Statistical Singing Voice Conversion with Direct Waveform Modification based on the Spectrum Differential

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 a novel statistical singing voice conversion (SVC) technique with direct waveform modification based on the spectrum differential that can convert voice timbre of a source singer into that of a target singer without using a vocoder to generate converted singing voice waveforms. SVC makes it possible to convert singing voice characteristics of an arbitrary source singer into those of an arbitrary target singer. However, speech quality of the converted singing voice is significantly degraded compared to that of a natural singing voice due to various factors, such as analysis and modeling errors in the vocoder-based framework. To alleviate this degradation, we propose a statistical conversion process that directly modifies the signal in the waveform domain by estimating the difference in the spectra of the source and target singers' singing voices. The differential spectral feature is directly estimated using a differential Gaussian mixture model (GMM) that is analytically derived from the traditional GMM used as a conversion model in the conventional SVC. The experimental results demonstrate that the proposed method makes it possible to significantly improve speech quality in the converted singing voice while preserving the conversion accuracy of singer identity compared to the conventional SVC.

Index Terms: singing voice, statistical voice conversion, vocoder, Gaussian mixture model, differential spectral compensation

1. Introduction

The 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 difficulty in changing it widely (e.g. changing their own voice timbre into that of another singer) owing to physical constraints in speech production. If it would be possible for singers to freely control their voice timbre beyond their physical constraints, it will 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. To flexibly control the synthesized singing voice as the users want, there has also been proposed a technique capable of automatically adjusting parameters of the singing voice synthesis engine so that the variation of power and pitch in the synthesized singing voice is similar to that of the given user's natural singing voice [4, 5]. Although these technologies using singing voice synthesis engines are very effective to produce the singing voices desired by the users, it is essentially difficult to directly convert singers' singing voices in realtime.

Several singing voice conversion methods have been proposed to make it possible for a singer to sing a song with the desired voice timbre beyond their own physical constraints. One of the typical methods is singing voice morphing between singing voices of different singers or different singing styles [6] using the speech analysis/synthesis framework [7], which can only be applied to singing voice samples of the same song. To convert a singer's voice timbre in any song, statistical voice conversion (VC) techniques [8, 9] have been successfully applied to singing voice conversion. This singing VC (SVC) method makes it possible to convert a source singer's singing voice into another target singer's singing voice [10, 11]. 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 [12, 13] have also been successfully applied to SVC [14] to develop a more flexible SVC system capable of achieving conversion between arbitrary source and target singers even if a parallel data set is not available.

Although SVC has great potential to bring a new singing styles to singers, there remain several problems to be solved. One of the biggest problems is that speech quality of the converted singing voice is significantly degraded compared to that of the natural singing voice. Conventional SVC uses a vocoder to generate a waveform of the converted singing voice from the converted acoustic features. Consequently, speech quality of the converted singing voice suffers from various errors, such as F_0 extraction errors, modeling errors in spectral parameterization, and oversmoothing effects often observed in the converted acoustic features. It is essential to address these issues to allow for practical use of SVC.

In this paper, we propose a SVC method that can perform SVC without the waveform generation process based on a vocoder. In conventional SVC, spectral envelope, F_0, and aperiodic components are extracted from the source singer's singing voice and converted to the target singer's singing voice. However, in intra-gender SVC, it is not always necessary to convert F_0 values of the source singer to those of the target because both singers often sing on key. Moreover, the conversion of the aperiodic components usually causes only a small impact on the converted singing voice. Therefore, it is expected that only spectral conversion is sufficient to achieve acceptable quality in intra-gender SVC. Based on this idea, in the proposed SVC method, we focus only on converting the spectral envelope. The waveform of the source singer is directly modified with a digital filter that uses the time-varying difference in the spectral envelope between the source and target singer's singing voices. This spectrum differential is statistically estimated from the spectral envelop of the source singer. It is shown from results of subjective experimental evaluation that the proposed SVC method significantly improves speech quality of the converted singing voice compared to the conventional SVC methods.

2. Statistical singing voice conversion (SVC)

SVC consists of a training process and a conversion process. In the training process, a joint probability density function of acoustic features of the source and target singers' singing voices is modeled with a Gaussian mixture model (GMM) using a parallel data set in the same manner as in statistical VC for normal voices [11]. As the acoustic 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. Their joint probability density modeled by the GMM is given by

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

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal 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.

In the conversion process, the source singer's singing voice is converted into the target singer's singing voice using maximum likelihood estimation of speech parameter trajectory with the GMM [9]. Time sequence vectors of the source features and the target features are denoted as $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$, 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 $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ is determined as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) \text{ subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \quad (2)$$

where $\boldsymbol{W}$ is a transformation matrix to expand the static feature vector sequence into the joint static and dynamic feature vector sequence [15]. The conditional probability density function $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$ is analytically derived from the GMM of the joint probability density given by Eq. (1). To alleviate the over-smoothing effects that usually make the converted singing voice sound muffled, global variance (GV) [9] is also considered to compensate the variation of converted feature vector sequence.

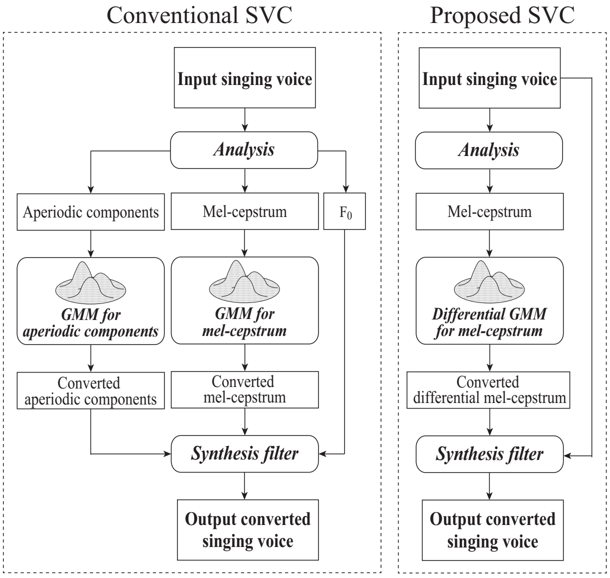

Figure 1: Conversion processes of conventional SVC (in Section 2) and proposed SVC methods (in Section 3).

3. SVC based on differential spectral compensation

Figure 1 shows the conversion processes of the conventional and proposed SVC methods. In the proposed method, the difference of the spectral features of the source and target singers is estimated from the source singer's spectral features using a differential GMM (DIFFGMM) modeling the joint probability density of the source singer's spectral features and the difference in the spectral features. Voice timbre of the source singer is converted into that of the target singer by directly filtering an input natural singing voice of the source singer with the converted spectral feature differential. The proposed SVC method doesn't need to generate excitation signals, which are needed in vocoder-based waveform generation. Therefore, the converted singing voice is free from various errors usually observed in the traditional SVC, such as F_0 extraction errors, unvoiced/voiced decision errors, spectral parameterization errors caused by liftering on the mel-cepstrum, and so on. On the other hand, the excitation parameters can not be converted in the proposed SVC method.

The DIFFGMM is analytically derived from the traditional GMM (in Eq. (1)) used in the conventional SVC. 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 $2D$-dimensional joint static and dynamic feature vector between the source and the differential features is given by

$$\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{D}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t - \boldsymbol{X}_t \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t \end{bmatrix}, \quad (3)$$

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{I} & \boldsymbol{I} \end{bmatrix}, \quad (4)$$

where $\boldsymbol{A}$ is a transformation matrix that transforms the joint feature vector between the source and target features into that of the source and difference features. $\boldsymbol{I}$ denotes the identity matrix. Applying the transformation matrix to the traditional

GMM in Eq. (1), the DIFFGMM is derived as follows:

$$P\left(\boldsymbol{X}_t, \boldsymbol{D}_t|\boldsymbol{\lambda}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t\\ \boldsymbol{D}_t\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)}\\ \boldsymbol{\mu}_m^{(D)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XD)}\\ \boldsymbol{\Sigma}_m^{(DX)} & \boldsymbol{\Sigma}_m^{(DD)}\end{bmatrix}\right), \quad (5)$$

$$\boldsymbol{\mu}_m^{(D)} = \boldsymbol{\mu}_m^{(Y)} - \boldsymbol{\mu}_m^{(X)}, \quad (6)$$

$$\boldsymbol{\Sigma}_m^{(XD)} = \boldsymbol{\Sigma}_m^{(DX)\top} = \boldsymbol{\Sigma}_m^{(XY)} - \boldsymbol{\Sigma}_m^{(XX)}, \quad (7)$$

$$\boldsymbol{\Sigma}_m^{(DD)} = \boldsymbol{\Sigma}_m^{(XX)} + \boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(XY)} - \boldsymbol{\Sigma}_m^{(YX)}. \quad (8)$$

The converted differential feature vector is determined in the same manner as described in Section 2. In this paper, the GV is not considered in the proposed SVC method based on the spectrum differential.

4. Experimental evaluation

4.1. Experimental conditions

We evaluated speech quality and singer identity of the converted singing voices to compare the conventional SVC and the proposed SVC. 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.

STRAIGHT [16] was used to extract spectral envelopes, which were parameterized to the 1-24th, 1-32th, and 1-40th mel-cepstral coefficients as spectral features. As the source excitation features for the conventional SVC, we used F_0 and aperiodic components in five frequency bands, i.e., 0-1, 1-2, 2-4, 4-6, and 6-8 kHz, which were also extracted by STRAIGHT [17]. The frame shift was 5 ms. The mel log spectrum approximation (MLSA) filter [18] was used as the synthesis filter in both the conventional and proposed methods.

We used 80 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 for the mel-cepstral coefficients was 128 and for the aperiodic components was 64.

Two preference tests were conducted. Speech quality of the converted singing voices was evaluated in the first preference test. The converted singing voice samples of the conventional SVC and the proposed SVC for the same phrase were presented to listeners in random order. The listeners selected which sample had better sound quality. On the other hand, the conversion accuracy of singer identity of the converted singing voices was evaluated in the other preference test. 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 SVC and the proposed SVC 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 8 and each listener evaluated 24 sample pairs in each order setting of the mel-cepstral coefficients. All listeners don't specialize in audio and they were allowed to replay each sample pair as many times as necessary.

4.2. Experimental results

Figure 2 indicates the results of the preference test for the speech quality. The proposed SVC makes it possible to generate the converted speech with better speech quality than the conventional SVC in any order setting of the mel-cepstral coefficients. This is assumed that the proposed SVC is free from various errors caused in the vocoder-based waveform generation, such as F_0 extraction errors or spectral modeling errors caused by liftering.

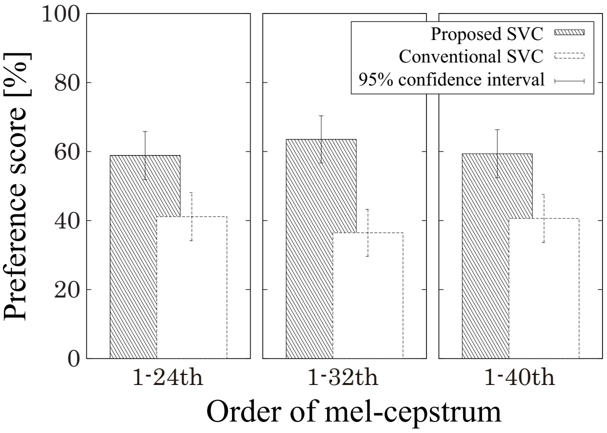

Figure 2: Evaluation of speech quality.

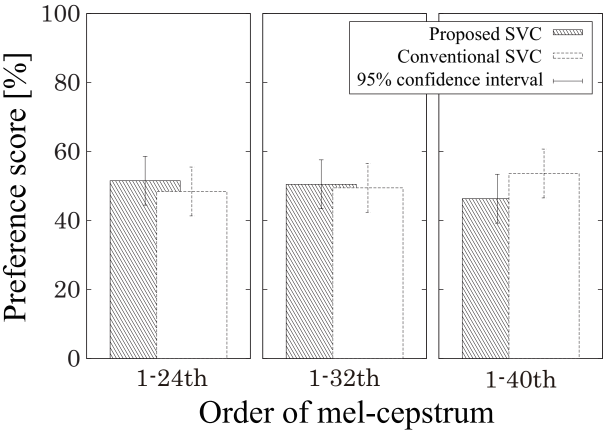

Figure 3: Evaluation of singer identity.

Figure 3 indicates the results of the preference test for the singer identity. The conversion accuracy of the singer identity of the proposed SVC is not statistically significantly different from that of the conventional SVC in any order setting of the mel-cepstral coefficients. This result suggests that the aperiodic components have little effect on the singer identity in singing voices, and even if the proposed SVC cannot convert the excitation features, the conversion accuracy of the singer identity still remains equivalent to that of the conventional SVC.

These results demonstrate that the proposed SVC is capable of converting the voice timbre with higher speech quality while causing no degradation in the conversion accuracy of singer identity compared to the conventional SVC. Note that the GV is considered in the conventional SVC while not considered in the proposed SVC.

4.3. Comparison of the converted spectral features

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

Source mel-cepstral coefficients extracted from the source

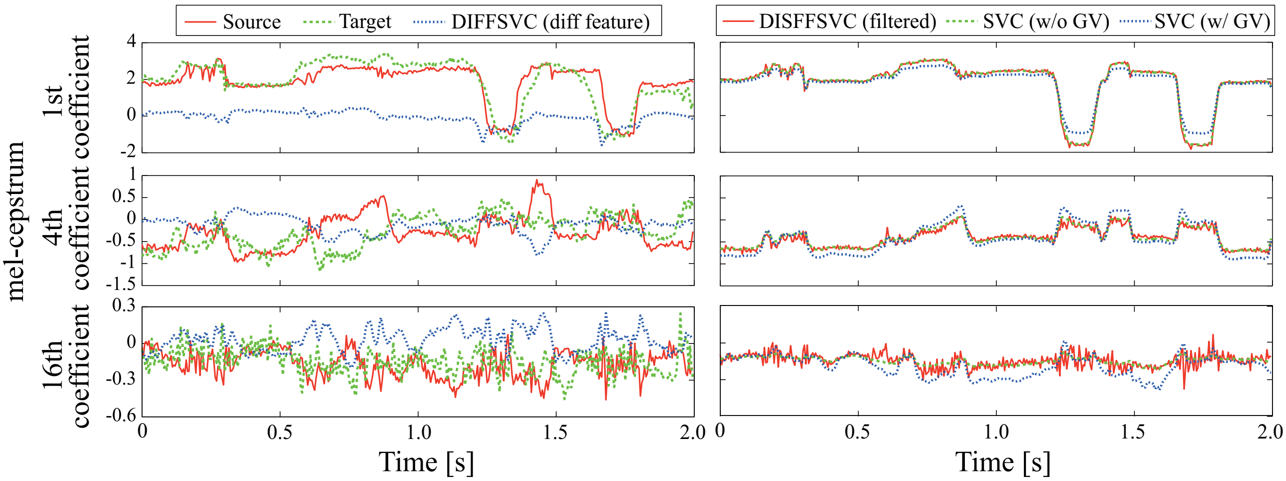

Figure 4: Example of trajectories of spectral feature sequence. Note that duration of "Target" trajectories is different from the other trajectories.

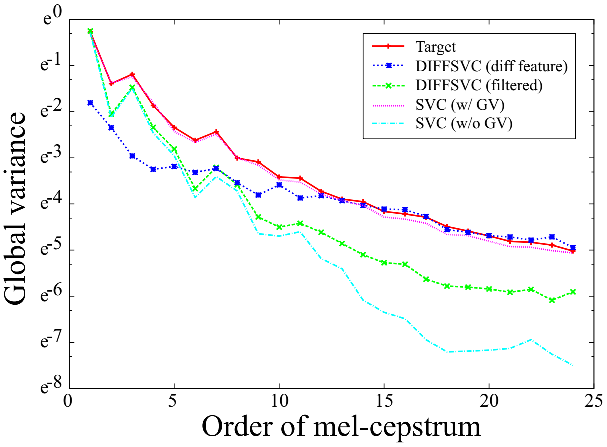

Figure 5: GVs of several mel-cepstral sequences.

singer's natural singing voice

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

DIFFSVC (diff feature) differences of mel-cepstral coefficients estimated with the differential GMM in the proposed SVC

DIFFSVC (filtered) mel-cepstral coefficients extracted from the singing voice converted in the proposed SVC

SVC (w/ GV) mel-cepstral coefficients estimated with the conventional GMM considering the GV

SVC (w/o GV) mel-cepstral coefficients estimated with the conventional GMM not considering the GV

Figure 4 shows trajectories of the mel-cepstral coefficients in each sample. It can be observed from "Source" and "Target" that higher-order mel-cepstral coefficients tend to have rapidly varying fluctuations. In other words, high modulation-frequency components tend to be larger as the order of the mel-cepstral coefficient is higher. On the other hand, such rapidly varying fluctuations are not observed in the trajectory of higher-order mel-cepstral coefficients of the "SVC (w/o GV)." They are still not observed even if considering the GV in "SVC (w/ GV)" although the GVs of higher-order mel-cepstral coefficients are recovered well. Therefore, these fluctuations are not modeled very well in SVC based on the conventional GMM. On the other hand, these fluctuations are still observed in "DIFFSVC (filtered)." Note that they do not appear in the estimated trajectories of the differences of mel-cepstral coefficients "DIFFSVC (diff feature)," which are estimated with the differential GMM in the proposed SVC. However, in the proposed SVC, the source singing voices are directly filtered to generate the converted singing voices. Therefore, these fluctuations observed in the source singing voices are still kept in the singing voices converted by the proposed SVC "DIFFSVC (filtered)." It is possible that the quality improvement is yielded by the proposed SVC because it generates converted trajectories having these fluctuations similar to those in natural singing voices.

Figure 5 shows the GVs calculated from trajectories of mel-cepstral coefficients. As reported in the previous work [9], the GVs of the converted mel-cepstral coefficients tend to be smaller in "SVC (w/o GV)" and this tendency is clearly observed especially in higher-order mel-cepstral coefficients, but the GVs are recovered by "SVC (w/ GV)," being almost equivalent to those of the target "Target." On the other hand, the GVs of the mel-cepstral coefficients in the proposed method "DIFFSVC (filtered)" tend to be smaller than those of the target. This tendency can also be observed in Figure 4. Note that the GV is not considered in the proposed method in this paper. It is expected that naturalness of the singing voices converted by the proposed SVC can be further improved by considering the GV so that the GVs of the filtered mel-cepstral coefficients are close to those of the target.

5. Conclusions

In order to improve quality of singing voice conversion (SVC), we proposed SVC with direct waveform modification based on the spectrum differential. The experimental results demonstrated that the proposed SVC makes it possible to convert voice timbre of a source singer into that of a target singer with higher speech quality compared to conventional SVC. In future work, we plan to implement a conversion algorithm consider in the global variance for the proposed method to further improve quality of the converted singing voice.

6. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 26280060 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*, 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] M. Morise, M. Onishi, H. Kawahara, and H. Katayose, "v. morish' 09: A morphing-based singing design interface for vocal melodies," *Proc. ICEC*, pp. 185–190, Sept. 2009.

[7] H. Ye and S. Young, "High quality voice morphing," *Proc. ICASSP*, vol. 1, pp. I–9–12, May 2004.

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

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "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] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and system straight," *Proc. MAVEBA*, Sept. 2001.

[18] 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.