Probabilistic Feature Mapping Based on Trajectory HMMs

Heiga Zen, Yoshihiko Nankaku, Keiichi Tokuda

Department of Computer Science and Engineering, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya, 466–8555 Japan

{zen,nankaku,tokuda}@sp.nitech.ac.jp

Abstract

This paper proposes a feature mapping algorithm based on the trajectory GMM or trajectory HMM. Although the GMM or HMM-based feature mapping algorithm works effectively, its conversion quality sometimes degrades due to the inappropriate dynamic characteristics caused by the frame-by-frame conversion. While the use of dynamic features can alleviate this problem, it also introduces an inconsistency between training and mapping. The proposed algorithm can solve this inconsistency while keeping the benefits of the use of dynamic features, and offers an entire sequence-level transformation rather than the frame-by-frame conversion. Experimental results in voice conversion show that the proposed algorithm outperforms the conventional one both in objective and subjective tests.

Index Terms: trajectory HMM, voice conversion

1. Introduction

During these last years, a probabilistic feature mapping algorithm based on Gaussian mixture models (GMMs) or hidden Markov models (HMMs) has been used in various speech applications [1–3]. In this algorithm, GMM or HMM-based mapping functions are estimated using pairs of source and target feature vectors. The resulting mapping functions make it possible to continuously transform of any sample of the source into that of target. Although this algorithm works effectively, its conversion performance is still insufficient. One of major factors which deteriorate the conversion quality could be inappropriate dynamic characteristics caused by the frame-by-frame conversion. To alleviate this problem, Toda et al. introduced the dynamic feature constraints [4]. The use of dynamic features makes it possible to convert a source feature vector sequence to a target one while satisfying the statistics of both static and dynamic characteristics. However, it also introduces an inconsistency between training and mapping: In probabilistic feature mapping, first joint probability density functions (PDFs) of source and target feature vector sequences are modeled by a probabilistic model [5]. Second, conditional PDFs of a target feature vector sequence for a given source one are obtained from the joint PDFs.[1] Finally, the converted target parameter vector sequence is determined so as to minimize its mean-square error (MMSE) [5] or maximize its likelihood (ML) [4] based on the conditional PDFs. In the conventional mapping algorithm using dynamic features, the relationship between static and dynamic features[2] is ignored on both the joint and conditional PDF estimation stages but utilized on the mapping stage [4, 6]. Ignoring these interdependencies allows inconsistency between the static and dynamic feature vector sequences when the HMM or GMM is used as a generative model in the obvious way. Zen et al. reported that this inconsistency degraded the quality of HMM-based speech synthesis [7]. It suggests that this inconsistency also degrades the performance of GMM or HMM-based feature mapping algorithm with dynamic features.

Recently, a trajectory model, derived from the HMM by imposing the explicit relationships between static and dynamic features, was proposed [7]. This model, named trajectory HMM, can overcome the frame-wise conditional independence assumption of state output probabilities and piecewise constant statistics of the HMM without any additional parameters.

In this paper, we propose a probabilistic feature mapping algorithm based on the trajectory GMM or trajectory HMM. The proposed algorithm can solve the inconsistency introduced by the use of dynamic features while keeping its benefits, and offers an entire sequence-level transformation rather than the frame-by-frame conversion. We evaluate the proposed algorithm in voice conversion, which is one of the most successful applications of probabilistic feature mapping.

This paper is organized as follows: Section 2 describes the joint probability modeling by the trajectory GMM or trajectory HMM. Section 3 derives mapping algorithms based on the MMSE and ML criteria. Section 4 shows objective and subjective experimental results in voice conversion. Concluding remarks and future plans are given in the final section.

2. Probability Density Function

Source and target static feature vector sequences, $\boldsymbol{x}$ and $\boldsymbol{y}$, are written as

$$\boldsymbol{x} = \left[\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_T^\top\right]^\top, \quad \boldsymbol{y} = \left[\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_T^\top\right]^\top, \tag{1}$$

where $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are the $M_{\boldsymbol{x}}$ and $M_{\boldsymbol{y}}$-dimensional source and target static feature vectors at the t-th frame, respectively, and T is the total number of frames. In this paper, we model the joint probability of $\boldsymbol{x}$ and $\boldsymbol{y}$ by a trajectory GMM or a set of trajectory HMMs [7] as follows:

$$P\left(\boldsymbol{z} \mid \lambda\right) = \sum_{\forall \boldsymbol{q}} P\left(\boldsymbol{q} \mid \lambda\right) P\left(\boldsymbol{z} \mid \boldsymbol{q}, \lambda\right), \tag{2}$$

$$P\left(\boldsymbol{z} \mid \boldsymbol{q}, \lambda\right) = \mathcal{N}\left(\boldsymbol{z}\,;\, \bar{\boldsymbol{z}}_{\boldsymbol{q}}, \boldsymbol{P}_{\boldsymbol{q}}\right), \tag{3}$$

$$P\left(\boldsymbol{q} \mid \lambda\right) = \begin{cases} \displaystyle\prod_{t=1}^{T} w_{q_t}, & \text{trajectory GMM} \\ \displaystyle\pi_{q_1} \prod_{t=2}^{T} a_{q_{t-1} q_t}, & \text{trajectory HMM} \end{cases} \tag{4}$$

[1] Modeling conditional PDFs directly has also been proposed [6].

[2] Generally, dynamic features are calculated as regression coefficients from static features of their neighboring frames.

$$\boldsymbol{z} = \left[\boldsymbol{z}_1^\top, \ldots, \boldsymbol{z}_T^\top\right]^\top, \quad \boldsymbol{z}_t = \left[\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top\right]^\top, \tag{5}$$

where λ denotes the set of model parameters, $\boldsymbol{z}$ is the MT-dimensional joint static feature vector sequence, $\boldsymbol{z}_t$ is the the M-dimensional joint static feature vector at the t-th frame, $M = M_{\boldsymbol{x}} + M_{\boldsymbol{y}}$ is the dimensionality of joint static feature vectors, $\boldsymbol{q} = \{q_1, \ldots, q_T\}$ is a Gaussian component sequence, $q_t \in \{1, \ldots, N\}$ is the Gaussian component at the t-th frame, N is the total number of Gaussian components in the model set, w_i is the mixture prior probability of the i-th Gaussian component, π_i is the initial probability of the i-th Gaussian component, and a_{ij} is the transition probability from the i-th Gaussian component to the j-th one. In Eq. (3), $\bar{\boldsymbol{z}}_{\boldsymbol{q}}$ and $\boldsymbol{P}_{\boldsymbol{q}}$ are the $MT \times 1$ mean vector and the $MT \times MT$ covariance matrix for $\boldsymbol{q}$, respectively. They are given by

$$\boldsymbol{R}_{\boldsymbol{q}} \bar{\boldsymbol{z}}_{\boldsymbol{q}} = \boldsymbol{r}_{\boldsymbol{q}}, \tag{6}$$

$$\boldsymbol{R}_{\boldsymbol{q}} = \boldsymbol{W}^\top \boldsymbol{\Omega}_{\boldsymbol{q}} \boldsymbol{W} = \boldsymbol{P}_{\boldsymbol{q}}^{-1}, \tag{7}$$

$$\boldsymbol{r}_{\boldsymbol{q}} = \boldsymbol{W}^\top \boldsymbol{\Omega}_{\boldsymbol{q}} \boldsymbol{\mu}_{\boldsymbol{q}}, \tag{8}$$

$$\boldsymbol{\mu}_{\boldsymbol{q}} = \left[\boldsymbol{\mu}_{q_1}^\top, \ldots, \boldsymbol{\mu}_{q_T}^\top\right]^\top, \tag{9}$$

$$\boldsymbol{\mu}_i = \left[\boldsymbol{\mu}_i^{(\boldsymbol{x})\top}, \boldsymbol{\mu}_i^{(\boldsymbol{y})\top}\right]^\top, \tag{10}$$

$$\boldsymbol{\Omega}_{\boldsymbol{q}} = \mathrm{diag}\left[\boldsymbol{\Omega}_{q_1} \ldots, \boldsymbol{\Omega}_{q_T}\right], \tag{11}$$

$$\boldsymbol{\Omega}_i = \begin{bmatrix} \boldsymbol{\Omega}_i^{(\boldsymbol{xx})} & \boldsymbol{\Omega}_i^{(\boldsymbol{xy})} \\ \boldsymbol{\Omega}_i^{(\boldsymbol{yx})} & \boldsymbol{\Omega}_i^{(\boldsymbol{yy})} \end{bmatrix}, \tag{12}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Omega}_i$ are the $3M \times 1$ mean vector and the $3M \times 3M$ inverse covariance (precision) matrix associated with the i-th Gaussian component, respectively, and $\boldsymbol{W}$ is a $3MT \times MT$ window matrix which appends dynamic features (delta and delta-delta features) to $\boldsymbol{z}$. Equations (2)–(4) show that the trajectory GMM and trajectory HMM can be interpreted as a MT-dimensional GMM whose mixture weights are given by products of the mixture prior probabilities and the transition probabilities over time, respectively. It should be noted that the intra and inter-frame covariance matrix of Eq. (3) is generally full. Therefore, the trajectory GMM and trajectory HMM can model both intra-frame and inter-frame dependencies between source and target static feature vector sequences without increasing the number of model parameters compared to the GMM and HMM with the same model topology, respectively.

3. Mapping

3.1. Conditional probability

We can rewrite Eq. (3) as follows:

$$P\left(\boldsymbol{z} \mid \boldsymbol{q}, \lambda\right) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}; \begin{bmatrix} \bar{\boldsymbol{x}}_{\boldsymbol{q}} \\ \bar{\boldsymbol{y}}_{\boldsymbol{q}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{xx})} & \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{xy})} \\ \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yx})} & \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yy})} \end{bmatrix}\right), \tag{13}$$

where

$$\boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{xx})} = \boldsymbol{C}_{\boldsymbol{q}}^{(\boldsymbol{xx})^{-1}}, \quad \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yy})} = \boldsymbol{C}_{\boldsymbol{q}}^{(\boldsymbol{yy})^{-1}}, \tag{14}$$

$$\boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{xy})} = -\boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{xx})^{-1}} \boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{xy})} \boldsymbol{C}_{\boldsymbol{q}}^{(\boldsymbol{yy})^{-1}} = \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yx})^\top}, \tag{15}$$

$$\boldsymbol{C}_{\boldsymbol{q}}^{(\boldsymbol{xx})} = \boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{xx})} - \boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{xy})} \boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{yy})^{-1}} \boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{yx})}, \tag{16}$$

$$\boldsymbol{C}_{\boldsymbol{q}}^{(\boldsymbol{yy})} = \boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{yy})} - \boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{yx})} \boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{xx})^{-1}} \boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{xy})}, \tag{17}$$

$$\boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{xx})} = \boldsymbol{W}^{(\boldsymbol{x})^\top} \boldsymbol{\Omega}_{\boldsymbol{q}}^{(\boldsymbol{xx})} \boldsymbol{W}^{(\boldsymbol{x})}, \tag{18}$$

$$\boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{yy})} = \boldsymbol{W}^{(\boldsymbol{y})^\top} \boldsymbol{\Omega}_{\boldsymbol{q}}^{(\boldsymbol{yy})} \boldsymbol{W}^{(\boldsymbol{y})}, \tag{19}$$

$$\boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{xy})} = \boldsymbol{W}^{(\boldsymbol{x})^\top} \boldsymbol{\Omega}_{\boldsymbol{q}}^{(\boldsymbol{xy})} \boldsymbol{W}^{(\boldsymbol{y})} = \boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{yx})^\top}, \tag{20}$$

$$\boldsymbol{\Omega}_{\boldsymbol{q}}^{(\boldsymbol{xx})} = \mathrm{diag}\left[\boldsymbol{\Omega}_{q_1}^{(\boldsymbol{xx})}, \ldots, \boldsymbol{\Omega}_{q_T}^{(\boldsymbol{xx})}\right], \tag{21}$$

$$\boldsymbol{\Omega}_{\boldsymbol{q}}^{(\boldsymbol{yy})} = \mathrm{diag}\left[\boldsymbol{\Omega}_{q_1}^{(\boldsymbol{yy})}, \ldots, \boldsymbol{\Omega}_{q_T}^{(\boldsymbol{yy})}\right], \tag{22}$$

$$\boldsymbol{\Omega}_{\boldsymbol{q}}^{(\boldsymbol{xy})} = \mathrm{diag}\left[\boldsymbol{\Omega}_{q_1}^{(\boldsymbol{xy})}, \ldots, \boldsymbol{\Omega}_{q_T}^{(\boldsymbol{xy})}\right] = \boldsymbol{\Omega}_{\boldsymbol{q}}^{(\boldsymbol{yx})^\top}, \tag{23}$$

and $\boldsymbol{W}^{(\boldsymbol{x})}$ and $\boldsymbol{W}^{(\boldsymbol{y})}$ are $3M_{\boldsymbol{x}}T \times M_{\boldsymbol{x}}T$ and $3M_{\boldsymbol{y}}T \times M_{\boldsymbol{y}}T$ window matrices, respectively. The mean vectors of Eq. (13), $\bar{\boldsymbol{x}}_{\boldsymbol{q}}$ and $\bar{\boldsymbol{y}}_{\boldsymbol{q}}$, are given as follows:

$$\bar{\boldsymbol{x}}_{\boldsymbol{q}} = \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{xx})} \left(\boldsymbol{r}_{\boldsymbol{q}}^{(\boldsymbol{x})} - \boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{xy})} \boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{yy})-1} \boldsymbol{r}_{\boldsymbol{q}}^{(\boldsymbol{y})}\right), \tag{24}$$

$$\bar{\boldsymbol{y}}_{\boldsymbol{q}} = \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yy})} \left(\boldsymbol{r}_{\boldsymbol{q}}^{(\boldsymbol{y})} - \boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{yx})} \boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{xx})-1} \boldsymbol{r}_{\boldsymbol{q}}^{(\boldsymbol{x})}\right), \tag{25}$$

where

$$\boldsymbol{r}_{\boldsymbol{q}}^{(\boldsymbol{x})} = \boldsymbol{W}^{(\boldsymbol{x})^\top} \left(\boldsymbol{\Omega}_{\boldsymbol{q}}^{(\boldsymbol{xx})} \boldsymbol{\mu}_{\boldsymbol{q}}^{(\boldsymbol{x})} + \boldsymbol{\Omega}_{\boldsymbol{q}}^{(\boldsymbol{xy})} \boldsymbol{\mu}_{\boldsymbol{q}}^{(\boldsymbol{y})}\right), \tag{26}$$

$$\boldsymbol{r}_{\boldsymbol{q}}^{(\boldsymbol{y})} = \boldsymbol{W}^{(\boldsymbol{y})^\top} \left(\boldsymbol{\Omega}_{\boldsymbol{q}}^{(\boldsymbol{yy})} \boldsymbol{\mu}_{\boldsymbol{q}}^{(\boldsymbol{y})} + \boldsymbol{\Omega}_{\boldsymbol{q}}^{(\boldsymbol{yx})} \boldsymbol{\mu}_{\boldsymbol{q}}^{(\boldsymbol{x})}\right), \tag{27}$$

$$\boldsymbol{\mu}_{\boldsymbol{q}}^{(\boldsymbol{x})} = \left[\boldsymbol{\mu}_{q_1}^{(\boldsymbol{x})^\top}, \ldots, \boldsymbol{\mu}_{q_T}^{(\boldsymbol{x})^\top}\right]^\top, \tag{28}$$

$$\boldsymbol{\mu}_{\boldsymbol{q}}^{(\boldsymbol{y})} = \left[\boldsymbol{\mu}_{q_1}^{(\boldsymbol{y})^\top}, \ldots, \boldsymbol{\mu}_{q_T}^{(\boldsymbol{y})^\top}\right]^\top. \tag{29}$$

As a result, the conditional PDF of $\boldsymbol{y}$ for given $\boldsymbol{x}$ and λ can be expressed as

$$P\left(\boldsymbol{y} \mid \boldsymbol{x}, \lambda\right) = \sum_{\forall \boldsymbol{q}} P\left(\boldsymbol{q} \mid \boldsymbol{x}, \lambda\right) P\left(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{q}, \lambda\right), \tag{30}$$

$$P\left(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{q}, \lambda\right) = \mathcal{N}\left(\boldsymbol{y}; \tilde{\boldsymbol{y}}_{\boldsymbol{q}}, \tilde{\boldsymbol{P}}_{\boldsymbol{q}}^{(\boldsymbol{yy})}\right), \tag{31}$$

$$\tilde{\boldsymbol{y}}_{\boldsymbol{q}} = \bar{\boldsymbol{y}}_{\boldsymbol{q}} + \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yx})} \boldsymbol{C}_{\boldsymbol{q}}^{(\boldsymbol{xx})} \left(\boldsymbol{x} - \bar{\boldsymbol{x}}_{\boldsymbol{q}}\right), \tag{32}$$

$$\tilde{\boldsymbol{P}}_{\boldsymbol{q}}^{(\boldsymbol{yy})} = \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yy})} - \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yx})} \boldsymbol{C}_{\boldsymbol{q}}^{(\boldsymbol{xx})} \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{xy})}. \tag{33}$$

3.2. MMSE-based mapping

The MMSE-based mapping [5] determines the estimated target static feature vector sequence $\hat{\boldsymbol{y}}$ as follows:

$$\hat{\boldsymbol{y}} = E\left[\boldsymbol{y} \mid \boldsymbol{x}\right] = \int P\left(\boldsymbol{y} \mid \boldsymbol{x}, \lambda\right) \boldsymbol{y}\, d\boldsymbol{y} \tag{34}$$

$$= \int \sum_{\forall \boldsymbol{q}} \gamma_{\boldsymbol{q}} \cdot P\left(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{q}, \lambda\right) \boldsymbol{y}\, d\boldsymbol{y} \tag{35}$$

$$= \sum_{\forall \boldsymbol{q}} \gamma_{\boldsymbol{q}} \cdot \tilde{\boldsymbol{y}}_{\boldsymbol{q}} \tag{36}$$

$$= \sum_{\forall \boldsymbol{q}} \gamma_{\boldsymbol{q}} \cdot \left(\boldsymbol{A}_{\boldsymbol{q}} \boldsymbol{x} + \boldsymbol{b}_{\boldsymbol{q}}\right), \tag{37}$$

where $E[\cdot]$ means expectation, $\gamma_{\boldsymbol{q}}$ is a posterior probability of $\boldsymbol{q}$ for the given $\boldsymbol{x}$, and $\boldsymbol{A}_{\boldsymbol{q}}$ and $\boldsymbol{b}_{\boldsymbol{q}}$ are given as follows:

$$\boldsymbol{A}_{\boldsymbol{q}} = \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yx})} \boldsymbol{C}_{\boldsymbol{q}}^{(\boldsymbol{xx})}, \tag{38}$$

$$\boldsymbol{b}_{\boldsymbol{q}} = \bar{\boldsymbol{y}}_{\boldsymbol{q}} - \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yx})} \boldsymbol{C}_{\boldsymbol{q}}^{(\boldsymbol{xx})} \bar{\boldsymbol{x}}_{\boldsymbol{q}}. \tag{39}$$

The estimated static feature vector sequence $\hat{\boldsymbol{y}}$ is defined as the weighted sum of the conditional mean vectors $\tilde{\boldsymbol{y}}_{\boldsymbol{q}}$, where the posterior probabilities of $\boldsymbol{q}$ are used as weights.

3.3. ML-based mapping

The ML-based mapping [4] determines the target parameter $\hat{\boldsymbol{y}}$ so as to maximize $P(\boldsymbol{y} \mid \boldsymbol{x}, \lambda)$ using the EM algorithm. Specifically, we maximize an auxiliary function of a current target static feature vector sequence $\boldsymbol{y}$ and an updated one $\hat{\boldsymbol{y}}$ defined by

$$\mathcal{Q}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sum_{\forall \boldsymbol{q}} P(\boldsymbol{q} \mid \boldsymbol{x}, \boldsymbol{y}, \lambda) \log P(\hat{\boldsymbol{y}}, \boldsymbol{q} \mid \boldsymbol{x}, \lambda). \quad (40)$$

The mapped target static feature vector sequence maximizing the auxiliary function is given by

$$\hat{\boldsymbol{y}} = \left(\sum_{\forall \boldsymbol{q}} \gamma'_{\boldsymbol{q}} \cdot \tilde{\boldsymbol{P}}_{\boldsymbol{q}}^{(\boldsymbol{yy})-1} \right)^{-1} \sum_{\forall \boldsymbol{q}} \gamma'_{\boldsymbol{q}} \cdot \tilde{\boldsymbol{P}}_{\boldsymbol{q}}^{(\boldsymbol{yy})-1} \tilde{\boldsymbol{y}}_{\boldsymbol{q}} \quad (41)$$

$$= \boldsymbol{A}' \boldsymbol{x} + \boldsymbol{b}', \quad (42)$$

where $\gamma'_{\boldsymbol{q}}$ is a posterior probability of $\boldsymbol{q}$ for given $\boldsymbol{x}$ and $\boldsymbol{y}$, and $\boldsymbol{A}'$ and $\boldsymbol{b}'$ are as follows:

$$\boldsymbol{A}' = \left(\sum_{\forall \boldsymbol{q}} \gamma'_{\boldsymbol{q}} \cdot \tilde{\boldsymbol{P}}_{\boldsymbol{q}}^{(\boldsymbol{yy})-1} \right)^{-1} \sum_{\forall \boldsymbol{q}} \gamma'_{\boldsymbol{q}} \cdot \tilde{\boldsymbol{P}}_{\boldsymbol{q}}^{(\boldsymbol{yy})-1} \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yx})} \boldsymbol{C}_{\boldsymbol{q}}^{(\boldsymbol{xx})} \quad (43)$$

$$\boldsymbol{b}' = \left(\sum_{\forall \boldsymbol{q}} \gamma'_{\boldsymbol{q}} \cdot \tilde{\boldsymbol{P}}_{\boldsymbol{q}}^{(\boldsymbol{yy})-1} \right)^{-1} \cdot \sum_{\forall \boldsymbol{q}} \gamma'_{\boldsymbol{q}} \cdot \tilde{\boldsymbol{P}}_{\boldsymbol{q}}^{(\boldsymbol{yy})-1} \left(\bar{\boldsymbol{y}}_{\boldsymbol{q}} - \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yx})} \boldsymbol{C}_{\boldsymbol{q}}^{(\boldsymbol{xx})} \bar{\boldsymbol{x}}_{\boldsymbol{q}} \right). \quad (44)$$

Note that under the Viterbi approximation with the same $\boldsymbol{q}$, $\hat{\boldsymbol{y}}$ by the ML-based mapping is identical to that by the MMSE-based one ($\hat{\boldsymbol{y}} = \tilde{\boldsymbol{y}}_{\boldsymbol{q}}$).

3.4. Discussions

The proposed algorithm incorporates the relationship between static and dynamic features explicitly in whole process: joint PDF training, conditional PDF estimation, and feature mapping. Therefore, we can solve the inconsistency between training and mapping. Zen et al. have reported that solving this inconsistency improved the quality of HMM-based speech synthesis [7]. It suggests that the proposed technique may also improve the quality of feature mapping.

The transformation matrices, $\boldsymbol{A}_{\boldsymbol{q}}$ and $\boldsymbol{A}'$, are generally full. Hence, the proposed algorithm offers an entire sequence-level transformation rather than the frame-by-frame conversion.

The proposed algorithm is similar to the articulatory-acoustic modeling algorithm proposed by Le et al. [8]. This algorithm models source and target features using two independent data streams:[3] interactions between source and target feature vector sequences are represented implicitly by the posterior probability of $\boldsymbol{q}$. On the other hand, the proposed algorithm models these interactions explicitly via the state output PDFs.

[3] In this case, the estimated target static vector sequence is determined as $\hat{\boldsymbol{y}} = \boldsymbol{P}_{\boldsymbol{q}}^{(\boldsymbol{yy})} \boldsymbol{r}_{\boldsymbol{q}}^{(\boldsymbol{y})}$ because $\boldsymbol{R}_{\boldsymbol{q}}^{(\boldsymbol{yx})}$ becomes a zero matrix.

4. Experiment

4.1. Experimental conditions

To evaluate the performance of the proposed algorithm, voice conversion experiments on the CMU ARCTIC database were performed. The first 593 sentences uttered by female speakers CLB and SLT were used for training (CLB: source, SLT: target). The last 40 sentences, which were not included in the training data, were used for evaluation. Speech signals were sampled at a rate of 16 kHz and windowed by a 25 ms Blackman window with a 5 ms shift, and then 24-th order mel-cepstral coefficients were obtained by a mel-cepstral analysis technique [9]. For each utterance, $\log F_0$ values were also extracted by ESPS get_f0. Static feature vectors of source and target speech consisted of 25 mel-cepstral coefficients including the zeroth coefficient. To obtain the joint vector sequences, we performed dynamic time warping (DTW) between the source and target vector sequences using an Euclidean norm including both static and dynamic features. In this paper, we calculated dynamic features as $\Delta \boldsymbol{x}_t = 0.5(\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1})$, $\Delta^2 \boldsymbol{x}_t = \boldsymbol{x}_{t+1} - 2\boldsymbol{x}_t + \boldsymbol{x}_{t-1}$. Unlike the common DTW method, we needed look-ahead [7] dynamic programming (DP) because dynamic features at each frame depended on the static features of both preceding and succeeding one frames. Using DTW, each target vector sequence was aligned to the corresponding source one. Therefore, the number of frames of the joint vector sequence was equal to that of the source one. Each joint vector was augmented with its dynamic features (150 dimensions in total).

The three-state left-to-right structure with no-skip was used for HMMs and trajectory HMMs to model 41 phonemes including silence (123 states in total). Each state had a single Gaussian distribution for its state output PDF. To compare the performance of each algorithm with the close number of model parameters, we trained a GMM and a trajectory GMM with 128 Gaussian components. The diagonal structure was used for $\{\boldsymbol{\Omega}_i^{(\boldsymbol{xx})}, \boldsymbol{\Omega}_i^{(\boldsymbol{xy})}, \boldsymbol{\Omega}_i^{(\boldsymbol{yx})}, \boldsymbol{\Omega}_i^{(\boldsymbol{yy})}\}$. After training the GMM and HMMs in the standard way based on the ML criterion, the trajectory GMM and trajectory HMMs were re-estimated by the Viterbi training [7] using the GMM and HMMs as their initial models, respectively. Here, we optimized Cholesky factors of $\{\boldsymbol{\Omega}_i\}$ rather than $\{\boldsymbol{\Omega}_i\}$ themselves because we can easily keep these matrices positive definite. The limited memory BFGS method was used for optimization.

For mapping, we used the Viterbi approximation and did not iterate the EM algorithm.[4] In this experiment, $\log F_0$ values were converted by a simple linear conversion used in [4]. To avoid the effect of recognition errors, correct transcriptions were given for HMM and trajectory HMM-based mapping. From the converted mel-cepstral coefficients and $\log F_0$ values, a speech waveform was synthesized through the MLSA filter [9, 10] with pulse or noise excitation.

4.2. Experimental results

The mel-cepstral distortion [4] between the target and converted mel-cepstral coefficients in the evaluation set was used as the objective evaluation measure. In [4], the GMM-based mapping algorithm with dynamic features [4] has been shown to obtain superior performance than the GMM-based mapping algorithm without dynamic features [5] on the MOCHA database.

[4] It was found in preliminary experiments that more EM iterations and avoiding Viterbi approximation improved the likelihoods, as measured by the mapping models, but did not improve the quality of mapped speech.

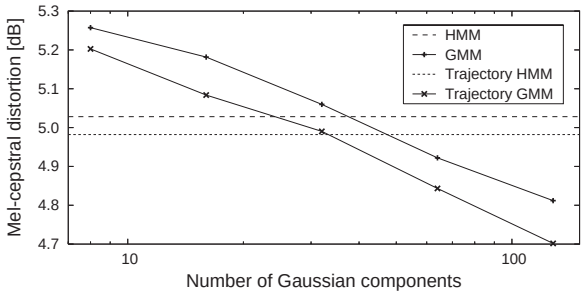

Figure 1: Average mel-cepstral distortions between the target and converted mel-cepstral coefficients in the evaluation set. Without conversion, average mel-cepstral distortion was 6.204.

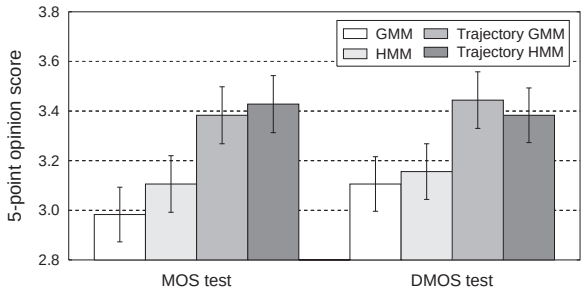

Figure 2: Mean opinion scores and differential mean opinion scores of converted speech by the conventional and proposed algorithms. Error bars show 95% confidence intervals.

Therefore, we focus on the performance comparison between the proposed trajectory GMM or trajectory HMM-based mapping algorithm and the GMM or HMM-based mapping with dynamic features in [4]. Figure 1 plots the mel-cepstral distortions. It can be seen from the figure that the proposed algorithm constantly outperformed the conventional one. This is because the inconsistency between training and mapping was solved in the proposed algorithm. Compared with the HMM and trajectory HMM-based mapping, the GMM and trajectory GMM-based ones achieved better objective scores, respectively. We expect that the use of single Gaussian distributions for state output PDFs of the HMMs and trajectory HMMs caused this phenomenon.

Two subjective listening tests were conducted to evaluate the speech quality and speaker similarity. The first test compared the naturalness of converted speech by the mean opinion score (MOS) test method. The second one evaluated the speaker similarity between the analysis–synthesized target speech and converted speech by the differential MOS (DMOS) test method. The subjects were 12 Japanese graduate students, all of whom completed both tests. Fifteen sentences were randomly chosen from the evaluation sentences for each test of each subject. Samples were presented in a random order for each test sentence. Before starting the test, the subjects listened to speech samples of one sentence to become familiar with the task. This sentence was randomly chosen for each subject and excluded from the actual test. In the MOS test, after listening to each test sample, the subjects were asked to assign it a five-point naturalness score (5: natural – 1: poor). In the DMOS test, after listening to the analysis–synthesized target speech and each test sample, the subjects were asked to assign it a five-point similarity score (5: very similar – 1: very far). Both experiments were carried out in a soundproof room using headphones.

Figure 2 plots the experimental results. It can be seen from the figures that the proposed algorithm achieved better subjective scores than the conventional ones in both the MOS and DMOS tests.

5. Conclusions

This paper proposed the probabilistic feature mapping algorithm based on the trajectory GMM or trajectory HMM. This algorithm can solve the inconsistency of the conventional GMM or HMM-based mapping algorithm using dynamic features, and offers the entire sequence-level transformation rather than the frame-by-frame conversion. Experimental results in voice conversion showed that the proposed algorithm outperformed the conventional GMM or HMM-based one both in the objective and subjective tests.

Future work includes improving the quality of converted speech using global variance [4], investigating the proposed algorithms on other applications, and integrating the DTW procedure into the model definition like DPGMM [11].

6. Acknowledgments

The authors would like to thank Yosuke Uto and Dr. Tomoki Toda for their helpful comments. This work was partly supported by the MEXT e-Society project, the Hori information science promotion foundation, and the Grant-in-Aid for Scientific Research (No. 1880009) of JSPS.

7. References

[1] Y. Stylianou, O. Capṕe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. SAP*, vol. 6, no. 2, pp. 131–142, 1998.

[2] S. Hiroya and M. Honda, “Estimation of articulatory movements from speech acoustics using an HMM-based speech production model,” *IEEE Trans. SAP*, vol. 12, no. 2, pp. 175–185, 2004.

[3] X.-D. Cui, M. Afify, and Y. Gao, “MMSE-based stereo feature stochastic mapping for noise robust speech recognition,” in *Proc. ICASSP*, 2008, pp. 4077–4080.

[4] T. Toda, A. Black, and K. Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” *IEEE Trans. ASLP*, vol. 15, no. 8, pp. 2222–2235, 2007.

[5] A. Kain and M. Macon, “Spectral voice conversion for text-to-speech synthesis,” in *Proc. ICASSP*, 1998, pp. 285–288.

[6] K. Richmond, “Trajectory mixture density network with multiple mixtures for acoustic-articulatory inversion,” in *Proc. NOLISP*, 2007, pp. 67–70.

[7] H. Zen, K. Tokuda, and T. Kitamura, “Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic features,” *Computer Speech & Language*, vol. 21, no. 1, pp. 153–173, 2007.

[8] L. Zhang and S. Renals, “Acoustic-articulatory modelling with the trajectory HMM,” *IEEE Signal Processing Letters*, vol. 15, pp. 245–248, 2008.

[9] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” in *Proc. ICASSP*, 1992, pp. 137–140.

[10] S. Imai, “Cepstral analysis synthesis on the mel frequency scale,” in *Proc. ICASSP*, 1983, pp. 93–96.

[11] Y. Nankaku, K. Nakamura, T. Toda, and K. Tokuda, “Spectral conversion based on statistical models including time-frequency matching,” in *Proc. ISCA SSW6*, 2007, pp. 333–338.