Evaluation of a Singing Voice Conversion Method Based on Many-to-Many Eigenvoice Conversion

Hironori Doi[1], *Tomoki Toda*[1], *Tomoyasu Nakano*[2], *Masataka Goto*[2], *Satoshi Nakamura*[1]

[1]Graduate School of Information Science,
Nara Institute of Science and Technology, Nara, Japan
[2]National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki, Japan
[1]{hironori-d, tomoki, s-nakamura}@is.naist.jp, [2]{t.nakano, m.goto}@aist.go.jp

Abstract

In this paper, we evaluate our proposed singing voice conversion method from various perspectives. To enable singers to freely control their voice timbre of singing voice, we have proposed a singing voice conversion method based on many-to-many eigenvoice conversion (EVC) that enables to convert the voice timbre of an arbitrary source singer into that of another arbitrary target singer using a probabilistic model. Furthermore, to easily develop training data consisting of multiple parallel data sets between a single reference singer and many other singers, a technique for efficiently and effectively generating the parallel data sets from nonparallel singing voice data sets of many singers using a singing-to-singing synthesis system have been proposed. However, we have never conducted sufficient investigations into the effectiveness of these proposed methods. In this paper, we conduct both objective and subjective evaluations to carefully investigate the effectiveness of proposed methods. Moreover, the differences between singing voice conversion and speaking voice conversion are also analyzed. Experimental results show that our proposed method succeeds in enabling people to control their own voice timbre by using only an extremely small amount of the target singing voice.

Index Terms: singing voice, voice conversion, eigenvoice conversion, singing-to-singing synthesis, performance evaluation

1. Introduction

Range of singing voice timbre that can be produced by individual singers is limited by physical constraints. To produce a singing voice beyond physical constraints, many approaches have been studied. One of the most popular approaches is the use of singing synthesis systems, which generate a singing voice from several pieces of information such as lyrics and the musical score. Among them, 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 Sinsy [2] is popular in Japan. 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 [3]. VocaListener [3], which is the system used for the estimation part of singing-to-singing synthesis, 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. However, it is still difficult to generate singing voices with arbitrary and desired voice timbre.

To make it possible for people to directly sing with a different specific voice timbre, and thus overcome physical constraints, singing voice conversion has been proposed [4]. Statistical voice conversion (VC) techniques [5, 6, 7] are used to convert the singing voice timbre of a source singer into that of a target singer. In this technique, 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 *a parallel data set*, that consists of pairs of songs of the 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 for any song while keeping the linguistic information of the lyrics unchanged. Moreover, real-time singing voice conversion can also be achieved using the low-delay conversion algorithm [8].

Towards realizing a more flexible singing voice conversion technique, we have proposed a singing voice conversion method [9] based on many-to-many eigenvoice conversion (EVC) [10]. Many-to-many EVC is a technique of converting from the voice of an arbitrary source singer into that of an arbitrary target singer. An eigenvoice GMM (EV-GMM) [11] is trained in advance using 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 timbre to that of its given voice samples in a text-independent (lyrics-independent) manner. Furthermore, we have proposed a technique for efficiently and effectively generating parallel data sets using a singing-to-singing synthesis system to artificially generate singing voices of the reference singer.

In this paper, we describe our proposed methods [9] and evaluate their effectiveness. A comparison between VC and EVC based singing voice conversion is conducted from various perspectives. Moreover, to analyze the differences between speaking voice and singing voice in voice conversion, we conduct comparison between singing voice conversion using EV-GMM trained from speaking voice and from singing voice.

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

In this section, we describe singing voice conversion method based on many-to-many EVC and training data generation using singing-to-singing synthesis system.

2.1. Training data generation

The development of parallel data sets consisting of singing voice pairs of the single reference singer and many prestored target singers is laborious work. To address this issue, we have artificially generated 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 [12]. 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. Thus, this training data generation approach can efficiently and effectively develop parallel data sets without recording singing voices of the reference singer.

2.2. Training process

As acoustic features of the reference singer and the s^{th} prestored target singer, 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 spectral 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:

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

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

where $\boldsymbol{w}^{(s)} = [w^{(s)}(1), \cdots, w^{(s)}(J)]^\top$ is the target-speaker-dependent weight parameter for controlling target voice timbre. $\lambda^{(EV)}$ 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. To alleviate the degradation of performance of EV-GMM caused by effects of acoustic variation of the many prestored target singers, the EV-GMM is trained by speaker adaptive training (SAT) [13, 14] using multiple parallel data sets consisting of utterance pairs of a reference and many prestored target singers.

2.3. Adaptation and conversion process

In the adaptation process, the EV-GMM is adapted to an arbitrary source singer and an arbitrary target singer by independently estimating the singer-dependent weight parameter using a few singing voice samples. The weight parameter for source singer $\hat{\boldsymbol{w}}^{(i)}$ is estimated by maximum a posteriori (MAP) [15, 16] as

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

where $\lambda^{(\omega)}$ is a model parameter set consisting of the mean vector $\boldsymbol{\mu}^{(w)}$ and the covariance matrix $\boldsymbol{\Sigma}^{(\omega)}$. This model parameter set is trained in advance using a set of weight parameters estimated for individual prestored target singer. $\boldsymbol{Y}_t^{(i)}$ is the acoustic features of the given source singer's voice at frame t. The balance between $P\left(\boldsymbol{w}|\lambda^{(w)}\right)$ and $\prod_{t=1}^{T} P\left(\boldsymbol{Y}_t^{(i)}|\lambda^{(EV)}, \boldsymbol{w}\right)$ is controlled by the hyperparameter τ. The weight parameter for the target singer $\hat{\boldsymbol{w}}^{(o)}$ is estimated in the same manner. On the other hand, our proposed method allows user to freely control voice timbre of the converted singing voice by manipulating the target singer's weight parameters.

Then, the joint probability density of the acoustic features between the source singer's voice and the target singer's voice is derived as

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

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 [6] is adopted. Note that real-time singing voice conversion is also achieved by using the low-delay conversion algorithm [8].

3. Experimental evaluations

To demonstrate effectiveness of our proposed method and investigate the differences between singing voice conversion and speaking voice conversion, four types of conversion model were compared.

VC conventional singing voice conversion based on VC [6]

EVC-human proposed singing voice conversion based on many-to-many EVC with conventional training data generation using a human voice as the reference singer's voice

EVC-synth proposed singing voice conversion based on many-to-many EVC with training data generation using singing-to-singing synthesis

EVC-speaking conventional many-to-many EVC for a speaking voice

3.1. Experimental conditions

In this evaluation, only the spectral feature is converted in all conversion methods because the voice timbre strongly depends on the spectral feature. The 1^{th} to 24^{th} mel-cepstral coefficients were used as a spectral feature. STRAIGHT analysis [17] was employed to extract these coefficients from singing voices. F_0 and the aperiodic components of the source singer are directly used to synthesize the converted singing voice. The shift length was 5 ms and the sampling frequency was 16000 Hz.

We used the solo singing voices of 30 Japanese songs in the RWC Music Database [12] as the prestored target singing voices to train EV-GMM. The phoneme balance was not considered in these songs. For EVC-human, the solo singing voices of one male singer were used as the singing voices of the reference singer. For EVC-synth, singing voices synthesized using the singing-to-singing synthesis system *VocaListener* with a singer database called *Hatsune Miku* [18] based on Vocaloid2 were used as the reference singer. The number of basis vectors of the EV-GMMs was set to 29 and the number of mixture components of the EV-GMMs was set to 128. On the other hand, in EVC-speaking, we used parallel data sets of a single reference male speaker and 152 prestored target speakers to train the EV-GMM. These speakers were from the Japanese Newspaper Article Sentence (JNAS) database. Each prestored target speaker uttered one of seven subsets. Each subset consists of 50 phonetically balanced sentences. The EV-GMM for spectral conversion was trained from 152 parallel data sets consisting of the recorded reference speaking voices and the prestored target speaking voices. The number of basis vectors of the EV-GMMs was set to 151 and the number of mixture components of the EV-GMMs was set to 128.

For the adaptation and testing of the EV-GMMs and for the training and testing of the GMM, we selected two Japanese songs from the RWC Music Database (RWC-MDB-P-2001 No.46 and No.76), which were not included in the above 30 songs. Then, 5 singers (four male singers and one female singer) sang these two songs. Thus, as adaptation/training data and test data, we prepared 10 songs consisting of two songs sung by each singer. As the training data for the VC-based method and the adaptation data for the EVC-based methods, 2, 4, 8, 16, 32, or 64% of the sung parts of songs sung by the source and target singers was used, then, the remaining 36% of data was used for the test. The GMM and EV-GMMs were prepared for all combinations of the source and target singers. Thus, for each method, 20 conversion models (10 models × 2 song) were prepared. The weight parameters of the source and target singer were independently estimated using the spectral features from the source and target singing voice samples. The hyperparameter of MAP adaptation shown in eq. (3) was preliminarily optimized in each method. In this evaluation, it was set to 250, 1000, and 100 for EVC-human, EVC-synth, and EVC-speaking, respectively. For VC, 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 the spectral conversion accuracy was maximized in the test data.

3.2. Objective evaluation

We evaluated two conditions of song setting: 1) the same-song condition, where the same song 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 and the test process. Figure 1 shows mel-cepstral distortion as a function of the amount of the singing voice adaptation data used in the EVC-based methods or the amount of parallel data of the singing voice pairs used in the VC-based method under the same-song condition. Figure 2 shows those under the different-song condition. In fig. 1 and 2, horizontal axis represents percentage of data that is used for training or adaptation from the sung parts of songs.

Under the same-song condition, when using a small amount of training/adaptation data, EVC-speaking is the best, EVC-human is the next, EVC-synth is the next, and VC is the worst in conversion accuracy. Although EVC-speaking exhibits the highest conversion accuracy, the differences from EVC-human are not so large even if the amount of training data for EVC-speaking is significantly larger than that for EVC-human. When using a large amount of training/adaptation data, VC is the best, EVC-speaking is the next, EVC-human is the next, and the EVC-synth is the worst in conversion accuracy. Note that the EVC-based methods do not require the use of parallel data in the adaptation, in contrast to VC.

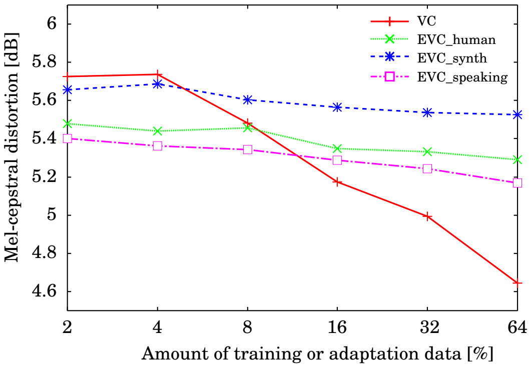

Figure 1: *Mel-cepstral distortion as a function of amount of target singing voice data (i.e., singing voice pairs in VC-based method or singing voice adaptation data in EVC-based methods) under the same-song condition.*

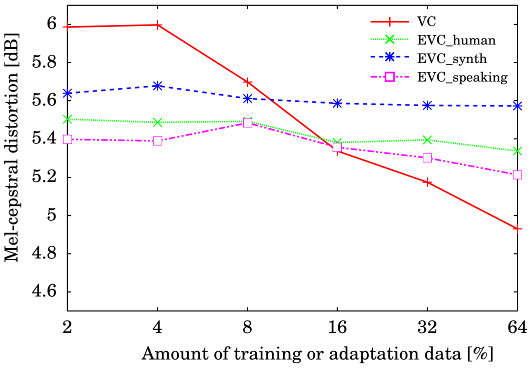

Figure 2: *Mel-cepstral distortion as a function of amount of target singing voice data (i.e., singing voice pairs in VC-based method or singing voice adaptation data in EVC-based methods) under the different-song condition.*

Under the different-song condition, VC has much lower conversion accuracy than under the same-song condition. This is because the voice timbre of the singing voice of a singer significantly changes depending on the song. On the other hand, it is observed that the EVC-based methods reduce this degradation. Since the EV-GMM is trained with many singers' voices, it is more robust against variations of the singing voice timbre.

3.3. Subjective evaluation

We conducted an opinion test on the naturalness of the singing voice and a preference test on singer individuality. The opinion was expressed using a five point scale (i.e., 1 (very poor) to 5 (excellent)). In this test, 10 listeners heard 16 types of converted singing voice sample, then they judged the naturalness 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 with more similar singer individuality to the target singing voice sample. The preference test was performed under the different-song condition because of its greater realism than same-song condition. In this tests, 9 listeners evaluated eight types of the singing voice generated under the different-song condition for all combinations of 2% or 64% of training/adaptation data and four types of conversion method.

Figure 3 shows the result of the opinion test on the naturalness of the singing voice. Under the same condition, VC us-

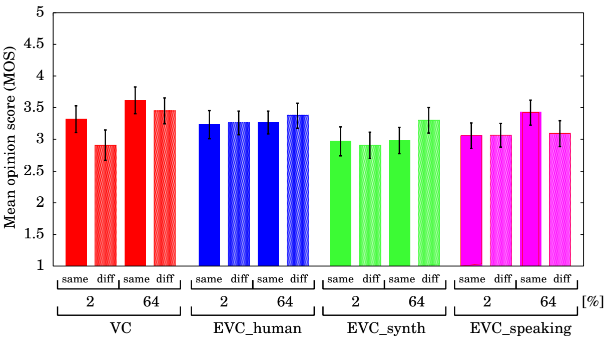

Figure 3: *Result of opinion test on naturalness.*

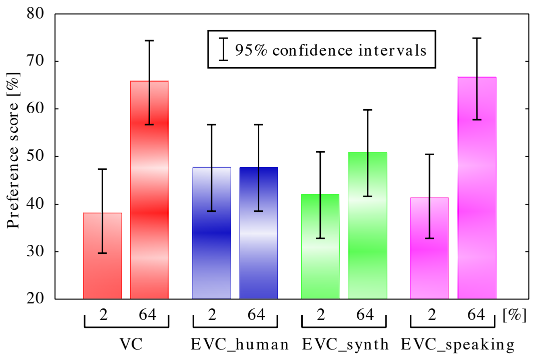

Figure 4: *Result of preference test on singer individuality under the different-song condition.*

ing 2% training data shows similar naturalness to that of EVC-human using 2% adaptation data in contrast to objective evaluation. On the other hand, the naturalness of EVC-speaking is not higher than that of other methods when using a small amount of adaptation data. This result suggests that it is difficult for EV-GMM trained with speaking voice to generate converted singing voice having high naturalness even if a large amount of speaking voice is available as training data. Other results show similar tendency to that observed in the result of the objective evaluation.

Figure 4 shows the result of the preference test on 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. When using a small amount of training/adaptation data, EVC-human is the best, EVC-synth is the next, EVC-speaking is the next, and VC is the worst in preference score of singer individuality. On the other hand, when using a large amount of training/adaptation data, VC and EVC-speaking show higher preference score of singer individuality than other methods. Note that VC requires the parallel data set of the source and target singers and the canonical EV-GMM of EVC-speaking is trained with significantly larger amount of training data than that of EVC-human and EVC-synth.

3.3.1. Comparison of each EV-GMMs

Figure 5 shows the cumulative distribution of occupancies of the canonical EV-GMM of EVC based methods. These individual mixture component occupancies have been calculated from all parallel data set in training process with SAT. In this figure, we can see that the occupancies of EVC-human and EVC-synth are more biased than that of EVC-speaking. Although the EV-GMM needs to model wide varieties of acoustic features of all prestored target speakers, this result shows that some mixture

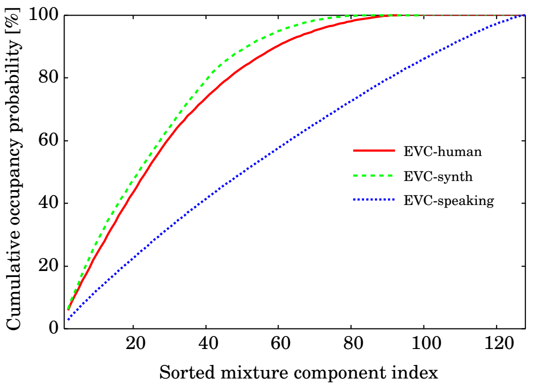

Figure 5: *Cumulative occupancy probability for all parallel data set using several models.*

components of EVC-human and EVC-synth model only acoustic features of a part of prestored target speakers. Consequently, it is expected that phonemic information and speaker individuality were not separated well in them. It is possible that this issue causes degradation of conversion performance.

The above results suggest that 1) the proposed EVC-human yields better conversion performance than VC when a small amount of singing voice data of the source and target singers is available, 2) the conversion performance of the proposed EVC-synth is slightly degraded than EVC-human, 3) since these proposed methods are robust against variations of the singing voice timbre often observed between different songs, they work reasonably well even when different songs are used in the adaptation and conversion processes, 4) the occupancies of individual mixture component of EV-GMM in EVC-human and EVC-synth are more biased than those in EVC-speaking, and then, this causes degradation of conversion accuracy for singer individuality, 5) the differences between a speaking voice and singing voice strongly affects to naturalness of converted singing voice. Based on these results, to more correctly control the voice timbre, it is necessary to train EV-GMM from larger parallel data sets considering phoneme balance. And then, it is expected that training data generation using singing-to-singing synthesis is significantly effective to construct them.

4. Conclusion

In this paper, we evaluated our proposed singing voice conversion methods. Our proposed methods are capable of converting the singing voice timbre 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 training data generation 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 methods enable the effective conversion of a singing voice between an arbitrary singer pair even when using only several seconds of their singing voices as adaptation data. We plan to construct larger parallel data sets considering phoneme balance and further improve conversion performance.

5. Acknowledgements

This work was supported in part by a JSPS KAKENHI Grant Number 22680016 and JST OngaCREST project. The authors are grateful to professor Hideki Kawahara of Wakayama University, Japan, for permission to use the STRAIGHT analysis-synthesis method.

6. References

[1] H. Kenmochi and H. Ohshita, “VOCALOID – Commercial 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] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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,” *APSIPA ASC 2012*, Dec. 2013.

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

[11] 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.

[12] 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.

[13] T. Anastasakos, J. McDonough, S. R., and J. Makhoul, “A compact model for speaker-adaptive training,” *Proc. ICSLP*, vol. 2, pp. 1137–1140, 1996.

[14] 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.

[15] G.-L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[16] D. Tani, T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano, “Maximum a posteriori adaptation for many-to-many eigenvoice conversion,” *Proc. INTERSPEECH*, pp. 1461–1464, Sept. 2008.

[17] 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.

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