Simultaneous Conversion of Duration and Spectrum Based on Statistical Models Including Time-Sequence Matching

Kaori Yutani1, Yosuke Uto1, Yoshihiko Nankaku1, Tomoki Toda2, Keiichi Tokuda1

1Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, Aichi, 466-8555 Japan
2Nara Institute of Science and Technology, 8916 Takayama, Ikoma, Nara, 630-0101 Japan
{yutani, uto, nankaku, ri, tokuda}@sp.nitech.ac.jp, tomoki@is.naist.jp

Abstract
This paper describes a simultaneous conversion technique of duration and spectrum based on a statistical model including time-sequence matching. Conventional GMM-based approaches cannot perform spectral conversion taking account of speaking rate because it assumes one to one frame matching between source and target features. However, speaker characteristics may appear in speaking rates. In order to perform duration conversion, we attach duration models to each mixture of 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 the subjective DMOS test, the proposed method is superior to the conventional method.

Index Terms: voice conversion, GMM, duration conversion

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 is derived based on the single objective function.

In this conventional GMM-based method, GMMs are trained under an assumption that source and target feature sequences have the same length, because GMMs are 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 Section 3 explains the conventional voice conversion technique based on GMM and DPGMM, respectively. A method of duration and spectrum conversion 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 GMM [3]. Let a vector \( O_i \) be a joint feature vector of the source one \( O_i^{(1)} \) and the target one \( O_i^{(2)} \) at time \( t \), where \( \top \) denotes transposition of the vector. An alignment between two feature sequences is obtained by the Dynamic Programing (DP) matching. In the GMM-based voice conversion, the vector sequence \( O = [O_1, O_2, \ldots, O_T] \top \) is modeled by GMM to learn a relation between source and target features. The output probability of \( O \) given GMM \( \lambda \) can be written as follows:

\[
P(O | \lambda) = \prod_{t=1}^{T} \sum_{i=1}^{M} w_i \mathcal{N}(O_t | \mu_i, \Sigma_i)
\]

where

\[
\mu_i = \begin{bmatrix} \mu_{i(1)} \\ \mu_{i(2)} \end{bmatrix}, \Sigma_i = \begin{bmatrix} \Sigma_{i(1,1)} & \Sigma_{i(1,2)} \\ \Sigma_{i(2,1)} & \Sigma_{i(2,2)} \end{bmatrix}
\]

and \( M \) means the number of mixtures, \( w_i = P(i | \lambda) \) is the mixture weight of the i-th component, \( \mu_i \) and \( \Sigma_i \) 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 maximum likelihood spectral conversion, the optimal converted feature sequence \( O^{(2)} = [O^{(2)1}_1, O^{(2)1}_2, \ldots, O^{(2)1}_T] \top \) given a source feature sequence \( O^{(1)} = [O^{(1)1}_1, O^{(1)1}_2, \ldots, O^{(1)1}_T] \top \) is obtained by maximizing the following conditional distribution:

\[
P(O^{(2)} | O^{(1)}, \lambda) = \sum_{m} \left[ P(m | O^{(1)}, \lambda) \prod_{t=1}^{T} P(O^{(2)t} | O^{(1)t}, m_t, \lambda) \right]
\]
where $m = [m_1, m_2, \cdots, m_T]$ is a mixture index sequence. The conditional distribution can also be written as GMM, and its output probability distribution is presented as follows:

$$P(O(2) | O(1), m_t = i, \lambda) = \mathcal{N}(O(2); E_i(t), D_i)$$ \hspace{1cm} (4)

where

$$E_i(t) = \mu_i^{(2)} + \Sigma_i^{(2)}(1)(O_i^{(1)} - \mu_i^{(1)})$$ \hspace{1cm} (5)

$$D_i = \Sigma_i^{(2)} - \Sigma_i^{(2)}(1)(\Sigma_i^{(1)})^{-1}\Sigma_i^{(2)}(1)$$ \hspace{1cm} (6)

Since the equation (3) includes latent variables, the optimal sequence of $O(2)$ 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 Q-function (auxiliary function):

$$Q(O(2), O(1)) = \sum_m \left[ P(O(2), m | O(1), \lambda) \ln P(O(2), m | O(1), \lambda) \right]$$ \hspace{1cm} (7)

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

$$\hat{O}(2) = \left( D^{-1} \right)^{-1} D^{-1} E$$ \hspace{1cm} (8)

where

$$D^{-1} = \text{diag} \left[ D_1^{-1}, D_2^{-1}, \cdots, D_T^{-1} \right]$$ \hspace{1cm} (9)

$$D_i^{-1} = \sum_{i=1}^M \gamma_i(i) D_i^{-1}$$ \hspace{1cm} (10)

$$D^{-1} E = \left[ D^{-1} E_1^T, D^{-1} E_2^T, \cdots, D^{-1} E_T^T \right]^T$$ \hspace{1cm} (11)

$$D_i^{-1} E_i = \sum_{i=1}^M \gamma_i(i) D_i^{-1} E_i(i)$$ \hspace{1cm} (12)

$$\gamma_i(i) = P(m_t = i | O(1), O(2), \lambda)$$ \hspace{1cm} (13)

### 3. Spectral Conversion Based on DPGMM

In the DPGMM-based method, we define the likelihood function $P(O(1), O(2) | \lambda)$ including the structure of sequence matching. The simultaneous optimization is performed for DP matching and training of model parameters based on the ML criterion. The advantage of the DPGMM is to directly represent two different length sequences $O^{(1)} = \left[ O_1^{(1)}, O_2^{(1)}, \cdots, O_t^{(1)} \right]^T$ and $O^{(2)} = \left[ O_1^{(2)}, O_2^{(2)}, \cdots, O_t^{(2)} \right]^T$. The likelihood function of observation sequences $O = \{ O^{(1)}, O^{(2)} \}$ is written as follows:

$$P(O | \lambda) = \sum_{m,a} \left[ P(m | \lambda) P(O^{(1)} | m, \lambda) \times P(a | \lambda) P(O^{(2)} | O^{(1)}, m, a, \lambda) \right]$$ \hspace{1cm} (14)

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

$$P(m | \lambda) = \prod_{t(1)}^{T(1)} P(m_{t(1)} | \lambda)$$ \hspace{1cm} (15)

$$P(O^{(1)} | m, \lambda)$$ \hspace{1cm} (16)

$$P(O^{(2)} | O^{(1)}, m, a, \lambda)$$ \hspace{1cm} (18)

where

$$C_i = \begin{bmatrix} \hat{\mu}_i \\ C_i \end{bmatrix}, \hat{O}^{(1)}(1) = \begin{bmatrix} 1 \\ O^{(1)}_1 \end{bmatrix}^T$$ \hspace{1cm} (19)

The model parameters of DPGMM are summarized as follows:

- $w = \{a_{t(1)} | 1 \leq t \leq M\}$ : the mixture weights of the GMM which generate the source feature sequence $O^{(1)}$, where $a_{t(1)} = P(m_{t(1)} = i | \lambda)$ is the probability of $i$-th mixture.

- $B^{(1)} = \{b_{t(1)}^{(1)} | 0 \leq t \leq T^{(1)}\}$ : the output probability distributions of source feature $O^{(1)}$, where $b_{t(1)}^{(1)} = P(O^{(1)}_{t(1)} | m_{t(1)} = i, \lambda)$ is the probability of source feature vector $O^{(1)}_{t(1)}$ at $i$-th mixture and which is assumed to be a Gaussian distribution: $\mathcal{N}(O^{(1)}_{t(1)}; \mu^{(1)}_{t(1)}, \Sigma^{(1)}_{t(1)})$ where $\mu^{(1)}_{t(1)}$ and $\Sigma^{(1)}_{t(1)}$ are the mean vector and covariance matrix, respectively.

- $c = \{c_{n} | 1 \leq n \leq N\}$ : the transition probabilities of the sequence matching where $c_{n}$ indicates the probability $P(a_{n+1} = a_{n+1} | a_{n}, \lambda)$. This parameter corresponds to the cost function in the DP matching.

- $B^{(2)} = \{b_{t(2)}^{(2)} | 0 \leq t \leq T^{(2)}\}$ : the output distributions of the target feature $O^{(2)}$, where $b_{t(2)}^{(2)} = P(O^{(2)}_{t(2)} | O^{(1)}_{t(1)}, m_{t(1)} = i, \lambda)$ is the probability of the target feature vector $O^{(2)}_{t(2)}$ given the corresponding source feature vector $O^{(1)}_{t(1)}$ at $i$-th mixture. This conditional distribution is assumed to be a Gaussian distribution: $\mathcal{N}(O^{(2)}_{t(2)}; C_{t(2)} O^{(1)}_{t(1)} + \hat{\mu}_i, \Sigma_i)$ where $\hat{\mu}_i$ and $\Sigma_i$ are the mean vector and the covariance matrix, respectively.

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

1. A mixture index sequence $m$ is determined according to the weight $P(m | \lambda)$.

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

1073
3. The frame matching between $O^{(1)}$ and $O^{(2)}$ is determined according to $P(o^{(1)} | \lambda)$. 
4. The target feature sequence $O^{(2)}$ is generated according to the conditional Gaussian distribution $P(O^{(2)} \mid O^{(1)}, m, a, \lambda)$, given the source feature sequence.

The parameters of DPGMM can be estimated via variational EM algorithm [4]. In the conversion process, the converted feature sequence $O^{(2)}$ can be obtained by maximizing a lower bound of the likelihood. The optimal sequence is given as the following equation:

$$
\bar{O}^{(2)}_{t(2)} = \left( \sum_{i=1}^{T(1)} \sum_{m=1}^{M} \gamma_{t(2)}^{(1)}(i) \delta(t^{(1)}, t^{(2)}) \bar{C}_{i} \tilde{O}_{t(1)}^{(1)} \right)^{-1} \times \left( \sum_{i=1}^{T(1)} \sum_{m=1}^{M} \gamma_{t(1)}^{(1)}(i) \delta(t^{(1)}, t^{(2)}) \bar{C}_{i} \tilde{O}_{t(1)}^{(1)} \right) \tag{20}
$$

where $\delta(u, v) = 1$ if $u = v$, $\delta(u, v) = 0$ otherwise, and $\gamma_{t(1)}^{(1)}$ denotes the expectation of a mixture index $m_{t(1)}$ with respect to posterior distribution:

$$
\gamma_{t(1)}^{(1)}(i) = P(m_{t(1)} = i \mid O^{(1)}, O^{(2)}, \lambda) = \sum_{m, a} P(m, a \mid O^{(1)}, O^{(2)}, \lambda) \delta(m_{t(1)}, i) \tag{21}
$$

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. (20)), because the Markovian transition probability $P(o^{(1)} | \lambda)$ 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 $d_s = [d_{s(1)}, d_{s(2)}]^T$ 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 $N(d_s \mid \nu_s, \Phi_s)$ with the mean vector $\nu_s$ and the covariance matrix $\Phi_s$, and each component of these parameters are defined as follows:

$$
\nu_s = \begin{bmatrix} \nu_{s(1)}^{(1)} \\ \nu_{s(1)}^{(2)} \end{bmatrix}, \Phi_s = \begin{bmatrix} \phi_{s(1)}^{(1, 1)} & \phi_{s(1)}^{(1, 2)} \\ \phi_{s(2)}^{(1, 1)} & \phi_{s(2)}^{(2, 2)} \end{bmatrix} \tag{22}
$$

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

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

2. Generate duration vectors $d_s$, $s = 1, \ldots, S$ from $m$ and $a$ obtained in step 1.

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

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

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

$$
\bar{d}^{(2)} = \nu^{(2)} + \frac{\phi^{(2, 1)}}{\phi^{(2, 2)}} (\bar{d}^{(1)} - \nu^{(1)}) \tag{23}
$$

3. A 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

$$
\bar{O}^{(2)}_{t(2)} = \left( \sum_{i=1}^{T(1)} \sum_{m=1}^{M} \gamma_{t(1)}^{(1)}(i) \delta(\bar{a}_{t(2)}, t^{(1)}) \bar{C}_{i} \tilde{O}_{t(1)}^{(1)} \right)^{-1} \times \left( \sum_{i=1}^{T(1)} \sum_{m=1}^{M} \gamma_{t(1)}^{(1)}(i) \delta(\bar{a}_{t(2)}, t^{(1)}) \bar{C}_{i} \tilde{O}_{t(1)}^{(1)} \right) \tag{24}
$$
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

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” mean the proposed methods with linear and nonlinear duration conversion, respectively. “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, 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.

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. Fifteen sentence were used for the evaluation set, and the number of listeners was 15.

Figure 5 shows the results of the DMOS test. 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 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 spectral information.

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 converts a speaking rate dependently on spectral information. In the experiments, it is confirmed that the proposed method achieved a higher performance than the conventional GMM-based approaches. A simultaneous optimization of DPGMM and duration models will be a future work.

7. References