An Investigation of Acoustic Features for Singing Voice Conversion based on Perceptual Age

Kazuhiro Kobayashi[1], *Hironori Doi*[1], *Tomoki Toda*[1], *Tomoyasu Nakano*[2], *Masataka Goto*[2], *Graham Neubig*[1], *Sakriani Sakti*[1], *Satoshi Nakamura*[1]

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

Abstract

In this paper, we investigate the acoustic features that can be modified to control the perceptual age of a singing voice. Singers can sing expressively by controlling prosody and vocal timbre, but the varieties of voices that singers can produce are limited by physical constraints. Previous work has attempted to overcome this limitation through the use of statistical voice conversion. This technique makes it possible to convert singing voice characteristics of an arbitrary source singer into those of an arbitrary target singer. However, it is still difficult to intuitively control singing voice characteristics by manipulating parameters corresponding to specific physical traits, such as gender and age. In this paper, we focus on controlling the perceived age of the singer and, as a first step, perform an investigation of the factors that play a part in the listener's perception of the singer's age. The experimental results demonstrate that 1) the perceptual age of singing voices corresponds relatively well to the actual age of the singer, 2) speech analysis/synthesis processing and statistical voice conversion processing don't cause adverse effects on the perceptual age of singing voices, and 3) prosodic features have a larger effect on the perceptual age than spectral features.

Index Terms: singing voice, voice conversion, perceptual age, spectral and prosodic features, subjective evaluations.

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 characteristics such as voice timbre to some degree, they usually have difficulty in changing their own voice characteristics widely, (e.g. changing them into those of another singer's singing voice) owing to physical constraints in speech production. If it would be possible for singers to freely control voice characteristics beyond these physical constraints, it will open up entirely new ways for singers to express themselves.

In previous research, a number of techniques have been proposed to change the characteristics of singing voices. One typical method is singing voice conversion (VC) based on speech morphing in the speech analysis/synthesis framework [1]. This method makes it possible to independently morph several acoustic parameters, such as spectral envelope, F_0, and duration, between singing voices of different singers or different singing styles. One of the limitations of this method is that the morphing can only be applied to singing voice samples of the same song.

To make it possible to more flexibly change of singing voice characteristics, statistical VC techniques [2, 3] have been successfully applied to convert the source singer's singing voice into another target singer's singing voice [4, 5]. In this method, 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, keeping the linguistic information of the lyrics unchanged. Furthermore, to develop a more flexible singing VC system, eigenvoice conversion (EVC) techniques [6] have been applied to singing VC [7]. In a singing VC system based on many-to-many EVC [8], which is one particular variety of EVC, an initial conversion model called the canonical eigenvoice GMM (EV-GMM) is trained in advance using multiple parallel data sets including song pairs of a single reference singer and many other singers. The EV-GMM is adapted into arbitrary source and target singers by automatically estimating a few adaptive parameters from the given singing voice samples of those singers. Although this system is also capable of flexibly changing singing voice characteristics by manipulating the adaptive parameters even if no target singing voice sample is available, it is difficult to achieve the desired singing voice characteristics, because it is hard to predict the change of singing characteristics caused by the manipulation of each adaptive parameter.

In the area of statistical parametric speech synthesis [9], there have been several attempts at developing techniques for manually controlling voice quality of synthetic speech by manipulating intuitively controllable parameters corresponding to specific physical traits, such as gender and age. Nose et al. [10] proposed a method for controlling speaking styles in synthetic speech with multiple regression hidden Markov models (HMM). Tachibana et al. [11] extended this method to control voice quality of synthetic speech using a voice quality control vector assigned to expressive word pairs describing voice quality, such as "warm – cold" and "smooth – non-smooth". A similar method has also been proposed in statistical VC [12]. Although these methods have only been applied to voice quality control of normal speech, it is expected that they would also be effective for controlling singing voice characteristics.

In this paper, we focus on the perceptual age, or the age that a listener predicts the singer to be, of singing voices as one of the factors to intuitively describe the singing voice. For normal speech, there is some research investigating acoustic feature changes caused by aging. It has been reported that aperiodicity of excitation signals tends to increase with aging [13]. A perceptual age classification method to classify speech of elderly people and non-elderly people using spectral and prosodic features has also been developed [14]. On the other hand, the perceptual age of singing voices has not yet been studied deeply.

As fully understanding the acoustic features that contribute to the perceptual age of singing voices is essential to the development of VC techniques to modify a singer's perceptual age, in this paper we perform an investigation of the acoustic features that play a part in the listener's perception of the singer's age. We conduct several types of perceptual evaluation to investigate 1) how well the perceptual age of singing voices corresponds to the actual age of the singer, 2) whether or not singing VC processing causes adverse effects on the perceptual age of singing voices, and 3) whether spectral or prosodic features have a larger effect on the perceptual age.

2. Statistical singing voice conversion

Statistical singing VC (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 GMM using a parallel data set in the same manner as in statistical VC for normal voices [5]. 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. λ 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. A 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 with the GMM using maximum likelihood estimation of speech parameter trajectory [3]. 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 speech sound muffled, global variance (GV) [3] is also considered in conversion.

3. Investigation of acoustic features affecting perceptual age

In the traditional SVC [5, 7], only the spectral features such as mel-cepstrum are converted. It is straightforward to also convert the aperiodic components [16], which capture noise strength on each frequency band of the excitation signal, as in the traditional VC for natural voices [17]. If the perceptual age of singing voices is captured well by these acoustic features, it will make it possible to develop a real-time SVC system capable of controlling the perceptual age of singing voices by combining the voice quality control based on statistical VC [12] and real-time statistical VC techniques [18, 19]. On the other hand, if the perceptual age of singing voices is not captured well by these acoustic features, which mainly represent segmental features, the conversion of other acoustic features, such as prosodic features (e.g., F_0 pattern), will also be necessary. In such a case, the voice-quality control framework of HMM-based speech synthesis [10, 11] can be used in the SVC system to control the perceptual age of singing voices, although it is not straightforward to develop a real-time SVC system in this framework. Because the synthesis technique that must be used will change according to the acoustic features to be converted, it will be highly beneficial to make clear which acoustic features need to be modified to control the perceptual age of singing voices. To do so, we compare the perceptual age of natural singing voices with that of several types of synthesized singing voices by modifying acoustic features as shown in Table 1.

3.1. Analysis/synthesis with aperiodic components (w/ AC)

In the analysis/synthesis framework, a voice is first converted into parameters of the synthesis model described in Section 2, then simply re-synthesized into a waveform using these parameters without change. As analysis and synthesis are necessary steps in converting acoustic features of singing voices, we investigate the effects of distortion caused by analysis/synthesis on the perceptual age of singing voices. STRAIGHT [20] is a widely used high-quality analysis/synthesis method, so we use it to extract acoustic features consisting of mel-cepstrum, F_0, and aperiodic components.

3.2. Analysis/synthesis without aperiodic components (w/o AC)

As mentioned above, previous research [13] has shown that aperiodic components tend to change with aging in normal speech as mentioned above. We investigate the effects of aperiodic components on the perceptual age of singing voices. Analysis/synthesized singing voice samples are reconstructed from mel-cepstrum and F_0 extracted with STRAIGHT. In synthesis, only a pulse train with phase manipulation [20] instead of STRAIGHT mixed excitation [17] is used to generate voiced excitation signals.

3.3. Intra-singer SVC

In SVC, conversion errors are inevitable. For example, some detailed structures of acoustic features not well modeled by the GMM of the joint probability density and often disappear through the statistical conversion process. Therefore, the acous-

Table 1: Acoustic features of several types of synthesized singing voices.

Features	Analysis/synthesis (w/ AC)	Analysis/synthesis(w/o AC)	Intra-singer SVC	SVC
Mel-cepstrum	Source singer	Source singer	Converted to source singer	Converted to target singer
Aperiodic components	Source singer	None	Converted to source singer	Converted to target singer
Power, F_0, duration	Source singer	Source singer	Source singer	Source singer

tic space on which the converted acoustic features are distributed tends to be smaller than the acoustic space that of the natural acoustic features. We investigate the effect of the conversion errors caused by this acoustic space reduction on the perceptual age of singing voices by converting one singer's singing voice into the same singer's singing voice. This SVC process is called intra-singer SVC in this paper.

To achieve intra-singer SVC for a specific singer, we must create a GMM to model the joint probability density of the same singer's acoustic features, i.e., $P(\boldsymbol{X}_t, \boldsymbol{X}'_t|\boldsymbol{\lambda})$ where $\boldsymbol{X}_t$ and $\boldsymbol{X}'_t$ respectively show the source and target acoustic features of the same singer, needs to be developed. Note that $\boldsymbol{X}_t$ is different from $\boldsymbol{X}'_t$, they depend on each other, and both are identically distributed. This GMM is analytically derived from the GMM of the joint probability density of the acoustic features of the same singer and another reference singer, i.e., $P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda})$ where $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ respectively show the source feature vector of the same singer and that of the reference singer, by marginalizing out the acoustic features of the reference singer in the same manner as used in the many-to-many EVC [7, 8] as follows:

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

$$\boldsymbol{\Sigma}_m^{(XYX)} = \boldsymbol{\Sigma}_m^{(XY)} \boldsymbol{\Sigma}_m^{(YY)^{-1}} \boldsymbol{\Sigma}_m^{(YX)}. \quad (4)$$

Using this GMM, intra-singer SVC is performed in the same manner as described in Section 2. The converted singing voice sample essentially has the same singing voice characteristics as those before the conversion although they suffer from conversion errors.

3.4. SVC

To investigate which acoustic features have a larger effect on the perceptual age of singing voices, segmental features or prosodic features, we use the SVC for converting only segmental features, such as mel-cepstrum and aperiodic components, of a source singer into those of a different target singer. The converted singing voice samples essentially have the segmental features of the target singer and the prosodic features, such as F_0 patterns, power patterns, and duration, of the source singer.

4. Experimental evaluation

4.1. Experimental conditions

In our experiments, we first investigated the correspondence between the perceptual age and the actual age of the singer. As test stimuli, we used all singing voices in the AIST humming database [21] consisting of singing voices of songs with Japanese lyrics sung by Japanese male and female amateur singers in their 20s, 30s, 40s, and 50s. The total number of the singers was 75. Each singer sang 25 songs. The length of each song was approximately 20 seconds. One Japanese male subject was asked to guess the age of each singing voice by listening to it.

In the second experiment, we investigate the acoustic features that affect the perceptual age of singing voices, by comparing the perceptual age of natural singing voices with that of each type of synthesized singing voice as shown in Table 1. Eight Japanese male subjects in their 20s assigned a perceptual age to each synthesized singing voice. To reduce the subjects' burden, one Japanese song (No. 39) that showed the highest correlation between the perceptual age and the actual age in the first evaluation was selected to be evaluated. Moreover, we selected 16 singers consisting of four singers (two male singers and two female singers) from each age group, i.e., their 20s, 30s, 40s, or 50s, who showed good correlation between the perceptual age and their actual age. The subjects were separated into two groups, A and B. The singers were also separated into two groups, A and B, so that one group always includes one male singer and one female singer in each age group. The subjects in each group evaluated only singing voices of the corresponding singer group.

The sampling frequency was set to 16 kHz. The 1st through 24th mel-cepstral coefficients extracted by STRAIGHT analysis 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. The frame shift was 5 ms.

As training data for the GMMs used in intra-singer SVC and SVC, we used 18 songs including the evaluation song (No. 39). In the intra-singer SVC, GMMs for converting the mel-cepstrum and aperiodic components were trained for each of the selected 16 singers. Another singer not included in these 16 singers was used as the reference singer to create each parallel data set for the GMM training. In the SVC, the GMMs for converting mel-cepstrum and aperiodic components were trained for all combinations of the source and target singer pairs in each singer group. The numbers of mixture components of each GMM were optimized experimentally.

4.2. Experimental results

Figure 1 shows the correlation between the perceptual age of natural singing voices and the actual age of the singer. Each point shows the actual age of one singer and the average of the perceptual ages over all different songs sung by the same singer. The correlation coefficient is 0.79. These results show quite high correlation between the perceptual age and the actual age.

Table 2 shows average values and standard deviations of differences between perceptual age of natural singing voices and each type of intra-singer synthesized singing voice: analysis/synthesis (w/ AC), analysis/synthesis (w/o AC) and the intra-singer SVC. The table also shows correlation coefficients between the perceptual age of natural and synthesized voices. From the results, we can see that in analysis/synthesis (w/ AC), the perceptual age difference is small and the correlation coefficient is very high. Therefore, distortion caused by analysis/synthesis processing does not affect the perceptual age. It can be observed from analysis/synthesis (w/o AC) that this re-

Table 2: Differences of the perceptual age between natural singing voices and each type of the synthesized singing voices.

Methods	Average	Standard deviation	Correlation coefficient
Analysis/synthesis (w/ AC)	0.77	3.57	0.96
Analysis/synthesis (w/o AC)	0.44	3.58	0.96
Intra-singer SVC	-0.50	7.25	0.85

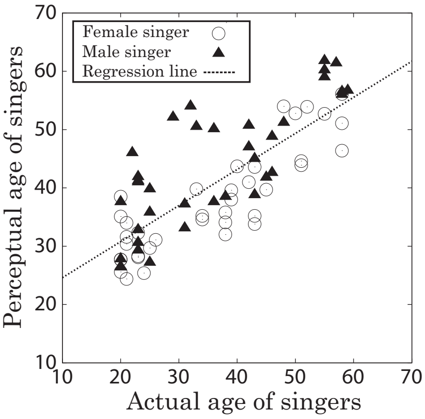

Figure 1: Correlation between singer's actual age and perceptual age.

sult does not change even if not using aperiodic components. Therefore, aperiodic components do not affect the perceptual age of singing voices. On the other hand, intra-singer SVC causes slightly larger differences between natural singing voices and the synthesized singing voices. Therefore, some acoustic cues to the perceptual age are removed through the statistical conversion processing. Nevertheless, the perceptual age differences are relatively small, and therefore, it is likely that important acoustic cues to the perceptual age are still kept in the converted acoustic features.

Figures 2 and 3 show a comparison between the perceptual age of singing voices generated by SVC and intra-singer SVC. In each figure, the vertical axis shows the perceptual age of converted singing voices by SVC (prosodic features: source singer, segmental features: target singer). The horizontal axis in Fig. 2 shows the perceptual age of singing voices generated by intra-singer SVC (prosodic features: source singer, segmental features: source singer) and that in Fig. 3 shows the perceptual age of singing voices generated by intra-singer SVC (prosodic features: target singer, segmental features: target singer). Therefore, if the prosodic features more strongly affect the perceptual age than the segmental features, a higher correlation will be observed in Fig. 2. If the segmental features more strongly affect the perceptual age than the prosodic features, a higher correlation will be observed in Fig. 3 than in Fig. 2. These figures demonstrate that 1) the segmental features affect the perceptual age but the effects are limited as shown in positive but weak correlation in Fig. 3 and 2) the prosodic features have a larger effect on the perceptual age than the segmental features.

5. Conclusions

In this paper, we have investigated the acoustic features that affect the perceptual age of singing voices. To factorize the effect of several acoustic features on the perceptual age of singing voices, several types of synthetic singing voices were constructed and evaluated. The experimental results have demonstrated that 1) statistical voice conversion processing has only a small effect on the perceptual age of singing voices and 2) the prosodic features more strongly affect the perceptual age than the segmental features. We plan to further study a conversion technique for controlling the perceptual age of singing voices.

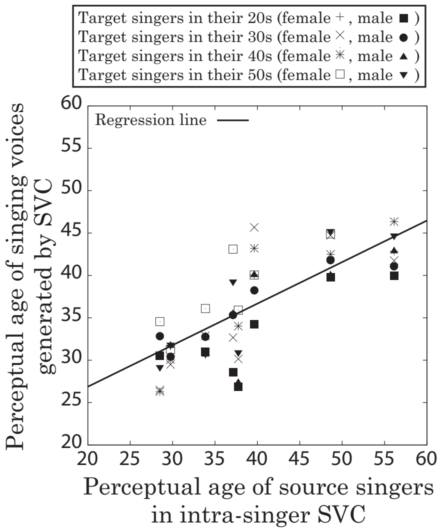

Figure 2: Correlation of perceptual age between singing voices generated by the intra-singer SVC and the SVC if setting horizontal axis to the perceptual age of the source singers.

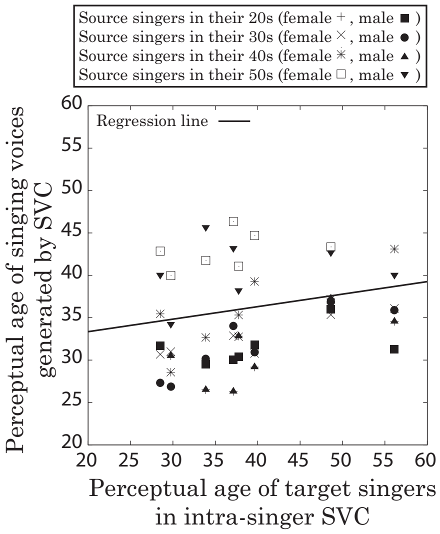

Figure 3: Correlation of perceptual age between singing voices generated by the intra-singer SVC and the SVC if setting horizontal axis to the perceptual age of the target singers.

6. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Number 22680016 and by the JST OngaCREST project.

7. References

[1] H. Kawahara and M. Morise, "Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown," *Proc. ICASSP*, pp. 5389–5392, Mar. 2012.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

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

[9] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[10] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis (speech and hearing)," *IEICE transactions on information and systems*, vol. 90, no. 9, pp. 1406–1413, Sep. 2007.

[11] M. Tachibana, T. Nose, J. Yamagishi, and T. Kobayashi, "A technique for controlling voice quality of synthetic speech using multiple regression HSMM," *Proc. INTERSPEECH*, pp. 2438–2441, Sept. 2006.

[12] 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.

[13] H. Kasuya, H. Yoshida, S. Ebihara, and H. Mori, "Longitudinal changes of selected voice source parameters," *Proc. INTERSPEECH*, pp. 2570–2573, Sept. 2010.

[14] N. Minematsu, M. Sekiguchi, and K. Hirose, "Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers," *Proc. ICASSP*, pp. 137–140, May. 2002.

[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, 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.

[17] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," *Proc. INTERSPEECH*, pp. 2266–2269, Sept. 2006.

[18] 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.

[19] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion," *Proc. INTERSPEECH*, Sept. 2012.

[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.

[21] 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, pp. 7–12, Aug. 2005.