Statistical Singing Voice Conversion based on Direct Waveform Modification with Global Variance

Kazuhiro Kobayashi, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan

{kazuhiro-k, tomoki, neubig, ssakti, s-nakamura}@is.naist.jp

Abstract

This paper presents techniques to improve the quality of voices generated through statistical singing voice conversion with direct waveform modification based on spectrum differential (DIFFSVC). The DIFFSVC method makes it possible to convert singing voice characteristics of a source singer into those of a target singer without using vocoder-based waveform generation. However, quality of the converted singing voice still degrades compared to that of a natural singing voice due to various factors, such as the over-smoothing of the converted spectral parameter trajectory. To alleviate this over-smoothing, we propose a technique to restore the global variance of the converted spectral parameter trajectory within the framework of the DIFFSVC method. We also propose another technique to specifically avoid over-smoothing at unvoiced frames. Results of subjective and objective evaluations demonstrate that the proposed techniques significantly improve speech quality of the converted singing voice while preserving the conversion accuracy of singer identity compared to the conventional DIFFSVC.
Index Terms: statistical singing voice conversion, direct waveform modification, spectral differential, global variance, Gaussian mixture model

1. Introduction

A singing voice is one of the most expressive components in music. In addition to pitch, dynamics, and rhythm, the linguistic information of the lyrics can be used by singers to express more varieties of expression than other music instruments. Although singers can also expressively control their voice timbre to some degree, they usually have a difficulty in changing it widely (e.g. changing their own voice timbre into that of another specific singer) owing to physical constraints in speech production. If singers could freely control their voice timbre beyond their physical constraints, it would open up entirely new ways for singers to express more varieties of expression.

Singing synthesis [1, 2, 3] has been a growing interest in computer-based music technology. Entering notes and lyrics to the singing synthesis engine, users (e.g., composers and singers) can easily produce a synthesized singing voice which has a specific singer's voice characteristics, different from those of the users. Previous work has proposed techniques to flexibly control the synthesized singing voice as the users want by automatically adjusting parameters of the singing synthesis engine so that the variation of power and pitch in the synthesized singing voice is similar to that of the given users' natural singing voice [4, 5]. Although these technologies using singing synthesis engines are effective to produce the singing voices desired by the users, it is essentially difficult to produce synthesized singing voices by controlling all singing voice components including lyrics on the fly.

Singing voice conversion (SVC), on the other hand, converts a source singer's singing voice into another target singer's singing voice [6, 7]. This makes it possible to produce the desired singing voices on the fly, enabling singers to sing songs with their desired voice timbre, not limited by physical constraints. One of the typical methods is based on statistical voice conversion (VC) techniques [8, 9]. A conversion model is trained in advance using acoustic features, which are extracted from a parallel data set of song pairs sung by the source and target singers. The trained conversion model makes it possible to convert the acoustic features of the source singer's singing voice into those of the target singer's singing voice in any song while keeping the linguistic information of the lyrics unchanged. Recently eigenvoice conversion (EVC) techniques [10, 11] have also been successfully applied to SVC [12] to develop more flexible SVC systems capable of achieving conversion between arbitrary source and target singers, even if a parallel data set is not available. However, speech quality of the singing voice converted by SVC is usually degraded compared to that of the natural singing voice due to various errors caused by not only the acoustic feature conversion process, but also the vocoding process for waveform generation.

To improve speech quality of the converted singing voice, we have proposed an SVC method with direct waveform modification based on spectrum differential (DIFFSVC) [13]. DIFFSVC can avoid using the vocoder framework in generation of the excitation signal by directly filtering an input singing voice waveform with a time sequence of spectral feature differentials estimated by a differential Gaussian mixture model (GMM) derived from the conventional GMM used in the standard SVC method. Although DIFFSVC is applicable to only situations in which pitch conversion is not necessary (such as in intra gender conversion), voice timbre of the input singing voice can be successfully converted into that of the target singer while achieving speech quality significantly higher than the standard SVC method. The direct waveform filtering tends to keep modulation components of the converted spectral parameter trajectory larger compared to those in the standard SVC with the vocoder-based waveform generation. However, they are still significantly smaller than those of natural spectral parameter trajectories as the converted spectral parameter trajectory tends to be excessively smoothed. This over-smoothing effect is well–known as a factor causing quality degradation in the synthesized sining voice.

In this paper, to alleviate the over-smoothing effect, we propose a parameter generation algorithm considering global variance (GV) for DIFFSVC. GV is a well-known feature to measure the over-smoothing effect [9]. To restore the GV of the converted spectral parameter trajectory, we modify the objective

function to determine a time sequence of the spectral differential. Additionally, we implement a process smoothing the spectral differential at unvoiced frames to avoid the over-smoothing effect at unvoiced sounds. We conduct subjective and objective evaluations, demonstrating that the proposed DIFFSVC method significantly improves speech quality of the converted singing voice compared to the conventional DIFFSVC method.

2. Statistical singing voice conversion with direct waveform modification (DIFFSVC)

DIFFSVC consists of a training process and a conversion process. In the training process, a joint probability density function of spectral features of a source singer and the differential between the source and target singers is modeled with a differential GMM, which is directly derived from a traditional GMM. As the spectral features of the source and target singers, we employ $2D$-dimensional joint static and dynamic feature vectors $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ of the source and $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ of the target consisting of D-dimensional static feature vectors $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ and their dynamic feature vectors $\Delta\boldsymbol{x}_t$ and $\Delta\boldsymbol{y}_t$ at frame t, respectively, where $\top$ denotes the transposition of the vector. As shown in [7], their joint probability density modeled by the GMM is given by

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right) \quad (1)$$

where $\mathcal{N}\left(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ denotes a Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The mixture component index is m. The total number of mixture components is M. $\boldsymbol{\lambda}$ is a GMM parameter set consisting of the mixture-component weight α_m, the mean vector $\boldsymbol{\mu}_m$, and the covariance matrix $\boldsymbol{\Sigma}_m$ of the m-th mixture component. The GMM is trained using joint vectors of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ in the parallel data set, which are automatically aligned to each other by dynamic time warping. Then, the differential GMM is analytically derived from the trained GMM by transforming the parameters. Let $\boldsymbol{D}_t = \left[\boldsymbol{d}_t^\top, \Delta\boldsymbol{d}_t^\top\right]^\top$ denote the static and dynamic differential feature vector, where $\boldsymbol{d}_t = \boldsymbol{y}_t - \boldsymbol{x}_t$. The joint probability density function of the source and differential spectral features is shown as follows:

$$P\left(\boldsymbol{X}_t, \boldsymbol{D}_t|\boldsymbol{\lambda}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{D}_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(D)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XD)} \\ \boldsymbol{\Sigma}_m^{(DX)} & \boldsymbol{\Sigma}_m^{(DD)}\end{bmatrix}\right) \quad (2)$$

$$\boldsymbol{\mu}_m^{(D)} = \boldsymbol{\mu}_m^{(Y)} - \boldsymbol{\mu}_m^{(X)} \quad (3)$$

$$\boldsymbol{\Sigma}_m^{(XD)} = {\boldsymbol{\Sigma}_m^{(DX)}}^\top = \boldsymbol{\Sigma}_m^{(XY)} - \boldsymbol{\Sigma}_m^{(XX)} \quad (4)$$

$$\boldsymbol{\Sigma}_m^{(DD)} = \boldsymbol{\Sigma}_m^{(XX)} + \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(XY)} - \boldsymbol{\Sigma}_m^{(YX)}. \quad (5)$$

In the conversion process, the converted spectral feature differential is estimated from the source singer's spectral features based on the differential GMM in the same manner as maximum likelihood estimation of speech parameter trajectory with the GMM [9]. The voice timbre of the source singer is converted into that of the target singer by directly filtering the speech waveform of the input natural singing voice with the converted spectral feature differential. Time sequence vectors of the source features and the spectrum feature differential are denoted as $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{D} = [\boldsymbol{D}_1^\top, \cdots, \boldsymbol{D}_T^\top]^\top$ where T is the number of frames included in the time sequence of the given source feature vectors. A time sequence vector of the converted static features $\hat{\boldsymbol{d}} = [\hat{\boldsymbol{d}}_1^\top, \cdots, \hat{\boldsymbol{d}}_T^\top]^\top$ is determined as follows:

$$\hat{\boldsymbol{d}} = \underset{\boldsymbol{d}}{\operatorname{argmax}}\, P(\boldsymbol{D}|\boldsymbol{X}, \boldsymbol{\lambda}) \text{ s.t. } \boldsymbol{D} = \boldsymbol{W}\boldsymbol{d} \quad (6)$$

$$P(\boldsymbol{D}|\boldsymbol{X}, \boldsymbol{\lambda}) = \prod_{t=1}^{T} \sum_{m=1}^{M} P(m|\boldsymbol{X}_t, \boldsymbol{\lambda}) P(\boldsymbol{D}_t|m, \boldsymbol{X}_t, \boldsymbol{\lambda}) \quad (7)$$

where $\boldsymbol{W}$ is a transformation matrix to expand the static feature vector sequence into the joint static and dynamic feature vector sequence [14] and the probability density function at frame t is given by

$$P\left(\boldsymbol{D}_t|m, \boldsymbol{X}_t, \boldsymbol{\lambda}\right) = \mathcal{N}\left(\boldsymbol{D}_t; \boldsymbol{E}_{m,t}^{(D)}, \boldsymbol{V}_m^{(D)}\right) \quad (8)$$

$$\boldsymbol{E}_{m,t}^{(D)} = \boldsymbol{\mu}_m^{(D)} + \boldsymbol{\Sigma}_m^{(DX)} {\boldsymbol{\Sigma}_m^{(XX)}}^{-1} \left(\boldsymbol{X}_t - \boldsymbol{\mu}_m^{(X)}\right) \quad (9)$$

$$\boldsymbol{V}_m^{(D)} = \boldsymbol{\Sigma}_m^{(DD)} - \boldsymbol{\Sigma}_m^{(DX)} {\boldsymbol{\Sigma}_m^{(XX)}}^{-1} \boldsymbol{\Sigma}_m^{(XD)}. \quad (10)$$

3. DIFFSVC considering global variance

In order to improve the speech quality of the converted singing voice in DIFFSVC, we propose two techniques: 1) restoration of the GV of the converted spectral feature trajectory and 2) smoothing of the converted spectral feature differential at unvoiced frames.

The GV of the target static feature vector over the time sequence is written as

$$\boldsymbol{v}(\boldsymbol{y}) = [v(1), v(2), \cdots, v(d), \cdots, v(D)]^\top \quad (11)$$

$$v(d) = \frac{1}{T} \sum_{t=1}^{T} \left(y_t(d) - \bar{y}(d)\right)^2 \quad (12)$$

$$\bar{y}(d) = \frac{1}{T} \sum_{\tau=1}^{T} y_\tau(d) \quad (13)$$

where $y_t(d)$ shows the d-th component of the target feature vector at frame t. The probability density function of the GV is modeled as follows:

$$P\left(\boldsymbol{v}(\boldsymbol{y}) \mid \boldsymbol{\lambda}^{(v)}\right) = \mathcal{N}\left(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(vv)}\right) \quad (14)$$

where $\boldsymbol{\lambda}^{(v)}$ is a parameter set of a Gaussian distribution for which the mean vector and covariance matrix are $\boldsymbol{\mu}^{(v)}$ and $\boldsymbol{\Sigma}^{(vv)}$, respectively. The converted feature differential trajectory is determined by maximizing a new objective function as follows:

$$\hat{\boldsymbol{d}} = \underset{\boldsymbol{d}}{\operatorname{argmax}}\, P(\boldsymbol{D}|\boldsymbol{X}, \boldsymbol{\lambda})^\omega P(\boldsymbol{v}(\boldsymbol{y}')|\boldsymbol{\lambda}^{(v)}) \text{ s.t. } \boldsymbol{D} = \boldsymbol{W}\boldsymbol{d} \quad (15)$$

where $\boldsymbol{y}' = [\boldsymbol{x} + \boldsymbol{d}]$ and the constant ω denotes a parameter for controlling the balance between the two likelihoods. The converted feature differential trajectory is iteratively updated by using the steepest descent method as follows:

$$\hat{\boldsymbol{d}}^{(i+1)-th} = \hat{\boldsymbol{d}}^{(i)-th} + \alpha \cdot \Delta\hat{\boldsymbol{d}}^{(i)-th} \quad (16)$$

where α is a step size parameter. The gradient vector $\Delta\hat{\boldsymbol{d}}^{(i)-th}$ is given by

$$\Delta \boldsymbol{d}^{(i)-th} = \left.\frac{\partial \mathcal{L}}{\partial \boldsymbol{d}}\right|_{\boldsymbol{d}=\boldsymbol{d}^{(i)-th}} \tag{17}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{d}} = \omega \left(-\boldsymbol{W}^\top {\boldsymbol{V}_m^{(D)}}^{-1} \boldsymbol{W}\boldsymbol{d} + \boldsymbol{W}^\top {\boldsymbol{V}_m^{(D)}}^{-1} \boldsymbol{E}_m^{(D)}\right) + \left[{\boldsymbol{v}'_1}^\top, {\boldsymbol{v}'_2}^\top, \cdots, {\boldsymbol{v}'_t}^\top, \cdots, {\boldsymbol{v}'_T}^\top\right]^\top \tag{18}$$

$$\boldsymbol{E}_m^{(D)} = \left[\boldsymbol{E}_{m_1,1}^{(D)}, \cdots, \boldsymbol{E}_{m_t,t}^{(D)}, \cdots, \boldsymbol{E}_{m_T,T}^{(D)}\right]^\top \tag{19}$$

$${\boldsymbol{V}_m^{(D)}}^{-1} = \mathrm{diag}\left[{\boldsymbol{V}_{m_1}^{(D)}}^{-1}, \cdots, {\boldsymbol{V}_{m_t}^{(D)}}^{-1}, \cdots, {\boldsymbol{V}_{m_T}^{(D)}}^{-1}\right] \tag{20}$$

$$\boldsymbol{v}'_t = \left[v'_t(1), v'_t(2), \cdots, v'_t(d), \cdots, v'_t(D)\right]^\top \tag{21}$$

$$v'_t(d) = -\frac{2}{T}\boldsymbol{p}^{(v)}(d)^\top \left(\boldsymbol{v}(\boldsymbol{y}') - \boldsymbol{\mu}^{(v)}\right)\left(y'_t(d) - \bar{y}'(d)\right) \tag{22}$$

where $\boldsymbol{p}^{(v)}(d)$ indicates the d-th column vector of the inverse matrix of $\boldsymbol{\Sigma}^{(vv)}$. An initial feature differential trajectory for the iterative update is determined by filtering in the conventional DIFFSVC as follows:

$$\hat{d}'_t(d) = \sqrt{\frac{\mu_v(d)}{v(d)}}(\hat{y}_t(d) - \overline{\hat{y}}(d)) + \overline{\hat{y}}(d) - x_t(d) \tag{23}$$

where $\hat{y}_t(d)$ indicates the converted feature at frame t determined by the conventional DIFFSVC and $\overline{\hat{y}}(d)$ indicates its average over a time sequence.

It has been reported that unvoiced consonants (e.g. /s/, /sh/) are less affected by speaker individuality compared to voiced sounds (e.g. /ae/, /n/) in normal speech [15]. Based on this finding, in order to alleviate the over-smoothing effect as much as possible, we minimize the amount of conversion at unvoiced frames by smoothing the converted feature differential at those frames. We implement this process on top of the previously described DIFFSVC with GV by modifying $\boldsymbol{E}_{m,t}^{(D)}$ and ${\boldsymbol{V}_m^{(D)}}^{-1}$ at unvoiced frames as follows:

$$\boldsymbol{E}_{m,t}^{(D)} = \left\{ \begin{array}{ll} \boldsymbol{0} & \text{(for static \& delta)} \end{array} \right. \tag{24}$$

$${\boldsymbol{V}_m^{(D)}}^{-1} = \left\{ \begin{array}{ll} \boldsymbol{0} & \text{(for static)} \\ {\boldsymbol{V}_m^{(\Delta D)}}^{-1} & \text{(for delta)} \end{array} \right. \tag{25}$$

where ${\boldsymbol{V}_m^{(\Delta D)}}^{-1}$ shows delta components of the inverse matrix of the covariance matrix in Eq. (8). These parameter modifications make the converted spectral feature differential smoothly vary at unvoiced frames. Note that we avoid updating the converted spectral feature differential at the unvoiced frames in Eq. (16).

4. Experimental evaluation

4.1. Experimental conditions

We evaluated speech quality and singer identity of the converted singing voices to compare the conventional and proposed DIFFSVC methods. We used singing voices of 21 Japanese traditional songs, which were divided into 152 phrases, where the duration of each phrase was approximately 8 seconds. 3 males and 3 females sang these phrases. The sampling frequency was set to 16 kHz.

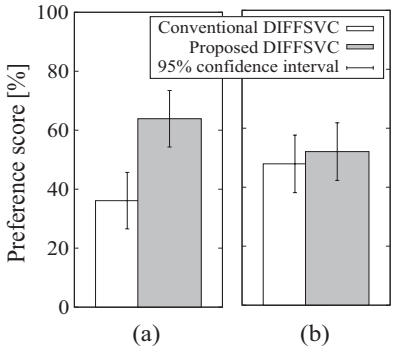

Figure 1: Results of preference test. (a) speech quality of converted singing voice, (b) conversion accuracy of singer individuality.

STRAIGHT [16] was used to extract spectral envelopes, which were parameterized to the 1-24th mel-cepstral coefficients as the spectral features. The frame shift was 5 ms. The mel log spectrum approximation (MLSA) filter [17] was used as the synthesis filter.

We used 80 randomly selected phrases for the GMM training and the remaining 72 phrases were used for evaluation. The speaker-dependent GMMs were separately trained for individual singer pairs determined in a round-robin fashion within intra-gender singers. The number of mixture components was 128.

Two preference tests were conducted. The first test evaluated speech quality of the converted singing voices. The converted singing voice samples of the conventional and proposed DIFFSVC methods for the same phrase were presented to listeners in random order. The listeners selected which sample had better sound quality. The second preference test evaluated the singer identity conversion accuracy. A natural singing voice sample of the target singer was presented to the listeners first as a reference. Then, the converted singing voice samples of the conventional and proposed DIFFSVC methods for the same phrase were presented in random order. The listeners selected which sample was more similar to the reference natural singing voice in terms of singer identity. The number of listeners was 6 and each listener evaluated 54 sample pairs. They were allowed to replay each sample pair as many times as necessary.

4.2. Subjective evaluation

Figure 1 (a) indicates the result of the preference test for the speech quality. The proposed DIFFSVC method generates the converted speech with better speech quality than the conventional DIFFSVC method. Figure 1 (b) indicates the result of the preference test for the singer identity. The conversion accuracy of the singer identity of the proposed DIFFSVC method is not significantly different from that of the conventional DIFFSVC method. Although the proposed DIFFSVC method avoids accurately converting spectral features at unvoiced frames, it still yields conversion accuracy of singer individuality almost equal to that of the conventional DIFFSVC method.

These results demonstrate that the proposed DIFFSVC method is capable of converting voice timbre with higher speech quality while causing no degradation in the conversion accuracy of singer identity compared to the conventional DIFFSVC method.

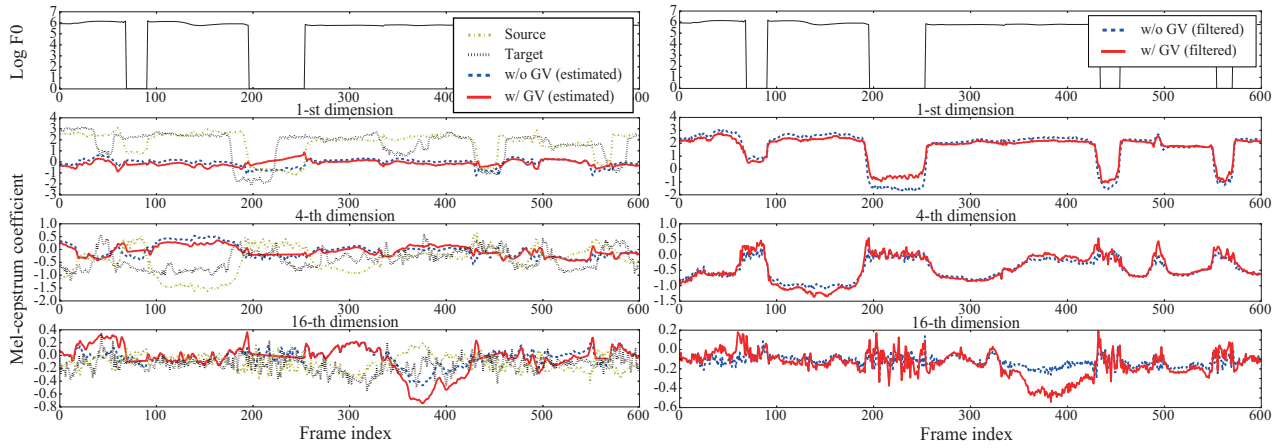

Figure 2: Example of trajectories of spectral feature sequences. Note that the duration of "Target" trajectories is different from the other trajectories.

4.3. Analysis of converted feature trajectories

To more deeply analyze what yields naturalness improvements in the proposed DIFFSVC method, we examine in detail the spectral feature trajectories of singing voices, which are given by

Source mel-cepstral coefficients extracted from the source singer's natural singing voice

Target mel-cepstral coefficients extracted from the target singer's natural singing voice

w/ GV (estimated) mel-cepstral coefficient differentials estimated with the proposed DIFFSVC method

w/ GV (filtered) mel-cepstral coefficients extracted from the singing voice converted in the proposed DIFFSVC method

w/o GV (estimated) mel-cepstral coefficient differentials estimated with the conventional DIFFSVC

w/o GV (filtered) mel-cepstral coefficients extracted from the singing voice converted in the conventional DIFFSVC method

Figure 2 shows the individual trajectories and the logarithmic F_0 trajectory. It can be observed from "Source" and "Target" that higher-order mel-cepstral coefficients tend to have rapidly varying fluctuations. It has been reported in [18] that these fluctuations are well modeled by the modulation spectrum and strongly affect speech quality of the converted speech. In the proposed method (w/ GV (estimated)), the converted feature differential trajectory is smoothly connected from the end of voiced segments to the start of voiced frames thanks to the proposed smoothing process at unvoiced frames. This yields a converted feature trajectory (w/ GV (filtered)) maintaining natural fluctuations at unvoiced frames. On the other hand, these fluctuations are obviously reduced in the conventional method (w/o GV (filtered)). We can also see that the GV of the converted feature trajectory at higher-order mel-cepstral coefficients is restored more effectively by the proposed method (w/ GV (filtered)) compared to the conventional method (w/o GV (filtered)). These results imply that the proposed method effectively approximates the target spectral fluctuations by using those of the source spectral trajectory and the GV of the target spectral trajectory.

Figure 3 shows the GVs calculated from several trajectories of mel-cepstral coefficients. The GV in the conventional

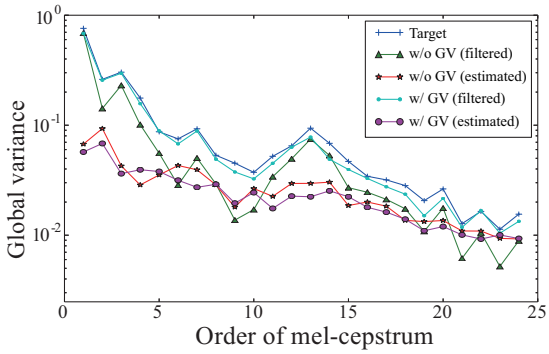

Figure 3: GVs of several mel-cepstral sequences.

method "w/o GV (filtered)" significantly decreases compared to that of "Target." On the other hand, the GV in the proposed method "w/ GV (filtered)" is close to that of "Target." This GV restoration yields significant improvements in speech quality of the converted singing voice. Note that the GV of the feature differential trajectories in the proposed method (w/ GV (estimated)) are still similar to those of the conventional method (w/o GV (estimated)). This shows the effectiveness of the proposed method modeling not the GV of the differential trajectory but the GV of the converted trajectory.

5. Conclusions

In order to improve quality of singing voice conversion based on direct waveform modification (DIFFSVC), we have proposed DIFFSVC considering global variance and smoothing of the conversion function at unvoiced frames. The experimental results have demonstrated that the proposed DIFFSVC method makes it possible to convert voice timbre of a source singer into that of a target singer with higher speech quality while not causing any adverse effects on the conversion accuracy of speaker identity compared to the conventional DIFFSVC method. In future work, we plan to apply the DIFFSVC framework to cross-gender conversion.

6. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Numbers: 26280060 and 15H02726, and by the JST OngaCREST project.

7. References

[1] H. Kenmochi and H. Ohshita, "VOCALOID – Commericial singing synthesizer based on sample concatenation," *Proc. INTERSPEECH*, pp. 4011–4012, Aug. 2007.

[2] K. Saino, M. Tachibana, and H. Kenmochi, "A singing style modeling system for singing voice synthesizers." *Proc. INTERSPEECH*, pp. 2894–2897, Sept. 2010.

[3] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, "Recent development of the HMM-based singing voice synthesis system - Sinsy," *SSW7*, pp. 211–216, Sept. 2010.

[4] T. Nakano and M. Goto, "VocaListener: A singing-to-singing synthesis system based on iterative parameter estimation," *Proc. SMC 2009*, pp. 343–348, July 2009.

[5] T. Nakano and M. Goto, "Vocalistener2: A singing synthesis system able to mimic a user's singing in terms of voice timbre changes as well as pitch and dynamics," *Proc. ICASSP*, pp. 453–456, May 2011.

[6] F. Villavicencio and J. Bonada, "Applying voice conversion to concatenative singing-voice synthesis," *Proc. INTERSPEECH*, pp. 2162–2165, Sept. 2010.

[7] Y. Kawakami, H. Banno, and F. Itakura, "GMM voice conversion of singing voice using vocal tract area function," *IEICE technical report. Speech (Japanese edition)*, vol. 110, no. 297, pp. 71–76, Nov. 2010.

[8] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. SAP*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[9] T. Toda, A. W. 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, Nov. 2007.

[10] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," *Proc. ICASSP*, pp. 1249–1252, Apr. 2007.

[11] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Many-to-many eigenvoice conversion with reference voice," *Proc. INTERSPEECH*, pp. 1623–1626, Sept. 2009.

[12] H. Doi, T. Toda, T. Nakano, M. Goto, and S. Nakamura, "Singing voice conversion method based on many-to-many eigenvoice conversion and training data generation using a singing-to-singing synthesis system," *Proc. APSIPA ASC*, Nov. 2012.

[13] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," *Proc. INTERSPEECH*, pp. 2514–2418, Sept. 2014.

[14] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *Proc. ICASSP*, pp. 1315–1318, June 2000.

[15] M. Sambur, "Selection of acoustic features for speaker identification," *IEEE Trans. ASSP*, vol. 23, no. 2, pp. 176–182, 1975.

[16] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, Apr. 1999.

[17] S. Imai, K. Sumita, and C. Furuichi, "Mel log spectrum approximation (mlsa) filter for speech synthesis," *Electronics and Communications in Japan (Part I: Communications)*, vol. 66, no. 2, pp. 10–18, 1983.

[18] S. Takamichi, T. Toda, A. Black, and S. Nakamura, "Modulation spectrum-based post-filter for gmm-based voice conversion," *APSIPA ASC*, Dec. 2014.