Incorporating Global Variance in the Training Phase of GMM-based Voice Conversion

Hsin-Te Hwang[1,3], Yu Tsao[2], Hsin-Min Wang[3], Yih-Ru Wang[1], Sin-Horng Chen[1]
[1]Dept. of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan
[2]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[3]Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: hwanght@iis.sinica.edu.tw, yu.tsao@citi.sinica.edu.tw, whm@iis.sinica.edu.tw, yrwang@cc.nctu.edu.tw, schen@mail.nctu.edu.tw

Abstract—**Maximum likelihood-based trajectory mapping considering global variance (MLGV-based trajectory mapping) has been proposed for improving the quality of the converted speech of Gaussian mixture model-based voice conversion (GMM-based VC). Although the quality of the converted speech is significantly improved, the computational cost of the online conversion process is also increased because there is no closed form solution for parameter generation in MLGV-based trajectory mapping, and an iterative process is generally required. To reduce the online computational cost, we propose to incorporate GV in the training phase of GMM-based VC. Then, the conversion process can simply adopt ML-based trajectory mapping (without considering GV in the conversion phase), which has a closed form solution. In this way, it is expected that the quality of the converted speech can be improved without increasing the online computational cost. Our experimental results demonstrate that the proposed method yields a significant improvement in the quality of the converted speech comparing to the conventional GMM-based VC method. Meanwhile, comparing to MLGV-based trajectory mapping, the proposed method provides comparable converted speech quality with reduced computational cost in the conversion process.**

I. INTRODUCTION

Voice conversion (VC) is a technique that transforms a source speaker's voice to that of a specific target speaker [1-11]. The overall VC includes two parts, namely spectral and prosody conversions. This study focuses on spectral conversion (SC). Many SC approaches have been proposed in the past. Among them, GMM-based SC is a successful one. A typical framework of GMM-based SC consists of training and conversion phases. In the training phase, a parallel speech corpus is first prepared. Then, a preprocessing step is performed to time-align the spectral feature sequences of the source and target feature vectors. Finally, a Gaussian mixture model (GMM) is employed to model the feature vectors of the source speaker [1] or the joint feature vectors of the source and target speakers [2]. In the conversion phase, least squares (LS)-based mapping [1], minimum mean square error-based mapping (MMSE-based mapping) [2], maximum likelihood-based trajectory mapping (ML-based trajectory mapping) [3], or maximum mutual information-based trajectory mapping (MMI-based trajectory mapping) [6] is then adopted to convert a source feature sequence to a target feature sequence in a frame by frame [1, 2, 9] or sequence by sequence fashion [3, 6-8, 10, 11]. Finally, a vocoding technique is employed to reconstruct the speech waveform from the converted feature sequence.

Although GMM-based SC has been proven effective, several issues remain unsolved, e.g., over-smoothing [3-9], time independent mapping [3, 10], and limited flexibility [11]. In this study, we focus our attention on handling the over-smoothing problem in the GMM-based SC. The over-smoothing problem occurs when the spectra are smoothed overly, thereby degrading the quality of the converted speech. Several methods have been proposed to handle this problem [3-9]. In this study, we focus on the global variance (GV)-based method [3]. Toda *et al.* [3] handle the over-smoothing problem by incorporating the global variance (GV) into the ML-based trajectory mapping (referred to as MLGV-based mapping) process. Although the quality of the converted speech is significantly improved, the computational cost is also increased accordingly due to the iterative process for parameter generation. To overcome this problem, we propose to integrate GV in the training phase of GMM-based VC and use simple ML-based trajectory mapping instead of complex MLGV-based mapping in the online conversion phase.

In the training phase of the proposed method, the model parameters are optimized with a new objective function, which consists of conditional PDF and GV likelihoods. The GV likelihood works as a penalty term to make the GV of the converted static feature sequence closer to that of the target (natural) one. Experimental results demonstrate that after updating the model parameters by the proposed method, the GV of the converted feature sequence (obtained by ML-based trajectory mapping) is closer to that of the natural one. Accordingly, the quality of the converted speech is dramatically improved, and is almost as good as that of MLGV-based mapping; but with reduced computational cost in the conversion process.

The remainder of this paper is organized as follows. Section 2 reviews the conventional GMM-based SC with ML-based trajectory mapping. Section 3 describes the proposed method. Section 4 presents our experimental results. Finally, the conclusion is given in the last section.

II. GMM-BASED SPECTRAL CONVERSION

A. Training a JDGMM

Let $\mathbf{X}=\left[\mathbf{X}_1^{\mathrm{T}},\cdots,\mathbf{X}_t^{\mathrm{T}},\cdots,\mathbf{X}_T^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\mathbf{Y}=\left[\mathbf{Y}_1^{\mathrm{T}},\cdots,\mathbf{Y}_t^{\mathrm{T}},\cdots,\mathbf{Y}_T^{\mathrm{T}}\right]^{\mathrm{T}}$ be the source and target feature vectors, respectively; both feature vectors consist of *T* feature frames; $\mathbf{X}_t=[\mathbf{x}_{S_t}^{\mathrm{T}},\mathbf{x}_{D_t}^{\mathrm{T}}]^{\mathrm{T}}$ and $\mathbf{Y}_t=[\mathbf{y}_{S_t}^{\mathrm{T}},\mathbf{y}_{D_t}^{\mathrm{T}}]^{\mathrm{T}}$ are the 2*D*-dimensional source and target feature vectors consisting of *D* static and *D* dynamic feature vectors at frame *t*. The superscript T denotes the vector transposition. A joint density GMM (JDGMM) is employed to model the joint feature vector $\mathbf{Z}_t=[\mathbf{X}_t^{\mathrm{T}},\mathbf{Y}_t^{\mathrm{T}}]^{\mathrm{T}}$ as

$$P(\mathbf{Z}_t\mid\Theta^{(\mathbf{Z})})=\sum_{m=1}^{M}\omega_m N(\mathbf{Z}_t;\boldsymbol{\mu}_m^{(\mathbf{Z})},\boldsymbol{\Sigma}_m^{(\mathbf{Z})})\ , \qquad (1)$$

where ω_m is the prior probability of the *m*th mixture component; $\boldsymbol{\mu}_m^{(\mathbf{Z})}=[(\boldsymbol{\mu}_m^{(\mathbf{X})})^{\mathrm{T}},(\boldsymbol{\mu}_m^{(\mathbf{Y})})^{\mathrm{T}}]^{\mathrm{T}}$ and $\boldsymbol{\Sigma}_m^{(\mathbf{Z})}=\begin{bmatrix}\boldsymbol{\Sigma}_m^{(\mathbf{XX})} & \boldsymbol{\Sigma}_m^{(\mathbf{XY})}\\ \boldsymbol{\Sigma}_m^{(\mathbf{YX})} & \boldsymbol{\Sigma}_m^{(\mathbf{YY})}\end{bmatrix}$ are the mean vector and covariance matrix of the *m*th mixture component. The covariance matrices $\boldsymbol{\Sigma}_m^{(\mathbf{XX})}$, $\boldsymbol{\Sigma}_m^{(\mathbf{XY})}$, $\boldsymbol{\Sigma}_m^{(\mathbf{YX})}$, and $\boldsymbol{\Sigma}_m^{(\mathbf{YY})}$ are usually diagonal. The parameter set $\Theta^{(\mathbf{Z})}=\{\omega_m,\boldsymbol{\mu}_m^{(\mathbf{Z})},\boldsymbol{\Sigma}_m^{(\mathbf{Z})}\}_{m=1,2,\cdots,M}$ can be estimated via the expectation maximization (EM) algorithm.

B. ML-Based Trajectory Mapping

In the conversion phase, the ML-based trajectory mapping [3] is adopted to obtain the converted static feature sequence $\hat{\mathbf{y}}$ as

$$\hat{\mathbf{y}}=\arg\max P(\mathbf{Y}\mid\mathbf{X},\Theta^{(\mathbf{Y}|\mathbf{X})}),\ \text{s.t. } \mathbf{Y}=\mathbf{W}\mathbf{y}\ , \qquad (2)$$

where $\mathbf{W}$ is a 2*DT*-by-*DT* weighting matrix (given in [3]) for computing the joint static and dynamic features; $P(\mathbf{Y}\mid\mathbf{X},\Theta^{(\mathbf{Y}|\mathbf{X})})$ is the conditional probability density function (PDF) given as

$$P(\mathbf{Y}\mid\mathbf{X},\Theta^{(\mathbf{Y}|\mathbf{X})})=\sum_{\text{all }\mathbf{m}}P(\mathbf{m}\mid\mathbf{X},\Theta^{(\mathbf{Y}|\mathbf{X})})P(\mathbf{Y}\mid\mathbf{X},\mathbf{m},\Theta^{(\mathbf{Y}|\mathbf{X})}),\ (3)$$

where $\mathbf{m}=[m_1,\cdots,m_t,\cdots,m_T]$ is an arbitrary mixture component sequence. (3) can be further represented as

$$P(\mathbf{Y}\mid\mathbf{X},\Theta^{(\mathbf{Y}|\mathbf{X})})=\prod_{t=1}^{T}\sum_{m=1}^{M}P(m\mid\mathbf{X}_t,\Theta^{(\mathbf{Y}|\mathbf{X})})P(\mathbf{Y}_t\mid\mathbf{X}_t,m,\Theta^{(\mathbf{Y}|\mathbf{X})})\ , \qquad (4)$$

where

$$P(m\mid\mathbf{X}_t,\Theta^{(\mathbf{Y}|\mathbf{X})})=\frac{\omega_m N(\mathbf{X}_t;\boldsymbol{\mu}_m^{(\mathbf{X})},\boldsymbol{\Sigma}_m^{(\mathbf{XX})})}{\sum_{l=1}^{M}\omega_l N(\mathbf{X}_t;\boldsymbol{\mu}_l^{(\mathbf{X})},\boldsymbol{\Sigma}_l^{(\mathbf{XX})})}, \qquad (5)$$

$$P(\mathbf{Y}_t\mid\mathbf{X}_t,m,\Theta^{(\mathbf{Y}|\mathbf{X})})\ =N(\mathbf{Y}_t;\boldsymbol{\mu}_{m,t}^{(\mathbf{Y}|\mathbf{X})},\boldsymbol{\Sigma}_m^{(\mathbf{Y}|\mathbf{X})})\ . \qquad (6)$$

The mean vector and the covariance matrix of the conditional PDF, $P(\mathbf{Y}_t\mid\mathbf{X}_t,m,\Theta^{(\mathbf{Z})})$, are given as

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

$$\boldsymbol{\Sigma}_m^{(\mathbf{Y}|\mathbf{X})}=\boldsymbol{\Sigma}_m^{(\mathbf{YY})}-\mathbf{A}_m\boldsymbol{\Sigma}_m^{(\mathbf{XX})}\mathbf{A}_m^{\mathrm{T}}\ , \qquad (8)$$

where the transformation matrix $\mathbf{A}_m$ and the bias vector $\mathbf{b}_m$ are given as

$$\mathbf{A}_m=\boldsymbol{\Sigma}_m^{(\mathbf{YX})}\boldsymbol{\Sigma}_m^{(\mathbf{XX})-1},\ \mathbf{b}_m=\boldsymbol{\mu}_m^{(\mathbf{Y})}-\mathbf{A}_m\boldsymbol{\mu}_m^{(\mathbf{X})}\ . \qquad (9)$$

The parameter set $\Theta^{(\mathbf{Y}|\mathbf{X})}=\{\omega_m,\boldsymbol{\mu}_m^{(\mathbf{X})},\boldsymbol{\Sigma}_m^{(\mathbf{XX})},\mathbf{A}_m,\mathbf{b}_m,\boldsymbol{\Sigma}_m^{(\mathbf{Y}|\mathbf{X})}\}_{m=1}^{M}$ can be directly obtained from the parameter set, $\Theta^{(\mathbf{Z})}$, of the ML-trained JDGMM. In practical implementation, a suboptimal solution is often employed to ML-based trajectory mapping [3]. In this case, the conditional PDF in (2), is approximated as $P(\mathbf{Y}\mid\mathbf{X},\Theta^{(\mathbf{Y}|\mathbf{X})})\approx P(\hat{\mathbf{m}}\mid\mathbf{X},\Theta^{(\mathbf{Y}|\mathbf{X})})P(\mathbf{Y}\mid\mathbf{X},\hat{\mathbf{m}},\Theta^{(\mathbf{Y}|\mathbf{X})})$, where the mixture component sequence is determined by $\hat{\mathbf{m}}=\arg\max P(\mathbf{m}\mid\mathbf{X},\Theta^{(\mathbf{Y}|\mathbf{X})})$. As a result, the converted static feature sequence $\hat{\mathbf{y}}$ can be obtained by substituting the approximated conditional PDF into (2) as

$$\begin{aligned}\hat{\mathbf{y}}_{\hat{\mathbf{m}}}&=\arg\max P(\hat{\mathbf{m}}\mid\mathbf{X},\Theta^{(\mathbf{Y}|\mathbf{X})})P(\mathbf{Y}\mid\mathbf{X},\hat{\mathbf{m}},\Theta^{(\mathbf{Y}|\mathbf{X})})\\ &\text{s.t. }\mathbf{Y}=\mathbf{W}\mathbf{y}\end{aligned}\ . \qquad (10)$$

Finally, by solving (10), the converted static feature sequence is given by

$$\hat{\mathbf{y}}_{\hat{\mathbf{m}}}=(\mathbf{W}^{\mathrm{T}}\mathbf{D}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})-1}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{D}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})-1}\mathbf{E}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})}\ , \qquad (11)$$

where

$$\begin{aligned}\mathbf{E}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})}&=[(\boldsymbol{\mu}_{\hat{m},1}^{(\mathbf{Y}|\mathbf{X})})^{\mathrm{T}},\cdots,(\boldsymbol{\mu}_{\hat{m},t}^{(\mathbf{Y}|\mathbf{X})})^{\mathrm{T}},\cdots,(\boldsymbol{\mu}_{\hat{m},T}^{(\mathbf{Y}|\mathbf{X})})^{\mathrm{T}}]^{\mathrm{T}}\\ \mathbf{D}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})-1}&=\mathrm{daig}[\boldsymbol{\Sigma}_{\hat{m},1}^{(\mathbf{Y}|\mathbf{X})-1},\cdots,\boldsymbol{\Sigma}_{\hat{m},t}^{(\mathbf{Y}|\mathbf{X})-1},\cdots,\boldsymbol{\Sigma}_{\hat{m},T}^{(\mathbf{Y}|\mathbf{X})-1}]\end{aligned}\ ; \qquad (12)$$

$\boldsymbol{\mu}_{\hat{m},t}^{(\mathbf{Y}|\mathbf{X})}$ and $\boldsymbol{\Sigma}_{\hat{m},t}^{(\mathbf{Y}|\mathbf{X})-1}$ can be calculated by (7)-(9), respectively, with determined $\hat{m}$.

III. THE PROPOSED TRAINING METHOD

A. The Objective Function

The goal of the proposed training method is to integrate the GV into the training phase of GMM-based VC. The proposed objective function is defined as

$$\begin{aligned}L_{PROP}&=L_{CPDF}+\omega\cdot L_{GV},\\ L_{CPDF}&=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{T_n}\log P(\hat{\mathbf{m}}_n\mid\mathbf{X}_n,\Theta^{(\mathbf{Y}|\mathbf{X})})P(\mathbf{Y}_n\mid\mathbf{X}_n,\hat{\mathbf{m}}_n,\Theta^{(\mathbf{Y}|\mathbf{X})}),\\ L_{GV}&=\frac{1}{N}\sum_{n=1}^{N}\log P(v(\mathbf{y}_n)\mid\hat{\mathbf{m}}_n,\Theta^{(\mathbf{Y}|\mathbf{X})},\Theta^{(v)}),\end{aligned} \qquad (13)$$

where the power weight ω controls the balance between the two log-scaled likelihoods L_{CPDF} and L_{GV} ; *n* denotes the *n*th utterance; *N* is the total number of training utterances; T_n is the total number of frames of the *n*th utterance; $\hat{\mathbf{m}}_n$ is the mixture component sequence of the *n*th utterance determined by the source feature vectors; $v(\mathbf{y})=[v(1),\cdots,v(d),\cdots,v(D)]^{\mathrm{T}}$ (omitting utterance index *n* in (13) hereafter for clarity) is a GV vector of the target static feature sequence and is calculated in an utterance by utterance manner as

$$v(d)=\frac{1}{T}\sum_{t=1}^{T}\left((y_t(d)-\langle y(d)\rangle)\right)^2 , \tag{14}$$

$$\langle y(d)\rangle=\frac{1}{T}\sum_{t=1}^{T}y_t(d), \tag{15}$$

where $v(d)$ is the GV of the dth dimension; $v(\mathbf{y})$ is modeled by a single Gaussian as

$$P(v(\mathbf{y})\,|\,\hat{\mathbf{m}},\Theta^{(\mathbf{Y}|\mathbf{X})},\Theta^{(v)})=N(v(\mathbf{y});v(\hat{\mathbf{y}}_{\hat{\mathbf{m}}}),\boldsymbol{\Sigma}^{(vv)}) . \tag{16}$$

The covariance matrix $\boldsymbol{\Sigma}^{(vv)}$ of the GV probability density can be obtained using the GVs of the target feature sequences calculated from individual utterances in the training data as described in [3]. Note that the converted feature sequence $\hat{\mathbf{y}}_{\hat{\mathbf{m}}}$ is defined as the result of suboptimal ML-based trajectory mapping given in (11), and the mean vector of the GV probability density is defined as the GV of the converted static feature sequence. As can be seen in (13), the proposed training method is employed to update the parameter set $\Theta^{(\mathbf{Y}|\mathbf{X})}$ such that the conditional PDF (L_{CPDF}) and the GV (L_{GV}) likelihoods are maximized. It is worth noting that the GV likelihood works as a penalty term to make the GV of the converted static feature sequence $v(\hat{\mathbf{y}}_{\hat{\mathbf{m}}})$ closer to that of the target (natural) one.

B. Implementation

In the implementation, we first prepare an ML-trained JDGMM with the parameter set, $\Theta^{(\mathbf{Z})}$. Then, the mixture component $\hat{m}$ at frame t is determined by the source feature vector as $\hat{m}=\arg\max P(m\,|\,\mathbf{X}_t,\Theta^{(\mathbf{Y}|\mathbf{X})})$. Finally, the updated parameter set $\Theta^{(\mathbf{Y}|\mathbf{X})}$ can be obtained by maximizing the proposed objective function (13).

The proposed training method adopts a gradient-based approach to update the model parameters. In this study, we only update the transformation and bias parameters, i.e., $\{\mathbf{A}_m,\mathbf{b}_m\}_{m=1}^{M}$; while the remaining parameters are kept the same as those of the conditional probability density function (PDF), $P(\mathbf{Y}\,|\,\mathbf{X},\Theta^{(\mathbf{Y}|\mathbf{X})})$, derived from the ML-trained JDGMM set. For simplicity, the parameter set $\{\mathbf{A}_m,\mathbf{b}_m\}$ is denoted by $\boldsymbol{\Phi}_m$ in the following discussion. The gradient-based approach updates $\boldsymbol{\Phi}_m$ by

$$\Phi_m(l+1)=\Phi_m(l)+\varepsilon\frac{1}{N}\sum_{n=1}^{N}\frac{\partial L'_{PROP}}{\partial\Phi_m(l)}, \tag{17}$$

where l denotes the lth iteration number, ε is the step size, and

$$\begin{aligned}
L'_{PROP}&=L'_{CPDF}+\omega\cdot L'_{GV},\\
L'_{CPDF}&=\frac{1}{T}\log P(\hat{\mathbf{m}}\,|\,\mathbf{X},\Theta^{(\mathbf{Y}|\mathbf{X})})P(\mathbf{Y}\,|\,\mathbf{X},\hat{\mathbf{m}},\Theta^{(\mathbf{Y}|\mathbf{X})}),\\
L'_{GV}&=\log P(v(\mathbf{y})\,|\,\hat{\mathbf{m}},\Theta^{(\mathbf{Y}|\mathbf{X})},\Theta^{(v)}).
\end{aligned} \tag{18}$$

Here, we denote the vector $\mathbf{a}_m$ as the diagonal elements of the transformation matrix $\mathbf{A}_m$. Then, by applying the partial derivative on (18) with respect to $\mathbf{a}_m$ and $\mathbf{b}_m$, we have

$$\begin{aligned}
\frac{\partial L'_{PROP}}{\partial\mathbf{a}_m}&=\mathbf{P}_{\hat{\mathbf{m}}}\left[\frac{1}{T}\mathbf{U}_{\hat{\mathbf{m}}}\circ\mathbf{X}+\omega(\mathbf{V}_{\hat{\mathbf{m}}}^{\mathrm{T}}\mathbf{Q}_{\hat{\mathbf{m}}})^{\mathrm{T}}\circ\mathbf{X}\right],\\
\frac{\partial L'_{PROP}}{\partial\mathbf{b}_m}&=\mathbf{P}_{\hat{\mathbf{m}}}\left[\frac{1}{T}\mathbf{U}_{\hat{\mathbf{m}}}+\omega(\mathbf{V}_{\hat{\mathbf{m}}}^{\mathrm{T}}\mathbf{Q}_{\hat{\mathbf{m}}})^{\mathrm{T}}\right],
\end{aligned} \tag{19}$$

where $\mathbf{P}_{\hat{\mathbf{m}}}$ is a $2D$-by-$2DT$ matrix whose elements are 0 or 1 determined according to the mixture component sequence $\hat{\mathbf{m}}$; $\mathbf{U}_{\hat{\mathbf{m}}}$ is a $2DT$-by-1 vector given by

$$\begin{aligned}
\mathbf{U}_{\hat{\mathbf{m}}}&=\left[\mathbf{U}_{\hat{\mathbf{m}},1}{}^{\mathrm{T}},\cdots,\mathbf{U}_{\hat{\mathbf{m}},t}{}^{\mathrm{T}},\cdots,\mathbf{U}_{\hat{\mathbf{m}},T}{}^{\mathrm{T}}\right]^{\mathrm{T}},\\
\mathbf{U}_{\hat{\mathbf{m}},t}&=\left[\mathbf{U}_{\hat{\mathbf{m}},t}(1),\cdots,\mathbf{U}_{\hat{\mathbf{m}},t}(d),\cdots,\mathbf{U}_{\hat{\mathbf{m}},t}(2D)\right]^{\mathrm{T}},\\
\mathbf{U}_{\hat{\mathbf{m}},t}(d)&=\frac{y_t(d)-\mu_{m,t}^{(\mathbf{Y}|\mathbf{X})}(d)}{\left(\sigma_m^{(\mathbf{Y}|\mathbf{X})}(d)\right)^2},
\end{aligned} \tag{20}$$

where d represents the dth dimension; $y_t(d)$ is the target feature vector at frame t; $\mu_{m,t}^{(\mathbf{Y}|\mathbf{X})}(d)$ and $\sigma_m^{(\mathbf{Y}|\mathbf{X})}(d)$ are the mean and standard deviation values of the conditional PDF, $P(\mathbf{Y}_t\,|\,\mathbf{X}_t,m,\Theta^{(\mathbf{Z})})$ at frame t, respectively; $\mathbf{V}_{\hat{\mathbf{m}}}$ is a DT-by-1 vector given by

$$\mathbf{V}_{\hat{\mathbf{m}}}(d)=-\frac{2}{T}\left(\hat{\mathbf{y}}_{\hat{\mathbf{m}}}(d)-\langle\hat{\mathbf{y}}_{\hat{\mathbf{m}}}(d)\rangle\right)\left(v(\hat{\mathbf{y}}_{\hat{\mathbf{m}}})-v(\mathbf{y})\right)^{\mathrm{T}}\mathbf{p}_v(d), \tag{21}$$

where $\mathbf{p}_v(d)$ is the dth column vector of $\mathbf{P}_v=\Sigma^{(vv)-1}$; and $\mathbf{Q}_{\hat{\mathbf{m}}}$ is a DT-by-$2DT$ matrix given as $\mathbf{Q}_{\hat{\mathbf{m}}}=(\mathbf{W}^{\mathrm{T}}\mathbf{D}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})-1}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{D}_{\hat{\mathbf{m}}}^{(\mathbf{Y}|\mathbf{X})-1}$. Based on (17)-(21), we can update $\{\mathbf{A}_m,\mathbf{b}_m\}_{m=1}^{M}$ iteratively.

IV. Experiments

A. Experimental Setup

This section describes the experimental setup for evaluating the proposed training method. For comparison, three GMM-based SC systems with different training and conversion approaches were implemented and tested. The first system used the proposed method that considered GV in the training phase and adopted suboptimal ML-based trajectory mapping (10) in the conversion phase (denoted as Prop+ML). The second system used the conventional GMM training method and adopted suboptimal MLGV-based trajectory mapping [3] in conversion (denoted as Conv+MLGV). The third system used the conventional GMM training method and adopted suboptimal ML-based trajectory mapping (10) in conversion (denoted as Conv+ML). In other words, the Prop+ML system incorporates GV in the training phase, the Conv+MLGV system incorporates GV in the conversion phase, while the Conv+ML system does not consider GV. The three GMM-

TABLE I
Conversion accuracy of the three GMM-based SC systems. The MCD before the conversion is 9.37 dB.

Methods	Conv+ML	Conv+MLGV	Prop+ML
MCD [dB]	5.12	5.67	5.50

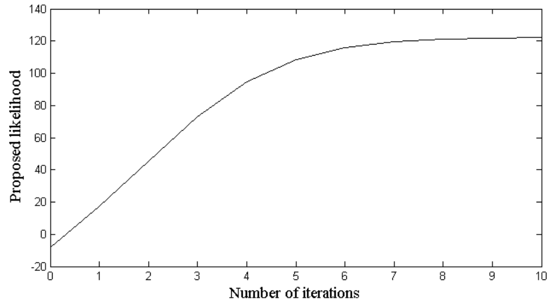

Fig. 1: The log-scaled likelihood curve of the proposed training method on the training set.

based SC systems were evaluated on a parallel Mandarin speech corpus. This corpus consisted of two speakers, one female and one male. Eighty parallel sentences were selected from both speakers. Among the 80 sentences, 40 sentences were used to establish the conversion system and the remaining 40 sentences were used to perform conversion and conduct evaluation.

Speech signals were firstly recorded in a 20kHz sampling rate, and then down-sampled to 16kHz. The resolution per sample was 16 bits. The spectral features were the first through 24 Mel-cepstral coefficients. STRAIGHT [13] was employed for the analysis-synthesis method. A dynamic time warping (DTW) algorithm was performed within each syllable boundary to obtain a joint feature vector sequence in the training phase. The number of Gaussian mixtures in each JDGMM set was 64. The maximum number of iterations in the proposed training method was set to 10. The step size ε in (17) was empirically determined. The power weight ω in (18) was set to 2. We reported both the objective and subjective evaluation results on the female to male SC task.

B. Objective Evaluations

We conducted the objective evaluation in terms of the conversion accuracy and the degree of the over-smoothing of the converted Mel-cepstra for comparing the three GMM-based SC systems. To evaluate the conversion accuracy, the Mel-cepstral distortion (MCD), $D_{MCD}(y_t,\hat{y}_t)$, was used to compute the difference of the target and converted Mel-cepstra in the evaluation set, which is given by

$$D_{MCD}(y_t,\hat{y}_t)\ [\text{dB}] = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}(y_t^d - \hat{y}_t^d)^2}\ , \qquad (22)$$

where y_t and $\hat{y}_t$ are the target and converted static feature vectors, respectively. A lower MCD value indicates a more accurate conversion.

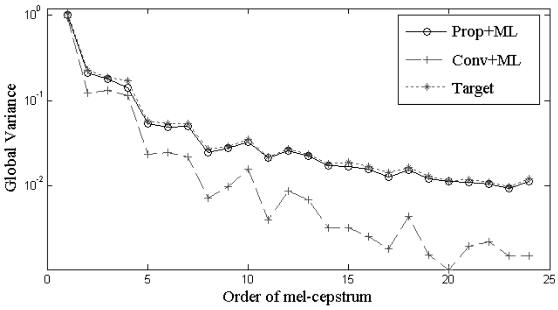

Fig. 2: The means of GV of the converted Mel-cepstra over all training utterances.

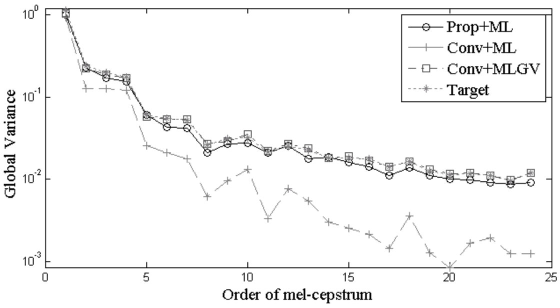

Fig. 3: The means of GV of the converted Mel-cepstra over all test utterances.

Table 1 shows the MCD results of the three GMM-based SC systems. From Table 1, we observe that Prop+ML and Conv+MLGV give higher MCD values than Conv+ML. This result implies that integration of GV in the training phase (Prop+ML) and conversion phase (Conv+MLGV) may decrease the conversion accuracy, thereby degrading the quality of the converted speech. However, previous studies have reported that the MCD values may not perfectly reflect the real subjective evaluation results when introducing the GV for GMM-based SC [3, 9]. Similar results are also observed in this study, which will be shown later in the subjective test.

To have a further analysis, Fig. 1 demonstrates the log-scaled likelihood (13) as a function of the number of iterations of the proposed training method, where the log-scaled likelihood at the 0th iteration is obtained by using the ML-trained JDGMM as the initial model parameters. It can be seen that the log-scaled likelihood consistently increases, and then saturates after nine iterations. The result shows that the training is quite successful.

Next, we evaluate the degree of the over-smoothing of the converted Mel-cepstra given by the three GMM-based SC systems. The GV measurement, computed by (14)-(15), was adopted to compare the GVs of the converted Mel-cepstra and the natural Mel-cepstra of the target speech (named the Target GV hereafter). Fig. 2 compares the GVs of the converted Mel-cepstra obtained by Conv+ML and Prop+ML and the Target GV on the training set. It can be seen that the GV obtained by Conv+ML is smaller than that obtained by Prop+ML and the Target GV, in particular for higher order Mel-cepstra.

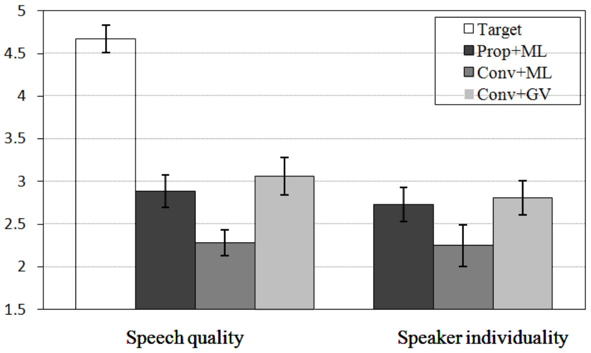

Fig. 4: Subjective test results of the speech converted by Prop+ML, Conv+ML, and Conv+GV. Error bars indicate 95% confidence intervals. “Target” shown in the speech quality evaluation denotes the analysis-synthesized target speech.

Moreover, the GV obtained by Prop+ML is almost equal to the Target GV. This result implies that the converted Mel-cepstra obtained by the conventional ML-trained JDGMM with ML-based trajectory mapping would be smoothed overly, and the over-smoothing problem can be effectively alleviated by the proposed training method.

The results for the same GV evaluation on the evaluation set are shown in Fig. 3. Similar results can be found as compared to Fig. 2, which is evaluated on the training set. In addition to Conv+ML and Prop+ML, we have also evaluated the GV obtained by Conv+MLGV for comparison. It can be seen that the GV obtained by Conv+MLGV is very close to the Target GV while the GV obtained by Prop+ML obviously differs from the Target GV. This result indicates that the proposed training method with ML-based trajectory mapping is comparable with the conventional MLGV-based trajectory mapping in terms of GV evaluation; but with reduced computational cost in the conversion process as mentioned earlier.

C. *Subjective Evaluations*

We conducted a formal listening test on the evaluation set and used the mean opinion score (MOS) method to evaluate the speech quality and speaker individuality. The results for three kinds of converted speech (by Prop+ML, Conv+ML, and Conv+MLGV) and the analysis-synthesized target speech are shown in Fig. 4. The evaluation was performed by eight subjects; Twenty five test sentences were randomly chosen from the test set for the eight subjects. Samples were presented in random order for the 25 test sentences. Since this study is focused on spectral conversion, the same simple linear transformation-based F_0 conversion was applied for all systems.

From Fig. 4, we can see that Prop+ML and Conv+MLGV obviously outperform Conv+ML on both speech quality and speaker individuality tests. This result indicates that GV plays an important cue to speech quality and speaker individuality. Incorporating GV in either training or conversion phase of SC systems effectively overcomes the over-smoothing problem. This result is consistent with previous studies, which employed GV for VC [3, 9, 10] and speech synthesis [12] in the online speech generation phase. Comparing the two GV-based SC systems, we can see that Prop+ML almost performs as well as Conv+MLGV on both speech quality and speaker individuality while requires a significantly lower computational cost in the conversion process. The current Prop+ML performance achieved can be improved by handling the overtraining problem and/or choosing better training parameters. This is worth of further study.

V. Conclusions

In this paper, we have proposed a new training method to incorporate GV into the training phase of the GMM-based VC. The major advantage of the proposed method is that the online computational cost in the conversion process has been significantly reduced while the quality of the converted speech can be maintained comparable with that using MLGV-based trajectory mapping. Our next step is to integrate GV into the trajectory GMM [10] for GMM-based VC, inspired by the idea of integrating GV into the trajectory hidden Markov model (HMM) for HMM-based speech synthesis [12].

Acknowledgment

The authors would like to thank Prof. H. Kawahara of Wakayama University, Japan, for the permission to use the STRAIGHT method.

References

[1] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp.131-142, Mar. 1998.

[2] A. Kain, and M. W. Macon, “Spectral voice conversion for text-to-speech synthesis,” *Proc. ICASSP*, 1998, vol. 1, pp. 285-288.

[3] T. Toda, A.W. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” *IEEE Trans. Audio, Speech, Lang., Process.*, vol. 15, no. 8, pp. 2222-2235, Nov. 2007.

[4] D. Erro, A. Moreno, and A. Bonafonte, “Voice conversion based on weighted frequency warping,” *IEEE Trans. Audio, Speech, Lang., Process.*, vol. 18, no. 5, pp. 922-931, July. 2010.

[5] E. Godoy, O. Rosec, and T. Chonavel, “Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora,” *IEEE Trans. Audio, Speech, Lang., Process*, vol. 20, no. 4, pp. 1313-1323, May. 2012.

[6] H. T. Hwang, Y. Tsao, H. M. Wang, Y. R. Wang and S. H. Chen, “A study of mutual information for GMM-based spectral conversion,” *Proc. INTERSPEECH*, 2012.

[7] H. T. Hwang, Y. Tsao, H. M. Wang, Y. R. Wang and S. H. Chen, “Exploring mutual information for GMM-based spectral conversion,” *Proc. ISCSLP*, 2012, pp. 50-54.

[8] H. T. Hwang, Y. Tsao, H. M. Wang, Y. R. Wang and S. H. Chen, “Alleviating the Over-Smoothing Problem in GMM-Based Voice Conversion with Discriminative Training,” *Proc. INTERSPEECH*, 2013.

[9] H. Benisty and D. Malah, “Voice conversion using GMM with enhanced global variance”, *Proc. INTERSPEECH*, 2011, pp. 669-672.

[10] H. Zen, Y. Nankaku, and K. Tokuda, “Continuous stochastic feature mapping based on trajectory HMMs,” *IEEE Trans. Audio, Speech, Lang., Process.*, vol. 19, no. 2, pp. 417-430, Feb. 2011.

[11] D. Saito, S. Watanabe, A. Nakamura, and N. Minematsu, "Statistical voice conversion based on noisy channel model," *IEEE Trans. Audio, Speech, Lang., Process.,* vol. 20, no. 6, pp. 1784-1794, Aug. 2012.

[12] T. Toda, and S. Young, "Trajectory Training considering global variance for HMM-based speech synthesis," *Proc. ICASSP*, 2009, pp. 4025-4028.

[13] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3-4, pp.187-207, 1999.