Gender-dependent Spectrum Differential Models for Perceived Age Control based on Direct Waveform Modification in Singing Voice Conversion

Kazuhiro Kobayashi*, Tomoki Toda*, Tomoyasu Nakano†, Masataka Goto†,
Graham Neubig*, Sakriani Sakti* and Satoshi Nakamura*
*Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)
Takayama 8916–5, Ikoma, Nara, 630–0192 Japan
†National Institute of Advanced Industrial Science and Technology (AIST)
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568 Japan

***Abstract*—The perceived age of a singing voice, which is the age of the singer as perceived by the listener, is one of the intuitively understandable measures to describe voice characteristics of the singing voice. Singers can sing expressively by controlling voice timbre to some extent but the varieties of voice timbre that singers can produce are limited by physical constraints. To overcome this limitation, previous work has proposed statistical voice timbre control technique based on the perceived age. This technique makes it possible to control the perceived age of singing voice while retaining singer individuality by the use of statistical voice conversion (SVC) with a multiple-regression Gaussian mixture model (MR-GMM). However, the range of controllable perceived age is limited and speech quality of the converted singing voice is significantly degraded compared to that of a natural singing voice. In this paper, we propose a method for perceived age control using direct waveform modification based on spectrum differential and gender-dependent modeling. The experimental results show that the proposed method makes the range of controllable perceived age wider and quality of converted singing voice higher compared to the conventional method.**

I. Introduction

The singing voice is one of the most expressive components in music. In addition to pitch, dynamics, and rhythm, singers can express various expressions by using the linguistic information of the lyrics. However, singers usually have difficulty in changing their voice timbre widely due to physical constraints in speech production. If it would be possible to freely control voice timbre of singers beyond their physical constraints, it will open up entirely new forms of expression in music.

Several techniques to control the voice timbre of the singing voice have been proposed. One approach is based on speech morphing [1] in the speech analysis/synthesis framework [2]. Another approach is based on statistical voice conversion techniques [3], [4]. A Gaussian mixture model (GMM) and an eigenvoice GMM (EV-GMM) [5] have been successfully applied to singing voice conversion (SVC) that converts the source singer's timbre into another target singer's timbre [6], [7], [8]. In particular, SVC with the EV-GMM makes it possible to convert singing voice timbre of an arbitrary source singer into that of an arbitrary target singer in any song by automatically adapting voice timbre control parameters to the given singing voices of those singers. However, it is still difficult to achieve the desired timbre if no target singer's singing voice is available because it is hard to predict the change of voice timbre caused by manually manipulating each adaptive parameter.

In our previous work [9], we have proposed a method for controlling one aspect of the singing voice, perceived age, based on a multiple-regression GMM (MR-GMM) [10]. However, the range of controllable perceived age is still limited. This is possibly caused by using a single MR-GMM to model all singers' voice timbre because it has been reported that spectral variations caused by aging are different between male and female speakers [11], [12]. Additionally, quality of the converted singing voice tends to be significantly degraded compared to that of a natural singing voice. The use of a vocoder to synthesize the converted singing voice is one of the biggest factors causing this degradation. To address the degradation problem by the use of the vocoder, a statistical conversion method based on direct waveform modification has been proposed [13]. The statistical conversion method avoids the degradation by directly filtering a waveform of the input singing voice based on time-varying spectrum differential that estimated with the traditional GMM.

In this paper, we improve the controllability of perceived age and the quality of the converted singing voice in our previously proposed voice timbre control technique based on the perceived age. To make the range of controllable perceived age wider, we propose a voice timbre control technique with gender-dependent MR-GMMs that can more accurately model the spectral variations caused by a change of the perceived age in each gender. Furthermore, we also apply the statistical conversion method based on direct waveform modification to the voice timbre control technique based on perceived age. It is shown from results of subjective evaluation that the proposed methods significantly improve performance of the perceived age control compared to the conventional method.

II. Voice timbre control based on perceived age while retaining singer individuality

A. Training process

1) Training of the MR-GMM: The MR-GMM is trained using multiple parallel data sets consisting of the reference singer's singing voices and many pre-stored target singers' singing voices. The joint probability density of 2D-dimensional joint static and dynamic feature vectors modeled by the MR-GMM is given by

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

$$\boldsymbol{\mu}_m^{(Y)}(s) = \boldsymbol{b}_m^{(Y)} w^{(s)} + \overline{\boldsymbol{\mu}}_m^{(Y)}, \quad (2)$$

978-616-361-823-8 © 2014 APSIPA

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. The vectors $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ and $\boldsymbol{Y}_t^{(s)} = [\boldsymbol{Y}_t^{(s)\top}, \Delta\boldsymbol{Y}_t^{(s)\top}]^\top$ are joint static and delta feature vectors of the reference singer and the s-th pre-stored target singer, which are automatically aligned by dynamic time warping for their corresponding singing voices. The vectors $\boldsymbol{b}_m^{(Y)}$ and $\overline{\boldsymbol{\mu}}_m^{(Y)}$ indicate the representative vector and bias vector, respectively. The value $w^{(s)}$ indicates the perceived age score of the s-th pre-stored target singer, which is manually assigned to each pre-stored target singer. The notation $\boldsymbol{\lambda}^{(MR)}$ indicates an MR-GMM parameter set consisting of mixture-dependent parameters, such as the mixture-component weight α_m, the mean vector $\boldsymbol{\mu}_m^{(X)}$, the representative vector $\boldsymbol{b}_m^{(Y)}$, the bias vector $\overline{\boldsymbol{\mu}}_m^{(Y)}$ and the covariance matrix $\boldsymbol{\Sigma}_m$ of the m-th mixture component.

$\boldsymbol{\lambda}^{(MR)}$ is a MR-GMM parameter set consisting of the mixture-component weight α_m, the mean vector $\boldsymbol{\mu}_m$, the representative vector $\boldsymbol{b}_m^{(Y)}$, the bias vector $\overline{\boldsymbol{\mu}}_m^{(Y)}$, and the covariance matrix $\boldsymbol{\Sigma}_m$ of the m-th mixture component.

To easily create the MR-GMMs for various source singers (i.e., users), the framework of the many-to-many SVC [8] is applied to the MR-GMM for the reference singer.

The joint probability density of many-to-many MR-GMM is given by

$$\begin{aligned}
&P\left(\boldsymbol{Y}_t^{(i)}, \boldsymbol{Y}_t^{(o)} | \boldsymbol{\lambda}^{(MR)}, w^{(i)}, w^{(o)}\right) \\
&= \sum_{m=1}^{M} P\left(m|\boldsymbol{\lambda}^{(MR)}\right) \int P\left(\boldsymbol{Y}_t^{(i)} | \boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(MR)}, w^{(i)}\right) \\
&\qquad P\left(\boldsymbol{Y}_t^{(o)} | \boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(MR)}, w^{(o)}\right) P\left(\boldsymbol{X}_t | m, \boldsymbol{\lambda}^{(MR)}\right) \mathrm{d}\boldsymbol{X}_t \\
&= \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{Y}_t^{(i)} \\ \boldsymbol{Y}_t^{(o)} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^{(Y)}(i) \\ \boldsymbol{\mu}_m^{(Y)}(o) \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(YXY)} \\ \boldsymbol{\Sigma}_m^{(YXY)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}\right), \quad (3)
\end{aligned}$$

$$\boldsymbol{\Sigma}_m^{(YXY)} = \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)^{-1}} \boldsymbol{\Sigma}_m^{(XY)}, \quad (4)$$

where $w^{(i)}$ and $w^{(o)}$ indicate the perceived age scores of the source singer and target singers, respectively. The source and target mean vectors, $\boldsymbol{\mu}_m^{(Y)}(i)$ and $\boldsymbol{\mu}_m^{(Y)}(o)$ are given by Eq. (2). An MR-GMM capable of converting each source singer can be easily created by adapting the perceived age score of the source singer.

2) Adaptation of MR-GMMs to a specific source singer: To implement the perceived age control while retaining source singer's individuality, we modify the representative form of the target mean vectors. The perceived age score of the target singing voice $w^{(o)}$ can be represented by that of the source singer $w^{(i)}$ and a perceived age score differential Δw, i.e., $w^{(o)} = w^{(i)} + \Delta w$. Then, the target mean vector of the m-th mixture component is given by

$$\boldsymbol{\mu}_m^{(Y)}(o) = \boldsymbol{b}_m^{(Y)} w^{(i)} + \overline{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{b}_m^{(Y)} \Delta w, \quad (5)$$

where $\boldsymbol{b}_m^{(Y)} w^{(i)} + \overline{\boldsymbol{\mu}}_m^{(Y)}$ can be regarded as an approximated form of the source mean vector. Therefore, it is modified as follows:

$$\boldsymbol{\mu}_m^{(Y)}(o) \quad \simeq \quad \hat{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{b}_m^{(Y)} \Delta w, \quad (6)$$

where $\hat{\boldsymbol{\mu}}_m^{(Y)}$ is the maximum likelihood estimate estimated by the use of parallel data of source and reference singers. Consequently, the source and target mean vectors of the modified MR-GMM for the source singer are represented by the source mean vectors $\hat{\boldsymbol{\mu}}_m^{(Y)}$, the representative vectors $\boldsymbol{b}_m^{(Y)}$, and the perceived age score differential Δw.

B. Conversion process

In the conversion process, the perceived age score differential Δw is manually set to a desired value. Then, the source singer's singing voice is converted into a perceived age controlled singing voice using maximum likelihood estimation of the speech parameter trajectory with the modified MR-GMM [4].

Time sequence vectors of the source and converted features are denoted as $\boldsymbol{Y}(i) = [\boldsymbol{Y}_1^\top(i), \cdots, \boldsymbol{Y}_T^\top(i)]^\top$ and $\boldsymbol{Y}(o) = [\boldsymbol{Y}_1^\top(o), \cdots, \boldsymbol{Y}_T^\top(o)]^\top$, 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 $\hat{\boldsymbol{y}}(o) = [\hat{\boldsymbol{y}}_1^\top(o), \cdots, \hat{\boldsymbol{y}}_T^\top(o)]^\top$ is determined as follows:

$$\begin{aligned}
\hat{\boldsymbol{y}}(o) = \operatorname*{argmax}_{\boldsymbol{y}(o)} & P(\boldsymbol{Y}(o) | \boldsymbol{Y}(i), \boldsymbol{\lambda}^{(MR)}, \hat{\boldsymbol{\mu}}_m^{(Y)}, \Delta w) \\
& \text{subject to } \boldsymbol{Y}(o) = \boldsymbol{W}\boldsymbol{y}(o), \quad (7)
\end{aligned}$$

where $\boldsymbol{W}$ is a transformation matrix to expand the static feature vector sequence into the joint static and dynamic feature vector sequence [14]. The conditional probability density function $P(\boldsymbol{Y}(o) | \boldsymbol{Y}(i), \boldsymbol{\lambda}^{(MR)}, \hat{\boldsymbol{\mu}}_m^{(Y)}, \Delta w)$ is analytically derived from the modified MR-GMM. To alleviate the oversmoothing effects that usually make the converted singing voice sound muffled, global variance (GV) [4] is also considered.

III. Perceived age control based on gender-dependent MR-GMM with direct waveform modification

To improve controllability and quality of the converted speech of the conventional perceived age control method, we further implement two techniques, 1) gender-dependent MR-GMMs for more accurately capturing spectral variations depending on the perceived age and 2) direct waveform modification based on spectral differential.

A. Gender-dependent MR-GMM

In the conventional method, the MR-GMM is trained using multiple parallel data sets consisting of singing voice pairs of the reference singer and all pre-stored target singers including both male and female singers. On the other hand, in the gender-dependent modeling, two MR-GMMs are trained separately using the parallel data sets consisting of only male singers or female singers as the reference singer and the pre-stored target singers. To create the MR-GMM for the source singer, the corresponding gender-dependent MR-GMM is adapted to him/her to develop the modified MR-GMM in the same manner as described in Section II.

B. Perceived age control with direct waveform modification based on spectral differential

In the direct waveform modification based on spectral differential, the spectral feature differential between the source singing voice and the converted singing voice is directly estimated from the source singer's spectral features using a differential MR-GMM (DIFFMR-GMM), which is analytically derived from the modified MR-GMM. The joint probability density of the source singer's spectral features and the spectral feature differential related to perceived age control is modeled using the DIFFMR-GMM. Then, voice timbre of the source singer's singing voice is converted by directly filtering a waveform of the source singer's natural singing voice with the time-varying spectral feature differential determined with the DIFFMR-GMM. In this conversion process, the converted singing voice is free from various errors usually observed in the conventional waveform generation process with vocoder, such as F_0 extraction errors, unvoiced/voiced decision errors, spectral parameterization errors caused by liftering on the mel-cepstrum, and so on.

The DIFFMR-GMM is derived as follows. Let $\boldsymbol{D}_t = \left[\boldsymbol{d}_t^\top, \Delta\boldsymbol{d}_t^\top\right]^\top$ denote the joint static and delta differential feature vector, where $\boldsymbol{d}_t = \boldsymbol{y}_t(o) - \boldsymbol{y}_t(i)$. The 2D-dimensional joint static and delta feature vector between the source and the differential features is represented as linear transformation of the conventional joint feature vectors as follows:

$$\begin{bmatrix} \boldsymbol{Y}_t^{(i)} \\ \boldsymbol{D}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_t^{(i)} \\ \boldsymbol{Y}_t^{(o)} - \boldsymbol{Y}_t^{(i)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{I} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{Y}_t^{(i)} \\ \boldsymbol{Y}_t^{(o)} \end{bmatrix}, \tag{8}$$

where $\boldsymbol{I}$ denotes the identity matrix. Applying this linear transform to the modified MR-GMM, the DIFFMR-GMM is derived as follows:

$$P\left(\boldsymbol{Y}_t^{(i)}, \boldsymbol{D}_t | \boldsymbol{\lambda}^{(MR)}, \hat{\boldsymbol{\mu}}_m^{(Y)}, \Delta w\right)$$

$$= \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{Y}_t^{(i)} \\ \boldsymbol{D}_t \end{bmatrix}; \begin{bmatrix} \hat{\boldsymbol{\mu}}_m^{(Y)} \\ \boldsymbol{b}_m^{(Y)} \Delta w \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(DYD)} \\ \boldsymbol{\Sigma}_m^{(DYD)} & \boldsymbol{\Sigma}_m^{(DD)} \end{bmatrix} \right), \tag{9}$$

$$\boldsymbol{\Sigma}_m^{(DYD)} = \boldsymbol{\Sigma}_m^{(YXY)} - \boldsymbol{\Sigma}_m^{(YY)}, \boldsymbol{\Sigma}_m^{(DD)} = 2(\boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(YXY)}). \tag{10}$$

In the conversion process, the converted differential feature vector is determined in the same manner as described in Sect. II-B. In this paper, the GV is not considered in the conversion process based on the spectrum differential.

IV. Experimental evaluations

A. Experimental conditions

We used the AIST humming database [15] consisting of songs with Japanese lyrics sung by Japanese male and female amateur singers in their 20s, 30s, 40s, and 50s. The sampling frequency was set to 16 kHz. The 1st through 24th mel-cepstral coefficients extracted by STRAIGHT analysis [16] were used as spectral features. As the source excitation features, 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 analysis [17]. The frame shift was 5 ms. The mel log spectrum approximation filter [18] was used as the synthesis filter in both the conventional waveform generation with vocoder and the proposed direct waveform modification.

In the training of the gender-independent MR-GMM, we used parallel data sets of a female reference singer in her 20s and 56 pre-stored target singers including 28 males and 28 females in their 20s, 30s, 40s and 50s. In the training of the gender-dependent MR-GMMs, we separately used a female and male reference singer in their 20s and 28 male or 28 female pre-stored target singers. The number of mixture components of each MR-GMM was 256 for the spectral feature and 128 for the aperiodic components. The number of training songs was 23 for each singer. The duration of each song was approximately 20 seconds. The perceived age score for each singer was determined as an average score of the singer rated by 8 subjects in their 20s [9].

In this evaluation, we define several methods follow:

- SVC (I): with gender-independent MR-GMM
- SVC (D): with gender-dependent MR-GMM
- DIFFSVC (D): with gender-dependent DIFFMR-GMM

The converted singing voice samples were generated by settings of the perceived age score differential to -60, -30, 0, 30, and 60. The number of source singers was 16, who were not included in the pre-stored target singers. We used P039 as an evaluation song.

First, we evaluated perceived age controllability. Eight subjects were divided into two groups, and the 16 evaluation singers were divided into two groups so that one group always included one male singer and one female singer in each age group. Each subject evaluated the converted singing voices from only one group of the evaluation singers. Subjects were asked to assign the perceived age to each converted singing voice by listening to it in random order.

In the second experiment, we evaluated the quality of the converted singing voice using a mean opinion score (MOS). The number of subjects and evaluation singers were the same as in the first experiment. The subjects rated the quality of the converted singing voice using a 5–point scale: "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad.

In the final experiment, we conducted an XAB test on the singer individuality to compare SVC (I) and DIFFSVC (D). The subjects and evaluation singers were separated into two groups in the same manner as the first experiment. A pair of singing voices converted by SVC (I) and by DIFFSVC (D) for the same singer and a setting of the perceived age score was presented to the subjects after presenting the natural singing voice as a reference. Then, they were asked which singing voice sounded more similar to the reference in terms of the singer individuality.

B. Experimental Results

Figure 1 shows the relationship between settings of the perceived age differential and actually perceived age of the converted singing voice in each method. We can see that the perceived age varies almost linearly according to a change of the settings of the perceived age differential from -60 to 60. Moreover, a range of the perceived age of the converted singing voice becomes wider by using SVC (D) and DIFFSVC (D) compared to SVC (I). This indicates that the gender-dependent model is effective for accurately modeling spectral variations depending on the perceived age.

Figure 2 indicates the results of the opinion test on the quality. We can see that DIFFSVC (D) tends to significantly improve quality of the converted singing voices compared to SVC (I) and SVC (D). Although the quality is greatly degraded in the conventional method SVC (I) as the perceived age score differential is set to larger or smaller values, this quality degradation is effectively alleviated by the proposed method DIFFSVC (D).

Figure 3 indicates the result of the XAB test on the singer individuality. DIFFSVC (D) better or equally retains singer individuality in these perceived age settings compared to the conventional method SVC (I). We can see that as a change of the perceived age differential setting is larger, the difference between DIFFSVC (D) and SVC (I) becomes smaller.

These results suggest that 1) the gender-dependent MR-GMM is effective for improving the perceived age controllability, and 2) the direct waveform modification technique with spectral differential significantly improves quality of the converted singing voice.

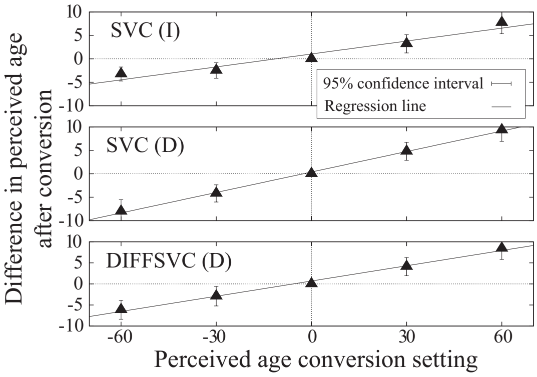

Fig. 1. Setting and actual differential in perceived age after conversion.

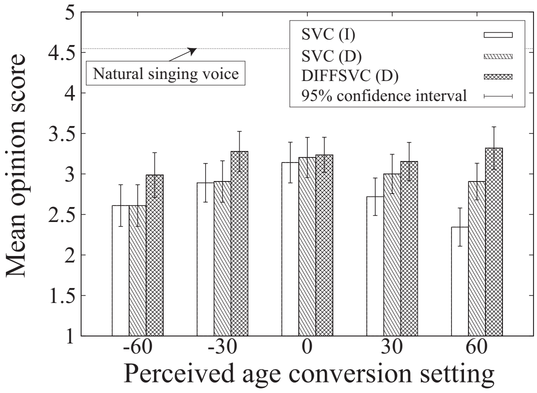

Fig. 2. Mean opinion score of speech quality.

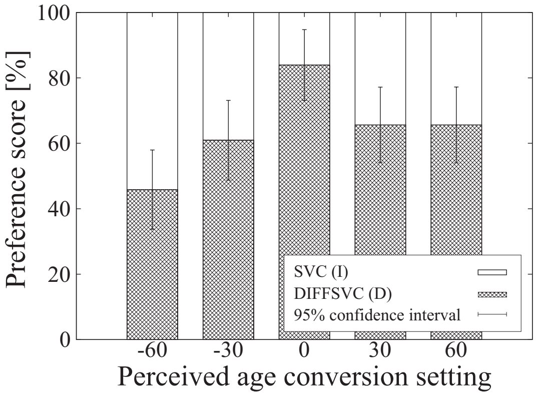

Fig. 3. Comparing singer individuality.

V. Conclusions

To improve performance of our previously proposed perceived age control technique based on multiple-regression Gaussian mixture models (MR-GMM), we have successfully implemented a gender-dependent modeling technique and a direct waveform modification technique with spectral differential. The experimental results have demonstrated that 1) the proposed method makes a range of the controllable perceived age wider and 2) it also enables to significantly improve quality of the converted singing voice. In future work, we will consider adaptive techniques for an arbitrary source singer by the use of several singing voices.

Acknowledgment

This work was supported in part by JSPS KAKENHI Grant Numbers: 26280060 and 25280061 and by the JST OngaCREST project.

References

[1] 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.

[2] 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, 2009.

[3] 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.

[4] 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.

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

[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] 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.

[9] K. Kobayashi, T. Toda, H. Doi, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura, "Voice timbre control based on perceived age in singing voice conversion," *IEICE Trans. Inf. Syst.*, vol. E97-D, no. 6, pp. 1419–1428, June 2014.

[10] K. Ohta, T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano, "Adaptive voice-quality control based on one-to-many eigenvoice conversion," *Proc. INTERSPEECH*, pp. 2158–2161, Sept. 2010.

[11] W. Endres, W. Bambach, and G. Flsser, "Voice spectrograms as a function of age, voice disguise, and voice imitation," *The Journal of the Acoustical Society of America*, vol. 49, no. 6B, pp. 1842–1848, 1971.

[12] S. E. Linville and J. Rens, "Vocal tract resonance analysis of aging voice using long-term average spectra," *Journal of Voice*, vol. 15, no. 3, pp. 323–330, 2001.

[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*, 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. Goto and T. Nishimura, "AIST humming database: Music database for singing research," *IPSJ SIG Notes (Technical Report) (Japanese edition)*, vol. 2005-MUS-61-2, no. 82, pp. 7–12, Aug. 2005.

[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] 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.