Speaker-adaptive speech synthesis based on eigenvoice conversion and language-dependent prosodic conversion in speech-to-speech translation. In *Proc. INTERSPEECH*, pp. 2769–2772, Florence, Italy, Aug. 2011.

[4] S. Aryal and R. G.-Osuna. Can voice conversion be used to reduce non-native accents? In *Proc. ICASSP*, Florence, Italy, May 2014.

[5] T. Nakashika, T. Takiguchi, and Y. Ariki. Voice conversion in time-invariant speaker independent space. In *Proc. ICASSP*, pp. 7939–7943, Frorence, Italy, May 2014.

[6] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li. Exemplar-based voice conversion using non-negative spectrogram deconvolution. In *Proc. 8th ISCA SSW*, Catalunya, Spain, Aug. 2013.

[7] E. Helander, H. Silen, T. Virtanen, , and M. Gabbouj. Voice conversion using dynamic kernel partial least squares regression. *IEEE Trans.*, Vol. 20, No. 3, pp. 806–817, Mar. 2012.

[8] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 2, pp. 131–142, Mar. 1988.

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

[10] K. Tokuda, T. Kobayashi, and S. Imai. Speech parameter generation from HMM using dynamic features. In *Proc. ICASSP*, pp. 660–663, Detroit, U.S.A, May 1995.

[11] T. Toda, T. Muramatsu, and H. Banno. Implementation of computationally efficient real-time voice conversion. In *Proc. INTERSPEECH*, Portland, Oregon, U.S., Sept. 2012.

[12] H. Zen, K. Tokuda, and T. Kitamura. Refomulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. *Computer Speech and Language*, Vol. 21, No. 1, pp. 153–173, Jan. 2007.

[13] K. Tokuda, Y Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura. Speech synthesis based on hidden Markov models. *Proceedings of the IEEE*, Vol. 101, No. 5, pp. 1234–1252, 2013.

[14] T. Toda and S. Young. Trajectory training considering global variance for HMM-based speech synthesis. In *Proc. ICASSP*, pp. 4025–4028, Taipei, Taiwan, Aug. 2009.

[15] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*, pp. 1315–1318, Istanbul, Turkey, June 2000.

[16] H. Hwang, Y. Tsao, H. Wang, Y. Wang, and S. Chen. Incorporating global variance in the training phase of GMM-based voice conversion. In *Proc. APSIPA*, pp. 1–6, Kaohsiung, Taiwan, Oct. 2013.

[17] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. A postfilter to modify modulation spectrum in HMM-based speech synthesis. In *Proc. ICASSP*, pp. 290–294, Florence, Italy, May 2014.

[18] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory. In *Proc. INTERSPEECH*, pp. 1076–1079, Brisbane, Australia, Sep. 2008.

[19] H. Zen, Y. Nankaku, and K. Tokuda. Continuous stochastic feature mapping based on trajectory HMMs. *IEEE Trans.*, Vol. 19, pp. 417–430, Jan. 2011.

[20] R. Drullman, J .M. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. of America*, Vol. 95, pp. 2670–2680, 1994.

[21] S. Thomas, S. Ganapathy, and H. Hermansky. Phoneme recgnition using spectral envelop and modulation frequency features. In *Proc. ICASSP*, pp. 4453–4456, Taipei, Taiwan, April 2009.

[22] J. Kominek and A. W Black. The CMU ARCTIC speech databases for speech synthesis research. In *Tech. Rep. CMU-LTI-03-177*, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, U.S.A, 2003.

[23] H. Kawahara, Jo Estill, and O. Fujimura. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT ” . In *MAVEBA 2001*, pp. 1–6, Firentze, Italy, Sept. 2001.

[24] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation. In *Proc. INTERSPEECH*, pp. 2266–2269, Pittsburgh, U.S.A., Sept. 2006.

[25] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne. 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 Commun.*, Vol. 27, No. 3–4, pp. 187–207, 1999.

[26] T. Toda, Y. Ohtani, and K. Shikano. One-to-many and many-to-one voice conversion based on eigenvoices. In *Proc. ICASSP*, pp. 1249–1252, Hawaii, U.S.A., Apr. 2007.

[27] Y. Ohtani, T. Toda, H. Saruwatari, and S. Shikano. Many-to-many eigenvoice conversion with reference voice. In *Proc. INTERSPEECH*, pp. 1623–1626, Brington U.K., Sep. 2009.