社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2008－SLP－74 (14)
2008 / 12 / 10

時系列マッチングを含む統計モデルを用いた継続長およびスペクトルの同時変換

油谷 かおり† 南角 吉彦† 戸田 智基†† 徳田 恵一†

† 名古屋工業大学大学院 工学研究科 創成シミュレーション工学専攻
〒 466–8555 名古屋市 昭和区 御器所町
†† 奈良先端科学技術大学院大学 情報科学研究科 〒 630–0192 奈良県生駒市高山町 8916–5

あらまし 本稿では時系列マッチングを含む統計モデルに基づいた継続長およびスペクトルの同時変換手法を提案する．これまで声質変換の主な手法としては，ガウス混合モデル (GMM) に基づくスペクトル変換法が用いられてきた．この手法では，元話者・目標話者間のフレームの対応を一対一と仮定しているため，発話速度を考慮したスペクトル変換を行うことができない．しかし，話者性は発話速度にも表れると考えられる．そこで本研究では継続長変換を行うため，時系列マッチングを含む統計モデル (DPGMM) を適用する．DPGMM は長さの異なる 2 つの系列を直接表現するため，継続長およびスペクトルの同時変換が可能となる．提案法では，DPGMM の各混合要素に継続長モデルを付加し，非線形かつスペクトル情報に依存した継続長変換を行う．

キーワード 声質変換，継続長変換，GMM

Simultaneous Transformation of Duration and Spectrum Using Statistical Models Including Time-Sequence Matching

Kaori YUTANI†, Yoshihiko NANKAKU†, Tomoki TODA††, and Keiichi TOKUDA†

† Department of Computer Science and Engineering, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya, 466–8555 Japan
†† Graduate School of Information Science, Nara Institute of Technology
8916–5, Takayama-cho, Ikoma-shi, Nara, 630–0192 Japan

Abstract This paper describes a simultaneous conversion technique of duration and spectrum based on a statistical model including time-sequence matching. The conventional GMM-based approach cannot perform spectral conversion taking account of speaking rates because it assumes one to one frame matching between source and target features. However, speaker characteristics may also appear in speaking rates. In order to perform duration conversion, we attach duration models to statistical models including time-sequence matching (DPGMM). Since DPGMM can represent two different length sequences directly, the conversion of spectrum and duration can be performed within an integrated framework. In the proposed technique, each mixture component of DPGMM has different duration transformation functions, therefore durations are converted nonlinearly and dependently on spectral information. In a subjective DMOS test, the proposed method is superior to the conventional method.

Key words Voice conversion, Duration conversion, GMM

1. Introduction

Voice conversion is a technique for converting a certain speaker's voice into another speaker's voice. It can modify speech characteristics using conversion rules statistically extracted from a small amount of data [1]. One of typical spectral conversion frameworks is based on a Gaussian Mixture Model (GMM) [2]. This method realizes the continuous mapping based on the soft clustering. A more accurate formulation of spectral conversion based on ML (Maximum Likelihood) criterion has been presented [3]. The ML-based conversion is a sophisticated technique because all processes

in the algorithm are derived based on the single objective function.

In the conventional GMM-based method, a GMM is trained under an assumption that source and target feature sequences have the same length, because a GMM is trained using joint feature vectors which are references of mapping rules, and the Dynamic Programming (DP) matching between source and target feature sequences is conducted prior to the training of GMMs. Because of this, it cannot take account of the correlation of duration between source and target features. To overcome this problem, we apply statistical models including time-sequence matching (DPGMM) [4]. The likelihood function of this model can directly deal with two different length sequences, in which a frame alignment between two sequences is represented by discrete hidden variables. It can perform modeling of duration correlations between source and target features. In the proposed voice conversion technique, we can convert a speaking rate nonlinearly and dependently on spectral information by attaching duration models to each mixture of DPGMM.

The paper is organized as follows. Section 2 and 3 explain the conventional voice conversion technique based on GMM and DPGMM, respectively. Simultaneous conversion of duration and spectrum is presented in Section 4 and experimental results are reported in Section 5. Finally, conclusions and future works are given in Section 6.

2. Spectral Conversion Based on GMM

To convert spectral feature sequences of a source speaker to that of a target speaker, the joint probability of two features are modeled by a GMM [3]. Let a vector $\boldsymbol{O}_t = \left[\boldsymbol{O}_t^{(1)\top}, \boldsymbol{O}_t^{(2)\top}\right]^\top$ be a joint feature vector of the source one $\boldsymbol{O}_t^{(1)}$ and the target one $\boldsymbol{O}_t^{(2)}$ at time t, where $\cdot^\top$ denotes transposition of a vector. An alignment between two feature sequences is obtained by the Dynamic Programing (DP) matching. In the GMM-based voice conversion, the joint feature vector sequence $\boldsymbol{O} = \left[\boldsymbol{O}_1{}^\top, \boldsymbol{O}_2{}^\top, \cdots, \boldsymbol{O}_T{}^\top\right]^\top$ is modeled by a GMM to learn the relation between source and target features. The output probability of $\boldsymbol{O}$ given GMM $\boldsymbol{\lambda}$ is defined as follows:

$$P(\boldsymbol{O} \mid \boldsymbol{\lambda}) = \prod_{t=1}^{T} \left[\sum_{i=1}^{M} w_i \mathcal{N}(\boldsymbol{O}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\right] \tag{1}$$

where

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^{(1)} \\ \boldsymbol{\mu}_i^{(2)} \end{bmatrix}, \boldsymbol{\Sigma}_i = \begin{bmatrix} \boldsymbol{\Sigma}_i^{(1,1)} & \boldsymbol{\Sigma}_i^{(1,2)} \\ \boldsymbol{\Sigma}_i^{(2,1)} & \boldsymbol{\Sigma}_i^{(2,2)} \end{bmatrix} \tag{2}$$

and M means the number of mixtures, $w_i = P(i \mid \boldsymbol{\lambda})$ is the mixture weight of the i-th component, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean vector and covariance matrix, respectively. These model parameters are estimated via the Expectation Maximization (EM) algorithm.

2.1 Maximum Likelihood Spectral Conversion

In the ML spectral conversion, the optimal converted feature sequence $\boldsymbol{O}^{(2)} = \left[\boldsymbol{O}_1^{(2)\top}, \boldsymbol{O}_2^{(2)\top}, \cdots, \boldsymbol{O}_T^{(2)\top}\right]^\top$ given a source feature sequence $\boldsymbol{O}^{(1)} = \left[\boldsymbol{O}_1^{(1)\top}, \boldsymbol{O}_2^{(1)\top}, \cdots, \boldsymbol{O}_T^{(1)\top}\right]^\top$ is obtained by maximizing the following conditional distribution:

$$\begin{aligned} &P(\boldsymbol{O}^{(2)} \mid \boldsymbol{O}^{(1)}, \boldsymbol{\lambda}) \\ &= \sum_{\boldsymbol{m}} \left[P(\boldsymbol{m} \mid \boldsymbol{O}^{(1)}, \boldsymbol{\lambda}) \prod_{t=1}^{T} P(\boldsymbol{O}_t^{(2)} \mid \boldsymbol{O}_t^{(1)}, m_t, \boldsymbol{\lambda})\right] \end{aligned} \tag{3}$$

where $\boldsymbol{m} = [m_1, m_2, \cdots, \cdots, m_T]$ is a mixture index sequence. The conditional distribution can also be written as GMM, and its output probability distribution is presented as follows:

$$P(\boldsymbol{O}_t^{(2)} \mid \boldsymbol{O}_t^{(1)}, m_t = i, \boldsymbol{\lambda}) = \mathcal{N}\left(\boldsymbol{O}_t^{(2)}; \boldsymbol{E}_i(t), \boldsymbol{D}_i\right) \tag{4}$$

where

$$\boldsymbol{E}_i(t) = \boldsymbol{\mu}_i^{(2)} + \boldsymbol{\Sigma}_i^{(2,1)} \boldsymbol{\Sigma}_i^{(1,1)-1} \left(\boldsymbol{O}_t^{(1)} - \boldsymbol{\mu}_i^{(1)}\right) \tag{5}$$

$$\boldsymbol{D}_i = \boldsymbol{\Sigma}_i^{(2,2)} - \boldsymbol{\Sigma}_i^{(2,1)} \boldsymbol{\Sigma}_i^{(1,1)-1} \boldsymbol{\Sigma}_i^{(1,2)} \tag{6}$$

Since equation (3) includes latent variables, the optimal sequence of $\boldsymbol{O}^{(2)}$ is estimated via the EM algorithm. The EM algorithm is an iterative method for approximating the maximum likelihood estimation. It maximizes the expectation of the complete data log-likelihood so called $\mathcal{Q}$-function (auxiliary function):

$$\begin{aligned} &\mathcal{Q}(\boldsymbol{O}^{(2)}, \hat{\boldsymbol{O}}^{(2)}) \\ &= \sum_{\boldsymbol{m}} \left[P(\boldsymbol{O}^{(2)}, \boldsymbol{m} \mid \boldsymbol{O}^{(1)}, \boldsymbol{\lambda}) \ln P(\hat{\boldsymbol{O}}^{(2)}, \boldsymbol{m} \mid \boldsymbol{O}^{(1)}, \boldsymbol{\lambda})\right] \end{aligned} \tag{7}$$

Taking the derivative of the $\mathcal{Q}$-function, the spectral sequence $\hat{\boldsymbol{O}}^{(2)}$ which maximizes the $\mathcal{Q}$-function is given by

$$\hat{\boldsymbol{O}}^{(2)} = \left(\overline{\boldsymbol{D}^{-1}}\right)^{-1} \overline{\boldsymbol{D}^{-1}\boldsymbol{E}} \tag{8}$$

where

$$\overline{\boldsymbol{D}^{-1}} = \mathrm{diag}\left[\overline{\boldsymbol{D}_1^{-1}}, \overline{\boldsymbol{D}_2^{-1}}, \cdots, \overline{\boldsymbol{D}_T^{-1}}\right] \tag{9}$$

$$\overline{\boldsymbol{D}_t^{-1}} = \sum_{i=1}^{M} \gamma_t(i) \boldsymbol{D}_i^{-1} \tag{10}$$

$$\overline{\boldsymbol{D}^{-1}\boldsymbol{E}} = \left[\overline{\boldsymbol{D}^{-1}\boldsymbol{E}_1}^\top, \overline{\boldsymbol{D}^{-1}\boldsymbol{E}_2}^\top, \cdots, \overline{\boldsymbol{D}^{-1}\boldsymbol{E}_T}^\top\right]^\top \tag{11}$$

$$\overline{\boldsymbol{D}^{-1}\boldsymbol{E}_t} = \sum_{i=1}^{M} \gamma_t(i) \boldsymbol{D}_i^{-1} \boldsymbol{E}_t(i) \tag{12}$$

$$\gamma_t(i) = P(m_t = i \mid \boldsymbol{O}_t^{(1)}, \boldsymbol{O}_t^{(2)}, \boldsymbol{\lambda}) \tag{13}$$

3. Spectral Conversion Based on DPGMM

3.1 Definition of Model Structure

In the DPGMM-based method [4], we define the likelihood function $P(\boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)} \mid \boldsymbol{\lambda})$ including the structure of sequence matching. The simultaneous optimization of the DP matching and training of model parameters is performed based on the ML criterion. The advantage of the DPGMM is to directly represent two different length sequences $\boldsymbol{O}^{(1)} = \left[\boldsymbol{O}_1^{(1)}, \boldsymbol{O}_2^{(1)}, \cdots, \boldsymbol{O}_{T^{(1)}}^{(1)}\right]^\top$ and $\boldsymbol{O}^{(2)} = \left[\boldsymbol{O}_1^{(2)}, \boldsymbol{O}_2^{(2)}, \cdots, \boldsymbol{O}_{T^{(2)}}^{(2)}\right]^\top$. The likelihood function of observation sequences $\boldsymbol{O} = \left\{\boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)}\right\}$ is written as follows:

$$P(\boldsymbol{O} \mid \boldsymbol{\lambda}) = \sum_{\boldsymbol{m},\boldsymbol{a}} \left[P(\boldsymbol{m} \mid \boldsymbol{\lambda}) P(\boldsymbol{O}^{(1)} \mid \boldsymbol{m}, \boldsymbol{\lambda}) \times P(\boldsymbol{a} \mid \boldsymbol{m}, \boldsymbol{\lambda}) P(\boldsymbol{O}^{(2)} \mid \boldsymbol{O}^{(1)}, \boldsymbol{m}, \boldsymbol{a}, \boldsymbol{\lambda})\right] \quad (14)$$

where $\boldsymbol{m} = [m_1, m_2, \cdots, m_{T^{(1)}}]$ is a mixture index sequence and its element $m_{t^{(1)}}$ means the mixture index of the observation $\boldsymbol{O}^{(1)}$ at time $t^{(1)}$. The variable $\boldsymbol{a} = [a_1, a_2, \cdots, a_{T^{(2)}}]$ represents the temporal matching between source and target feature sequences and $a_{t^{(2)}} \in \left\{1, \cdots, T^{(1)}\right\}$ indicates the frame number of source sequence $\boldsymbol{O}^{(1)}$ which corresponds to the $t^{(2)}$-th frame of target sequence $\boldsymbol{O}^{(2)}$. Each element of the complete data likelihood is defined as follows:

$$P(\boldsymbol{m} \mid \boldsymbol{\lambda}) = \prod_{t^{(1)}=1}^{T^{(1)}} P(m_{t^{(1)}} \mid \boldsymbol{\lambda}) \quad (15)$$

$$P(\boldsymbol{O}^{(1)} \mid \boldsymbol{m}, \boldsymbol{\lambda}) = \prod_{t^{(1)}=1}^{T^{(1)}} \mathcal{N}\left(\boldsymbol{O}_{t^{(1)}}^{(1)}; \boldsymbol{\mu}_{m_{t^{(1)}}}^{(1)}, \boldsymbol{\Sigma}_{m_{t^{(1)}}}^{(1)}\right) \quad (16)$$

$$P(\boldsymbol{a} \mid \boldsymbol{m}, \boldsymbol{\lambda}) = \prod_{t^{(2)}=1}^{T^{(2)}} P(a_{t^{(2)}} \mid a_{t^{(2)}-1}, m_{t^{(2)}-1} \boldsymbol{\lambda}) \quad (17)$$

$$P(\boldsymbol{O}^{(2)} \mid \boldsymbol{O}^{(1)}, \boldsymbol{m}, \boldsymbol{a}, \boldsymbol{\lambda}) = \prod_{t^{(2)}=1}^{T^{(2)}} \mathcal{N}\left(\boldsymbol{O}_{t^{(2)}}^{(2)}; \bar{\boldsymbol{C}}_{m_{a_{t^{(2)}}}} \bar{\boldsymbol{O}}_{a_{t^{(2)}}}^{(1)}, \bar{\boldsymbol{\Sigma}}_{m_{a_{t^{(2)}}}}\right) \quad (18)$$

where

$$\bar{\boldsymbol{C}}_i = \left[\begin{array}{cc} \bar{\boldsymbol{\mu}}_i & \boldsymbol{C}_i \end{array}\right], \bar{\boldsymbol{O}}_{t^{(1)}}^{(1)} = \left[\begin{array}{cc} 1 & \boldsymbol{O}_{t^{(1)}}^{(1)\top} \end{array}\right]^\top \quad (19)$$

The model parameters of DPGMM are summarized as follows:

- $\boldsymbol{w} = \{w_i \mid 1 \leqq i \leqq M\}$: the mixture weights of the GMM which generate the source feature sequence $\boldsymbol{O}^{(1)}$, where $w_i = P(m_{t^{(1)}} = i \mid \boldsymbol{\lambda})$ is the probability of i-th mixture.
- $\boldsymbol{B}^{(1)} = \left\{b_i^{(1)} \mid 1 \leqq i \leqq M\right\}$: the output probability distributions of source feature $\boldsymbol{O}^{(1)}$, where $b_i^{(1)} = P(\boldsymbol{O}_{t^{(1)}}^{(1)} \mid m_{t^{(1)}} = i, \boldsymbol{\lambda})$ is the probability of source feature vector $\boldsymbol{O}_{t^{(1)}}^{(1)}$ at i-th mixture and which is assumed to be a Gaussian distribution: $\mathcal{N}\left(\boldsymbol{O}_{t^{(1)}}^{(1)}; \boldsymbol{\mu}_i^{(1)}, \boldsymbol{\Sigma}_i^{(1)}\right)$ where $\boldsymbol{\mu}_i^{(1)}$ and $\boldsymbol{\Sigma}_i^{(1)}$ are the mean vector and covariance matrix, respectively.
- $\boldsymbol{c} = \left\{c_n^{(i)} \mid 1 \leqq n \leqq N\right\}$: the transition probabilities of the sequence matching where $c_n^{(i)}$ indicates the probability $P(a_{t^{(2)}} = a_{t^{(2)}-1} + n \mid a_{t^{(2)}-1}, m_{t^{(2)}-1} = i)$. This parameter corresponds to the cost function in the DP matching.
- $\boldsymbol{B}^{(2)} = \left\{b_i^{(2)} \mid 1 \leqq i \leqq M\right\}$: the output distributions of the target feature $\boldsymbol{O}^{(2)}$, where $b_i^{(2)} = P(\boldsymbol{O}_{t^{(2)}}^{(2)} \mid \boldsymbol{O}_{t^{(1)}}^{(1)}, m_{t^{(1)}} = i, a_{t^{(2)}} = t^{(1)})$ is the probability of the target feature vector $\boldsymbol{O}_{t^{(2)}}^{(2)}$ given the corresponding source feature vector $\boldsymbol{O}_{t^{(1)}}^{(1)}$ at i-th mixture. This conditional distribution is assumed to be a Gaussian distribution: $\mathcal{N}\left(\boldsymbol{O}_{t^{(2)}}^{(2)}; \boldsymbol{C}_i \boldsymbol{O}_{t^{(1)}}^{(1)} + \bar{\boldsymbol{\mu}}_i, \bar{\boldsymbol{\Sigma}}_i\right)$ where $\bar{\boldsymbol{\mu}}_i$ and $\bar{\boldsymbol{\Sigma}}_i$ are the mean vector and the covariance matrix, respectively.

Using shorthand notation, the model is defined as $\boldsymbol{\lambda} = \left\{\boldsymbol{w}, \boldsymbol{c}, \boldsymbol{B}^{(1)}, \boldsymbol{B}^{(2)}\right\}$. Figure1 shows the model structure including time-sequence matching. The generative procedure is summarized as follows:

(1) The mixture index sequence $\boldsymbol{m}$ is determined according to the weight $P(\boldsymbol{m} \mid \boldsymbol{\lambda})$.

(2) The source feature sequence $\boldsymbol{O}^{(1)}$ is generated from Gaussian distribution $P(\boldsymbol{O}^{(1)} \mid \boldsymbol{m}, \boldsymbol{\lambda})$.

(3) The frame matching between $\boldsymbol{O}^{(1)}$ and $\boldsymbol{O}^{(2)}$ is determined according to $P(\boldsymbol{a} \mid \boldsymbol{\lambda})$.

(4) The target feature sequence $\boldsymbol{O}^{(2)}$ is generated according to the conditional Gaussian distribution $P(\boldsymbol{O}^{(2)} \mid \boldsymbol{O}^{(1)}, \boldsymbol{m}, \boldsymbol{a}, \boldsymbol{\lambda})$ given the source feature sequence.

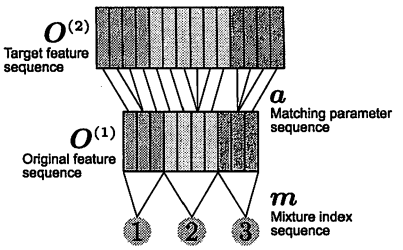


Fig. 1 Model structure including time-sequence matching

3.2 Training Algorithm

The parameters of DPGMM can be estimated via the EM algorithm. The $\mathcal{Q}$-function of DPGMM can be written as

$$\mathcal{Q}(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \sum_{\boldsymbol{m},\boldsymbol{a}} P(\boldsymbol{m}, \boldsymbol{a} \mid \boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)}, \boldsymbol{\lambda}) \times \ln P(\boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)}, \boldsymbol{m}, \boldsymbol{a} \mid \boldsymbol{\lambda}') \quad (20)$$

By maximizing the $\mathcal{Q}$-function, the re-estimation formula in the M-step are derived as follows:

$$w_i = \frac{1}{T^{(1)}} \sum_{t^{(1)}} \gamma_{t^{(1)}}^{(1)}(i) \quad (21)$$

$$\boldsymbol{\mu}_i^{(1)} = \frac{1}{N_i^{(1)}} \sum_{t^{(1)}} \gamma_{t^{(1)}}^{(1)}(i) \boldsymbol{O}_{t^{(1)}}^{(1)} \quad (22)$$

$$\boldsymbol{\Sigma}_i^{(1)} = \frac{1}{N_i^{(1)}} \sum_{t^{(1)}} \gamma_{t^{(1)}}^{(1)}(i) \times \left(\boldsymbol{O}_{t^{(1)}}^{(1)} - \boldsymbol{\mu}_i^{(1)}\right)\left(\boldsymbol{O}_{t^{(1)}}^{(1)} - \boldsymbol{\mu}_i^{(1)}\right)^\top \quad (23)$$

$$c_n^{(i)} = \frac{1}{N_i^{(2)}} \sum_{t^{(2)}} \sum_{t^{(1)}} \xi_{t^{(2)}}^{(2)}(t^{(1)}, n) \quad (24)$$

$$\bar{\boldsymbol{C}}_i = \left(\sum_{t^{(2)}} \sum_{t^{(1)}} \gamma_{t^{(2)}}^{(2)}(t^{(1)}, i) \boldsymbol{O}_{t^{(2)}}^{(2)} \bar{\boldsymbol{O}}_{t^{(1)}}^{(1)\top} \right) \times \left(\sum_{t^{(2)}} \sum_{t^{(1)}} \gamma_{t^{(2)}}^{(2)}(t^{(1)}, i) \bar{\boldsymbol{O}}_{t^{(1)}}^{(1)} \bar{\boldsymbol{O}}_{t^{(1)}}^{(1)\top} \right)^{-1} \quad (25)$$

$$\boldsymbol{\Sigma}_i^{(2)} = \frac{1}{N_i^{(2)}} \sum_{t^{(2)}} \sum_{t^{(1)}} \gamma_{t^{(2)}}^{(2)}(t^{(1)}, i) \times \left(\boldsymbol{O}_{t^{(2)}}^{(2)} - \bar{\boldsymbol{C}}_i \bar{\boldsymbol{O}}_{t^{(1)}}^{(1)}\right)\left(\boldsymbol{O}_{t^{(2)}}^{(2)} - \bar{\boldsymbol{C}}_i \bar{\boldsymbol{O}}_{t^{(1)}}^{(1)}\right)^\top \quad (26)$$

where $T^{(1)}$ means the total number of frames of source feature sequences, and $N_i^{(1)}$ and $N_i^{(2)}$ are the occupancy counts of i-th mixture which can be written as follows:

$$N_i^{(1)} = \sum_{t^{(1)}} \gamma_{t^{(1)}}^{(1)}(i), \quad N_i^{(2)} = \sum_{t^{(2)}} \sum_{t^{(1)}} \gamma_{t^{(2)}}^{(2)}(t^{(1)}, i) \quad (27)$$

Notation γ and ξ denote the expectations with respect to the posterior distribution over the hidden variables. These expectations are computed in the E-step by the following equations:

$$\gamma_{t^{(1)}}^{(1)}(i) = P(m_{t^{(1)}} = i \,|\, \boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)}, \boldsymbol{\lambda}) = \sum_{\boldsymbol{m}, \boldsymbol{a}} P(\boldsymbol{m}, \boldsymbol{a} \,|\, \boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)}, \boldsymbol{\lambda}) \delta(m_{t^{(1)}}, i) \quad (28)$$

$$\gamma_{t^{(2)}}^{(2)}(t^{(1)}, i) = P(a_{t^{(2)}} = t^{(1)} \,|\, \boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)}, \boldsymbol{\lambda}) = \sum_{\boldsymbol{m}, \boldsymbol{a}} P(\boldsymbol{m}, \boldsymbol{a} \,|\, \boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)}, \boldsymbol{\lambda}) \times \delta(m_{t^{(1)}, i}) \delta(a_{t^{(2)}}, t^{(1)}) \quad (29)$$

$$\xi_{t^{(2)}}^{(2)}(t^{(1)}, n) = P(a_{t^{(2)}-1} = t^{(1)}, a_{t^{(2)}} = t^{(1)} + n \,|\, \boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)}, \boldsymbol{\lambda}) = \sum_{\boldsymbol{m}, \boldsymbol{a}} P(\boldsymbol{m}, \boldsymbol{a} \,|\, \boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)}, \boldsymbol{\lambda}) \times \delta(a_{t^{(2)}-1}, t^{(1)}) \delta(a_{t^{(2)}}, t^{(1)} + n) \quad (30)$$

$\delta(\cdot)$ is the Kronecker delta function: $\delta(u, v) = 1$ if $u = v$, $\delta(u, v) = 0$ otherwise. If we compute expectations in the exact E-step directly according to equation (28)–(30), we need to take the summation over all the combinations of $\boldsymbol{m}$ and $\boldsymbol{a}$. Therefore the complexity of the E-step becomes $O(M^{T^{(1)}} T^{(2)T^{(1)}})$ and it is infeasible due to the number of hidden variables. To overcome this problem, the variational method is used for approximate ML estimation in probabilistic graphical models with hidden variables [5].

3.3 Variational approximation

Variational methods [5] have been used for approximate maximum likelihood estimation of probabilistic graphical models. We present a structure approximation in which the hidden variables representing mixture index sequences and time sequence matching are decoupled. An arbitrary distribution $\mathcal{Q}(\boldsymbol{m}, \boldsymbol{a})$ over the hidden variables is introduced to define a lower bound on the log-likelihood:

$$\begin{aligned} \ln P(\boldsymbol{O} \,|\, \boldsymbol{\lambda}) &= \ln \sum_{\boldsymbol{m}, \boldsymbol{a}} \mathcal{Q}(\boldsymbol{m}, \boldsymbol{a}) \frac{P(\boldsymbol{O}, \boldsymbol{m}, \boldsymbol{a} \,|\, \boldsymbol{\lambda})}{\mathcal{Q}(\boldsymbol{m}, \boldsymbol{a})} \\ &\geqq \sum_{\boldsymbol{m}, \boldsymbol{a}} \mathcal{Q}(\boldsymbol{m}, \boldsymbol{a}) \ln \frac{P(\boldsymbol{O}, \boldsymbol{m}, \boldsymbol{a} \,|\, \boldsymbol{\lambda})}{\mathcal{Q}(\boldsymbol{m}, \boldsymbol{a})} \\ &= \mathcal{F}(\mathcal{Q}, \boldsymbol{\lambda}) \end{aligned} \quad (31)$$

where we have applied Jensen's inequality.

3.4 ML-Based Spectral Conversion

In the conversion process, the converted feature sequence $\boldsymbol{O}^{(2)}$ can be obtained by maximizing a lower bound of the likelihood. The optimal sequence is given as the following equation:

$$\hat{\boldsymbol{O}}_{t^{(2)}}^{(2)} = \left(\sum_{t^{(1)}=1}^{T^{(1)}} \sum_{i=1}^{M} \gamma_{t^{(1)}}^{(1)}(i) \delta(t^{(1)}, t^{(2)}) \bar{\boldsymbol{\Sigma}}_i^{-1} \right)^{-1} \times \left(\sum_{t^{(1)}=1}^{T^{(1)}} \sum_{i=1}^{M} \gamma_{t^{(1)}}^{(1)}(i) \delta(t^{(1)}, t^{(2)}) \bar{\boldsymbol{\Sigma}}_i^{-1} \bar{\boldsymbol{C}}_i \bar{\boldsymbol{O}}_{t^{(1)}}^{(1)} \right) \quad (32)$$

Although the DPGMM can represent different length sequences of source and target features, one to one frame matching is assumed in the conversion process (Eq. (32)), because the Markovian transition probability $P(\boldsymbol{a} \,|\, \boldsymbol{m}, \boldsymbol{\lambda})$ is insufficient to convert durations.

4. Simultaneous Conversion of Duration and Spectrum

To convert a speaking rate, we define duration models attached to each mixture of DPGMM. A duration of s-th segment is represented by a joint duration vector $\boldsymbol{d}_s = \left[d_s^{(1)}, d_s^{(2)}\right]^\top$ which consists of source duration $d_s^{(1)}$ and target duration $d_s^{(2)}$. The segment means a period in which the same mixture component continues. Duration models are represented by 2-dimensional Gaussian distributions $\mathcal{N}(\boldsymbol{d}_s \,|\, \boldsymbol{\nu}_i, \boldsymbol{\Phi}_i)$ with the mean vector $\boldsymbol{\nu}_i$ and the covariance matrix $\boldsymbol{\Phi}_i$ and each component of these parameters are defined as follows:

$$\boldsymbol{\nu}_i = \begin{bmatrix} \nu_i^{(1)} \\ \nu_i^{(2)} \end{bmatrix}, \quad \boldsymbol{\Phi}_i = \begin{bmatrix} \phi_i^{(1,1)} & \phi_i^{(1,2)} \\ \phi_i^{(2,1)} & \phi_i^{(2,2)} \end{bmatrix} \quad (33)$$

Figure 2 shows an overview of training duration models and the procedure is summarized as follows:

(1) Determine the mixture index sequence $\boldsymbol{m}$ and frame matching $\boldsymbol{a}$ so as to maximize the posterior probability $P(\boldsymbol{m}, \boldsymbol{a} \,|\, \boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)}, \lambda)$.

(2) Generate duration vectors $\boldsymbol{d}_s, s = 1, \cdots, S$ from $\boldsymbol{m}$ and $\boldsymbol{a}$ obtained in step 1.

(3) Estimate duration models for each mixture component using the corresponding duration vectors.

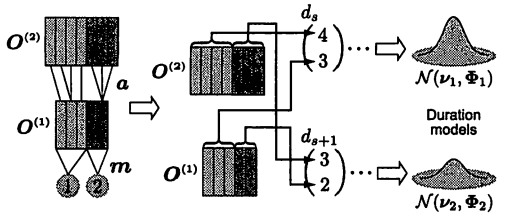


Fig. 2 Training of duration models

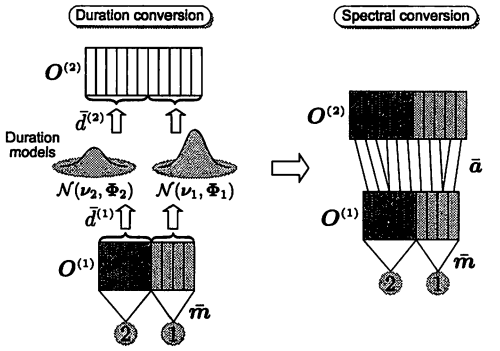


Fig. 3 Duration conversion

The simultaneous conversion of duration and spectrum is performed based on DPGMM with duration models. An overview of duration conversion is shown in Figure 3 and the procedure is summarized as follows:

(1) Determine the mixture index sequence $\bar{m}$ which maximizes posterior probability $P(\boldsymbol{m} \mid \boldsymbol{O}^{(1)}, \lambda)$ given an input feature sequence.

(2) Extract source duration $\bar{d}^{(1)}$ from the mixture index sequence $\bar{m}$ and convert it into the target duration $\bar{d}^{(2)}$ using the following equation:

$$\bar{d}^{(2)} = \nu_i^{(2)} + \frac{\phi_i^{(2,1)}}{\phi_i^{(1,1)}} \left(\bar{d}^{(1)} - \nu_i^{(1)} \right) \tag{34}$$

(3) The matching sequence $\bar{a}$ is determined using duration $\bar{d}^{(1)}$ and $\bar{d}^{(2)}$. Frame matching within each segment is determined at even intervals.

The voice conversion taking account of a speaking rate is performed by converting spectrum based on the matching sequence $\bar{a}$ which are obtained by the above procedure. The converted feature sequence is obtained as

$$\hat{\boldsymbol{O}}_{t^{(2)}}^{(2)} = \left(\sum_{t^{(1)}=1}^{T^{(1)}} \sum_{i=1}^{M} \gamma_{t^{(1)}}^{(1)}(i) \delta(\bar{a}_{t^{(2)}}, t^{(1)}) \bar{\boldsymbol{\Sigma}}_i^{-1} \right)^{-1} \times \left(\sum_{t^{(1)}=1}^{T^{(1)}} \sum_{i=1}^{M} \gamma_{t^{(1)}}^{(1)}(i) \delta(\bar{a}_{t^{(2)}}, t^{(1)}) \bar{\boldsymbol{\Sigma}}_i^{-1} \bar{\boldsymbol{C}}_i \bar{\boldsymbol{O}}_{t^{(1)}}^{(1)} \right) \tag{35}$$

In the proposed method, each mixture component of DPGMM has different transformation function of duration, therefore durations are converted nonlinearly and dependently on spectral information.

5. Experiments

5.1 Experimental conditions

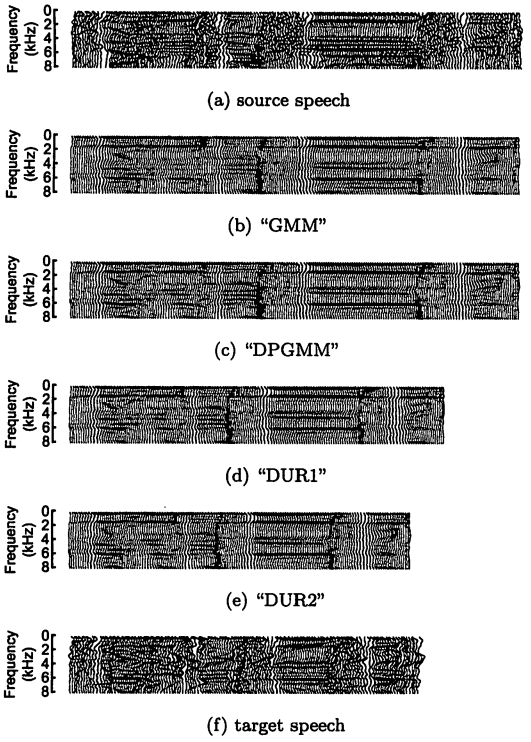
Voice conversion experiments on the ATR Japanese speech database were conducted. Two male speakers were selected as a source and a target speaker (source:MTK target:MYI). The target speaker has a more rapid speaking rate than the source speaker. Ten sentences uttered by the both speakers were used for training and 50 sentences were used for evaluation. The speech data were down-sampled from 20kHz to 16kHz, windowed at a 5-ms frame rate using a 25-ms Blackman window, and parameterized into 24 mel-cepstral coefficients excepting the zero-th coefficients and their first order derivative were used as the dynamic features. The number of mixtures are four.

Figure 4 shows the comparison of spectrum for a Japanese sentence "*muzukashii*" which is not included in the training data. The notation "GMM" and "DPGMM" indicate the conventional methods based on GMM and DPGMM, respectively. "DUR1" and "DUR2" represent the proposed methods with linear and nonlinear duration conversion, respectively. The transition probabilities of the sequence matching are assumed to be independent on the mixture index sequence in "DUR1" and "DUR2." (That is $P(\boldsymbol{a} \mid \boldsymbol{m}, \boldsymbol{\lambda}) \Rightarrow P(\boldsymbol{a} \mid \boldsymbol{\lambda})$.) "DUR1" uses only one linear transformation (Gaussian distribution) and it is equivalent to a special case of "DUR2" in which the parameters of duration models are shared among all mixture components. From Figure 4, it can be seen that the speaking rate of the conventional methods ("GMM" and "DPGMM") are similar to that of the source speech. However, the converted spectrum of the proposed methods ("DUR1" and "DUR2") are more rapid than that of the source speech. Furthermore, although the speaking rate of "DUR1" was converted by a constant ratio, "DUR2" locally changed the speaking rate dependently on spectral information.

5.2 The effectiveness of duration conversion

A DMOS (Differential Mean Opinion Score) test was performed for evaluating the similarity between the target and converted speech in speaker characteristics. The opinion score was set to a 5-point scale.

Figure 5 shows the results of the DMOS test. The subjects were 15 Japanese graduate students. Fifty sentences were randomly chosen from the evaluation sentences. Comparing the proposed methods with duration conversion ("DUR1" and "DUR2") and the conventional methods without duration conversion ("GMM" and "DPGMM"), the proposed methods are superior to the conventional methods. This means that the duration conversion is effective for improving



(a) source speech

(b) "GMM"

(c) "DPGMM"

(d) "DUR1"

(e) "DUR2"

(f) target speech

Fig. 4 Comparison of spectrum for a phrase "*muzukashii*"

the similarity in the converted speech. Furthermore, comparing "DUR1" and "DUR2," "DUR2" could obtain a higher score than "DUR1." It is confirmed that the nonlinear conversion using DPGMM can accurately convert durations because of the dependency on spectrum information.

5.3 The effectiveness of mixture dependent transition probabilities

We also conducted the DMOS test for evaluating the effectiveness of the transition probabilities of the sequence matching which depends on the mixture index sequences. Twenty sentence were used for the evaluation set, and the number of listeners was 10. Figure 6 plots the result of the DMOS test. "DUR2" is the same approach in the previous test and it uses the mixture independent transition probability ($P(\boldsymbol{a} \mid \boldsymbol{\lambda})$). "DURM" means the proposed method that uses the mixture dependent transition probabilities ($P(\boldsymbol{a} \mid \boldsymbol{m}, \boldsymbol{\lambda})$). It can be seen from the figure that no significant difference between "DUR2" and "DURM" is observed. However, the effectiveness of the duration conversion is shown similarly in Figure 5.

6. Conclusion

This paper has proposed a simultaneous conversion method of duration and spectrum based on statistical models including time-sequence matching. The proposed technique

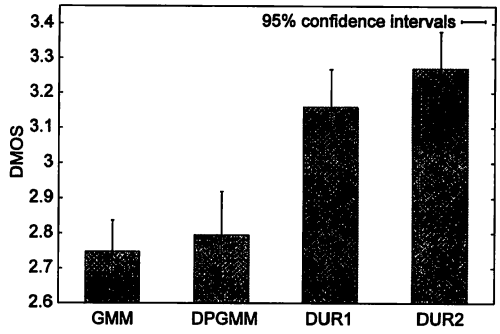


Fig. 5 Result of DMOS test for the comparing speaker *similarity* between with and without duration conversion

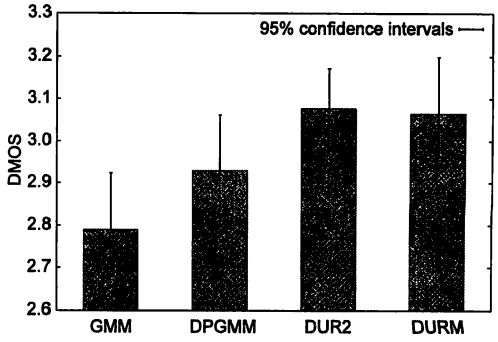


Fig. 6 Result of DMOS test for evaluating the effectiveness of mixture dependent transision probabilities

converts a speaking rate dependently on spectral information. In the experiments, it was confirmed that the proposed method achieved a higher performance than the conventional GMM-based approaches.

Simultaneous optimization of DPGMM and duration models will be a future work.

Acknowledgement The authors would like to thank Dr. Akinobu Lee for his helpful comments and discussions.

Reference

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. Jpn.*, vol. 11, no. 2, pp.71–76, 1990.

[2] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Proc. of IEEE Trans. Speech and Audio Processing*, vol. 6, No. 2, pp. 131–142, 1998.

[3] T. Toda, A.W. Black, and K. Tokuda, "Spectral conversion based on maximun likelihood estimation considering global variance of converted parameter," *Proc. of ICASSP*, vol. 1, pp.9–12, Mar. 2005.

[4] Y. Nankaku, K. Nakamura, T. Toda, and K. Tokuda, "Spectral conversion based on statistical models including time-sequence matching," *Proc. of ISCA Speech Synthesis Workshop*, pp. 333–338, Aug. 2007.

[5] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An Introduction to Variational Methods for Graphical Models," Machine Learning, vol. 37, pp. 183–233, Jan. 1997.