Low-Delay Voice Conversion based on Maximum Likelihood Estimation of Spectral Parameter Trajectory

Takashi Muramatsu¹, Yamato Ohtani¹, Tomoki Toda¹, Hiroshi Saruwatari¹, and Kiyohiro Shikano†

†Graduate School of Information Science, NARA Institute of Science and Technology
8916-5, Takayama-cho, Ikoma-shi, Nara 630-0192 JAPAN
E-mail: ‡{takashi-mu,yamato-o,tomoki,sawatari,shikano}@is.naist.jp

Abstract

As typical voice conversion methods, two spectral conversion processes have been proposed: 1) the frame-based conversion that converts spectral parameters frame by frame and 2) the trajectory-based conversion that converts all spectral parameters over an utterance simultaneously. The former process is capable of real-time conversion but it sometimes causes inappropriate spectral movements. On the other hand, the latter process provides the converted spectral parameters exhibiting proper dynamic characteristics but a batch process is inevitable. To achieve the real-time conversion process considering spectral dynamic characteristics, we propose a time-recursive conversion algorithm based on maximum likelihood estimation of spectral parameter trajectory. Experimental results show that the proposed method achieves the low-delay conversion process, e.g., only one frame delay, while keeping the conversion performance comparably high to that of the conventional trajectory-based conversion.

Index Terms: speech synthesis, voice conversion, Gaussian mixture model, maximum likelihood estimation, time-recursive algorithm.

1. INTRODUCTION

Voice conversion (VC) is a technique for converting the source speech feature to the target one while keeping linguistic information. There are many VC applications for speech communication such as bandwidth expansion for telephone speech [1] and body-transmitted speech enhancement [2]. It is no doubt that a real-time conversion process is essential in these applications.

As a typical statistical approach to VC, the conversion method based on the Gaussian mixture model (GMM) has been proposed. In this method, the joint probability density of source and target features is modeled by a GMM, which is trained beforehand using a parallel data set consisting of several sentence movements with inappropriate dynamic characteristics due to ignoring the feature correlation between frames. Consequently, the converted speech quality is often deteriorated. On the other hand, the latter method provides the converted spectral parameter sequence exhibiting proper dynamic characteristics by considering dynamic features of the converted parameters. This method causes significant quality improvements of the converted speech. However, it does not allow the real-time conversion process because the source features over an utterance need to be converted simultaneously for considering the feature correlation between frames.

In order to achieve the real-time conversion process considering dynamic characteristics of the converted parameters, we propose a time-recursive conversion algorithm based on MLE of spectral parameter trajectory. This method is inspired by the vector generation algorithm for HMM-based speech synthesis [5] and the vector quantization algorithm for speech coding [6]. Experimental results demonstrate that the proposed method achieves the low-delay conversion process while keeping the conversion performance comparably high to that of the conventional trajectory-based conversion.

This paper is organized as follows. In Section 2, we describe the conventional VC methods. In Section 3, we describe the proposed low-delay VC method. In Section 4, we provide the experimental results. Finally, conclusions are given in Section 5.

2. Conventional GMM-Based Spectral Conversion Methods

2.1. Training of GMM

D-dimensional source and target feature vectors at frame t are denoted by \( \mathbf{z}_t \) and \( \mathbf{y}_t \), respectively. The joint probability density of the source and target feature vectors is modeled as follows:

\[
P(\mathbf{z}_t | \lambda) = \sum_{m=1}^{M} w_m N(\mathbf{z}_t; \mu_m, \Sigma_m),
\]

where \( \mathbf{z}_t \) is a joint vector \( [\mathbf{z}_t^T, \mathbf{y}_t^T]^T \). The superscript \( ^T \) indicates transposition. The mixture component index is \( m \). The total number of mixture components is \( M \). The weight of the \( m \)-th mixture component is \( w_m \). \( N(\mathbf{z}; \mu, \Sigma) \) denotes the normal distribution with mean vector \( \mu \) and covariance matrix \( \Sigma \).

A parameter set of GMM, \( \lambda \), consists of weights, mean vectors and covariance matrices for individual mixture components. The mean vector \( \mu_m \) and the covariance matrix \( \Sigma_m \) of the \( m \)-th mixture component are defined as

\[
\mu_m = \begin{bmatrix} \mu_{m}^{(x)} \\ \mu_{m}^{(y)} \end{bmatrix}, \quad \Sigma_m = \begin{bmatrix} \Sigma_{m}^{(xx)} & \Sigma_{m}^{(xy)} \\ \Sigma_{m}^{(yx)} & \Sigma_{m}^{(yy)} \end{bmatrix},
\]

where \( m = 1, 2, \ldots, M \).
where $\mu^{(m)}_t$ and $\mu^{(m)}_t$ are the mean vectors of the $m$-th mixture component for the source and the target, respectively. The matrices $\Sigma^{(m)}_t$ and $\Sigma^{(m)}_t$ are the covariance matrices of the $m$-th mixture component for the source and target, respectively. The matrices $\Sigma^{(m)}_t$ and $\Sigma^{(m)}_t$ are the cross-covariance matrices of the $m$-th mixture component for the source and target, respectively. These parameters are estimated with the EM algorithm using the joint vectors extracted from the training data [7].

### 2.2. Frame-Based Conversion

The conditional probability density of $y_t$, given $x_t$, is represented as follows:

$$P(y_t|x_t, \lambda^{(z)}) = \sum_{m=1}^{M} P(m|x_t, \lambda^{(z)}) P(y_t|m, \lambda^{(z)}),$$

where

$$P(m|x_t, \lambda^{(z)}) = \frac{w_m \mathcal{N}(x_t; \mu^{(m)}_t, \Sigma^{(m)}_t)}{\sum_{m=1}^{M} w_m \mathcal{N}(x_t; \mu^{(m)}_t, \Sigma^{(m)}_t)},$$

and

$$P(y_t|m, \lambda^{(z)}) = \mathcal{N}(y_t; \mu^{(m)}_t, \Sigma^{(m)}_t).$$

The mean vector $E^{(m)}_t$ and the covariance matrix $D^{(m)}_t$ of the $m$-th conditional probability distribution are written as

$$E^{(m)}_t = \mu^{(m)}_t + \Sigma^{(m)}_t \Sigma^{(m)}_t^{-1} (x_t - \mu^{(m)}_t),$$

$$D^{(m)}_t = \Sigma^{(m)}_t - \Sigma^{(m)}_t \Sigma^{(m)}_t^{-1} \Sigma^{(m)}_t.$$

The MMSE-based conversion [3] is performed as follows:

$$\hat{y}_t = \sum_{m=1}^{M} P(m|x_t, \lambda^{(z)}) E^{(m)}_{t},$$

where $\hat{y}_t$ is the converted target feature vector. This method converts spectral parameters frame by frame.

### 2.3. Trajectory-Based Conversion

The target feature vector is defined as a $2D$-dimensional vector $Y_t$ consisting of $D$-dimensional static and dynamic features at frame $t$. A relationship between the time sequences $Y = [Y^T_1, Y^T_2, \ldots, Y^T_T]$ and $\bar{y} = \bar{y}_1, \bar{y}_2, \ldots, \bar{y}_T$ is presented as $Y = W \bar{y}$ where

$$W = \begin{bmatrix} w_1, w_2, \ldots, w_T \end{bmatrix}^T,$$

$$w_t = \begin{bmatrix} 0_{2D \times (t-2)D}, w_t, 0_{2D \times (T-t)D} \end{bmatrix},$$

and

$$w = \begin{bmatrix} 0_{D \times D} & I_{D \times D} & -gI_{D \times D} \end{bmatrix}.$$

As the source feature vector $X_t$, the feature vector including both static and dynamic features [4] or the concatenated feature vector from multiple frames [8] is employed. This time sequence is written as $X = [X^T_1, X^T_2, \ldots, X^T_T]$.

Using joint vectors $Z_t = [X^T_t, Y^T_t]^T$, the GMM parameter set $\lambda^{(z)}$ of the joint probability density $P(Z_t|\lambda^{(z)})$ is trained beforehand by adopting the conventional training framework [7].

In the conversion process, the converted feature vector sequence $\hat{y}$ is determined by maximizing the likelihood function $P(Y|X, \lambda^{(z)})$ as follows:

$$\hat{y} = \arg \max_y P(Y|X, \lambda^{(z)})$$

$$\approx \arg \max_y P(\hat{m}|X, \lambda^{(z)}) P(Y|X, \hat{m}, \lambda^{(z)}).$$

Figure 1: Schematic image of the trajectory-based conversion process.

where the suboptimal mixture component sequence $\hat{m}$ is determined by maximizing the posterior probability $P(\hat{m}|X, \lambda^{(z)})$. The ML estimate of $\hat{y}$ is given by solving the following linear equation

$$R\hat{y} = r,$$

where

$$R = \hat{W}^T D(m)_m^{-1} W, \quad \hat{W} = \begin{bmatrix} E^{(m)}_1, \ldots, E^{(m)}_T \end{bmatrix}^T, \quad \hat{W} = \begin{bmatrix} D^{(m)}_1^{-1}, \ldots, D^{(m)}_T^{-1} \end{bmatrix}^T.$$

The mean vector $E^{(m)}_t$ and the covariance matrix $D^{(m)}_t$ are calculated based on $\lambda^{(z)}$. This paper uses only the diagonal elements of $D^{(m)}_t$.

Figure 1 shows the schematic image of the trajectory-based conversion process. Note that the matrix $P$ is generally full. It is necessary to store statistics of the conditional probabilities over the time sequence shown by Eqs. 14 and 17 for determining $\hat{y}$.

### 3. Proposed Low-Delay Conversion Method

We propose the time-recursive conversion algorithm by modifying the trajectory-based conversion in a manner similar to that described in [5] [6]. The following matrix $W^{(t-1)}$ is obtained by replacing the elements of $W$, $w_t$ with $W$ from time $t$ to $T$.

$$\hat{W}^{(t-1)} = \begin{bmatrix} w_1, \ldots, w_{t-1}, w_t, w_{t+1}, \ldots, w_T \end{bmatrix}^T,$$

where

$$\hat{w}_t = \begin{bmatrix} 0_{2D \times (t-2)D}, w_t, 0_{2D \times (T-t)D} \end{bmatrix},$$

$$\hat{w} = \begin{bmatrix} 0_{D \times D} & I_{D \times D} & 0_{D \times D} \end{bmatrix}.$$

In this case, the set of equations corresponding to Eq. (13) can be written as

$$\hat{R}^{(t-1)} \hat{y}^{(t-1)} = \hat{y}^{(t-1)},$$

where

$$\hat{R}^{(t-1)} = \hat{W}^{(t-1)} D^{(m)}_m^{-1} \hat{W}^{(t-1)},$$

$$\hat{y}^{(t-1)} = \hat{W}^{(t-1)} D^{(m)}_m^{-1} E^{(m)}_m,$$

$$\hat{y}^{(t-1)} = \begin{bmatrix} \hat{y}_1^{(t-1)}, \ldots, \hat{y}_T^{(t-1)} \end{bmatrix}^T.$$
The matrix $\mathbf{w}$ is defined so that $\mathbf{W}^{(t)}$ is a $D$-by-$D$ block diagonal matrix. When $\mathbf{w}_t$ is replaced with $\mathbf{w}_t$, the following equation is obtained

$$\mathbf{R}^{(t)} \mathbf{y}^{(t)} = \mathbf{y}^{(t)},$$

where $\mathbf{w}^{(t)} = \begin{bmatrix} \mathbf{w}_1^T, \cdots, \mathbf{w}_{t-1}^T, \mathbf{w}_t^T, \mathbf{w}_{t+1}^T, \cdots, \mathbf{w}_T^T \end{bmatrix}^T$. (26)

Using the above equations, the following equations are derived:

$$\mathbf{R}^{(t)} = \mathbf{R}^{(t-1)} + v_t^T \mathbf{E}_m^{(Y)} v_t,$$

$$\mathbf{y}^{(t)} = \mathbf{y}^{(t-1)} + v_t^T \mathbf{D}_m^{(Y)} \mathbf{E}_m^{(Y)},$$

$$v_t = v_{t-1} - \mathbf{w}_t.$$ (29)

Using the matrix inversion lemma as in the derivation of the RLS algorithm [9], $\mathbf{y}^{(t)}$ and $\mathbf{R}^{(t)} = \mathbf{R}^{(t-1)}$ are calculated recursively from $\mathbf{y}^{(t-1)}$ and $\mathbf{R}^{(t-1)} = \mathbf{R}^{-1}$, respectively. All elements of $\mathbf{y}^{(t)}$ are replaced by updating $\mathbf{w}_t$ with $\mathbf{w}_t$. Note that $\mathbf{y}^{(0)}$ is equal to the static parts of $\mathbf{E}_m^{(Y)}$. By replacing $\mathbf{w}_t$ from time 1 to $T$, the final updated vector $\mathbf{y}^{(T)}$ is obtained, which corresponds to the solution of Eq. (13), $\mathbf{y}$.

Table 1 shows the time-recursive algorithm derived by further introducing the sliding window concept of the RLS algorithm to the above recursive algorithm. In this algorithm, the sizes of the recursively updated parameters, $\mathbf{y}^{(t)}$ and $\mathbf{R}^{(t)}$, are limited to LD-by-LD, and LD-by-LD, respectively. Therefore, the converted vectors obtained by this algorithm are suboptimal. Equations (A.1) and (A.2) are the key parts of this algorithm. Old elements are thrown out by using the matrix $\mathbf{J}_a$ written as

$$\mathbf{J}_a = \begin{bmatrix} 0_{(L-1)D \times D} & \mathbf{I}_{(L-1)D \times (L-1)D} & 0_{0 \times (L-1)D} \\ 0_{0 \times D} & 0_{D \times D} & \mathbf{I}_{D \times (L-1)D} \end{bmatrix},$$

and then new components are added by using the $LD \times LD$ matrix $\mathbf{P}^{(t)}$ and the $LD \times 1$ vector $\mathbf{y}^{(t)}$, which are respectively written as

$$\mathbf{P}^{(t)} = \text{diag} \left[ 0_{0 \times D}, \cdots, 0_{0 \times D}, \mathbf{R}_t^{-1} \right],$$

$$\mathbf{y}^{(t)} = \begin{bmatrix} 0, \cdots, 0, (\mathbf{R}_t^{-1} \mathbf{r}_t)^T \end{bmatrix}^T.$$ (32)

The new components, the $D \times D$ matrix $\mathbf{R}_t$ and the $D \times 1$ vector $\mathbf{r}_t$ are written as

$$\mathbf{R}_t = \mathbf{J}_a \mathbf{w}_t^T \mathbf{D}_m^{(Y)} \mathbf{w}_t \mathbf{J}_a^T,$$

$$\mathbf{r}_t = \mathbf{J}_a \mathbf{w}_t^T \mathbf{D}_m^{(Y)} \mathbf{E}_m^{(Y)},$$

where the $2D \times LD$ matrices $\mathbf{w}_t$, $\mathbf{w}_t^T$, and $\mathbf{w}_t$, and the $D \times LD$ matrix $\mathbf{J}_a$ are respectively written as follows:

$$\mathbf{w}_t = \begin{bmatrix} 0_{0 \times D}, \cdots, 0_{0 \times D}, \mathbf{w} \end{bmatrix},$$

$$\mathbf{w}_t = \begin{bmatrix} 0_{0 \times D}, \cdots, 0_{0 \times D}, \mathbf{w} \end{bmatrix},$$

$$\mathbf{v}_t = \mathbf{w}_t - \mathbf{w}_t,$$

$$\mathbf{J}_a = \begin{bmatrix} 0_{0 \times D}, \cdots, 0_{0 \times D}, \mathbf{I}_{D \times D} \end{bmatrix}.$$ (38)

Figure 2 shows the schematic image of this conversion process with $D = 1$. The value $L$ is the number of frames stored in this algorithm, which is determined by

$$L = \max(D, 1) + 1,$$

where $D$ is the length of frame delay allowed in the spectral conversion. The converted vector at time $t$, $\mathbf{y}_t$, is obtained as

$$\mathbf{y}_t = \mathbf{y}_{t-D},$$

where $\mathbf{y}_{t-D}$ is a vector contained in $\mathbf{y}^{(t)}$ written as

$$\mathbf{y}^{(t)} = \begin{bmatrix} \mathbf{y}_{t-L}^T, \cdots, \mathbf{y}_{t-1}^T \end{bmatrix}^T.$$ (41)

It is noted that this time-recursive algorithm is valid if the cross-covariance values of static and dynamic features in $\mathbf{D}_m^{(Y)}$ are zeros. If using full elements of $\mathbf{D}_m^{(Y)}$, it is necessary to derive the time-recursive algorithm based on a replacement of elements of $\mathbf{D}_m^{(Y)}$ instead of $\mathbf{W}^{(t)}$ in a manner similar to that described in [5].

4. Experimental Evaluations

The performance of the proposed method is evaluated in the framework of body-transmitted VC [2]. There have been proposed voice conversion frameworks for improving the quality of several types of body-transmitted speech. In this paper, we conduct the conversion from body-transmitted ordinary speech (BTOS) into natural speech. This conversion system is very effective for achieving noise-robust speech communication.

4.1. Experimental Conditions

We simultaneously recorded BTOS and natural speech uttered by the same speaker using a NAM microphone [2], which is one of the body-conductive microphones, and an air-conductive microphone. Sampling frequency was set to 8 kHz. The 0-th through 16-th mel-cepstral coefficients were used as a spectral feature. The frame shift was set to 5 ms. In order to compensate missing acoustic features of BTOS due to the body transmission effects as described in [2], we used the spectral segment vector as the source feature, which was constructed by
Consequently, the proposed method achieves a very low-delay voice. Table 2 shows the number of training/evaluation sentences in each speaker. The conversion model was trained for each speaker separately. The number of mixture components was set to 64.

4.2. Objective Evaluations

We used the mel-cepstral distortion between the target and converted mel-cepstra as the objective evaluation measure. Figure 3 shows the mel-cepstral distortions in the frame-based, trajectory-based and proposed methods, respectively. The proposed method outperforms the conventional frame-based, trajectory-based and proposed methods, respectively.

4.3. Subjective Evaluations

We conducted a preference test on speech quality. We evaluated converted speech by two conventional methods (the frame-based and the trajectory-based) and by the proposed methods when setting the delay value to 0, 1, and 5, respectively. A pair of the converted voices from two different methods was randomly presented to the five listeners, and then they were asked which voice sounded more natural.

Figure 4 shows the preference scores of individual methods with confidence intervals (95%). The converted speech quality of the proposed method with no delay is almost equal to that in the trajectory-based method when setting the delay value to around 5, even only one frame (i.e., 5 ms). Consequently, the proposed method achieves a very low-delay frame-by-frame conversion process while keeping the converted speech quality comparably high to that of the trajectory-based conversion.

5. Conclusions

In this paper, we proposed a spectral conversion method for high-quality and low-delay VC. The experimental results show that the proposed method of spectral conversion by the time-recursive algorithm archives high-quality and low-delay conversion. We plan to reduce the computational cost of the conversion process and the required memory size to develop a small run-time VC system for embedded applications.

6. Acknowledgments

This work was supported in part by MEXT Grant-in-Aid for Young Scientists (A).

7. References