Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory

Tomoki Toda, *Member, IEEE*, Alan W. Black, *Member, IEEE*, and Keiichi Tokuda, *Member, IEEE*

***Abstract*—In this paper, we describe a novel spectral conversion method for voice conversion (VC). A Gaussian mixture model (GMM) of the joint probability density of source and target features is employed for performing spectral conversion between speakers. The conventional method converts spectral parameters frame by frame based on the minimum mean square error. Although it is reasonably effective, the deterioration of speech quality is caused by some problems: 1) appropriate spectral movements are not always caused by the frame-based conversion process, and 2) the converted spectra are excessively smoothed by statistical modeling. In order to address those problems, we propose a conversion method based on the maximum-likelihood estimation of a spectral parameter trajectory. Not only static but also dynamic feature statistics are used for realizing the appropriate converted spectrum sequence. Moreover, the over-smoothing effect is alleviated by considering a global variance feature of the converted spectra. Experimental results indicate that the performance of VC can be dramatically improved by the proposed method in view of both speech quality and conversion accuracy for speaker individuality.**

***Index Terms*—Dynamic feature, global variance, Gaussian mixture model (GMM), maximum-likelihood estimation (MLE), voice conversion (VC).**

I. Introduction

VOICE conversion (VC) is a potential technique for flexibly synthesizing various types of speech. This technique can modify nonlinguistic information such as voice characteristics while keeping linguistic information unchanged. A statistical feature mapping process is often employed in VC. A mapping function is trained in advance using a small amount of training data consisting of utterance pairs of source and target voices. The resulting mapping function allows the conversion of any sample of the source into that of the target without any linguistic features such as phoneme transcription. A typical VC application is speaker conversion [1], in which the voice of a certain speaker (source speaker) is converted to sound like that of another speaker (target speaker). Because no linguistic features are used, this conversion framework can straightforwardly be extended to cross-language speaker conversion [2], [3] to realize target speakers' voices in various languages by applying the mapping function trained in a certain language into the conversion process in another language. There are many other VC applications such as conversion from narrow-band speech to wide-band speech for telecommunication [4], [5], modeling of speech production [6], [7], acoustic-to-articulatory inversion mapping [8], [9], body-transmitted speech enhancement [10]–[12], and a speaking aid [13], [14]. In this paper, we describe spectral conversion algorithms for speaker conversion, which can also be applied to various other applications.

Many statistical approaches to VC have been studied since the late 1980s [15]. Abe *et al.* [1] proposed a codebook mapping method based on hard clustering and discrete mapping. The converted feature vector $\hat{\boldsymbol{y}}_t$ at frame t is determined by quantizing the source feature vector $\boldsymbol{x}_t$ to the nearest centroid vector of the source codebook and substituting it with a corresponding centroid vector $\boldsymbol{c}_m^{(y)}$ of the mapping codebook as follows:

$$\hat{\boldsymbol{y}}_t = \boldsymbol{c}_m^{(y)}. \tag{1}$$

The large quantization error due to hard clustering is effectively reduced by adopting fuzzy vector quantization (VQ) [16] that realizes soft clustering. Continuous weights $w_{m,t}^{(x)}$ for individual clusters are determined at each frame according to the source feature vector. The converted feature vector is defined as a weighted sum of the centroid vectors of the mapping codebook as follows:

$$\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} w_{m,t}^{(x)} \boldsymbol{c}_m^{(y)} \tag{2}$$

where M is the number of centroid vectors. Moreover, more variable representations of the converted feature vector are achieved by modeling a difference vector between the source and target feature vectors [17] as follows:

$$\hat{\boldsymbol{y}}_t = \boldsymbol{x}_t + \sum_{m=1}^{M} w_{m,t}^{(x)} \left(\boldsymbol{c}_m^{(y)} - \boldsymbol{c}_m^{(x)} \right). \tag{3}$$

In this method, a very strong correlation between those two vectors is assumed. In order to directly model the correlation between them, Valbret *et al.* [18] proposed a conversion method using linear multivariate regression (LMR), i.e., continuous mapping based on hard clustering, as follows:

$$\hat{\boldsymbol{y}}_t = \boldsymbol{A}_m \boldsymbol{x}_t + \boldsymbol{b}_m \tag{4}$$

where $\boldsymbol{A}_m$ and $\boldsymbol{b}_m$ are regression parameters. There are many methods other than those mentioned above, e.g., conversion methods based on speaker interpolation [19] and neural networks [20]. As the most popular method, Stylianou *et al.* [21] proposed a conversion method with a Gaussian mixture

Manuscript received January 15, 2007; revised August 1, 2007. This work was supported in part by a MEXT Grant-in-Aid for Young Scientists (A). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Abeer Alwan.

T. Toda is with the Graduate School of Information Science, Nara Institute of Science and Technology, Nara 630-0192, Japan (e-mail: tomoki@is.naist.jp).

A.W. Black is with the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: awb@cs.cmu.edu).

K. Tokuda is with the Graduate School of Engineering, Nagoya Institute of Technology, Nagoya 466-8555, Japan (e-mail: tokuda@ics.nitech.ac.jp).

Digital Object Identifier 10.1109/TASL.2007.907344

model (GMM) that realizes continuous mapping based on soft clustering as follows:

$$\hat{\boldsymbol{y}}_t = \sum_{m=1}^{M} w_{m,t}^{(x)} \left(\boldsymbol{A}_m \boldsymbol{x}_t + \boldsymbol{b}_m \right). \tag{5}$$

This mapping method is reasonably effective. However, the performance of the conversion is still insufficient. The converted speech quality is deteriorated by some factors, e.g., spectral movement with inappropriate dynamic characteristics caused by the frame-by-frame conversion process and excessive smoothing of converted spectra [22], [23].

We propose spectral conversion based on the maximum-likelihood estimation (MLE) of a spectral parameter trajectory. In order to realize appropriate spectral movements, we consider the feature correlation between frames by applying a parameter generation algorithm with dynamic features [24]–[26], which works very well in hidden Markov model (HMM)-based speech synthesis [27], [28], to the GMM-based mapping. This idea makes it possible to estimate an appropriate spectrum sequence in view of not only static but also dynamic characteristics. Furthermore, in order to address the oversmoothing problem of the converted spectra, we consider the global variance (GV) of the converted spectra over a time sequence as a novel features-capturing characteristic of the parameter trajectory. This idea effectively models missing information in the conventional frameworks of statistical conversion. Results of objective and subjective evaluations demonstrate that the proposed method successfully causes dramatic improvements in both the converted speech quality and the conversion accuracy for speaker individuality. We present further details of the conversion method, more discussions, and more evaluations than described in our previous work [23].

The paper is organized as follows. In Section II, we describe the conventional GMM-based mapping method. In Section III, we describe the proposed conversion method considering dynamic features and the GV. In Section IV, experimental evaluations are presented. Finally, we summarize this paper in Section V.

II. Conventional GMM-Based Mapping

A. Probability Density Function

Let $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ be D-dimensional source and target feature vectors at frame t, respectively. The joint probability density of the source and target feature vectors is modeled by a GMM as follows:

$$P\left(\boldsymbol{z}_t|\boldsymbol{\lambda}^{(z)}\right) = \sum_{m=1}^{M} w_m \mathcal{N}\left(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}\right) \tag{6}$$

where $\boldsymbol{z}_t$ is a joint vector $\left[\boldsymbol{x}_t^\top,\ \boldsymbol{y}_t^\top\right]^\top$. The notation $^\top$ denotes transposition of the vector. The mixture component index is m. The total number of mixture components is M. The weight of the mth mixture component is w_m. The normal distribution with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is denoted as $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. A parameter set of the GMM is $\boldsymbol{\lambda}^{(z)}$, which consists of weights, mean vectors, and the covariance matrices for individual mixture components. The mean vector $\boldsymbol{\mu}_m^{(z)}$ and the covariance matrix $\boldsymbol{\Sigma}_m^{(z)}$ of the mth mixture component are written as

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \quad \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix} \tag{7}$$

where $\boldsymbol{\mu}_m^{(x)}$ and $\boldsymbol{\mu}_m^{(y)}$ are the mean vector of the mth mixture component for the source and that for the target, respectively. The matrices $\boldsymbol{\Sigma}_m^{(xx)}$ and $\boldsymbol{\Sigma}_m^{(yy)}$ are the covariance matrix of the mth mixture component for the source and that for the target, respectively. The matrices $\boldsymbol{\Sigma}_m^{(xy)}$ and $\boldsymbol{\Sigma}_m^{(yx)}$ are the cross-covariance matrices of the mth mixture component for the source and the target, respectively. These covariance matrices, $\boldsymbol{\Sigma}_m^{(xx)}$, $\boldsymbol{\Sigma}_m^{(xy)}$, $\boldsymbol{\Sigma}_m^{(yx)}$, and $\boldsymbol{\Sigma}_m^{(yy)}$, are diagonal in this study.

The GMM is trained with the EM algorithm using the joint vectors, which are automatically aligned by dynamic time warping (DTW), in a training set. This training method provides estimates of the model parameters robustly compared with the least-squares estimation [21], particularly when the amount of training data is small [29].

B. Mapping Function

The conditional probability density of $\boldsymbol{y}_t$, given $\boldsymbol{x}_t$, is also represented as a GMM as follows:

$$P\left(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) = \sum_{m=1}^{M} P\left(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) P\left(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}\right) \tag{8}$$

where

$$P\left(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) = \frac{w_m \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)}\right)}{\sum_{n=1}^{M} w_n \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_n^{(x)}, \boldsymbol{\Sigma}_n^{(xx)}\right)} \tag{9}$$

$$P\left(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}\right) = \mathcal{N}\left(\boldsymbol{y}_t; \boldsymbol{E}_{m,t}^{(y)}, \boldsymbol{D}_m^{(y)}\right). \tag{10}$$

The mean vector $\boldsymbol{E}_{m,t}^{(y)}$ and the covariance matrix $\boldsymbol{D}_m^{(y)}$ of the mth conditional probability distribution are written as

$$\boldsymbol{E}_{m,t}^{(y)} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}\right) \tag{11}$$

$$\boldsymbol{D}_m^{(y)} = \boldsymbol{\Sigma}_m^{(yy)} - \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} \boldsymbol{\Sigma}_m^{(xy)}. \tag{12}$$

In the conventional method [21], [29], the conversion is performed on the basis of the minimum mean-square error (mmse) as follows:

$$\begin{aligned} \hat{\boldsymbol{y}}_t &= E[\boldsymbol{y}_t|\boldsymbol{x}_t] \\ &= \int P\left(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) \boldsymbol{y}_t d\boldsymbol{y}_t \\ &= \int \sum_{m=1}^{M} P\left(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) P\left(\boldsymbol{y}_t|\boldsymbol{x}_t, m, \boldsymbol{\lambda}^{(z)}\right) \boldsymbol{y}_t d\boldsymbol{y}_t \\ &= \sum_{m=1}^{M} P\left(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}\right) \boldsymbol{E}_{m,t}^{(y)} \end{aligned} \tag{13}$$

where $E[\cdot]$ means the expectation and $\hat{\boldsymbol{y}}_t$ is the converted target feature vector. Note that this mapping function has the same form as in (5) with $P(m|\boldsymbol{x}_t, \boldsymbol{\lambda}^{(z)}) = w_{m,t}^{(x)}$, $\boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} =$

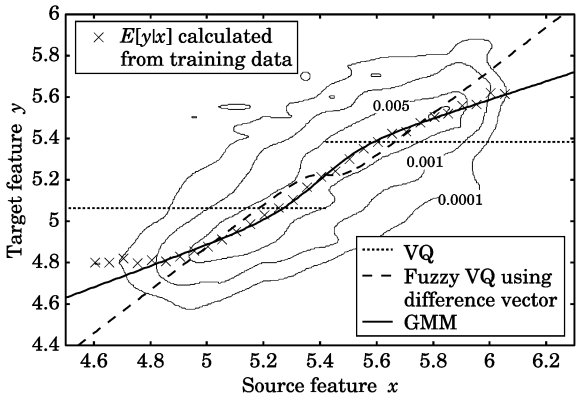

Fig. 1. Mapping functions on the joint feature space. The contour lines show normalized frequency distribution of training data samples.

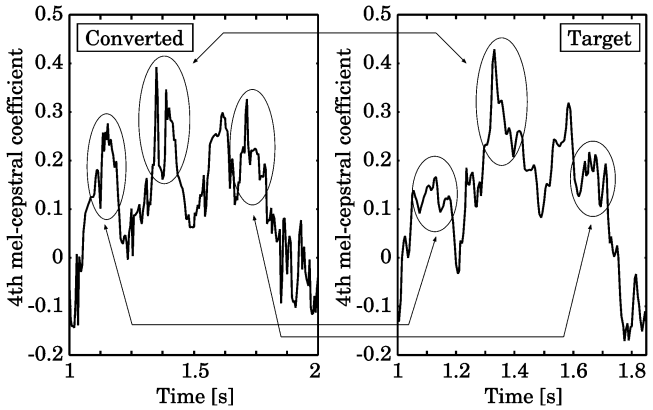

Fig. 2. Example of converted and natural target parameter trajectories. Different local patterns are observed in ellipses. Note that phoneme duration of the converted speech is different from that of the target.

$\boldsymbol{A}_m$, and $\boldsymbol{\mu}_m^{(y)} - \boldsymbol{\Sigma}_m^{(yx)}\boldsymbol{\Sigma}_m^{(xx)^{-1}}\boldsymbol{\mu}_m^{(x)} = \boldsymbol{b}_m$. In each mixture component, the conditional target mean vector for the given source feature vector is calculated by a simple linear conversion based on the correlation between the source and target feature vectors, as shown in (11). The converted feature vector is defined as the weighted sum of the conditional mean vectors, where the posterior probabilities, in (9), of the source vector belonging to each one of the mixture components are used as weights.

Fig. 1 shows some mapping functions on a joint space of one-dimensional source and target features: that in the codebook mapping method ("VQ") shown by (1), that in the fuzzy VQ mapping method using the difference vector ("Fuzzy VQ using difference vector") shown by (3), and that in the GMM-based mapping method ("GMM") shown by (13). The number of centroids or the number of Gaussian mixture components is set to 2. We also show values of the conditional expectation $E[y|x]$ calculated directly from the training samples. The mapping function in the codebook mapping method is discontinuous because of hard clustering. The mapping function to which the fuzzy VQ and the difference vector are applied nearly approximates the conditional expectation. However, its accuracy is not high enough. Compared with those, the GMM-based mapping function is much closer to the conditional expectation because of the direct modeling of the correlation between the source and target features. Furthermore, it allows soft clustering based on the posterior probabilities of the GMM that can represent the probability distribution of features more accurately than the VQ-based algorithms by modeling the covariance. Therefore, the GMM-based mapping function has a reasonably high conversion accuracy.

C. Problems

Although the GMM-based mapping function works well, there still remain some problems to be solved. This paper is focused on two main problems, i.e., the time-independent mapping and the oversmoothing effect.

Fig. 2 shows an example of the parameter trajectory converted using the GMM-based mapping function and the natural target trajectory. Although these two trajectories seem similar, they sometimes have different local patterns. Such differences are often observed because the correlation of the target feature vectors between frames is ignored in the conventional mapping. In order to realize the appropriate converted spectrum sequence, it is necessary to consider the dynamic features of the parameter trajectory.

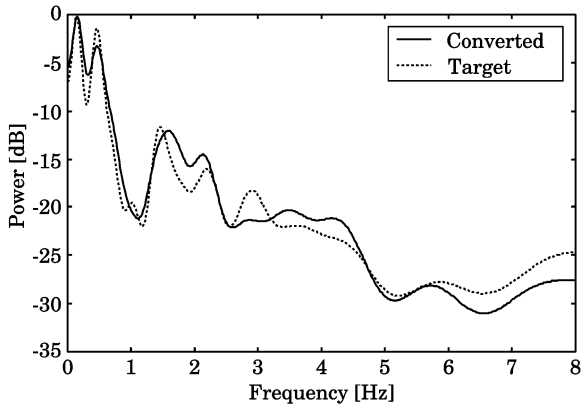

Fig. 3. Example of converted and natural target spectra.

Fig. 3 shows an example of the converted and natural target spectra. We can see that the converted spectrum is excessively smoothed compared with the natural one. The statistical modeling often removes the details of spectral structures. This smoothing undoubtedly causes error reduction of the spectral conversion. However, it also causes quality degradation of the converted speech because the removed structures are still necessary for synthesizing high-quality speech.

III. Proposed Spectral Conversion

Instead of the conventional frame-based conversion process, we propose the trajectory-based conversion process written as

$$\hat{\boldsymbol{y}} = f(\boldsymbol{x}) \tag{14}$$

where $f(\cdot)$ is a mapping function. The vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are time sequences of the source and target feature vectors, respectively, and are written as

$$\boldsymbol{x} = \left[\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_t^\top, \ldots, \boldsymbol{x}_T^\top\right]^\top \tag{15}$$

$$\boldsymbol{y} = \left[\boldsymbol{y}_1^\top, \boldsymbol{y}_2^\top, \ldots, \boldsymbol{y}_t^\top, \ldots, \boldsymbol{y}_T^\top\right]^\top. \tag{16}$$

The proposed framework simultaneously converts feature vectors in all frames over a time sequence.

In the following, we emphasize the proposed framework by introducing two main ideas: 1) the conversion considering the feature correlation between frames and 2) the conversion considering the GV.

A. Conversion Considering Dynamic Features

We use $2D$-dimensional source and target feature vectors $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ and $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ consisting of D-dimensional static and dynamic features at frame t.[1] Their time sequences are respectively written as

$$\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \ldots, \boldsymbol{X}_t^\top, \ldots, \boldsymbol{X}_T^\top\right]^\top \tag{17}$$

$$\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \ldots, \boldsymbol{Y}_t^\top, \ldots, \boldsymbol{Y}_T^\top\right]^\top. \tag{18}$$

The GMM $\boldsymbol{\lambda}^{(Z)}$ of the joint probability density $P(\boldsymbol{Z}_t|\boldsymbol{\lambda}^{(Z)})$ is trained in advance using joint vectors $\boldsymbol{Z}_t = \left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top$ by straightforwardly adopting the conventional training framework [29].

1) Likelihood Function: We perform the spectral conversion based on maximizing the following likelihood function:

$$\begin{aligned} P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right) &= \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right) P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}^{(Z)}\right) \\ &= \prod_{t=1}^{T} \sum_{m=1}^{M} P\left(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(Z)}\right) \\ &\quad \times P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(Z)}\right) \end{aligned} \tag{19}$$

where $\boldsymbol{m} = \{m_1, m_2, \ldots, m_t, \ldots, m_T\}$ is a mixture component sequence. The conditional probability density at each frame is modeled as a GMM. At frame t, the mth mixture component weight $P\left(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(Z)}\right)$ and the mth conditional probability distribution $P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(Z)}\right)$ are given by

$$P\left(m|\boldsymbol{X}_t, \boldsymbol{\lambda}^{(Z)}\right) = \frac{w_m \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_m^{(X)}, \boldsymbol{\Sigma}_m^{(XX)}\right)}{\sum_{n=1}^{M} w_n \mathcal{N}\left(\boldsymbol{X}_t; \boldsymbol{\mu}_n^{(X)}, \boldsymbol{\Sigma}_n^{(XX)}\right)} \tag{20}$$

$$P\left(\boldsymbol{Y}_t|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(Z)}\right) = \mathcal{N}\left(\boldsymbol{Y}_t; \boldsymbol{E}_{m,t}^{(Y)}, \boldsymbol{D}_m^{(Y)}\right) \tag{21}$$

where

$$\boldsymbol{E}_{m,t}^{(Y)} = \boldsymbol{\mu}_m^{(Y)} + \boldsymbol{\Sigma}_m^{(YX)} {\boldsymbol{\Sigma}_m^{(XX)}}^{-1} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}\right) \tag{22}$$

[1]We may also use delta-delta features.

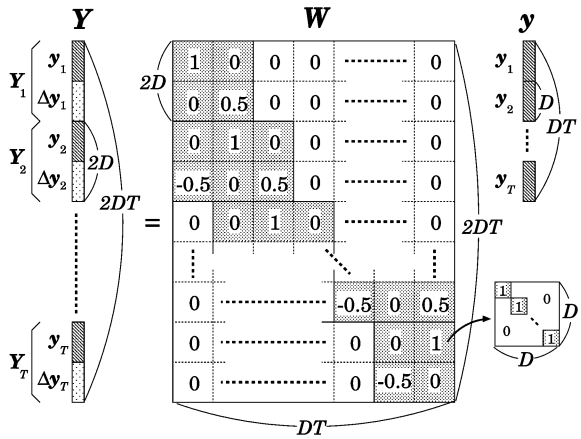

Fig. 4. Relationship between a sequence of the static feature vectors $\boldsymbol{y}$ and that of the static and dynamic feature vectors $\boldsymbol{Y}$ under $L_-^{(1)} = L_+^{(1)} = 1$, $w^{(1)}(-1) = -0.5$, $w^{(1)}(0) = 0$, and $w^{(1)}(1) = 0.5$ in (28).

$$\boldsymbol{D}_m^{(Y)} = \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(YX)} {\boldsymbol{\Sigma}_m^{(XX)}}^{-1} \boldsymbol{\Sigma}_m^{(XY)}. \tag{23}$$

Note that the conditional mean vector is represented as a linear conversion from the source feature vector, as described in the previous section.

2) MLE of Parameter Trajectory: A time sequence of the converted feature vectors is determined as follows:

$$\hat{\boldsymbol{y}} = \arg\max P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right). \tag{24}$$

In the same manner as the parameter generation algorithm from an HMM [24], [25], this determination is performed under the explicit relationship between a sequence of the static feature vectors $\boldsymbol{y}$ and that of the static and dynamic feature vectors $\boldsymbol{Y}$ represented as the following linear conversion (see also Fig. 4):

$$\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{25}$$

where $\boldsymbol{W}$ is the $2DT$-by-DT matrix written as

$$\boldsymbol{W} = \left[\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_t, \ldots, \boldsymbol{W}_T\right]^\top \otimes \boldsymbol{I}_{D\times D} \tag{26}$$

$$\boldsymbol{W}_t = \left[\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}\right] \tag{27}$$

$$\begin{aligned} \boldsymbol{w}_t^{(n)} = \Bigg[&\overbrace{0, \ldots, 0}^{1\text{st}}, w^{(n)}\left(-L_-^{(n)}\right), \ldots, \overbrace{w^{(n)}(0)}^{(t)\text{-th}}, \ldots, \\ &w^{(n)}\left(L_+^{(n)}\right), \overbrace{0, \ldots, 0}^{T\text{-th}} \Bigg]^\top, \quad n = 0, 1 \end{aligned} \tag{28}$$

where $L_-^{(0)} = L_+^{(0)} = 0$, and $w^{(0)}(0) = 1$. In this paper, we describe two solutions for (24).

EM algorithm: As described in [26], we iteratively maximize the following auxiliary function with respect to $\hat{\boldsymbol{y}}$

$$Q(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{\text{all } \boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\lambda}^{(Z)}\right) \log P\left(\hat{\boldsymbol{Y}}, \boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right). \tag{29}$$

The converted vector sequence maximizing the auxiliary function is given by

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)^{-1}}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)^{-1}} \boldsymbol{E}^{(Y)}} \tag{30}$$

where

$$\overline{\boldsymbol{D}^{(Y)^{-1}}} = \mathrm{diag}\left[\overline{\boldsymbol{D}_1^{(Y)^{-1}}}, \overline{\boldsymbol{D}_2^{(Y)^{-1}}}, \ldots, \overline{\boldsymbol{D}_t^{(Y)^{-1}}}, \ldots \overline{\boldsymbol{D}_T^{(Y)^{-1}}}\right] \tag{31}$$

$$\overline{\boldsymbol{D}^{(Y)^{-1}} \boldsymbol{E}^{(Y)}} = \left[\overline{\boldsymbol{D}_1^{(Y)^{-1}} \boldsymbol{E}_1^{(Y)}}^\top, \overline{\boldsymbol{D}_2^{(Y)^{-1}} \boldsymbol{E}_2^{(Y)}}^\top, \ldots, \overline{\boldsymbol{D}_t^{(Y)^{-1}} \boldsymbol{E}_t^{(Y)}}^\top, \ldots, \overline{\boldsymbol{D}_T^{(Y)^{-1}} \boldsymbol{E}_T^{(Y)}}^\top\right]^\top \tag{32}$$

$$\overline{\boldsymbol{D}_t^{(Y)^{-1}}} = \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(Y)^{-1}} \tag{33}$$

$$\overline{\boldsymbol{D}_t^{(Y)^{-1}} \boldsymbol{E}_t^{(Y)}} = \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(Y)^{-1}} \boldsymbol{E}_{m,t}^{(Y)} \tag{34}$$

$$\gamma_{m,t} = P(m|\boldsymbol{X}_t, \boldsymbol{Y}_t, \boldsymbol{\lambda}^{(Z)}). \tag{35}$$

The derivation of (30) is given in Appendix I. As an initial parameter sequence, the converted vector sequence determined under the following approximation with the suboptimum mixture sequence is effective.

Approximation with suboptimum mixture sequence: The likelihood function in (19) is approximated with a single mixture component sequence as follows:

$$P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right) \simeq P\left(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right) P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}^{(Z)}\right). \tag{36}$$

First, the suboptimum mixture component sequence $\hat{\boldsymbol{m}}$ is determined by

$$\hat{\boldsymbol{m}} = \mathrm{argmax}\, P\left(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right). \tag{37}$$

Then the approximated log-scaled likelihood function is written as

$$\mathcal{L} = \log P\left(\hat{\boldsymbol{m}}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right) P\left(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}^{(Z)}\right). \tag{38}$$

The converted static feature vector sequence $\hat{\boldsymbol{y}}$ that maximizes $\mathcal{L}$ under the constraint of (25) is given by

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)} \tag{39}$$

where

$$\boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)} = \left[\boldsymbol{E}_{\hat{m}_1,1}^{(Y)}, \boldsymbol{E}_{\hat{m}_2,2}^{(Y)}, \ldots, \boldsymbol{E}_{\hat{m}_t,t}^{(Y)}, \ldots, \boldsymbol{E}_{\hat{m}_T,T}^{(Y)}\right] \tag{40}$$

$$\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} = \mathrm{diag}\left[\boldsymbol{D}_{\hat{m}_1}^{(Y)^{-1}}, \boldsymbol{D}_{\hat{m}_2}^{(Y)^{-1}}, \ldots, \boldsymbol{D}_{\hat{m}_t}^{(Y)^{-1}}, \ldots, \boldsymbol{D}_{\hat{m}_T}^{(Y)^{-1}}\right]. \tag{41}$$

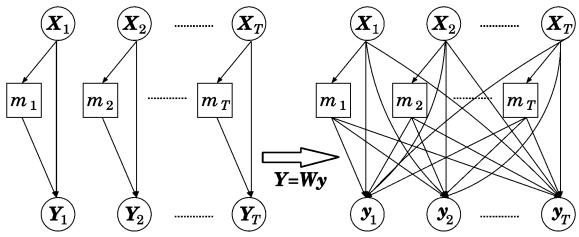

Fig. 5. Graphical representation of relationship between individual variables in conversion process. The frame-based conversion process shown on the left is changed into the trajectory-based one shown on the right by imposing the constraint of (25).

We omit the derivation of (39) because it is very similar to that of (30). The approximated solution effectively reduces the computational cost. Our preliminary experiment demonstrated that there were no large differences in the conversion accuracy between the approximated solution and the EM algorithm. Therefore, we adopt the approximated solution.

Fig. 5 shows a graphical representation of the relationship between individual variables in the conversion process. If we determine a time sequence of the converted static and dynamic feature vectors $\boldsymbol{Y}$ that maximizes the likelihood function of (38), the determination process at each frame is independent of that at the other frames, as shown in the left diagram in Fig. 5. In contrast, the proposed method yields a time sequence of only the converted static feature vectors $\boldsymbol{y}$ that maximizes the likelihood function under the condition of (25). As shown in (39), the $2DT$-by-$2DT$ covariance matrix $\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}}$ is converted into a DT-by-DT covariance matrix $\left(\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{W}\right)^{-1}$, which is generally full because $\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{W}$ is a band matrix. It effectively models interframe correlations of the target feature vectors. Consequently, the source feature vectors at all frames affects the determination of the converted feature vector at each frame, as shown in the right diagram in Fig. 5.[2]

3) Relationship With Conventional Mapping Method by (13): If dynamic features are not considered in the proposed conversion (i.e., $\boldsymbol{W}$ is set to the identity matrix $\boldsymbol{I}_{DT\times DT}$ and the GMM of the joint probability density is used on only static features $\boldsymbol{\lambda}^{(z)}$), the converted vector at frame t is given by

$$\hat{\boldsymbol{y}}_t = \left(\sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(y)^{-1}}\right)^{-1} \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{D}_m^{(y)^{-1}} \boldsymbol{E}_{m,t}^{(y)}. \tag{42}$$

The conversion process at each frame is independent of that at other frames [7]. Because the proposed conversion is based not on mmse but MLE, not only the mean vectors $\boldsymbol{E}_{m,t}^{(y)}$ but also the covariance matrices $\boldsymbol{D}_m^{(y)}$ of the conditional probability distributions affect the determination process. Those covariance matrices are used as weights in the weighted sum of the conditional mean vectors, as shown in (42). Namely, they are regarded as a kind of confidence measure for the conditional mean vectors from individual mixture components.

[2]Note that the resulting converted feature vector sequence is the same as the mean vector sequence of a trajectory model [30] derived from the conditional probability density distributions for both static and dynamic features by imposing an explicit relationship between those features.

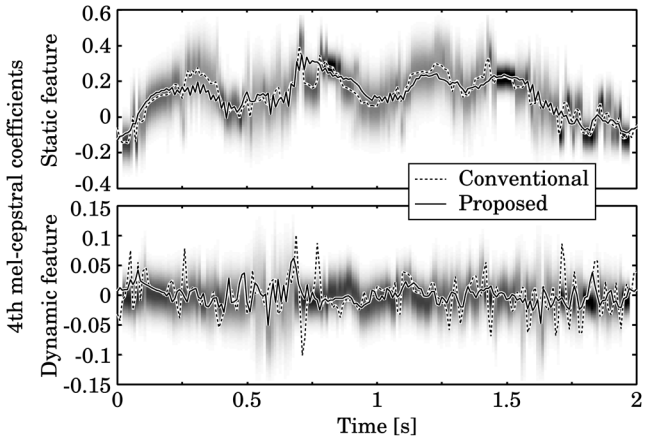

Fig. 6. Example of trajectories converted by the conventional mapping method (13) and by our proposed method (39). Intense black areas show high conditional probability densities.

Furthermore, we assume the conditions that 1) individual mixture components have the same conditional covariance matrix $\boldsymbol{D}^{(y)}$ which 2) has sufficiently large values to make the posterior probabilities $P\left(\boldsymbol{y}_t|m,\boldsymbol{x}_t,\boldsymbol{\lambda}^{(z)}\right)$ for individual mixture components equal each other. Namely, the influence of the conditional covariance matrices is disregarded. In such a case, the posterior probability is written as

$$\begin{aligned}\gamma_{m,t} &= P\left(m|\boldsymbol{x}_t,\boldsymbol{y}_t,\boldsymbol{\lambda}^{(z)}\right)\\ &= \frac{P\left(m|\boldsymbol{x}_t,\boldsymbol{\lambda}^{(z)}\right)P\left(\boldsymbol{y}_t|m,\boldsymbol{x}_t,\boldsymbol{\lambda}^{(z)}\right)}{\sum_{m=1}^{M}P\left(n|\boldsymbol{x}_t,\boldsymbol{\lambda}^{(z)}\right)P\left(\boldsymbol{y}_t|m,\boldsymbol{x}_t,\boldsymbol{\lambda}^{(z)}\right)}\\ &= P\left(m|\boldsymbol{x}_t,\boldsymbol{\lambda}^{(z)}\right). \end{aligned} \tag{43}$$

Then, $\hat{\boldsymbol{y}}_t$ is given by

$$\begin{aligned}\hat{\boldsymbol{y}}_t &= \left(\boldsymbol{D}^{(y)^{-1}}\sum_{m=1}^{M}\gamma_{m,t}\right)^{-1}\boldsymbol{D}^{(y)^{-1}}\sum_{m=1}^{M}\gamma_{m,t}\boldsymbol{E}_{m,t}^{(y)}\\ &= \sum_{m=1}^{M}P\left(m|\boldsymbol{x}_t,\boldsymbol{\lambda}^{(z)}\right)\boldsymbol{E}_{m,t}^{(y)}. \end{aligned} \tag{44}$$

Note that the converted vector is identical to that obtained by the conventional mmse-based mapping shown by (13). Therefore, the conventional mapping is regarded as an approximation of the proposed conversion method.

4) Effectiveness: Fig. 6 shows an example of the trajectories converted by the conventional mapping[3] and the proposed method on a time sequence of the conditional probability density functions. Note that each trajectory on the dynamic feature is derived from that on the static feature. Inappropriate dynamic characteristics are observed at some parts on the trajectory converted by the conventional method. On the other hand, the proposed method yields a converted trajectory with both appropriate static and appropriate dynamic characteristics. We can see an interesting example in the figure: there are some inconsistencies between a sequence of the conditional probability densities for the static feature and that for the dynamic feature at around 0.7 to 0.8 s. Conditional mean values for the static feature rapidly change but those for the dynamic feature remain at around zero. It is impossible to generate a converted trajectory with both static and dynamic features close to their respective conditional mean values. This situation arises particularly when using a context-independent conversion model such as a GMM. In such a case, our proposed method generates the converted trajectory based on more reliable conditional probability densities, of which the likelihoods are larger than the others, while keeping the likelihood reduction of the others to a minimum. In the above case, dynamic features of the resulting trajectory are close to conditional mean values, but its static features are not very close to conditional mean values. It is interesting that the resulting local pattern is similar to the target one shown in Fig. 2. This process is regarded to be a kind of smoothing of the conventional trajectory based on statistics of both static and dynamic features.[4]

[3]In order to consider the dynamic feature also in the conventional mapping method [21], [29], the converted feature vector was calculated as $\hat{\boldsymbol{y}}_t = E[\boldsymbol{y}_t|\boldsymbol{X}_t]$ on the basis of the GMM $\boldsymbol{\lambda}^{(Z)}$. Our experimental result demonstrated the conversion accuracy was almost the same as that when not considering the dynamic feature.

B. Conversion Considering GV

The GVs of parameter trajectories reconstructed in the conventional statistical modeling framework often differ significantly from the observed GVs of the target ones. In this section, we show how to model the variance more accurately by incorporating it directly into the objective function.

1) Global Variance: The GV of the target static feature vectors over a time sequence is written as

$$\boldsymbol{v}(\boldsymbol{y}) = [v(1), v(2), \ldots, v(d), \ldots, v(D)]^{\top} \tag{45}$$

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t(d) - \overline{y}(d)\right)^2 \tag{46}$$

$$\overline{y}(d) = \frac{1}{T}\sum_{t=1}^{T}y_t(d) \tag{47}$$

where $y_t(d)$ is the dth component of the target static feature vector at frame t. We calculate the GV utterance by utterance.

Fig. 7 shows time sequences of the third Mel-cepstral coefficients extracted from the natural target speech and from the converted speech. It can be observed that the GV of the converted Mel-cepstra is smaller than that of the target ones. The proposed trajectory-based conversion with MLE makes the generated trajectory close to the mean vector sequence of the conditional probability density functions. The conventional frame-based conversion with mmse shown by (13) essentially does this as well. The GV reduction is often observed because each mixture component is trained with multiple inventories from different contexts. Removed variance features are regarded to be noise in the statistical modeling of acoustic probability density.

2) Likelihood Function: We define a new likelihood function consisting of two probability density functions for a sequence of

[4]This process can also be regarded as Kalman filtering [31].

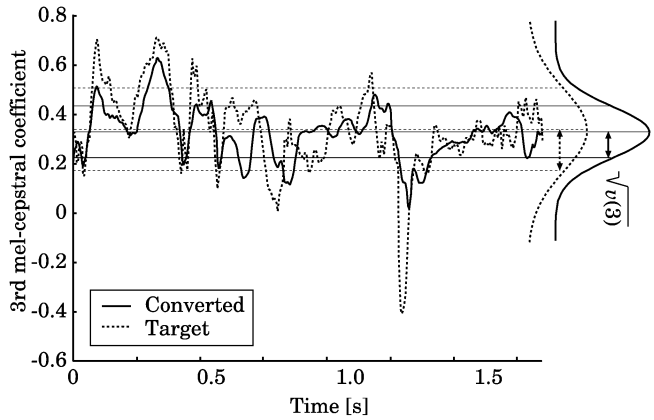

Fig. 7. Natural target and converted Mel-cepstrum sequences. Square root of GV of each sequence is shown by bidirectional arrows. Note that phoneme duration of the converted speech is different from that of the target.

the target static and dynamic feature vectors and for the GV of the target static feature vectors as follows:

$$P\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)},\boldsymbol{\lambda}^{(v)}\right)=P\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)}\right)^{\omega}P\left(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}\right) \tag{48}$$

where $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$ is modeled by a single Gaussian with the mean vector $\boldsymbol{\mu}^{(v)}$ and the covariance matrix $\boldsymbol{\Sigma}^{(vv)}$ as follows:

$$P\left(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}\right)=\mathcal{N}\left(\boldsymbol{v}(\boldsymbol{y});\boldsymbol{\mu}^{(v)},\boldsymbol{\Sigma}^{(vv)}\right). \tag{49}$$

The Gaussian distribution $\boldsymbol{\lambda}^{(v)}$ and the GMM $\boldsymbol{\lambda}^{(Z)}$ are independently trained using the training data.[5] The constant ω denotes the weight for controlling the balance between the two likelihoods. We set ω as the ratio of the number of dimensions between vectors $\boldsymbol{v}(\boldsymbol{y})$ and $\boldsymbol{Y}$, i.e., $1/(2T)$.

3) MLE of Parameter Trajectory: A time sequence of the converted feature vectors is determined as follows:

$$\hat{\boldsymbol{y}}=\operatorname{argmax}P\left(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)},\boldsymbol{\lambda}^{(v)}\right). \tag{50}$$

Note that the proposed likelihood is a function of $\boldsymbol{y}$. Namely, the proposed conversion process is performed under both the constraint of (25) and another constraint on the GV of the generated trajectory. The likelihood $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$ might be viewed as a penalty term for the reduction of the GV. We can determine $\hat{\boldsymbol{y}}$ in similar ways to those described previously.

EM algorithm: We iteratively maximize the following auxiliary function with respect to $\hat{\boldsymbol{y}}$

$$Q(\boldsymbol{Y},\hat{\boldsymbol{Y}})=\omega\sum_{\text{all }\boldsymbol{m}}P\left(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\lambda}^{(Z)}\right)\times\log P\left(\hat{\boldsymbol{Y}},\boldsymbol{m}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)}\right)+\log P\left(\boldsymbol{v}(\hat{\boldsymbol{y}})|\boldsymbol{\lambda}^{(v)}\right). \tag{51}$$

[5]We may also employ the conditional probability density of the GV given some features, such as the GV of the source feature vectors $\boldsymbol{v}(\boldsymbol{x})$ or the GV of the converted vectors $\boldsymbol{v}(\hat{\boldsymbol{y}})$ determined by (39). Our preliminary experiment demonstrated that there was little performance difference between the probability density estimated from the target training data and the conditional probability densities.

At each M-step, we iteratively update the converted parameter trajectory as follows:

$$\hat{\boldsymbol{y}}^{(i+1)\text{-th}}=\hat{\boldsymbol{y}}^{(i)\text{-th}}+\alpha\cdot\boldsymbol{\Delta}\hat{\boldsymbol{y}}^{(i)\text{-th}} \tag{52}$$

where α is the step-size parameter. When employing the steepest descent algorithm using the first derivative,[6] $\boldsymbol{\Delta}\hat{\boldsymbol{y}}^{(i)\text{-th}}$ is written as

$$\frac{\partial Q(\boldsymbol{Y},\hat{\boldsymbol{Y}})}{\partial\hat{\boldsymbol{y}}}=\omega\left(-\boldsymbol{W}^{\top}\overline{\boldsymbol{D}^{(Y)^{-1}}}\boldsymbol{W}\hat{\boldsymbol{y}}+\boldsymbol{W}^{\top}\overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}}\right)+\left[\boldsymbol{v}_1'^{\top},\boldsymbol{v}_2'^{\top},\ldots,\boldsymbol{v}_t'^{\top},\ldots,\boldsymbol{v}_T'^{\top}\right]^{\top} \tag{53}$$

$$\boldsymbol{v}_t'=[v_t'(1),v_t'(2),\ldots,v_t'(d),\ldots,v_t'(D)]^{\top} \tag{54}$$

$$v_t'(d)=-\frac{2}{T}\boldsymbol{p}_v^{(d)\top}\left(\boldsymbol{v}(\hat{\boldsymbol{y}})-\boldsymbol{\mu}_v\right)\left(\hat{y}_t(d)-\overline{\hat{y}}(d)\right) \tag{55}$$

where vector $\boldsymbol{p}_v^{(d)}$ is the dth column vector of $\boldsymbol{P}_v=\boldsymbol{\Sigma}^{(vv)^{-1}}$. The derivations of (53)–(55) are given in Appendix II.

Approximation with suboptimum mixture sequence: The computational cost is effectively reduced by approximating (48) with the suboptimum mixture component sequence. The approximated log-scaled likelihood function is given by

$$\mathcal{L}=\log\left\{P\left(\hat{\boldsymbol{m}}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)}\right)P\left(\boldsymbol{Y}|\boldsymbol{X},\hat{\boldsymbol{m}},\boldsymbol{\lambda}^{(Z)}\right)\right\}^{\omega}P\left(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}\right). \tag{56}$$

We iteratively update the converted parameter trajectory using the first derivative given by

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{y}}=\omega\left(-\boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}}\boldsymbol{W}\boldsymbol{y}+\boldsymbol{W}^{\top}\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}}\boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)}\right)+\left[\boldsymbol{v}_1'^{\top},\boldsymbol{v}_2'^{\top},\ldots,\boldsymbol{v}_t'^{\top},\ldots,\boldsymbol{v}_T'^{\top}\right]^{\top}. \tag{57}$$

We omit the derivation of (57) because it is very similar to that of (53). Our preliminary experiment demonstrated that the above approximated solution did not cause any significant quality degradation compared with the EM algorithm. Therefore, we adopt the approximated solution.

There are mainly two settings of the initial trajectory $\boldsymbol{y}^{(0)\text{-th}}$. One is to use the trajectory $\hat{\boldsymbol{y}}$ calculated by (39). The other is to use the trajectory $\hat{\boldsymbol{y}}'$ linearly converted from $\hat{\boldsymbol{y}}$ as follows:

$$\hat{y}_t'(d)=\sqrt{\frac{\mu_v(d)}{v(d)}}\left(\hat{y}_t(d)-\overline{\hat{y}}(d)\right)+\overline{\hat{y}}(d). \tag{58}$$

The trajectory $\hat{\boldsymbol{y}}$ maximizes the GMM likelihood $P(\boldsymbol{Y}|\boldsymbol{X},\hat{\boldsymbol{m}},\boldsymbol{\lambda}^{(Z)})$, while $\boldsymbol{y}'$ maximizes the GV likelihood $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$. In our preliminary experiment, we found that $\hat{\boldsymbol{y}}'$ usually has a larger value of the proposed likelihood than $\hat{\boldsymbol{y}}$ in the described setting of weight ω.

4) Effectiveness: Fig. 8 shows an example of the converted trajectories with/without the GV. By considering the GV, at a certain dimension, the trajectory movements are greatly emphasized, but at another dimension, they remain almost the same. The degree of emphasis varies between individual dimensions and frames, and it is automatically determined according to the

[6]We may employ the Newton–Raphson method using both the first and the second derivatives [32].

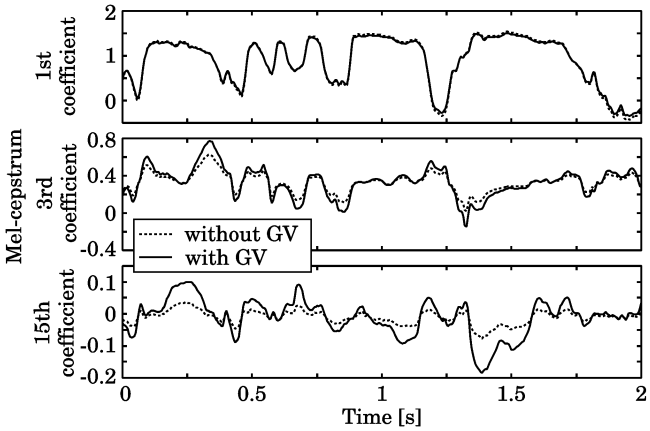

Fig. 8. Example of generated parameter trajectories with/without GV.

proposed likelihood. This process may be regarded as statistical postfiltering.

Using more mixture components for modeling the probability density also alleviates the oversmoothing effect. However, it also causes another problem of overtraining due to an increase in the number of model parameters, which often causes performance degradations for data samples not included in the training data. One of the advantages of considering the GV is that the number of parameters is kept almost equal to that when not considering the GV. In addition, since the proposed framework is based on the statistical process, it retains many advantages of the GMM-based mapping method, such as allowing model training or adaptation [33], [34] in a manner supported mathematically. Although it increases the computation cost because of employing the gradient method, the process is still sufficiently fast, as described in [35].

IV. Experimental Evaluations

A. Experimental Conditions

We conducted three kinds of experimental evaluations. First, in order to demonstrate the effectiveness of considering dynamic features, we compared the proposed trajectory-based conversion method not considering the GV shown by (39) with the conventional frame-based mapping method [21], [29] shown by (13). Second, in order to demonstrate the effectiveness of considering the GV in the proposed conversion method, we compared the method in which both dynamic features and the GV are considered with the method in which only dynamic features are considered, and spectral enhancement by postfiltering (PF), which is one of the most popular and effective enhancement techniques, was adopted instead of considering the GV. Finally, we conducted an evaluation of the total performance of the proposed VC system.

In the first and second experimental evaluations, we performed male-to-female and female-to-male VCs using the MOCHA database [36] consisting of 460 sentences for each of one male and one female speaker.[7] We selected 50 sentences at random as an evaluation set. We selected six training sets consisting of 10, 25, 50, 100, 200, and 400 sentences each from the remaining 410 sentences so that the diphone coverage of each set for all 460 sentences was maximized.[8] In order to measure only the performance of the spectral conversion, we synthesized the converted speech using the natural prosodic features automatically extracted from the target speech as follows: a time alignment for modifying the duration was performed with DTW, and then, at each frame, F_0 and total power of the converted linear spectrum were substituted with the aligned target values.

In the third experimental evaluation, we used speech data of four speakers from the CMU ARCTIC database [37], two male English speakers (bdl and rms) and two female English speakers (clb and slt), to evaluate the proposed VC system for a greater variety of speaker pairs. VC was performed for 12 speaker pairs. In each pair, 50 sentence pairs were used for training and the other 24 sentences were used for the test. The proposed spectral conversion considering the GV was employed. As for the prosodic features, only F_0 was converted as follows:

$$\hat{y}_t = \frac{\sigma^{(y)}}{\sigma^{(x)}}\left(x_t - \mu^{(x)}\right) + \mu^{(y)} \tag{59}$$

where x_t and $\hat{y}_t$ are a log-scaled F_0 of the source speaker and the converted one at frame t. Parameters $\mu^{(x)}$ and $\sigma^{(x)}$ are the mean and standard deviation of log-scaled F_0 calculated from the training data of the source speaker and $\mu^{(y)}$ and $\sigma^{(y)}$ are those of log-scaled F_0 of the target speaker.

We used the Mel-cepstrum as a spectral feature. The first through 24th Mel-cepstral coefficients were extracted from 16-kHz sampling speech data. The STRAIGHT analysis and synthesis method [38] were employed for spectral extraction and speech synthesis, respectively.

B. Effectiveness of Considering Dynamic Features

1) Objective Evaluations: The Mel-cepstral distortion between the target and converted Mel-cepstra in the evaluation set given by the following equation was used as the objective evaluation measure:

$$\text{Mel} - \text{CD [dB]} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}\left(mc_d^{(y)} - \hat{mc}_d^{(y)}\right)^2} \tag{60}$$

where $mc_d^{(y)}$ and $\hat{mc}_d^{(y)}$ are the dth coefficients of the target and converted Mel-cepstra, respectively.

Fig. 9 shows Mel-cepstral distortions in the evaluation set as a function of the number of training sentences. For each size of the training sets, we optimized the number of mixture components so that the Mel-cepstral distortion was minimized. Our proposed method (39) significantly outperforms the conventional method (13). This is because the proposed method realizes an appropriate parameter trajectory by considering the interframe correlation that is ignored in the conventional method.

The optimum number of mixture components reasonably increases as the size of the training set increases because a larger amount of training data allows the training of a more complex

[7]We use only the MOCHA speech data that include both speech and articulatory data.

[8]The resulting diphone coverage of each training set was 62.4, 81.7, 91.4, 97.0, 99.4, 99.8, and 99.8%.

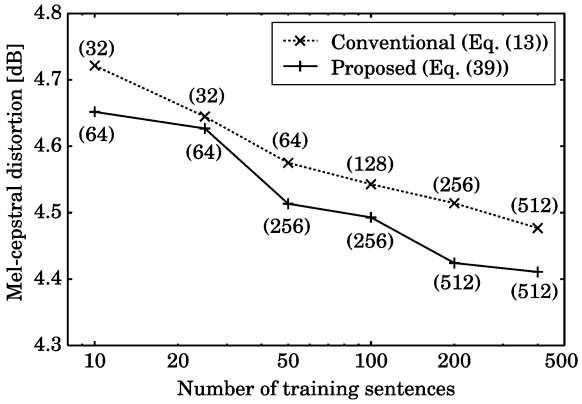

Fig. 9. Mel-cepstral distortion as a function of the number of training sentences. The distortion before the conversion is 7.30 [dB]. The numbers within parentheses indicate the optimum number of mixture components. Note that the GV is not considered in the proposed method.

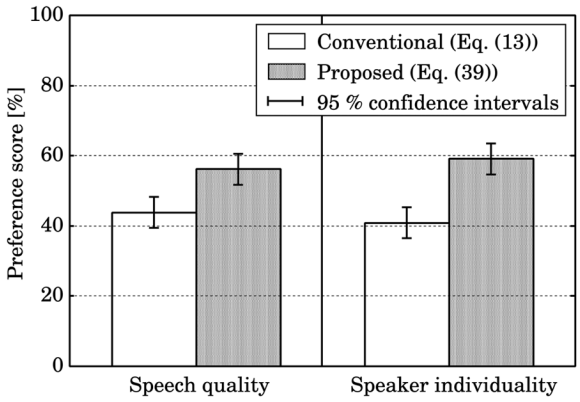

Fig. 10. Results of preference tests of speech quality and speaker individuality. Note that the GV is not considered in the proposed method.

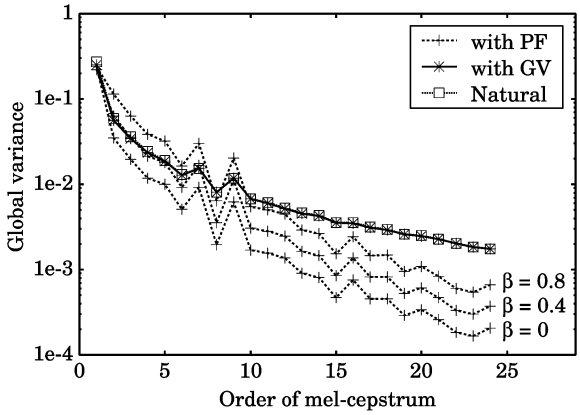

Fig. 11. GVs of several Mel-cepstrum sequences. These values show GV means over all converted voices.

conversion model. It is observed that the optimum number of mixture components in the proposed method tends to be larger than that in the conventional method. This is because a larger number of mixture components are needed to model the joint probability density of not only static but also dynamic features.

2) Subjective Evaluations: We conducted preference tests concerning speech quality and speaker individuality. In the test of speech quality, samples of speech converted by the conventional method (13) and by the proposed method (39) for each test sentence were presented to listeners in random order. Listeners were asked which sample sounded more natural. As the test of speaker individuality, an XAB test was conducted. The analysis–synthesized target speech was presented as X, and the speech converted by the conventional method and that converted by the proposed method were presented to listeners in random order as A and B. Note that both A and B are speech converted from the source speaker into the target speaker X. In each XAB set, speech samples of the same sentence were presented. Listeners were asked to choose which of A or B sounded more similar to X in terms of speaker individuality. The number of training sentences was 50. The number of mixtures was set to the optimum value shown in Fig. 9 for each conversion method. We used 25 sentences in the evaluation set. The number of listeners was ten.

Fig. 10 shows the results of the preference tests. It is observed that the proposed method yields converted speech with significantly better speech quality and a more similar personality to the target speaker compared with the conventional method. These results are consistent with the previous objective evaluation.

From the above results, it is demonstrated that the proposed trajectory-based conversion method significantly outperforms the conventional frame-based one [21], [29] in view of both objective and subjective measures even if the GV is not considered in the conversion process.

C. Effectiveness of Considering GV

1) Objective Evaluations: We conducted objective evaluations of the converted trajectories in terms of their GV characteristics and the GMM and GV likelihoods. We varied the coefficient β of the postfilter for the Mel-cepstrum [39] from 0.0 to 0.8, in which the converted spectra are increasingly emphasized as β increases. The number of training sentences was set to 50. The number of mixture components was set to 128.

Fig. 11 shows GVs of several Mel-cepstrum sequences: the converted Mel-cepstrum with the PF, that with the GV, and the natural Mel-cepstra of the target speech. It can be seen that the GV of the converted Mel-cepstra is evidently small when not considering the GV and not employing the postfilter (PF, $\beta = 0.0$). Postfiltering ($\beta > 0.0$) actually emphasizes the Mel-cepstral coefficients at a constant rate, except for the first coefficient [39]. Although it makes the GV large, GV characteristics of the emphasized Mel-cepstra are obviously different from those of the target. On the other hand, the method in which the GV is considered realizes the converted Mel-cepstra of which the GV is almost equal to the target GV.

Fig. 12 shows the logarithmic GMM likelihood $P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(Z)}\right)$ normalized by the number of frames. It is reasonable that the largest GMM likelihood is obtained when employing neither the GV nor the postfilter ($\beta = 0.0$), and it decreases when applying the postfilter or when considering the GV. An interesting point is that the GMM likelihood for the natural target trajectory is smaller than those for the converted trajectories. This implies that we do not necessarily estimate the converted trajectory that maximizes only the

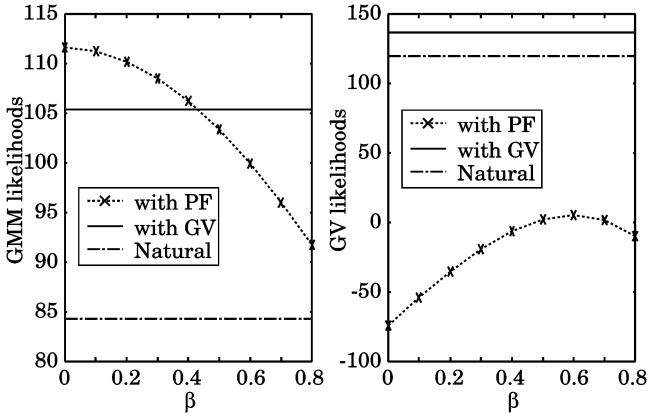

Fig. 12. Log-scaled GMM and GV likelihoods on Mel-cepstrum sequences as a function of postfilter coefficient β. The GMM likelihoods are normalized by the number of frames.

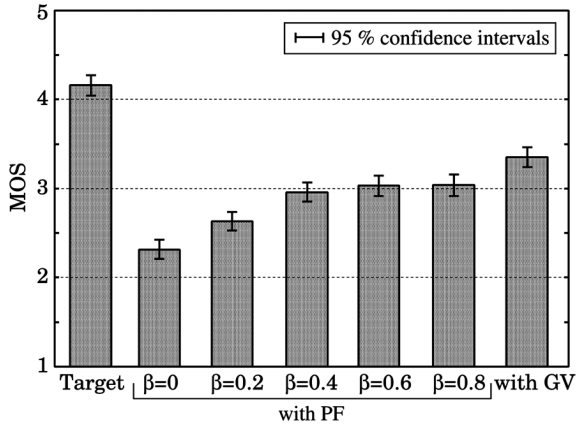

Fig. 13. Results of opinion test on speech quality. "Target" shows the result for analysis-synthesized target speech. Note that the method shown as "$\beta = 0$" of "with PF" is identical to that shown as "Proposed (39)" in Fig. 10.

GMM likelihood, though it seems reasonable to at least keep the likelihood larger than that for the natural target trajectory.

Fig. 12 also shows the logarithmic GV likelihood $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$. The GV likelihoods are very small when not considering the GV, because of the GV reduction shown in Fig. 11. Although they are recovered by postfiltering, the resulting likelihoods are still much smaller than that for the natural target. On the other hand, the method in which the GV is considered allows the converted trajectory for which the GV likelihood is sufficiently large.

Consequently, the conversion method with the GV makes both GMM and GV likelihoods exceed those for the target. These results demonstrate that the conversion method with the GV realizes more similar converted trajectories to the target one in view of satisfying a greater variety of characteristics than when not considering the GV or employing the postfilter.

2) Subjective Evaluations: We conducted an opinion test on speech quality and an XAB test on speaker individuality. In the opinion test, the opinion score was set to a five-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). The voices converted by the conversion method with the GV and those with the PF when varying the coefficient β were evaluated by each listener. In order to demonstrate speech quality under the assumption that spectral and prosodic features were converted perfectly, the analysis-synthesized target speech samples were also evaluated. More than ten samples, including both the analysis-synthesized target speech and each kind of converted speech, were presented to listeners before starting the test to make their scores more consistent. In the XAB test, the analysis-synthesized target speech was presented as X, and the GV-based and PF-based converted voices from the source into the target were presented in random order as A and B. Speech samples of the same sentence were presented as an XAB set. The postfilter coefficient β was varied again. Listeners were asked to choose which of A or B sounded more similar to X in terms of speaker individuality. We used 25 sentences in the evaluation set. The number of listeners was ten. The other experimental conditions were the same as described for the previous objective evaluation.

Fig. 13 shows the results of the opinion test. The converted speech quality is obviously improved when the GV is considered. Although the spectral enhancement by postfiltering also results in quality improvements, the improved quality is inferior to that attained by considering the GV. Because the proposed algorithm varies the emphasis rate according to the conditional probability density at each of the frames and dimensions, a more reasonable enhancement is achieved compared with postfiltering. It is interesting that the trend of MOS in the β change is similar to that of the GV likelihoods. This result implies that the GV is an important cue to speech quality.

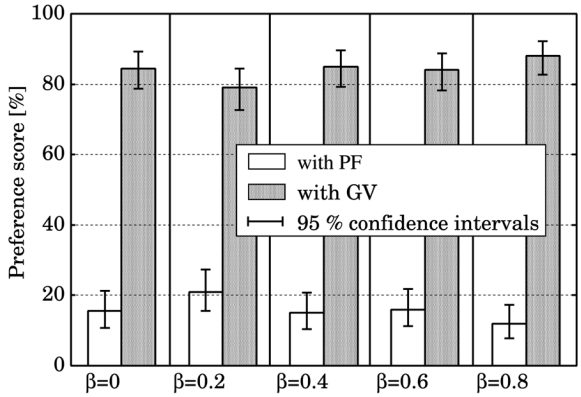

Fig. 14. Results of XAB test on speaker individuality. Note that the method shown as "$\beta = 0$" of "with PF" is identical to that shown as "Proposed (39)" in Fig. 10.

Fig. 14 shows the results of the XAB test. It is observed that the conversion method in which the GV is considered generates converted voices that have a much more similar personality to the target speaker compared with the PF-based conversion method. As described in the previous evaluations, the postfilter does not realize proper GV values. Thus, the improvements are caused by realizing the converted trajectory with the proper GV. It is possible that the GV feature itself contributes to the speaker individuality.

As a reference, an example of spectrum sequences of the target and converted voices is shown in Fig. 15. We can see that spectral peaks become much sharper when the GV is considered. Note that increasing the GV usually causes an increase in Mel-cepstral distortion between the converted trajectory and

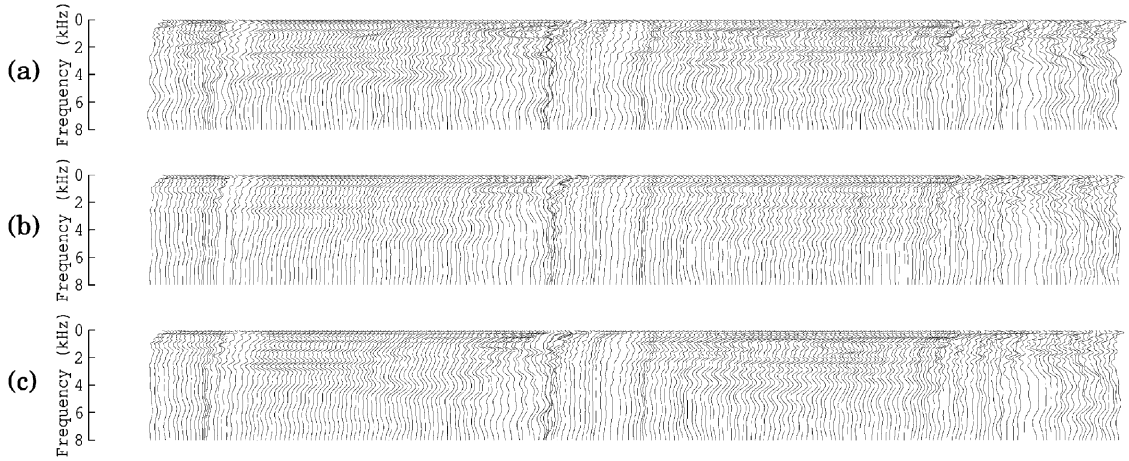

Fig. 15. Example of spectrum sequences of (a) natural target speech, (b) speech converted by the proposed method without GV, and (c) speech converted by the proposed method with GV, for the sentence fragment "farmers grow oats."

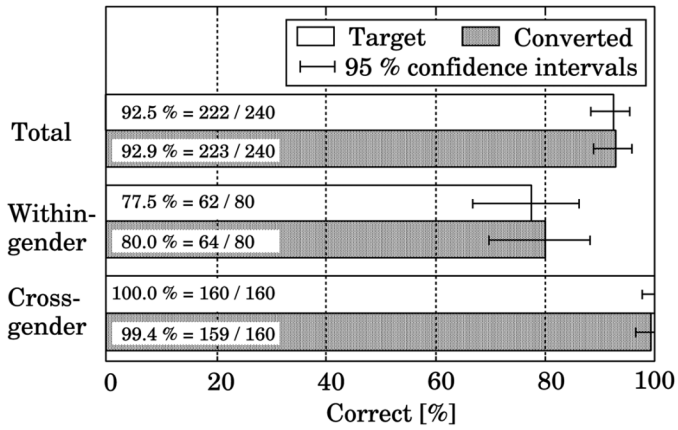

Fig. 16. Result of ABX test on speaker individuality. "Target" shows the result for analysis-synthesized target speech. "Converted" shows the result for converted speech by the proposed trajectory-based conversion method considering both dynamic features and the GV, which is identical to the method shown as "with GV" in Figs. 13 and 14.

the target one, e.g., 3.75 dB without the GV and 4.11 dB with the GV in Fig. 15. It is possible that the process of simply increasing the GV causes the quality degradation of the converted speech because it does not always make the converted sequence close to the natural one. The proposed algorithm increases the GV considering the GV likelihood while also considering the GMM likelihood to alleviate the quality degradation due to excessive increase of the GV. This process successfully results in dramatic improvements in view of both speech quality and the conversion accuracy for speaker individuality.

D. Total Performance Evaluation

We conducted an ABX test. Natural speech of the source speaker and that of the target speaker were presented to listeners in random order as A and B. Then, the converted speech was presented as X. In each ABX set, the same sentence was used for A and B, and a different one was used for X to prevent listeners from evaluating only a specific prosodic pattern of each utterance. Listeners were asked to judge whether utterance X sounded closer to utterance A or B in terms of speaker individuality. In order to show the performance of perfect spectral and prosodic voice conversion, the analysis-synthesized speech of the target speaker was also evaluated as X. The number of mixture components was set to 128. The number of listeners was ten.

Fig. 16 shows a result of the ABX test. The proposed VC system obviously is very effective.[9] It is not surprising that almost all the errors occurred in the within-gender conversion. It is noteworthy that the performance of the proposed VC system is comparable to that of the perfect VC system when it should be inferior to the perfect one because the proposed VC system does not carefully convert prosodic patterns. The performance difference between those two might be observed in another speaker pair whose prosodic characteristics are very different from each other. It might also be observed in a more challenging test such as a speaker recognizability test rather than the ABX text.

V. Conclusion

We proposed a spectral conversion method for VC based on maximum-likelihood estimation of a parameter trajectory. It was shown that the conventional frame-based mapping method based on the minimum mean square error [21], [29] is regarded as an approximation of the proposed conversion method. We emphasized the proposed framework by introducing two main ideas: 1) the conversion considering the feature correlation between frames for realizing appropriate spectral local patterns and 2) the conversion considering the global variance for alleviating the oversmoothing effect. Experimental results demonstrated that the proposed ideas can dramatically improve the conversion performance in view of both speech quality and the conversion accuracy for speaker individuality.

It is indispensable to continue to make progress in the conversion method to make VC practically applicable. It is worthwhile to convert the source features, such as residual signals [40]–[42] as well as spectral features. Prosodic conversion, such as F_0 conversion [43] and duration conversion, is also important to

[9]Some samples are available at http://spalab.naist.jp/~tomoki/IEEE/MLVC/index.html.

more accurately realize speaker personality. Moreover, it is desired to realize a more flexible training framework of the conversion model accepting nonparallel data [33], [41], [44]. It seems effective to apply model adaptation techniques developed in a speech recognition area to VC for realizing a novel VC framework [33], [34], [45].

APPENDIX I
DERIVATION OF (30)

The auxiliary function of (29) is written as

$$\begin{aligned}
&Q(\boldsymbol{Y},\hat{\boldsymbol{Y}}) \\
&\quad=\sum_{\text{all }\boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\lambda}^{(Z)}\right)\log P\left(\hat{\boldsymbol{Y}},\boldsymbol{m}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)}\right) \\
&\quad=\sum_{t=1}^{T}\sum_{m=1}^{M} P\left(m|\boldsymbol{X}_t,\boldsymbol{Y}_t,\boldsymbol{\lambda}^{(Z)}\right)\log P\left(\hat{\boldsymbol{Y}}_t,m|\boldsymbol{X}_t,\boldsymbol{\lambda}^{(Z)}\right) \\
&\quad=\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_{m,t}\left(-\frac{1}{2}\hat{\boldsymbol{Y}}_t^{\top}\boldsymbol{D}_m^{(Y)^{-1}}\hat{\boldsymbol{Y}}_t+\hat{\boldsymbol{Y}}_t^{\top}\boldsymbol{D}_m^{(Y)^{-1}}\boldsymbol{E}_{m,t}^{(Y)}\right) \\
&\qquad+\overline{K} \\
&\quad=\sum_{t=1}^{T}-\frac{1}{2}\hat{\boldsymbol{Y}}_t^{\top}\overline{\boldsymbol{D}_t^{(Y)^{-1}}}\hat{\boldsymbol{Y}}_t+\hat{\boldsymbol{Y}}_t^{\top}\overline{\boldsymbol{D}_t^{(Y)^{-1}}\boldsymbol{E}_t^{(Y)}}+\overline{K} \\
&\quad=-\frac{1}{2}\hat{\boldsymbol{Y}}^{\top}\overline{\boldsymbol{D}^{(Y)^{-1}}}\hat{\boldsymbol{Y}}+\hat{\boldsymbol{Y}}^{\top}\overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}}+\overline{K} \\
&\quad=-\frac{1}{2}\hat{\boldsymbol{y}}^{\top}\boldsymbol{W}^{\top}\overline{\boldsymbol{D}^{(Y)^{-1}}}\boldsymbol{W}\hat{\boldsymbol{y}}+\hat{\boldsymbol{y}}^{\top}\boldsymbol{W}^{\top}\overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}} \\
&\qquad+\overline{K} \qquad (61)
\end{aligned}$$

where $\overline{\boldsymbol{D}^{(Y)^{-1}}}$, $\overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}}$, $\overline{\boldsymbol{D}_t^{(Y)^{-1}}}$, $\overline{\boldsymbol{D}_t^{(Y)^{-1}}\boldsymbol{E}_t^{(Y)}}$, and $\gamma_{m,t}$ are given by (31)–(35), respectively. The constant $\overline{K}$ is independent of $\hat{\boldsymbol{y}}$. By setting the first derivative of the auxiliary function with respect to $\hat{\boldsymbol{y}}$ given by

$$\frac{\partial Q(\boldsymbol{Y},\hat{\boldsymbol{Y}})}{\partial \boldsymbol{y}}=-\boldsymbol{W}^{\top}\overline{\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}}}\boldsymbol{W}\boldsymbol{y}+\boldsymbol{W}^{\top}\overline{\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}}\boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)}} \qquad (62)$$

to zero, $\hat{\boldsymbol{y}}$ that maximizes the auxiliary function is determined as follows:

$$\hat{\boldsymbol{y}}=\left(\boldsymbol{W}^{\top}\overline{\boldsymbol{D}^{(Y)^{-1}}}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^{\top}\overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}}. \qquad (63)$$

APPENDIX II
DERIVATIONS OF (53)–(55)

The auxiliary function of (51) is written as

$$\begin{aligned}
Q(\boldsymbol{Y},\hat{\boldsymbol{Y}}) &=\omega\sum_{\text{all }\boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\lambda}^{(Z)}\right)\log P\left(\hat{\boldsymbol{Y}},\boldsymbol{m}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)}\right) \\
&\quad+\log P(\boldsymbol{v}(\hat{\boldsymbol{y}})|\boldsymbol{\lambda}^{(v)}) \\
&=\omega\mathcal{L}_1+\mathcal{L}_2 \qquad (64)
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{L}_1 &=\sum_{\text{all }\boldsymbol{m}} P\left(\boldsymbol{m}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\lambda}^{(Z)}\right)\log P\left(\hat{\boldsymbol{Y}},\boldsymbol{m}|\boldsymbol{X},\boldsymbol{\lambda}^{(Z)}\right) \\
&=-\frac{1}{2}\hat{\boldsymbol{y}}^{\top}\boldsymbol{W}^{\top}\overline{\boldsymbol{D}^{(Y)^{-1}}}\boldsymbol{W}\hat{\boldsymbol{y}}+\hat{\boldsymbol{y}}^{\top}\boldsymbol{W}^{\top}\overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}} \\
&\quad+\overline{K} \qquad (65) \\
\mathcal{L}_2 &=\log P(\boldsymbol{v}(\hat{\boldsymbol{y}})|\boldsymbol{\lambda}^{(v)}) \\
&=-\frac{1}{2}\boldsymbol{v}(\hat{\boldsymbol{y}})^{\top}\boldsymbol{\Sigma}^{(vv)^{-1}}\boldsymbol{v}(\hat{\boldsymbol{y}})+\boldsymbol{v}(\hat{\boldsymbol{y}})^{\top}\boldsymbol{\Sigma}^{(vv)^{-1}}\boldsymbol{\mu}^{(v)} \\
&\quad+\overline{K'}. \qquad (66)
\end{aligned}$$

The constants $\overline{K}$ and $\overline{K'}$ are independent of $\hat{\boldsymbol{y}}$. The derivative of $\mathcal{L}_1$ with respect to $\hat{\boldsymbol{y}}$ is given by (62). The derivative of $\mathcal{L}_2$ with respect to $\hat{\boldsymbol{y}}$ is given by

$$\begin{aligned}
\frac{\partial\mathcal{L}_2}{\partial\hat{\boldsymbol{y}}} &=\left[\frac{\partial\mathcal{L}_2}{\partial\hat{\boldsymbol{y}}_1}^{\top},\frac{\partial\mathcal{L}_2}{\partial\hat{\boldsymbol{y}}_2}^{\top},\ldots,\frac{\partial\mathcal{L}_2}{\partial\hat{\boldsymbol{y}}_t}^{\top},\ldots,\frac{\partial\mathcal{L}_2}{\partial\hat{\boldsymbol{y}}_T}^{\top}\right]^{\top} \qquad (67) \\
\frac{\partial\mathcal{L}_2}{\partial\hat{\boldsymbol{y}}_t} &=\left[\frac{\partial\mathcal{L}_2}{\partial\hat{y}_t(1)}^{\top},\frac{\partial\mathcal{L}_2}{\partial\hat{y}_t(2)}^{\top},\ldots,\right. \\
&\qquad\left.\frac{\partial\mathcal{L}_2}{\partial\hat{y}_t(d)}^{\top},\ldots,\frac{\partial\mathcal{L}_2}{\partial\hat{y}_t(D)}^{\top}\right]^{\top}=\boldsymbol{v}_t' \qquad (68) \\
\frac{\partial\mathcal{L}_2}{\partial\hat{y}_t(d)} &=\frac{\partial\mathcal{L}_2}{\partial v(d)}\frac{\partial v(d)}{\partial\hat{y}_t(d)}=v_t'(d) \qquad (69) \\
\frac{\partial\mathcal{L}_2}{\partial v(d)} &=-\boldsymbol{p}_v^{(d)^{\top}}(\boldsymbol{v}(\hat{\boldsymbol{y}})-\boldsymbol{\mu}_v) \qquad (70) \\
\frac{\partial v(d)}{\partial\hat{y}_t(d)} &=\frac{2}{T}\left\{\left(\hat{y}_t(d)-\overline{\hat{y}}(d)\right)-\frac{1}{T}\sum_{\tau=1}^{T}\left(\hat{y}_\tau(d)-\overline{\hat{y}}(d)\right)\right\} \\
&=\frac{2}{T}\left(\hat{y}_t(d)-\overline{\hat{y}}(d)\right) \qquad (71)
\end{aligned}$$

where vector $\boldsymbol{p}_v^{(d)}$ is the dth column vector of $\boldsymbol{P}_v=\boldsymbol{\Sigma}^{(vv)^{-1}}$. Consequently, the derivative of the auxiliary function with respect to $\hat{\boldsymbol{y}}$ is written as

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{Y},\hat{\boldsymbol{Y}})}{\partial\hat{\boldsymbol{y}}} &=\omega\left(-\boldsymbol{W}^{\top}\overline{\boldsymbol{D}^{(Y)^{-1}}}\boldsymbol{W}\hat{\boldsymbol{y}}+\boldsymbol{W}^{\top}\overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}}\right) \\
&\quad+\left[\boldsymbol{v}_1'^{\top},\boldsymbol{v}_2'^{\top},\ldots,\boldsymbol{v}_t'^{\top},\ldots,\boldsymbol{v}_T'^{\top}\right]^{\top} \qquad (72) \\
\boldsymbol{v}_t' &=\left[v_t'(1),v_t'(2),\ldots,v_t'(d),\ldots,v_t'(D)\right]^{\top} \qquad (73) \\
v_t'(d) &=-\frac{2}{T}\boldsymbol{p}_v^{(d)^{\top}}(\boldsymbol{v}(\hat{\boldsymbol{y}})-\boldsymbol{\mu}_v)\left(\hat{y}_t(d)-\overline{\hat{y}}(d)\right). \qquad (74)
\end{aligned}$$

ACKNOWLEDGMENT

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

REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. Jpn. (E)*, vol. 11, no. 2, pp. 71–76, 1990.
[2] M. Abe, K. Shikano, and H. Kuwabara, "Statistical analysis of bilingual speaker's speech for cross-language voice conversion," *J. Acoust. Soc. Amer.*, vol. 90, no. 1, pp. 76–82, 1991.
[3] M. Mashimo, T. Toda, H. Kawanami, K. Shikano, and N. Campbell, "Cross-language voice conversion evaluation using bilingual databases," *IPSJ J.*, vol. 43, no. 7, pp. 2177–2185, 2002.
[4] K. Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Proc. ICASSP*, Istanbul, Turkey, Jun. 2000, pp. 1847–1850.

[5] M. L. Seltzer, A. Acero, and J. Droppo, "Robust bandwidth extension of noise-corrupted narrowband speech," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 1509–1512.
[6] Y. Shiga and K. Simon, "Accurate spectral envelope estimation for articulation-to-speech synthesis," in *Proc. 5th ISCA Speech Synth. Workshop*, Pittsburgh, PA, Jun. 2004, pp. 19–24.
[7] T. Toda, A. W. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis," in *Proc. 5th ISCA Speech Synth. Workshop*, Pittsburgh, PA, Jun. 2004, pp. 31–36.
[8] K. Richmond, S. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Comput. Speech Lang.*, vol. 17, pp. 153–172, 2003.
[9] T. Toda, A. W. Black, and K. Tokuda, "Acoustic-to-articulatory inversion mapping with Gaussian mixture model," in *Proc. Interspeech*, Jeju, Korea, Oct. 2004, pp. 1129–1132.
[10] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X. Huang, "Air- and bone-conductive integrated microphones for robust speech detection and enhancement," in *Proc. ASRU*, St. Thomas, Virgin Islands, Dec. 2003, pp. 249–254.
[11] T. Toda and K. Shikano, "NAM-to-speech conversion with Gaussian mixture models," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005, pp. 1957–1960.
[12] M. Nakagiri, T. Toda, H. Saruwatari, and K. Shikano, "Improving body transmitted unvoiced speech with statistical voice conversion," in *Proc. Interspeech*, Pittsburgh, PA, Sep. 2006, pp. 2270–2273.
[13] A. Kain, X. Niu, J.-P. Hosom, Q. Miao, and J. van Santen, "Formant re-synthesis of dysarthric speech," in *Proc. 5th ISCA Speech Synth. Workshop*, Pittsburgh, PA, Jun. 2004, pp. 25–30.
[14] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech," in *Proc. Interspeech*, Pittsburgh, PA, Sep. 2006, pp. 1395–1398.
[15] H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: Control and conversion," *Speech Commun.*, vol. 16, no. 2, pp. 165–173, 1995.
[16] S. Nakamura and K. Shikano, "Speaker adaptation applied to HMM and neural networks," in *Proc. ICASSP*, Glasgow, U.K., May 1989, pp. 89–92.
[17] H. Matsumoto and Y. Yamashita, "Unsupervised speaker adaptation from short utterances based on a minimized fuzzy objective function," *J. Acoust. Soc. Jpn. (E)*, vol. 14, no. 5, pp. 353–361, 1993.
[18] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," *Speech Commun.*, vol. 11, no. 2–3, pp. 175–187, 1992.
[19] N. Iwahashi and Y. Sagisaka, "Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks," *Speech Commun.*, vol. 16, no. 2, pp. 139–151, 1995.
[20] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech Commun.*, vol. 16, no. 2, pp. 207–216, 1995.
[21] 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.
[22] T. Toda, H. Saruwatari, and K. Shikano, "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum," in *Proc. ICASSP*, Salt Lake City, UT, May 2001, pp. 841–844.
[23] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proc. ICASSP*, Philadelphia, PA, Mar. 2005, vol. 1, pp. 9–12.
[24] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. ICASSP*, Detroit, MI, May 1995, pp. 660–663.
[25] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," in *Proc. Eurospeech*, Madrid, Spain, Sep. 1995, pp. 757–760.
[26] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, Istanbul, Turkey, Jun. 2000, pp. 1315–1318.
[27] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *Proc. IEEE Workshop Speech Synth.*, Santa Monica, CA, Sep. 2002, pp. 227–230.
[28] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, Budapest, Hungary, Sep. 1999, pp. 2347–2350.
[29] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, Seattle, WA, May 1998, pp. 285–288.
[30] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajetory model by imposing explicit relationships between static and dynamic feature vector sequences," *Comput. Speech Lang.*, vol. 21, pp. 153–173, 2007.
[31] Y. Minami, E. McDermott, A. Nakamura, and S. Katagiri, "A theoretical analysis of speech recognition based on feature trajectory models," in *Proc. Interspeech*, Jeju, Korea, Oct. 2004, pp. 549–552.
[32] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans.*, vol. E90-D, no. 5, pp. 816–824, May 2007.
[33] A. Mouchtaris, J. V. der Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 952–963, May 2006.
[34] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," in *Proc. ICSLP*, Pittsburgh, PA, Sep. 2006, pp. 2446–2449.
[35] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the nitech HMM-based speech synthesis system for the Blizzard challenge 2005," *IEICE Trans.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.
[36] A. Wrench, "The MOCHA-TIMIT Articulatory Database," Queen Margaret Univ. College, Edinburgh, U.K., 1999 [Online]. Available: http://www.cstr.ed.ac.uk/artic/mocha.html
[37] J. Kominek and A. W. Black, "CMU ARCTIC Databases for Speech Synthesis," Lang. Technol. Inst., Carnegie Mellon Univ., Pittsburgh, PA, 2003 [Online]. Available: http://festvox.org/cmu_arctic/index.html
[38] 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.
[39] K. Koishida, K. Tokuda, T. Kobayashi, and S. Imai, "CELP coding based on Mel-cepstral analysis," in *Proc. ICASSP*, Detroit, MI, May 1995, pp. 33–36.
[40] A. Kain and M. W. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *Proc. ICASSP*, Salt Lake City, UT, May 2001, pp. 813–816.
[41] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1301–1312, Jul. 2006.
[42] D. Suendermann, A. Bonafonte, H. Ney, and H. Hoege, "A study on residual prediction techniques for voice conversion," in *Proc. ICASSP*, Philadelphia, PA, Mar. 2005, vol. 1, pp. 13–16.
[43] B. Gillett and S. King, "Transforming F_0 contours," in *Proc. Interspeech*, Geneva, Switzerland, Sep. 2003, pp. 101–104.
[44] D. Suendermann, H. Hoege, A. Bonafonte, H. Ney, A. W. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," in *Proc. ICASSP*, Toulouse, France, Mar. 2006, vol. 1, pp. 81–84.
[45] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," in *Proc. ICASSP*, Honolulu, HI, Apr. 2007, vol. 4, pp. 1249–1252.

Tomoki Toda (M'05) received the B.E. degree in electrical engineering from Nagoya University, Nagoya, Japan, in 1999 and the M.E. and Ph.D. degrees in engineering from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Nara, Japan, in 2001 and 2003, respectively.

From 2001 to 2003, he was an Intern Researcher at the ATR Spoken Language Translation Research Laboratories, Kyoto, Japan. He was a Research Fellow of the Japan Society for the Promotion of Science in Graduate School of Engineering, Nagoya Institute of Technology from 2003 to 2005. He was a Visiting Researcher at the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, from October 2003 to September 2004. He is currently an Assistant Professor in the Graduate School of Information Science, NAIST, and a Visiting Researcher at the ATR Spoken Language Communication Research Laboratories. His research interests include speech transformation, speech synthesis, speech analysis, and speech recognition.

Dr. Toda received the TELECOM System Technology Award for Students from the Telecommunications Advancement Foundation in 2003. He has been a member of the Speech and Language Technical Committee of the IEEE Signal Processing Society since January 2007. He is a member the ISCA, IEICE, and ASJ.

Alan W. Black (M'03) received the B.Sc. degree (hons) in computer science from Coventry University, Coventry, U.K., in 1984, the M.Sc. degree in knowledge-based systems from Edinburgh University, Edinburgh, U.K, in 1986 and the Ph.D. degree in computational linguistics from Edinburgh University in 1993.

He is an Associate Research Professor in the Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA. He previously worked in the Centre for Speech Technology Research, University of Edinburgh, Edinburgh, U.K., and before that at the ATR Spoken Language Translation Research Laboratories, Kyoto, Japan. He is one of the principal authors of the Festival Speech Synthesis System, the FestVox voice building tools and CMU Flite, a small footprint speech synthesis engine. He is also Chief Scientist and cofounder of the for-profit company Cepstral, LLC. Although his recent work primarily focuses on speech synthesis, he also works on small footprint speech-to-speech translation systems (Croatian, Arabic, and Thai), telephone-based spoken dialog systems, and speech for robots. In 2004, with Prof. K. Tokuda, he initiated the now annual Blizzard Challenge, the largest multisite evaluation of corpus-based speech synthesis techniques.

Prof. Black was a member of the IEEE Speech Technical Committee from 2004 to 2007 and is a member of the ISCA Advisory Council. He was Program Chair of the ISCA Speech Synthesis Workshop 2004, and was General Co-Chair of Interspeech 2006—ICSLP.

Keiichi Tokuda (M'89) received the B.E. degree in electrical and electronic engineering from the Nagoya Institute of Technology, Nagoya, Japan, in 1984 and the M.E. and Dr.Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1986 and 1989, respectively.

From 1989 to 1996, he was a Research Associate in the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. From 1996 to 2004, he was an Associate Professor in the Department of Computer Science, Nagoya Institute of Technology, where he is currently a Professor. He is also an Invited Researcher at the ATR Spoken Language Communication Research Laboratories, Kyoto, Japan, and was a Visiting Researcher at Carnegie Mellon University, Pittsburgh, PA, from 2001 to 2002. His research interests include speech coding, speech synthesis and recognition, and statistical machine learning.

Prof. Tokuda is a corecipient of the Paper Award and the Inose Award, both from the IEICE in 2001, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2001. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society. He is a member the ISCA, IEICE, IPSJ, ASJ, and JSAI.