Singing Voice Conversion Method Based on Many-to-Many Eigenvoice Conversion and Training Data Generation Using a Singing-to-Singing Synthesis System

Hironori Doi*, Tomoki Toda*, Tomoyasu Nakano†, Masataka Goto†, and Satoshi Nakamura*
* Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)
E-mail: {hironori-d, tomoki, s-nakamura}@is.naist.jp Tel: +81-743-72-5265
† National Institute of Advanced Industrial Science and Technology (AIST)
E-mail: {t.nakano, m.goto}@aist.go.jp

***Abstract*—The voice quality (identity) of singing voices is usually fixed in each singer. To overcome this limitation and enable singers to freely change their voice quality using signal-processing technologies, we propose a singing voice conversion method based on many-to-many eigenvoice conversion (EVC) that can convert the voice quality of an arbitrary source singer into that of another arbitrary target singer. Previous EVC-based methods required *parallel data* consisting of song pairs of a single reference singer and many prestored target singers for training a voice conversion model, but it was difficult to record such data. Our proposed method therefore uses a singing-to-singing synthesis system called *VocaListener* to generate parallel data by imitating singing voices of many prestored target singers with the system's singing voices. Experimental results show that our method succeeded in enabling people to sing a song with the voice quality of a different target singer even if only an extremely small amount of the target singing voice is available.**

I. Introduction

Singing voices are different from sounds of musical instruments because singing voices can usually convey the linguistic information of the lyrics as well as the pitch, dynamics, and voice quality. Although a singer can expressively control voice timbres to some degree, the singer usually has difficulty changing his or her voice quality (identity) of singing into that of another singer. This is due to physical constraints in speech production that limit expressive freedom when singing a song.

When using singing synthesis systems, people do not face such a limitation. Instead of using a physical body, people can artificially generate humanlike singing voices having different voice qualities by changing the singing synthesis parameters. Since 2007, many end users have started to use singing synthesis systems, such as Vocaloid2 [1] and Sinsy [2], to produce music, and the number of listeners enjoying synthesized singing voices has been increasing. Particularly in Japan, various compact discs that include synthesized vocal tracks have often appeared in the popular music chart [3]. In addition, several techniques that can change the timbre of a user's or synthesized singing voice have been proposed [4], [5], [6], [7]. However, it is still difficult to generate singing voices with arbitrary and desired voice qualities.

To make it possible for people to directly sing with a different specific voice quality, and thus overcome physical constraints, singing voice conversion has been proposed [8]. Statistical voice conversion techniques [9], [10], [11] are used to convert the singing voice quality of a source singer into that of a target singer. In this technique, a Gaussian mixture model (GMM) of the joint probability density of an acoustic feature between the source singer's singing voice and the target singer's singing voice is trained in advance using a special data set, called *parallel data*, that consists of song pairs of these two singers. The trained model is capable of converting 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.

Towards realizing a more flexible singing-voice conversion technique that enables singers to freely control the converted singing voice quality and is capable of rapidly adapting the conversion model to arbitrary singers, we propose a singing-voice conversion method based on two techniques: many-to-many eigenvoice conversion (EVC) [12] and training data generation using a singing-to-singing synthesis system [13]. Many-to-many EVC is a technique of conversion from the voice of an arbitrary source singer into the voice of an arbitrary target singer. An eigenvoice GMM (EV-GMM) [14] is trained in advance with multiple parallel data sets that consist of a single predefined singer, called a reference singer in this paper, and many prestored target singers. The EV-GMM is capable of easily adapting the source/target voice quality to that of a few of their given voice samples in a text-independent (lyrics-independent) manner. By using this flexible voice conversion technique, the proposed method enables any singer or end user to freely control their singing voice quality. Furthermore, to easily develop multiple parallel data sets from nonparallel singing voice data sets of many singers, we propose a technique for efficiently and effectively generating parallel data sets using a singing-to-singing synthesis system called *VocaListener* to artificially generate voices of the reference singer. In this paper, we conduct objective and subjective

experimental evaluations to demonstrate the effectiveness of the proposed methods.

This paper is organized as follows. In Section II, we describe the many-to-many EVC algorithm. In Section III, the singing-to-singing synthesis system used in this paper is explained. In Section IV, the proposed method of singing voice conversion is described. In Section V, experimental evaluations are described. Finally, we summarize this paper in Section VI.

II. Many-to-many EVC [12]

In this section, we describe many-to-many EVC as a technique for flexibly developing a conversion model for an arbitrary speaker pair. This technique consists of a training process, an adaptation process, and a conversion process.

A. Training process

As acoustic features of the reference speaker and the s^{th} target speaker, we employ two D-dimensional joint features, $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ and $\boldsymbol{Y}_t^{(s)} = [\boldsymbol{y}_t^{(s)\top}, \Delta\boldsymbol{y}_t^{(s)\top}]^\top$, consisting of D-dimensional static and dynamic features at frame t, respectively, where $\top$ denotes the transposition of the vector. The joint probability density of reference and target features is modeled with the EV-GMM as follows:

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}^{(s)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^{(s)\top}]^\top; \boldsymbol{\mu}_m^{(s)}, \boldsymbol{\Sigma}_m^{(X,Y)}\right) \quad (1)$$

$$\boldsymbol{\mu}_m^{(s)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{A}_m\boldsymbol{w}^{(s)} + \boldsymbol{b}_m \end{bmatrix}, \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}, \quad (2)$$

where $\boldsymbol{w}^{(s)} = [w^{(s)}(1), \cdots, w^{(s)}(J)]^\top$ is the target-speaker-dependent weight vector for controlling target voice quality. $\boldsymbol{\lambda}^{(EV)}$ is a canonical EV-GMM parameter set consisting of the weight α_m, the mean vector $\boldsymbol{\mu}_m^{(X)}$, the covariance matrix $\boldsymbol{\Sigma}_m^{(X,Y)}$, the bias vector $\boldsymbol{b}_m$, and the basis vectors $\boldsymbol{A}_m = [\boldsymbol{a}_m(1), \cdots, \boldsymbol{a}_m(J)]$ for the m^{th} mixture component, where the number of basis vectors is J. Acoustic features of an arbitrary target speaker are modeled by setting only $\boldsymbol{w}^{(s)}$ to the speaker's specific values. The other parameters are the same for all target speakers. The EV-GMM is trained by adaptive training using multiple parallel data sets consisting of utterance pairs of a reference and many prestored target speakers [15].

B. Adaptation and conversion process

In the adaptation process, the EV-GMM is adapted to an arbitrary source speaker and an arbitrary target speaker by independently estimating the speaker-dependent weight vector using a few speech samples. The weight vector for source speaker $\hat{\boldsymbol{w}}^{(i)}$ is estimated as

$$\hat{\boldsymbol{w}}^{(i)} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \prod_{t=1}^{T} \int P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(i)}|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}\right) d\boldsymbol{X}_t \quad (3)$$

where $\boldsymbol{Y}_t^{(i)}$ is the acoustic features of the given source speaker's voice at frame t. The weight vector for the target speaker $\hat{\boldsymbol{w}}^{(o)}$ is estimated in the same manner. Then, the joint probability density of the acoustic features between the source speaker's voice and the target speaker's voice is derived as

$$\begin{aligned} &P\left(\boldsymbol{Y}_t^{(i)}, \boldsymbol{Y}_t^{(o)}|\hat{\boldsymbol{w}}^{(i)}, \hat{\boldsymbol{w}}^{(o)}, \boldsymbol{\lambda}^{(EV)}\right) \\ &= \sum_{m=1}^{M} P(m|\boldsymbol{\lambda}^{(EV)}) \int P\left(\boldsymbol{Y}_t^{(i)}|\boldsymbol{X}_t, m, \boldsymbol{w}_t^{(i)}, \boldsymbol{\lambda}^{(EV)}\right) \\ &\quad P\left(\boldsymbol{Y}_t^{(o)}|\boldsymbol{X}_t, m, \boldsymbol{w}_t^{(o)}, \boldsymbol{\lambda}^{(EV)}\right) P\left(\boldsymbol{X}_t|m, \boldsymbol{\lambda}^{(EV)}\right) d\boldsymbol{X}_t \\ &= \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{y}^{(i)} \\ \boldsymbol{y}^{(o)} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^{(i)} \\ \boldsymbol{\mu}_m^{(o)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(YXY)} \\ \boldsymbol{\Sigma}_m^{(YXY)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}\right), \end{aligned} \quad (4)$$

where

$$\boldsymbol{\Sigma}_m^{(YXY)} = \boldsymbol{\Sigma}_m^{(YX)}\boldsymbol{\Sigma}_m^{(XX)^{-1}}\boldsymbol{\Sigma}_m^{(XY)}. \quad (5)$$

In the conversion process, the converted static feature sequence vector is estimated using the adapted EV-GMM. Maximum likelihood estimation considering dynamic features and a global variance [11] is adopted.

III. Singing-to-Singing synthesis

As the singing synthesis system, a text-to-singing approach, which synthesizes a singing voice from note-level score information of the melody with its lyrics, such as Vocaloid2 [1], and a speech-to-singing approach, which synthesizes a singing voice from speech samples of read lyrics by controlling acoustic features [16], have been proposed. Moreover, singing-to-singing synthesis, which automatically synthesizes a more naturally sounding singing voice by estimating the parameters of the text-to-singing system from a target singing voice, has been proposed [13]. VocaListener [13] is the system used for the estimation part of singing-to-singing synthesis. VocaListener estimates parameters of pitch and dynamics for the singing synthesis system so that the synthesized singing voice becomes more similar to the target singing voice. If a user's singing voice and the corresponding lyrics without any score information are available, VocaListener can synchronize them automatically to determine the musical note corresponding to each phoneme of the lyrics. Thus, the singing-to-singing synthesis system allows users to easily, speedily, and effectively synthesize the singing voice.

IV. Proposed singing voice conversion method

Figures 1 and 2 show the training process of a conventional method using a standard GMM that converts the source singer's voice into the target singer's voice and the training and adaptation process of the proposed method using the EV-GMM that converts an arbitrary source singer's voice into an arbitrary target singer's voice, respectively.

A. Singing voice conversion based on many-to-many EVC

As shown in Fig. 1, in the conventional method, the source and target singers need to sing the same songs for several minutes to generate parallel data for training the standard GMM. On the other hand, as shown in the right part in Fig. 2, in the proposed method, the GMM for the arbitrary source

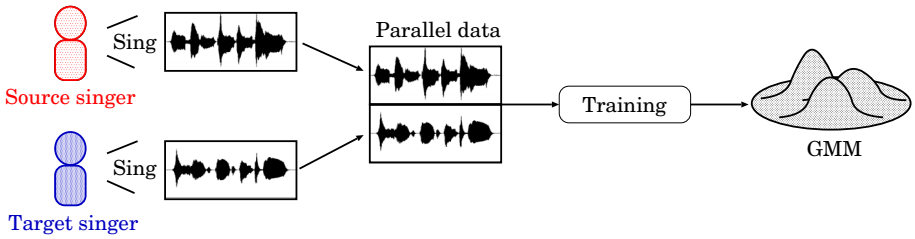

Fig. 1. *Training process of conventional singing voice conversion.*

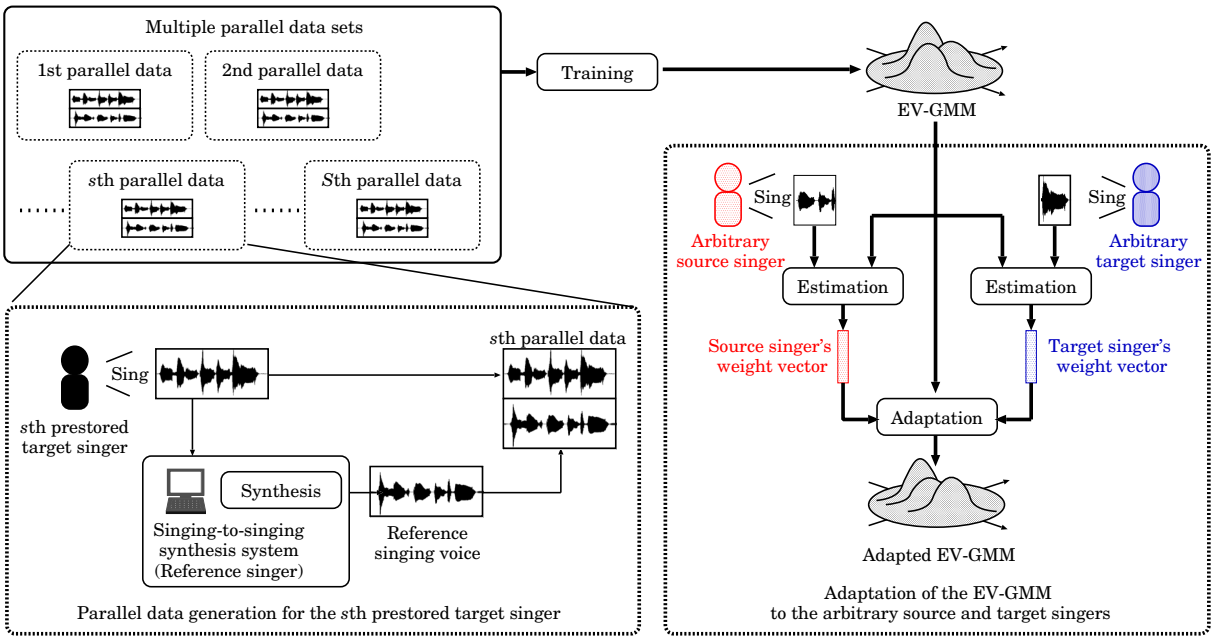

Fig. 2. *Training and adaptation process of proposed singing voice conversion.*

and target singers is easily built by adapting the EV-GMM to each of these singers using weight vectors that are separately estimated with only a small amount of their singing voices, e.g., for only several seconds. Moreover, since the adaptation of the EV-GMM is performed in a text-independent manner, the singing voices of any song are usable as adaptation data; i.e., the source and target singers need not sing the same song. Therefore, the proposed method effectively reduces the amount of effort required to prepare singing voices of the source and target singers. Note that real-time singing voice conversion is also achieved by using the low-delay conversion algorithm [17].

B. Training data generation with singing-to-singing synthesis

Although the proposed singing voice conversion method based on many-to-many EVC effectively and rapidly develops the voice conversion model for the arbitrary source and target singers, it needs to train the EV-GMM in advance using multiple parallel data sets consisting of singing voice pairs of the single reference singer and many prestored target singers. The development of these parallel data sets is laborious work. To address this issue, we artificially generate singing voices of the reference singer by applying a singing-to-singing synthesis system to singing voices of many prestored target singers. In this approach, we need to prepare only singing voices of multiple prestored target singers who need not sing the same song; these are available in existing databases, such as the RWC Music Database [18]. As shown in the left part in Fig. 2, for the singing voices of each prestored target singer, corresponding singing voices of the reference singer are artificially generated by using the singing-to-singing synthesis system. Then, a parallel data set is developed between singing voices of each prestored target singer and the artificially generated singing voices of the system's singer as the reference singer. Finally the EV-GMM is trained using many of the developed multiple parallel data sets. Thus, this training data generation approach can efficiently and effectively develop parallel data sets without recording singing voices of the reference singer.

In the singing-to-singing synthesis, some acoustic parameters, such as phoneme duration and vibrato, of the synthesized singing voice are automatically tuned so that they are similar to those of the given target singing voice. Therefore, the synthesized reference singing voice more closely corresponds to the target singing voice than a singing voice sung by a real singer. Moreover, the voice quality of the singing voices

generated by the singing-to-singing synthesis system is more consistent, regardless of the song, genre of music, or the singer's physical condition than that of the singing voice of a real singer. These are favorable properties for training the EV-GMM since, in the EV-GMM training algorithm, the reference singing voices are assumed to basically have consistent voice quality over all multiple parallel data sets.

V. Experimental evaluations

To demonstrate the effectiveness of the proposed singing voice conversion system, we conducted experimental evaluations.

A. Experimental conditions

We used solo singing voices of 30 Japanese songs in the RWC Music Database [18] as the prestored target singing voices. As the reference singing voices, we used singing voices synthesized using the singing-to-singing synthesis system *VocaListener* with a singer database called *Hatsune Miku* [19] based on Vocaloid2. As adaptation and test data of source/target singing voices, we used the original female solo singing voices of two Japanese songs in the RWC Music Database (RWC-MDB-P-2001 No.35 and No.71), which were not included in the above 30 songs, and also recorded singing voices of another female singer for the same two songs.

The 1st through 24th mel-cepstral coefficients were used as a spectral parameter. STRAIGHT analysis [20] was employed to extract these coefficients from singing voices. The shift length was 5 ms and the sampling frequency was 16000 Hz.

The EV-GMM for spectral conversion was trained from 30 parallel data sets consisting of the synthesized reference singing voices and the prestored target singing voices. The number of basis vectors of the EV-GMM was set to 29. The number of mixture components of the EV-GMM was set to 128. As a conventional approach for reference (i.e., singing voice conversion [8]), we also trained a standard GMM for spectral conversion using a parallel data set consisting of the source and target singing voices. The number of mixture components of the GMM was preliminarily optimized so that spectral conversion accuracy was maximized in the test data. As the training data for the conventional method and the adaptation data for the proposed method, 2, 4, 8, 16, 32, or 64% of the singing parts of a song (RWC-MDB-P2001 No.35) sung by the source and target singers were used, and then, the remaining 36% of data were used for the test. The total length of this song was 193 s including 116 s of singing parts. We evaluated two conditions of the song setting; 1) the same-song condition, where the same song (RWC-MDB-P2001 No.35) is used in both the training/adaptation process and the test process, and 2) the different-song condition, where different songs are used in the training/adaptation process (RWC-MDB-P2001 No.35) and the test process (RWC-MDB-P2001 No.71).

B. Objective evaluation

We evaluated the spectral conversion accuracy of the conventional singing voice conversion and the proposed singing voice conversion using mel-cepstral distortion between the converted and target mel-cepstra as an evaluation metric. Figure 3 shows mel-cepstral distortion as a function of the amount of singing voice adaptation data used in the proposed method or the amount of parallel data of singing voice pairs used in the conventional method. Under the same-song condition, we can see that the proposed method yields better spectral conversion accuracy than the conventional method when using a small amount of available data of the source/target singers (i.e., 2, 4, and 8%). Note that the proposed method does not require the use of parallel data in adaptation but the conventional method needs to use them.

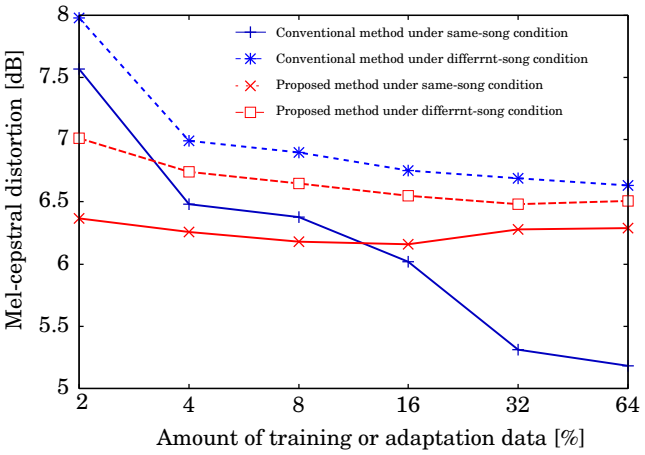

Fig. 3. *Mel-cepstral distortion as a function of amount of target singing voice data (i.e., singing voice pairs in conventional method or singing voice adaptation data in proposed method).*

Under the different-song condition, the conventional method shows much lower conversion accuracy than under the same-song condition. This is because the voice quality of the singing voice significantly changes depending on the song even if the same singer sings. On the other hand, it is observed that the proposed method reduces this degradation. Since the EV-GMM is trained with many singers' voices, it is more robust against variations of the singing voice quality.

C. Subjective evaluation

We conducted an opinion test of voice quality and a preference test of singer individuality. The opinion score in the opinion test was set to a 5-point scale (i.e., 1 (very poor) to 5 (excellent)). In this test, listeners heard each converted singing voice sample, then they judged the voice quality of each sample using the opinion score. In the preference test, listeners heard a target singing voice sample and two converted singing voice samples, then they chose the converted singing voice sample that had more similar singer individuality to the target singing voice sample. The preference test was independently performed under the same- or different-song condition. In these tests, 5 listeners evaluated 8 types of singing voices generated under various conditions consisting of all combinations of the following: 2% or 64% of training/adaptation data, the

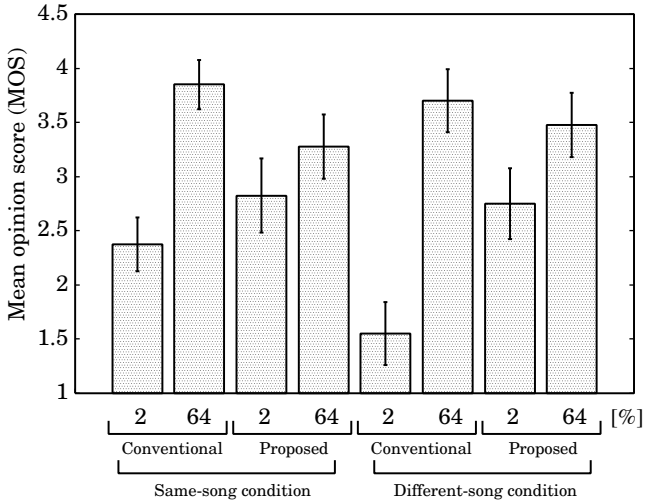

Fig. 4. *Result of opinion test of voice quality.*

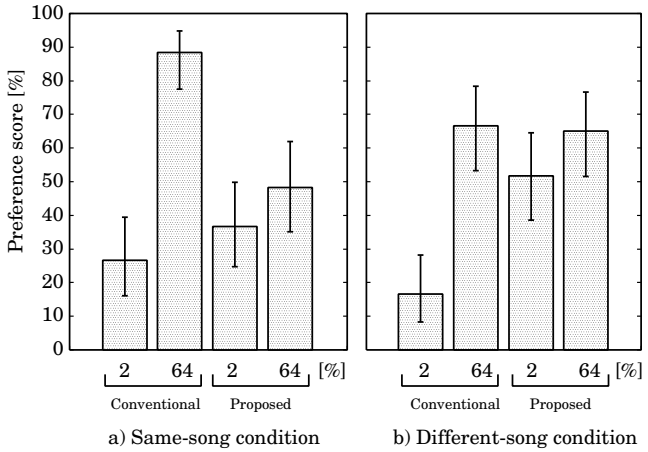

Fig. 5. *Result of preference test of singer individuality in a) same-song condition and b) different-song condition.*

conventional or proposed method, and the same- or different-song condition.

Figure 4 shows the result of the opinion test of voice quality. Under the same-song condition, the proposed method using only 2% adaptation data yields better voice quality than the conventional method using 2% parallel training data. This is because the proposed method effectively uses the prestored target singers' data to develop the conversion model for the new source and target singers. An increase in the adaptation data from 2% to 64% yields further voice quality improvements in the proposed method but it does not reach that in the conventional method using 64% parallel training data. Note that the proposed method does not use the parallel data set of the source and target singers, but the conventional method does.

Under the different-song condition, a decrease in the parallel training data from 64% to 2% in the conventional method causes much greater voice quality degradation than that under the same-song condition. This is because the acoustic characteristics of singing voices are considerably different between different songs. On the other hand, the proposed method yields almost the same voice quality as that under the same-song condition even when using only 2% adaptation data. As also observed in the objective evaluation, the proposed method is robust against variations of the singing voice quality. An increase in the adaptation data from 2% to 64% yields further voice quality improvements, and the voice quality is almost equal to that in the conventional method using 64% parallel training data. It is again worth noting that the proposed method does not use the parallel data but the conventional method does.

Figure 5 shows the results of the preference test of singer individuality. The preference score was calculated as the ratio of the number of samples selected as having better singer individuality to the number of samples presented to the listeners. We can see basically the same tendency as that observed in the results of the opinion test of voice quality. The proposed method also yields a better result in conversion accuracy for singer individuality when only 2% adaptation data is available compared with the conventional method using 2% parallel training data. This difference between the proposed and conventional methods is much larger under the different-song condition than under the same-song condition. An increase in the adaptation data from 2% to 64% in the proposed method also yields further improvements. Although the result of the proposed method is still significantly worse than the that of the conventional method when using 64% parallel training data under the same-song condition, they are almost equal under the different-song condition.

The above results suggest that 1) the proposed method yields better voice quality and conversion accuracy for singer individuality than the conventional method when a small amount of singing voice data of the source and target singers is available; 2) the proposed method is capable of effectively using nonparallel data of the source and target singers to rapidly develop the conversion model between these singers, and 3) since the proposed method is robust against variations of the singing voice quality often observed between different songs, it works reasonably well even when using different songs between the adaptation and conversion processes.

VI. Conclusions

We have presented a singing voice conversion method based on many-to-many EVC and training data generation using a singing-to-singing synthesis system. Our proposed method is capable of converting the singing voice quality of an arbitrary source singer into that of an arbitrary target singer by adapting a small number of adaptive parameters of a conversion model using an extremely small amount of source and target singing voice data. Moreover, our proposed method can alleviate the burden of having to record singing voices to develop parallel data sets, by using a singing-to-singing synthesis system. The experimental result demonstrated that the proposed method enables the effective conversion of a singing voice between an

arbitrary speaker pair even when using only several seconds of their singing voices as adaptation data.

VII. Acknowledgements

This work was supported in part by a MEXT Grant-in-Aid for Young Scientists (A). The authors are grateful to professor Hideki Kawahara of Wakayama University, Japan, for permission to use the STRAIGHT analysis-synthesis method.

References

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

[2] 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.

[3] H. Kenmochi, "VOCALOID and Hatsune Miku phenomenon in Japan," *Proc. InterSinging*, pp. 1–4, Oct. 2010.

[4] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno, "Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown," *Proc. ICASSP*, pp. 3905–3908, Apr. 2009.

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

[6] T. Nakano and M. Goto, "Vocalistener2: A singing synthesis sytem 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 2012.

[7] Y. Yoshida, R. Nishimura, T. Irino, and H. Kawahara, "Vowel-based voice conversion and its application to singing-voice manipulation," *Proc. AES 35th International Conference: Audio for Games*, no. 6, Feb. 2009.

[8] Y. Kawakami, H. Banno, and F. Itakura, "GMM voice conversion of singing voice using vocal tract area function," *IEICE technical report. Speech 110(297) (Japanese edition)*, pp. 71–76, Nov. 2010.

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

[10] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. ICASSP*, pp. 285–288, May 1998.

[11] 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.

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

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

[14] 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.

[15] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Adaptive training for voice conversion based on eigenvoices," *IEICE Trans. Inf. and Syst.*, vol. E93-D, no. 6, pp. 1589–1598, June 2010.

[16] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voice," *Proc. WASPAA*, pp. 215–218, Oct. 2007.

[17] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *Proc. INTERSPEECH*, pp. 1076–1079, Sept. 2008.

[18] M. Goto, T. Nishimura, H. Hashiguchi, and R. Oka, "RWC Music Database: Music genre database and musical instrument sound database," *Proc. ISMIR*, pp. 229–230, Oct. 2003.

[19] Crypton Future Media, "What is the "HATSUNE MIKU movement"?" 2012. [Online]. Available: http://www.crypton.co.jp/miku_eng

[20] 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.