REGRESSION APPROACHES TO PERCEPTUAL AGE CONTROL IN SINGING VOICE CONVERSION

Kazuhiro Kobayashi[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]

ABSTRACT

The perceptual age of a singing voice is the age of the singer as perceived by the listener, and is one of the notable characteristics that determines perceptions of a song. In this paper, we describe a novel voice timbre control technique based on the perceptual age for singing voice conversion (SVC). Singers can sing expressively by controlling prosody and voice timbre, but the varieties of voices that singers can produce are limited by physical constraints. Previous work has attempted to overcome the limitation through the use of statistical voice conversion. This technique makes it possible to convert singing voice timbre of an arbitrary source singer into that 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 develop a technique for controlling the voice timbre based on perceptual age that maintains the singer's individuality. The experimental results show that the proposed voice timbre control method makes it possible to change the singer's perceptual age while not having an adverse effect on the perceived individuality.

Index Terms— singing voice conversion, perceptual age, voice timbre control, regression approaches, singer's individuality.

1. INTRODUCTION

The singing voice is one of the most expressive components in music. In addition to pitch, dynamics, and rhythm, singers can express more varieties of expression than other music instruments 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 generate voice timbre beyond these physical constraints, it will open up entirely new forms of expression in music.

Singing synthesis systems such as Vocaloid [1] and Sinsy [2], which produce a synthesized singing voice from note-level score information of the melody with its lyrics have been successfully applied as new music instruments. VocaListener [3] has been proposed to synthesize more expressive singing voices, similar to those sung by actual singers. Towards manual control of expressive singing voices, Nose et al. applied style control techniques [4, 5] in statistical parametric speech synthesis with hidden Markov models (HMM) [6] to singing synthesis [7]. This technique makes it possible to control singing styles based on a word pair of "child-like – adult-like". These systems are used in place of actually singing, therefore it is difficult to directly control the singer's voice timbre in real-time.

A number of techniques to control voice timbre have been proposed in singing voice conversion (SVC). One typical method is SVC based on speech morphing [8] in the speech analysis/synthesis framework [9]. This method makes it possible to independently morph several acoustic parameters 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 more flexible SVC possible, statistical voice conversion (VC) techniques [10, 11] with Gaussian mixture model (GMM) have been successfully applied to convert the source singer's singing voice into another target singer's singing voice [12, 13]. These techniques make it possible to convert the spectral 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, several attempts have been performed in order to implement more expressive voice timbre control techniques (described in Section 2). However, it has still been difficult to control voice timbre based on intuitive parameters in SVC.

In this paper, we focus on the perceptual age, or the age that a listener predicts the singer to be, and develop a method to control this intuitively understandable parameter in SVC. In previous work, we have reported that both prosodic features (e.g., F_0 pattern) and spectral features have an effect on perceptual age, and prosodic features more strongly affect the perceptual age than spectral features but they also cause an adverse effect on the perceived singer's individuality [14]. In traditional SVC [13, 15], only the spectral features such as mel-cepstrum are converted. It is straightforward to develop a real-time SVC system capable of controlling the perceptual age of singing voices, applying statistical VC with multiple-regression GMM (MR-GMM) [16] and real-time statistical VC techniques [17, 18] to SVC. On the other hand, it is not straightforward to treat the prosodic features for controlling perceptual age in real-time SVC. Furthermore, the adverse effect on the perceived singer's individuality in the prosodic feature conversion needs to be alleviated to intuitively control only the perceptual age. Based on these facts, although controlling prosodic features makes it possible to change singer's perceptual age widely, we only treat spectral features as a acoustic cue in this paper.

In this paper, we apply statistical VC with MR-GMM to SVC to achieve the perceptual age control in SVC. The standard MR-GMM has difficulty maintaining the individuality of the source singer, because the subspace of the MR-GMM only expresses the average voice timbre of training singers. To solve this problem, we propose a novel voice timbre conversion method that modifies the MR-GMM to convert the singer's perceptual age while retaining singer's individuality in SVC.

2. RELATED WORK

SVC with GMM is only capable of converting acoustic features between a pair of trained source and target singers. To develop a more flexible SVC system, eigenvoice conversion (EVC) techniques [19]

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

have been applied to SVC [15]. In a SVC system based on many-to-many EVC [20], 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 timbre by manipulating the adaptive parameters even if no target singing voice sample is available, it is difficult to achieve the desired singing voice timbre, because it is hard to predict the change of singing timbre caused by the manipulation of each adaptive parameter.

3. STATISTICAL SINGING VOICE CONVERSION (SVC)

3.1. SVC with GMM

SVC with GMM is a technique that converts the voice timbre of a source singer into that of a target singer. SVC with GMM 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 [13]. 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

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

where $\mathcal{N}\left(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$ 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 [11]. 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:

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

where $\boldsymbol{W}$ is a transformation matrix to expand the static feature vector sequence into the joint static and dynamic feature vector sequence [21]. 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 oversmoothing effects that usually make the converted speech sound muffled, global variance (GV) [11] is also considered in conversion.

3.2. Many-to-Many SVC with EV-GMM

Many-to-many SVC is a technique to convert acoustic features of an arbitrary source singer into those of an arbitrary target singer [15]. Many-to-many SVC with EV-GMM also consists of a training process and a conversion process.

In the training process, the joint probability density of reference and target features is modeled with the EV-GMM as follows:

$$\begin{aligned} &P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{e}^{(s)}\right) \\ =&\sum_{m=1}^{M}\alpha_m\mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t\\ \boldsymbol{Y}_t^{(s)}\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}_m^{(X)}\\ \boldsymbol{\mu}_m^{(Y)}(s)\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)}\\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right), \end{aligned} \tag{3}$$

$$\boldsymbol{\mu}_m^{(Y)}(s) = \boldsymbol{A}_m\boldsymbol{e}^{(s)} + \boldsymbol{c}_m, \tag{4}$$

where $\boldsymbol{e}^{(s)} = [e^{(s)}(1), \cdots, e^{(s)}(J)]^\top$ is the target singer-dependent weight parameter for controlling target voice timbre. $\boldsymbol{\lambda}^{(EV)}$ is a parameter set of the canonical EV-GMM consisting of the bias vector $\boldsymbol{c}_m$, 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 in addition to the parameter set of the GMM. Acoustic features of an arbitrary target singer are modeled by setting only $\boldsymbol{e}^{(s)}$ to the singer's specific values.

In the conversion process, the joint probability density of the acoustic features between the source singer's voice $\boldsymbol{Y}_t^{(i)}$ and the target singer's voice $\boldsymbol{Y}_t^{(o)}$ is derived as

$$\begin{aligned} &P\left(\boldsymbol{Y}_t^{(i)}, \boldsymbol{Y}_t^{(o)}|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{e}^{(i)}, \boldsymbol{e}^{(o)}\right) \\ =&\sum_{m=1}^{M} P\left(m|\boldsymbol{\lambda}^{(EV)}\right)\int P\left(\boldsymbol{Y}_t^{(i)}|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(EV)}, \boldsymbol{e}^{(i)}\right) \\ &P\left(\boldsymbol{Y}_t^{(o)}|\boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(EV)}, \boldsymbol{e}^{(o)}\right) P\left(\boldsymbol{X}_t|m, \boldsymbol{\lambda}^{(EV)}\right)\mathrm{d}\boldsymbol{X}_t \\ =&\sum_{m=1}^{M}\alpha_m\mathcal{N}\left(\begin{bmatrix}\boldsymbol{Y}_t^{(i)}\\ \boldsymbol{Y}_t^{(o)}\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}_m^{(Y)}(i)\\ \boldsymbol{\mu}_m^{(Y)}(o)\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(YXY)}\\ \boldsymbol{\Sigma}_m^{(YXY)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right), \end{aligned} \tag{5}$$

$$\boldsymbol{\Sigma}_m^{(YXY)} = \boldsymbol{\Sigma}_m^{(YX)}\boldsymbol{\Sigma}_m^{(XX)^{-1}}\boldsymbol{\Sigma}_m^{(XY)}, \tag{6}$$

where $\boldsymbol{e}^{(i)}$ and $\boldsymbol{e}^{(o)}$ are adaptive parameters of the source and target singer. Using this many-to-many EV-GMM, the converted singing voice is generated in the same manner as described in Section 3.1.

4. SVC CONSIDERING PERCEPTUAL AGE

In this paper, we employ VC with MR-GMM [16] for developing an SVC system that can control perceptual age. We first apply the MR-GMM to SVC to make it possible to control voice timbre based on perceptual age. Then, we apply the conversion process of many-to-many SVC to SVC with MR-GMM to more flexibly develop the MR-GMM for an arbitrary source singer. Furthermore, we implement a perceptual age control technique to keep singer's individuality by changing the representative form of the target mean vector.

4.1. SVC with Multiple Regression GMM (MR-GMM)

SVC with MR-GMM is a voice timbre control technique that manipulates the voice timbre of a target singer by intuitive parameters. SVC with MR-GMM 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 many pre-stored target singers' singing voices are modeled with MR-GMM using parallel data sets

Table 1. Relationship of features and mean vectors of MR-GMM and Modified MR-GMM.

Method	Source features	Target feature	Target mean vector
MR-GMM	Singing voice, $w^{(o)}$	Average singing voice of the age of $w^{(o)}$	$\boldsymbol{b}_m^{(Y)} w^{(o)} + \overline{\boldsymbol{\mu}}_m^{(Y)}$
Modified MR-GMM	Singing voice of the age of $w^{(i)}$, Δw	Singing voice of the age of $(w^{(i)} + \Delta w)$	$\hat{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{b}_m^{(Y)} \Delta w$

in the same manner as in statistical VC with MR-GMM for normal speech [16]. The joint probability density modeled by the MR-GMM is given by

$$\begin{aligned} &P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{\lambda}^{(MR)}, w^{(s)}\right) \\ =&\sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t^{(s)}\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}(s)\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right). \end{aligned} \tag{7}$$

The mean vector of the s-th prestored singer is given by

$$\boldsymbol{\mu}_m^{(Y)}(s) = \boldsymbol{b}_m^{(Y)} w^{(s)} + \overline{\boldsymbol{\mu}}_m^{(Y)}, \tag{8}$$

where $\boldsymbol{b}_m^{(Y)}$ and $\overline{\boldsymbol{\mu}}_m^{(Y)}$ indicate the representative vector and bias vector respectively. $w^{(s)}$ indicates the s-th pre-stored target singer's perceptual age score, which is manually assigned for each pre-stored target singer.

In the conversion process, the perceptual age score is manually set to a desired value. Then, the source singer's singing voice is converted into the target singer's singing voice based on the perceptual age score with the MR-GMM in the same manner as the conversion process in Section 3.1.

4.2. MR-GMM implementation based on Many-to-Many SVC

To implement voice timbre control based on the perceptual age for an arbitrary source singer, we adapt many-to-many SVC [15] to SVC based on MR-GMM. The joint probability density of the many-to-many MR-GMM is as follows:

$$\begin{aligned} &P\left(\boldsymbol{Y}_t^{(i)}, \boldsymbol{Y}_t^{(o)} | \boldsymbol{\lambda}^{(MR)}, w^{(i)}, w^{(o)}\right) \\ &= \sum_{m=1}^{M} P\left(m | \boldsymbol{\lambda}^{(MR)}\right) \int P\left(\boldsymbol{Y}_t^{(i)} | \boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(MR)}, w^{(i)}\right) \\ &\quad P\left(\boldsymbol{Y}_t^{(o)} | \boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(MR)}, w^{(o)}\right) P\left(\boldsymbol{X}_t | m, \boldsymbol{\lambda}^{(MR)}\right) \mathrm{d}\boldsymbol{X}_t \\ &= \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{Y}_t^{(i)} \\ \boldsymbol{Y}_t^{(o)}\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(Y)}(i) \\ \boldsymbol{\mu}_m^{(Y)}(o)\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(YXY)} \\ \boldsymbol{\Sigma}_m^{(YXY)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right) \end{aligned} \tag{9}$$

$$\boldsymbol{\Sigma}_m^{(YXY)} = \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)^{-1}} \boldsymbol{\Sigma}_m^{(XY)}, \tag{10}$$

where $w^{(i)}$ and $w^{(o)}$ indicate the perceptual age score of the source singer and the target singers, respectively. Source and target mean vectors are given by Eq. (8).

In SVC with many-to-many MR-GMM, it is possible to use Eq. (8) to describe the source mean vectors $\boldsymbol{\mu}_m^{(Y)}(i)$ based on the perceptual age score of the source singer. However, accuracy of acoustic modeling by the MR-GMM tends to decrease because the acoustic characteristics of the source singer are not always modeled well on a subspace spanned by the basis vector. To develop a better MR-GMM for the source singer who is not included among the training singers of the MR-GMM, we assume an ideal condition that the parallel data of the source and reference singers, which are used in the training of MR-GMM is available. This condition is still practical in the development of an SVC system that can control perceptual age of the source singer. Using parallel data between the source and the reference singers, the source mean vector of the MR-GMM is updated according to the maximum likelihood criterion. Consequently, the source mean vector is given by

$$\boldsymbol{\mu}_m^{(Y)}(i) = \hat{\boldsymbol{\mu}}_m^{(Y)}, \tag{11}$$

where $\hat{\boldsymbol{\mu}}_m^{(Y)}$ is its maximum likelihood estimate.

4.3. Modified MR-GMM to Retain Singer Individuality

In SVC with many-to-many MR-GMM, it is possible to convert voice timbre of the source singer into voice timbre corresponding to the desired perceptual age score. However, the target mean vector given by Eq. (8) only expresses average voice timbre of several pre-stored target singers. Therefore, the converted singing voice doesn't express voice timbre of the source singer.

For the purpose of developing SVC based on perceptual age while retaining the source singer's individuality, we change the representative form of the target mean vector as follows:

$$\begin{aligned} \boldsymbol{\mu}_m^{(Y)}(o) &= \boldsymbol{b}_m^{(Y)} w^{(o)} + \overline{\boldsymbol{\mu}}_m^{(Y)} \\ &= \boldsymbol{b}_m^{(Y)} (w^{(i)} + \Delta w) + \overline{\boldsymbol{\mu}}_m^{(Y)} \\ &= \boldsymbol{b}_m^{(Y)} w^{(i)} + \overline{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{b}_m^{(Y)} \Delta w \\ &\simeq \hat{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{b}_m^{(Y)} \Delta w, \end{aligned} \tag{12}$$

where the perceptual age score of the target singing voice $w^{(o)}$ is represented by that of the source singing voice $w^{(i)}$ and the difference in perceptual age score Δw between them. In the modified representative form, the target mean vector is represented by the source mean vector $\hat{\boldsymbol{\mu}}_m^{(Y)}$ and the additional vector corresponding to a difference perceptual age score Δw. As the source mean vector $\hat{\boldsymbol{\mu}}_m^{(Y)}$ is directly used instead of its projection on the subspace $\boldsymbol{b}_m^{(Y)} w^{(i)} + \overline{\boldsymbol{\mu}}_m^{(Y)}$, it is expected that acoustic characteristics of the source singer's singing voice are preserved in this modified representative form.

5. EXPERIMENTAL EVALUATIONS

We define the conversion method in Section 4.2 as MR-GMM and in Section 4.3 as Modified MR-GMM in this experiment. Table 1 shows the source and target features of MR-GMM and Modified MR-GMM.

5.1. Experimental Conditions

We used the AIST humming database [22] consisting of sungs 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 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 [23]. The frame shift was 5 ms.

In the training of the MR-GMM, we prepared parallel data sets of a single female reference singer in her 20s and 27 male and 27 female singers in their 20s, 30s, 40s and 50s as pre-stored target singers not included in the 16 evaluation singers. The number of training sungs was 25 for each singer. The length of each song was approximately 20 seconds. The perceptual age score for each singer

was determined as an average score over 25 sungs of the singer rated by one male subject in his 20s [14]. The number of mixture components of the MR-GMM was 128 for spectral envelope and 32 for aperiodic components.

First, we evaluated the extent to which MR-GMM was able to perceivably change perceptual age. Eight male subjects in their 20s 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. We changed the difference perceptual age score in Eq. (12) into -60, -40, -20, 0, 20, 40 and 60. Subjects were asked to guess the age of each converted singing voice by listening to it in random order.

In the second experiment, we conducted an XAB test on the singer individuality of both MR-GMM and Modified MR-GMM. Subjects and evaluation singers were separated into two groups in the same manner as the first experiment. We changed perceptual age score in Eq. (12) into -60, -30, 30 and 60 in the Modified MR-GMM. In the MR-GMM, we varied the perceptual age score $\pm 30, 60$ in Eq. (8) from the average perceptual age results in each evaluation singer of intra-singer SVC (Δw = 0) in the prior experiment. A pair of songs generated by the MR-GMM and Modified MR-GMM of the same singer and the specified variation of perceptual age scores was presented to subjects after presenting the intra-singer SVC as a reference. Then, they were asked which voice sounded more similar to the reference in terms of the singer individuality.

In the final experiment, we evaluated the naturalness 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 perceptual age score was the same as for the second experiment. Subjects rated the naturalness of the converted sungs using a 5–point scale: "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad.

5.2. Experimental Results

Figure 1 shows the relationship between the age change setting and actual perceived perceptual age change in the Modified MR-GMM. For settings from -60 to 60, the perceptual age of the singer varied almost linearly. Especially, we can see the same tendency as observed in the investigation of segmental features shown in Figure 3 of [14], where changing segmental features from those of a 20-year-old singer to those of a 60-year-old singer resulted in a linear change of about 5 years. This indicates that the Modified MR-GMM is able to achieve a change in perceptual age similar to that achieved using natural spectral parameters of a singer in a different age group.

Figure 2 indicates the result of the XAB test for the singer individuality. We can see that as we make larger changes in perceptual age, the preference score of the Modified MR-GMM tends to decrease. However Modified MR-GMM has a higher preference score than the MR-GMM for each setting.

Figure 3 indicates the results of MOS test for the naturalness. This figure has the same tendency as displayed in Figure 2. The Modified MR-GMM has a higher MOS than the MR-GMM for each setting. The bias vectors of the Modified MR-GMM ($\hat{\boldsymbol{\mu}}_m^{(Y)}$ in Eq. (12)) model voice timbre of the source singer. On the other hand, those of the MR-GMM ($\overline{\boldsymbol{\mu}}_m^{(Y)}$ in Eq. (12)) model voice timbre of multiple pre-stored target singers. Therefore, over-smoothing effects of the MR-GMM tend to be larger than those of the Modified MR-GMM. Consequently, the naturalness of the sungs is also improved by using the Modified MR-GMM.

These results suggest that 1) the Modified MR-GMM enables to control the perceptual age of sungs relatively well, 2) the Modified MR-GMM is able to retain the singer individuality better than the MR-GMM during the perceptual age control, and 3) the Modified MR-GMM also generates better quality converted sungs compared with the MR-GMM.

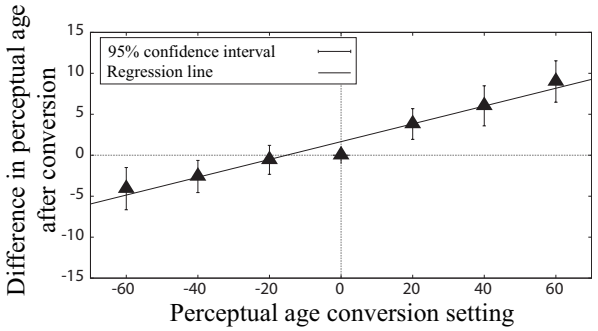

Fig. 1. Setting and actual differential in perceptual age after conversion.

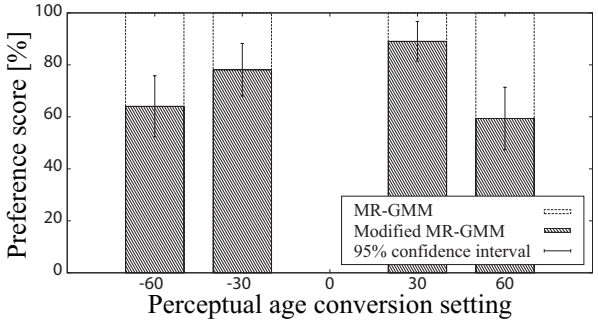

Fig. 2. Comparing singer individuality of MR-GMM and Modified MR-GMM converted singing voice.

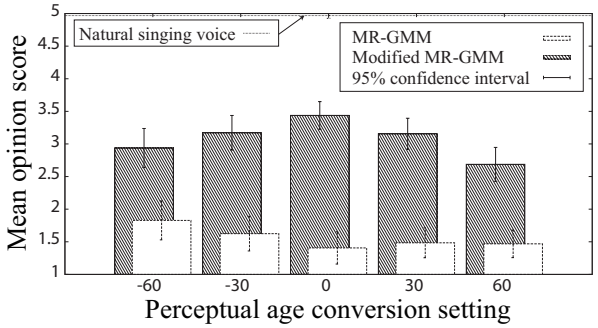

Fig. 3. Mean opinion score of MR-GMM and Modified MR-GMM.

6. CONCLUSION

In order to develop voice timbre control based on perceptual age, we have proposed a method for controlling perceptual age that retains the singer's individuality. A conventional voice timbre control technique based on multiple-regression Gaussian mixture models (MR-GMM) was not able to control the singer's perceptual age while retaining the singer's individuality. To solve this problem, we proposed a modified version of the MR-GMM that overcomes this problem. The experimental results have demonstrated that the proposed method makes it possible to control the singer's perceptual age while preserving individuality.

Acknowledgments

This work was supported in part by JSPS KAKENHI Grant Number 22680016 and by the JST OngaCREST project.

7. REFERENCES

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

[2] K. 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, "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 2011.

[4] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Information and Systems*, vol. E90-D, no. 9, pp. 1406–1413, Sep. 2007.

[5] 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.

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

[7] T. Nose, M. Kanemoto, T. Koriyama, and T. Kobayashi, "A style control technique for singing voice synthesis based on multiple-regression HSMM," *Proc. INTERSPEECH*, pp. 378–382, Aug. 2013.

[8] 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.

[9] 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.

[10] 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.

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

[13] 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.

[14] K. Kobayashi, H. Doi, T. Toda, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura, "An investigation of acoustic features for singing voice conversion based on perceptual age," *Proc. INTERSPEECH*, pp. 1057–1061, Aug. 2013.

[15] 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.

[16] 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.

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

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.