PAPER

Improvements of Voice Timbre Control Based on Perceived Age in Singing Voice Conversion

Kazuhiro KOBAYASHI[†a)], ***Nonmember***, **Tomoki TODA**[††], ***Member***, **Tomoyasu NAKANO**[†††], **Masataka GOTO**[†††], ***Nonmembers***, ***and*** **Satoshi NAKAMURA**[†], ***Member***

SUMMARY As one of the techniques enabling individual singers to produce the varieties of voice timbre beyond their own physical constraints, a statistical voice timbre control technique based on the perceived age has been developed. In this technique, the perceived age of a singing voice, which is the age of the singer as perceived by the listener, is used as one of the intuitively understandable measures to describe voice characteristics of the singing voice. The use of statistical voice conversion (SVC) with a singer-dependent multiple-regression Gaussian mixture model (MR-GMM), which effectively models the voice timbre variations caused by a change of the perceived age, makes it possible for individual singers to manipulate the perceived ages of their own singing voices while retaining their own singer identities. However, there still remain several issues; e.g., 1) a controllable range of the perceived age is limited; 2) quality of the converted singing voice is significantly degraded compared to that of a natural singing voice; and 3) each singer needs to sing the same phrase set as sung by a reference singer to develop the singer-dependent MR-GMM. To address these issues, we propose the following three methods; 1) a method using gender-dependent modeling to expand the controllable range of the perceived age; 2) a method using direct waveform modification based on spectrum differential to improve quality of the converted singing voice; and 3) a rapid unsupervised adaptation method based on maximum a posteriori (MAP) estimation to easily develop the singer-dependent MR-GMM. The experimental results show that the proposed methods achieve a wider controllable range of the perceived age, a significant quality improvement of the converted singing voice, and the development of the singer-dependnet MR-GMM using only a few arbitrary phrases as adaptation data.

key words: *statistical singing voice conversion, perceived age, gender-dependent modeling, direct waveform modification, unsupervised adaptation*

1. Introduction

Singers can express various singing expressions by using not only the linguistic information of the lyrics but also pitch, dynamics, and rhythm. Voice timbre can also be changed to some extent but it is essentially difficult to widely control due to physical constraints in speech production of the singers. Towards the development of new forms of singing expression in music, several techniques to widely control voice timbre of a singing voice beyond each singer's physical constraints have been proposed, such as a singing voice morphing technique [1] in the speech analysis/synthesis framework [2] or a singing voice conversion (SVC) technique [3]–[5] based on statistical voice conversion techniques [6], [7].

In our previous work, we have proposed a method for controlling voice timbre of the singing voice of a specific singer by manipulating its perceived age, which is the age of the singer as perceived by the listener [8]. This method makes it possible to use the perceived age as an intuitively understandable and controllable measure to describe converted singing voice characteristics. Such a voice timbre control process has been successfully implemented within the SVC framework [8], inspired by statistical voice conversion techniques based on a Gaussian mixture model (GMM) [6], [7] and voice quality control techniques based on a multiple-regression GMM (MR-GMM) [9]. Using multiple parallel data sets consisting of phrase pairs sung by a single reference singer and multiple pre-stored target singers, a singer-independent MR-GMM is first developed to model voice timbre variations caused by the perceived age differential. Then, it is further adapted to each singer (i.e., a user of the voice timbre control) by using only his/her corresponding parallel data set to develop a singer-dependent MR-GMM to model voice timbre variations likely observed in his/her singing voices.

Although our previously proposed method makes it possible for individual singers to manipulate the perceived ages of their own singing voices while retaining their own singer identities, there still remain several issues to be addressed. In this paper, we focus on the following three issues.

- **1) A controllable range of the perceived age is limited.** Significantly large quality degradation in the converted singing voice tends to be easily caused if setting the perceived age to ±5 ages from the original one. In the previous method, the single MR-GMM is used to model the voice timbre variation caused by a change of the perceived age assuming that it can be shared among all singers. On the other hand, it has been observed from results of analysis of normal voices that spectral variations caused by aging are different between male and female speakers [10], [11]. Therefore, it is expected that modeling accuracy of voice timbre variations in singing voices caused by a change of the perceived age is also improved by considering gender dependency, and it will expand the controllable range of

Manuscript received May 31, 2016.
Manuscript publicized July 21, 2016.
[†]The authors are with Nara Institute of Science and Technology (NAIST), Ikoma-shi, 630–0192 Japan.
[††]The author is with Information Technology Center, Nagoya University, Nagoya-shi, 464–8601 Japan.
[†††]The authors are with National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba-shi, 305–8568 Japan.
a) E-mail: kazuhiro-k@is.naist.jp
DOI: 10.1587/transinf.2016EDP7234

the perceived age.

- **2) Quality of the converted singing voice is significantly degraded compared to that of a natural singing voice.** The use of a vocoder [12] to synthesize the converted singing voice is one of the biggest factors causing this degradation because sound quality of the converted singing voice suffers from various errors in vocoding process, such as F_0 extraction errors and modeling errors in spectral parameterization. Moreover, an over-smoothing effect on the converted acoustic features is also a well-known factor causing the quality degradation [13], [14]. These issues are indeed hard to be addressed even by using high-quality vocoder systems [15]–[18].
- **3) Each singer always needs to sing the same phrase set as sung by the reference singer to develop the singer-dependent MR-GMM.** In the previous method, the singer-dependent MR-GMM parameters need to be estimated using a parallel data set consisting of the singer's singing voices corresponding to the reference singer's singing voices because the MR-GMM models a joint probability density function of those two singers' acoustic features. It is more convenient to develop a more flexible framework capable of using only a few phrases or accepting arbitrary phrases even if they are not the same as sung by the reference singer.

Towards the development of a better controllable, higher-quality, and more flexible framework compared to the previous one, we propose the following three methods to address the above three issues;

- **1) a method using gender-dependent MR-GMMs** that can more accurately model the spectral variations caused by a change of the perceived age in each gender to expand the controllable range of the perceived age;
- **2) a method using direct waveform modification based on spectrum differential** to improve quality of the converted singing voice by avoiding using vocoder in converted waveform generation; and
- **3) a rapid unsupervised adaptation method** based on maximum a posteriori (MAP) estimation [19]–[21] to easily develop the singer-dependent MR-GMM.

It is shown from results of several subjective evaluations that the proposed methods yield significant improvements in controllability of the perceived age, quality of the converted singing voices, and flexibility of the development of the singer-dependent MR-GMM. In this paper, we present further details of the proposed method, more discussions, and more evaluations than those in our previous work [22].

2. Voice Timbre Control Based on Perceived Age while Retaining Singer Individuality

2.1 Training Process

2.1.1 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 function of $2D$-dimensional joint static and dynamic feature vectors modeled by the MR-GMM is given by

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

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

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 $\boldsymbol{Y}_t(s) = [\boldsymbol{Y}_t^\top(s), \Delta\boldsymbol{Y}_t^\top(s)]^\top$ are joint static and delta feature vectors of the reference singer and the s-th pre-stored target singer at frame t, which are automatically aligned to each other by applying dynamic time warping to their corresponding singing voices. The vectors $\boldsymbol{b}_m^{(Y)}$ and $\overline{\boldsymbol{\mu}}_m^{(Y)}$ indicate a representative vector to capture voice timbre variations caused by a change of the perceived age and a bias vector to capture voice characteristics averaged over all pre-stored target singers, 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.

These MR-GMM parameters are trained as shown in Fig. 1. First, a singer-independent GMM is trained using all of the multiple parallel data sets. And then, only its target mean vectors are updated separately using individual parallel data sets to develop singer-dependent GMMs of the individual pre-stored target singers. The updated target mean vectors of each singer-dependent GMMs are extracted as the singer-dependent target mean vectors, and then, linear regression is performed using them and the corresponding perceived age scores among all pre-stored target singers to extract the representative vector and the bias vector. The other parameters of the MR-GMM are extracted from the singer-independent GMM.

To easily create the MR-GMMs for various source singers (i.e., users), the framework of the many-to-many SVC [5] is applied to the MR-GMM. The joint probability

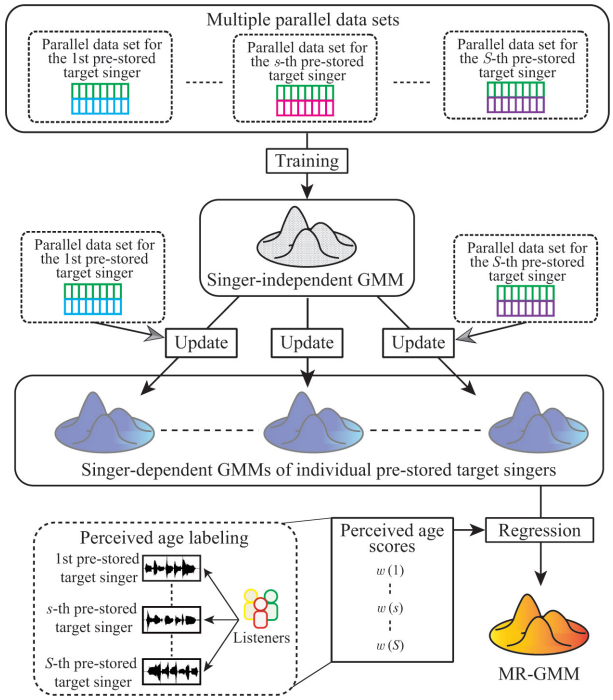

Fig. 1 Trainging process of the MR-GMM.

density function of many-to-many MR-GMM is analytically derived from that of the MR-GMM shown in Eq. (1), which is given by

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

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

where $w(i)$ and $w(o)$ indicate the perceived age scores of the source and target singers, respectively. The source and target mean vectors, $\overline{\boldsymbol{\mu}}_m^{(Y)}(i)$ and $\overline{\boldsymbol{\mu}}_m^{(Y)}(o)$, are modeled by the same subspace spanned by the representative vector and the bias vector as shown in Eq. (2). Therefore, the representation form of the target mean vectors can be reformulated by applying $w(o) = w(i) + \Delta w$, where the perceived age score of the target singing voice $w(o)$ can be represented by using that of the source singing voice $w(i)$ and the perceived age score differential Δw. The reformulated target mean vector of the m-th mixture component is given by

$$
\begin{aligned}
\overline{\boldsymbol{\mu}}_m^{(Y)}(o) &= \boldsymbol{b}_m^{(Y)}\left(w(i) + \Delta w\right) + \overline{\boldsymbol{\mu}}_m^{(Y)} \\
&= \boldsymbol{b}_m^{(Y)} w(i) + \overline{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{b}_m^{(Y)} \Delta w \\
&= \overline{\boldsymbol{\mu}}_m^{(Y)}(i) + \boldsymbol{b}_m^{(Y)} \Delta w. \quad (5)
\end{aligned}
$$

Namely, the target mean vectors of the many-to-many MR-GMM can be represented as the source mean vectors $\overline{\boldsymbol{\mu}}_m^{(Y)}(i)$, the representative vectors $\boldsymbol{b}_m^{(Y)}$, and the perceived age score differential Δw.

2.1.2 Adaptation of MR-GMM to a Specific Singer

To develop the voice timbre control system for each user, the many-to-many MR-GMM needs to be adapted to him/her. It is possible to do it by only adjusting the perceived age score $w(i)$ to adapt the source mean vectors $\overline{\boldsymbol{\mu}}_m^{(Y)}(i)$, which are represented by Eq. (2). However, modeling accuracy of the MR-GMM adapted by this approach is usually insufficient due to the limited representation of the adapted source mean vectors, which also affect the target mean vectors as shown in Eq. (5). To develop a better singer-dependent MR-GMM by more accurately adapting the many-to-many MR-GMM, the source mean vectors $\overline{\boldsymbol{\mu}}_m^{(Y)}(i)$ are directly updated by using parallel data sets between the reference singer and the singer to be adapted.

Let $\boldsymbol{Y}_t(k) = \left[\boldsymbol{Y}_t^{\top}(k), \Delta \boldsymbol{Y}_t^{\top}(k)\right]^{\top}$ denote the joint static and delta feature vector at frame t of the singer k to be adapted. The updated source mean vector set $\hat{\boldsymbol{\mu}}(k) = \{\hat{\boldsymbol{\mu}}_1(k), \cdots, \hat{\boldsymbol{\mu}}_M(k)\}$ is determined as the target mean vectors updated by maximizing the likelihood function of the MR-GMM given in Eq. (1) as follows:

$$
\hat{\boldsymbol{\mu}}(k) = \underset{\boldsymbol{\mu}(k)}{\operatorname{argmax}} \prod_{t=1}^{T} P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t(k)|\lambda^{(MR)}, \boldsymbol{\mu}(k)\right). \quad (6)
$$

This adaptation process is performed using the EM algorithm by maximizing the following auxiliary function:

$$
\begin{aligned}
Q\left(\boldsymbol{\mu}(k), \hat{\boldsymbol{\mu}}(k)\right) = \sum_{t=1}^{T} \sum_{m=1}^{M} P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t(k), \lambda^{(MR)}, \boldsymbol{\mu}_m(k)\right) \\
\log P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t(k), m|\lambda^{(MR)}, \hat{\boldsymbol{\mu}}_m(k)\right). \quad (7)
\end{aligned}
$$

The ML estimate of the m-th target mean vector $\hat{\boldsymbol{\mu}}_m(k)$ is calculated as follows:

$$
\begin{aligned}
\hat{\boldsymbol{\mu}}_m(k) = &\left\{\sum_{m=1}^{M} \Gamma_m \boldsymbol{P}_m^{(YY)}\right\}^{-1} \\
&\left\{\sum_{m'=1}^{M} \boldsymbol{P}_{m'}^{(YY)} \overline{\boldsymbol{Y}}_{m'}(k) + \boldsymbol{P}_{m'}^{(YX)}\left(\overline{\boldsymbol{X}}_{m'} - \Gamma_{m'} \boldsymbol{\mu}_{m'}^{(X)}\right)\right\}, \quad (8)
\end{aligned}
$$

where

$$
\Gamma_m = \sum_{t=1}^{T} P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t(k), \lambda^{(MR)}, \boldsymbol{\mu}_m(k)\right), \quad (9)
$$

$$
\overline{\boldsymbol{Y}}_m(k) = \sum_{t=1}^{T} P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t(k), \lambda^{(MR)}, \boldsymbol{\mu}_m(k)\right) \boldsymbol{Y}_t(k), \quad (10)
$$

$$
\overline{\boldsymbol{X}}_m = \sum_{t=1}^{T} P\left(m|\boldsymbol{X}_t, \boldsymbol{Y}_t(k), \lambda^{(MR)}, \boldsymbol{\mu}_m(k)\right) \boldsymbol{X}_t, \quad (11)
$$

$$
\begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{P}_m^{(XX)} & \boldsymbol{P}_m^{(XY)} \\ \boldsymbol{P}_m^{(YX)} & \boldsymbol{P}_m^{(YY)} \end{bmatrix}. \quad (12)
$$

The updated mean vectors $\hat{\boldsymbol{\mu}}_m(k)$ are applied to the many-to-many MR-GMM as follows:

$$\overline{\boldsymbol{\mu}}_m^{(Y)}(i) \simeq \hat{\boldsymbol{\mu}}_m(k), \tag{13}$$

$$\overline{\boldsymbol{\mu}}_m^{(Y)}(o) \simeq \hat{\boldsymbol{\mu}}_m(k) + \boldsymbol{b}_m^{(Y)}\Delta w. \tag{14}$$

These modification based on supervised adaptation using the parallel data set is effective for developing the singer-dependent MR-GMM capable of controlling the singer's perceived age while retaining the singer individuality.

2.2 Conversion Process

In the conversion process, the perceived age score differential Δw is manually set to a desired value. Then, a singing voice of the singer k is converted into his/her singing voice corresponding to the desired perceived age using maximum likelihood estimation of the speech parameter trajectory with the singer-dpendent MR-GMM [7].

Time sequence vectors of the source and converted features for the singer k are denoted as $\boldsymbol{Y}^{(i)}(k) = \left[\boldsymbol{Y}_1^{(i)}(k)^\top, \cdots, \boldsymbol{Y}_T^{(i)}(k)^\top\right]^\top$ and $\boldsymbol{Y}^{(o)}(k) = \left[\boldsymbol{Y}_1^{(o)}(k)^\top, \cdots, \boldsymbol{Y}_T^{(o)}(k)^\top\right]^\top$, where T is the number of frames over the given source feature vector sequence. A time sequence vector of the converted static features $\hat{\boldsymbol{y}}^{(o)}(k) = \left[\hat{\boldsymbol{y}}_1^{(o)}(k)^\top, \cdots, \hat{\boldsymbol{y}}_T^{(o)}(k)^\top\right]^\top$ is determined as follows:

$$\hat{\boldsymbol{y}}^{(o)}(k) = \underset{\boldsymbol{y}^{(o)}(k)}{\operatorname{argmax}} P\left(\boldsymbol{Y}^{(o)}(k)|\boldsymbol{Y}^{(i)}(k), \boldsymbol{\lambda}^{(MR)}, \hat{\boldsymbol{\mu}}(k), \Delta w\right)$$
$$\text{subject to } \boldsymbol{Y}^{(o)}(k) = \boldsymbol{W}\boldsymbol{y}^{(o)}(k), \tag{15}$$

where $\boldsymbol{W}$ is a transformation matrix to expand the static feature vector sequence into the joint static and dynamic feature vector sequence [23]. The conditional probability density function $P\left(\boldsymbol{Y}^{(o)}(k)|\boldsymbol{Y}^{(i)}(k), \boldsymbol{\lambda}^{(MR)}, \hat{\boldsymbol{\mu}}(k), \Delta w\right)$ is analytically derived from the singer-dependent MR-GMM for the singer k. To alleviate the over-smoothing effects that usually make the converted singing voice sound muffled, global variance (GV) [7] is also considered.

3. Proposed Techniques for Improving Voice Timbre Control Based on Perceived Age

To improve controllability of the perceived age, quality of the converted speech, and flexibility of the model development in the conventional voice timbre control method, we further implement three techniques, 1) gender-dependent MR-GMMs for more accurately capturing spectral variations depending on the perceived age, 2) direct waveform modification based on spectral differential, and 3) a rapid unsupervised adaptation method based on MAP estimation to easily develop the singer-dependent MR-GMM.

3.1 Gender-Dependent MR-GMM

Multiple parallel data sets used in the conventional training method of the MR-GMM consist of singing voice pairs of both male and female singers. To improve modeling accuracy of the MR-GMM on the voice timbre variations, we propose the gender-dependent modeling, inspired by the previous work showing that the voice timbre variations of normal voices caused by aging significantly depend on the gender [10], [11]. Two gender-dependent MR-GMMs are trained separately using the parallel data sets consisting of only male singers or female singers. And then, the singer-dependent MR-GMM for the specific singer is developed by adapting the corresponding gender-dependent MR-GMM to the singer in the same manner as described in Sect. 2.1.2. Note that not only the representative vectors but also the other parameters, such as the covariance matrices, are different between these two gender-dependent MR-GMMs.

3.2 Direct Waveform Modification Based on Spectral Differential

As a SVC framework without using vocoder-based waveform generation, we have proposed a direct waveform modification method based on spectral differential [24]. In this paper, this method is applied to the voice timbre control framework using the MR-GMM.

Figure 2 shows both conventional and proposed conversion processes. 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) modeling the joint probability density function of the source singer's spectral features and the spectral feature differential caused by the given perceived age 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 based on a differential MR-GMM (DIFFMR-GMM). The DIFFMR-GMM models the joint probability density function of the source singer's spectral features and the spectral feature differential caused by the given perceived age differential. This differential model can be analytically derived from the conventional singer-dependent MR-GMM by applying a simple linear transform to the conventional model. The source singer's spectral feature is converted into the spectral feature differential using the DIFFMR-GMM. Then, a waveform of the source singing voice is directly filtered with a time sequence of the estimated the spectral feature differentials. 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 analytically derived from the singer-dependent MR-GMM 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

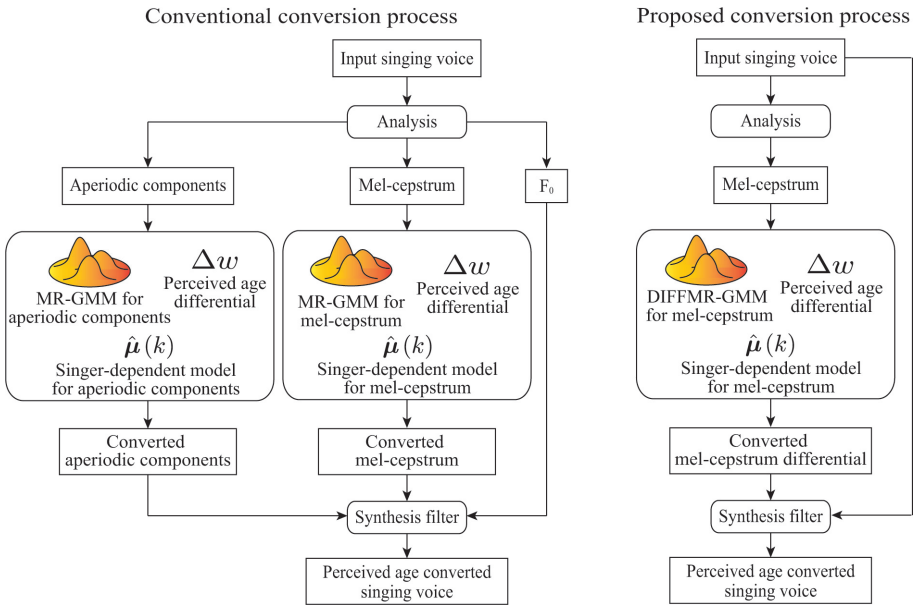

Fig. 2 Conventional and proposed conversion processes of perceived age control.

the differential features is represented as linear transformation of the original joint feature vectors as follows:

$$\begin{bmatrix} \boldsymbol{Y}_t(i) \\ \boldsymbol{D}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_t(i) \\ \boldsymbol{Y}_t(o) - \boldsymbol{Y}_t(i) \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\boldsymbol{I} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{Y}_t(i) \\ \boldsymbol{Y}_t(o) \end{bmatrix}, \quad (16)$$

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

$$P\left(\boldsymbol{Y}_t(i), \boldsymbol{D}_t | \boldsymbol{\lambda}^{(DIFFMR)}, \hat{\boldsymbol{\mu}}(Y), \Delta w\right)$$
$$= \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{Y}_t(i) \\ \boldsymbol{D}_t \end{bmatrix}; \begin{bmatrix} \hat{\boldsymbol{\mu}}_m^{(Y)} \\ \boldsymbol{b}_m^{(Y)} \Delta w \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(DYD)} \\ \boldsymbol{\Sigma}_m^{(DYD)} & \boldsymbol{\Sigma}_m^{(DD)} \end{bmatrix}\right), \quad (17)$$

$$\boldsymbol{\Sigma}_m^{(DYD)} = \boldsymbol{\Sigma}_m^{(YXY)} - \boldsymbol{\Sigma}_m^{(YY)}, \quad (18)$$

$$\boldsymbol{\Sigma}_m^{(DD)} = 2(\boldsymbol{\Sigma}_m^{(YY)} - \boldsymbol{\Sigma}_m^{(YXY)}). \quad (19)$$

In the conversion process, the converted differential feature vector is determined in the same manner as described in Sect. 2.2 except for not considering the GV[†].

3.3 Unsupervised Adaptation

To make it possible to reduce the amount of singing voices and also accept arbitrary phrases used as the adaptation data to develop the singer-dependent MR-GMM, we propose an unsupervised adaptation technique based on the MAP estimation. Figure 3 shows the conventional and proposed methods for developing the singer-dependent MR-GMM.

As the prior distribution for the MAP adaptation, the following Gaussian distribution is employed:

$$P\left(\boldsymbol{\mu} | \lambda^{(pri)}\right) = \prod_{m=1}^{M} \mathcal{N}\left(\boldsymbol{\mu}_m; \boldsymbol{\mu}_m^{(pri)}, \boldsymbol{\Sigma}_m^{(pri)}\right), \quad (20)$$

[†]We can also consider the GV in the conversion process based on the spectrum differential as presented in [25].

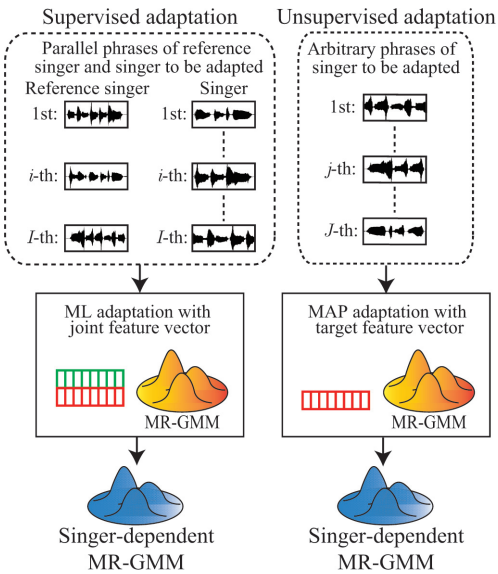

Fig. 3 Adaptation process of perceived age control based on singer-dependent MR-GMM.

where $\lambda^{(pri)}$ is a model parameter set consisting of the mean vectors $\boldsymbol{\mu}^{(pri)} = \left\{\boldsymbol{\mu}_1^{(pri)}, \cdots, \boldsymbol{\mu}_M^{(pri)}\right\}$ and the covariance matrices $\boldsymbol{\Sigma}^{(pri)} = \left\{\boldsymbol{\Sigma}_1^{(pri)}, \cdots, \boldsymbol{\Sigma}_M^{(pri)}\right\}$. This model parameter set is trained in advance using a set of the singer-dependent target mean vectors of all pre-stored target singers as follows:

$$\hat{\lambda}^{(pri)} = \underset{\lambda^{(pri)}}{\operatorname{argmax}} \prod_{s=1}^{S} P\left(\boldsymbol{\mu}^{(Y)}(s) | \lambda^{(pri)}\right), \quad (21)$$

where $\boldsymbol{\mu}^{(Y)}(s) = \left\{\boldsymbol{\mu}_1^{(Y)}(s), \cdots, \boldsymbol{\mu}_M^{(Y)}(s)\right\}$. For the given adaptation data, $\boldsymbol{Y}(k) = \left[\boldsymbol{Y}_1^{\top}(k), \cdots, \boldsymbol{Y}_T^{\top}(k)\right]^{\top}$, which denotes a time sequence of the feature vector of the singer k, the MAP

adaptation of the MR-GMM is conducted as follows:

$$\hat{\boldsymbol{\mu}}(k) = \underset{\boldsymbol{\mu}(k)}{\operatorname{argmax}} P\left(\boldsymbol{\mu}(k)|\lambda^{(pri)}\right)^{\tau} \int P\left(\boldsymbol{X}, \boldsymbol{Y}(k)|\lambda^{(MR)}, \boldsymbol{\mu}(k)\right) d\boldsymbol{X}$$
$$= \underset{\boldsymbol{\mu}(k)}{\operatorname{argmax}} P\left(\boldsymbol{\mu}(k)|\lambda^{(pri)}\right)^{\tau} \prod_{t=1}^{T} \int P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t(k)|\lambda^{(MR)}, \boldsymbol{\mu}(k)\right) d\boldsymbol{X}_t$$
$$= \underset{\boldsymbol{\mu}(k)}{\operatorname{argmax}} P\left(\boldsymbol{\mu}(k)|\lambda^{(pri)}\right)^{\tau} \prod_{t=1}^{T} P\left(\boldsymbol{Y}_t(k)|\lambda^{(MR)}, \boldsymbol{\mu}(k)\right). \quad (22)$$

where τ is a hyper-parameter controlling the balance between the prior distribution of mean vectors and the marginalized distribution $P\left(\boldsymbol{Y}(k)|\lambda^{(MR)}, \boldsymbol{\mu}(k)\right)$. The MAP estimate is determined using the EM algorithm by maximizing the following auxiliary function:

$$Q(\boldsymbol{\mu}(k), \hat{\boldsymbol{\mu}}(k)) = \tau \sum_{m=1}^{M} \log P\left(\hat{\boldsymbol{\mu}}_m(k)|\lambda^{(pri)}\right) + \sum_{t=1}^{T} \sum_{m=1}^{M} P\left(m|\boldsymbol{Y}_t(k), \lambda^{(MR)}, \boldsymbol{\mu}_m(k)\right) \log P\left(\boldsymbol{Y}_t(k), m|\lambda^{(MR)}, \hat{\boldsymbol{\mu}}_m(k)\right). \quad (23)$$

The MAP estimate is given by

$$\hat{\boldsymbol{\mu}}_m(k) = \left\{\tau \boldsymbol{\Sigma}_m^{(pri)^{-1}} + \Gamma_m \boldsymbol{\Sigma}_m^{(YY)^{-1}}\right\}^{-1} \cdot \left\{\tau \boldsymbol{\Sigma}_m^{(pri)^{-1}} \boldsymbol{\mu}_m^{(pri)} + \boldsymbol{\Sigma}_m^{(YY)^{-1}} \overline{\boldsymbol{Y}}_m(k)\right\}. \quad (24)$$

4. Experimental Evaluations

4.1 Overall Experimental Conditions

Table 1 indicates a simple description of the experimental condtions. We used the AIST humming database [26] consisting of phrases 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 [15] 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 [27]. The frame shift was 5 ms. The mel log spectrum approximation filter [28] was used as the synthesis filter in both the conventional waveform generation with vocoder and the proposed direct waveform modification.

Table 1 Experimental conditions.

Singing voice database	AIST humming database
Sampling frequency	16 [kHz]
Duration of one phrase	about 20 [s]
The number of training singers	28 males, 28 females
The number of evaluation singers	8 males, 8 females
The number of training data	23 phrases
The number of subjects	8

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. Each singer sung 23 phrases, where the duration of each phrase was approximately 20 seconds. The number of mixture components of each MR-GMM was 128 for the spectral feature and 64 for the aperiodic components. We have developed the singer-dependent MR-GMMs for 16 singers consisting of two male and two female singers in each age group (20s, 30s, 40s, and 50s), who were not included in the pre-stored target singers, and conducted voice timbre control evaluations for these singers. We used P039 as an evaluation phrase. The perceived age score for each singer was determined as an average score of the singer rated by 8 subjects in their 20s [8].

4.2 Experimental Evaluation of Gender-Dependent Modeling and Direct Waveform Modification

4.2.1 Experimental Conditions

To examine the effectiveness of two proposed techniques, the gender-dependent modeling and the direct waveform modification, singing voices converted by the following three methods were evaluated:

- SVC (GI): converted with the gender-independent MR-GMM
- SVC (GD): converted with the gender-dependent MR-GMM
- DIFFSVC (GD): converted with the gender-dependent DIFFMR-GMM and the direct waveform modification

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 training phrases for the development of the singer-dependent MR-GMM was 23 in each singer. Figure 4 indicates a method of dividing 16 evaluation singers into two groups. 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 was assigned one evaluation singer group in each evaluation in order to evaluate the evaluation singers of both genders and all age groups.

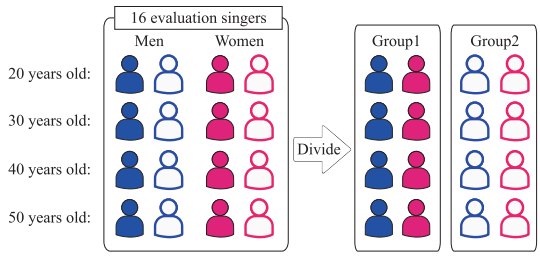

Fig. 4 Method for dividing 16 evaluation singers into two groups.

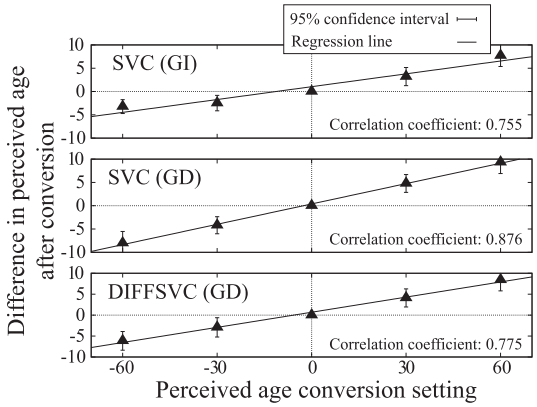

Fig. 5 Experimental result on perceived age controllability.

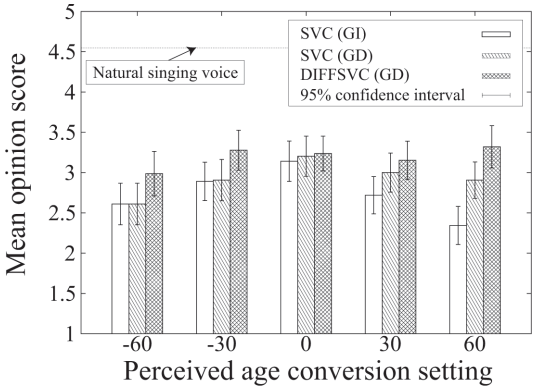

Fig. 6 Mean opinion score of speech quality.

First, we evaluated perceived age controllability. The number of converted singing voice of a evaluation singer was 15. Each subject evaluated the converted singing voices of 120 phrases from only one group of the evaluation singers. Subjects were asked to assign the perceived age to each converted singing voice sample 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). Each subject evaluated the natural and converted singing voices of evaluation singers. The number of evaluation phrases in each subject is 128. 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 the conventional method SVC (GI) and the proposed method DIFFSVC (GD). The evaluation singers were separated into two groups and each subject evaluated the converted singing voices from only one group in the same manner as the first experiment. A pair of singing voices converted by SVC (GI) and by DIFFSVC (GD) for the same singer with the same setting of the perceived age score differential 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. The number of evaluation pairs in each subject is 40.

4.2.2 Experimental Results

Figure 5 shows the relationship between the perceived age differentials given to the system to generate the converted singing voices and their perceived ages actually evaluated by the listeners. We can see that using the proposed gender-dependent models (SVC(GD) and DIFFSVC(GD)), the perceived age varies more linearly according to a change of the settings of the perceived age differential from -60 to 60 compared to the conventional gender-independent model (SVC(GI)). Moreover, a range of the perceived age of the converted singing voice becomes wider by using SVC (GD) and DIFFSVC (GD) compared to SVC (GI). These results indicate that voice timbre variations caused by the perceived age depend on the gender in singing voices and they are well modeled by using the proposed gender-dependent modeling technique.

Figure 6 indicates the results of the opinion test on the quality. We can see that DIFFSVC (GD) tends to significantly improve quality of the converted singing voices compared to SVC (GI) and SVC (GD). Although the quality is greatly degraded in the conventional method SVC (GI) as the perceived age score differential is set to larger or smaller values, this quality degradation is effectively alleviated by the proposed method DIFFSVC (GD) because the DIFSVC (GD) method can avoid the errors caused by spectrum parameterization and excitation generation. In comparison between SVC (GD) and SVC (GI), the speech quality of SVC (GD) is improved compared with that of SVC (GI) as the perceived age score differential is set to higher values (+30, +60). On the other hand, in terms of setting lower values ($-60, -30$), we can see that there is no significant difference between these methods. As shown in Fig. 5, the perceived age differential achieved by SVC (GI) tends to be smaller than that by SVC (GD) when setting the perceived age score differential to -60. This result implies that the resulting acoustic changes by SVC (GI) are smaller than those by SVC (GD) under such a setting, also making the quality degradation in SVC (GI) smaller. Even in such an unfair condition, SVC (GD) causes no quality degradation compared to SVC (GI).

Figure 7 indicates the result of the XAB test on the singer individuality. DIFFSVC (GD) better or equally retains singer individuality in any perceived age setting compared to the conventional method SVC (GI). We can see that as a change of the perceived age differential setting is larger, the difference between DIFFSVC (GD) and SVC (GI) becomes smaller. In particular, no difference is observed between them when setting the perceived age differential to -60 while the significant difference is still observed when setting it to 60. It is expected that this result is also caused by the resulting acoustic changes by SVC (GI) is smaller than SVC (GD) when setting the perceived age differential

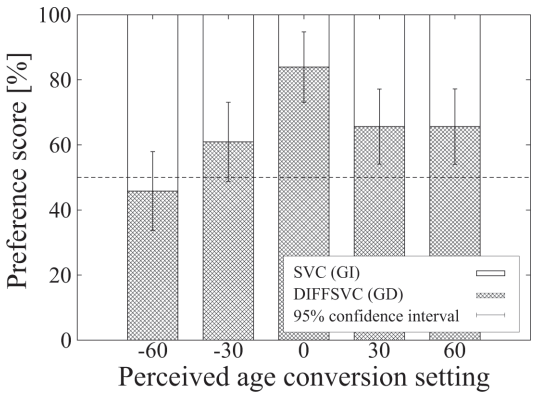

Fig. 7 Preference score on singer individuality.

to −60 as mentioned above.

These results suggest that 1) the gender-dependent modeling technique 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.

Although the proposed method DIFFSVC (GD) makes it possible to control the perceived age with higher speech quality compared to the conventional method SVC (GI) and SVC (GD) in Figs. 6, 7, there still remains the speech quality degradation compared to the natural singing voice. It is expected that this degradation is caused by insufficient modeling accuracy of the perceived age variations using the gender-dependent MR-GMM. Therefore, it is worthwhile to further improve the modeling accuracy.

4.3 Experimental Evaluation of Unsupervised Adaptation

4.3.1 Experimental Conditions

In this evaluation, we varied the number of the adaptation phrases as 1, 6, 12, and 22 in order to evaluate the effectiveness of the proposed unsupervised adaptation technique. The adaptation phrases are selected in order from the beginning of the index of singing voice database. In this evaluation, the ML estimation with parallel phrases was used as the supervised adaptation and the MAP estimation with only phrases of each evaluation singer was used as the unsupervised adaptation. The hyper-parameter τ for the MAP adaptation was manually set to 3.0 in the subjective evaluations.

First, we evaluated the modeling accuracy of the singer-dependent MR-GMMs developed with the adaptation approaches using Mahalanobis distance of their mean vectors to those of the singer-dependent MR-GMMs developed with the conventional supervised approach using 22 parallel phrases in each singer, which is calculated as

$$D(i) = \frac{1}{L}\sum_{l=1}^{L}\sum_{m=1}^{M}\alpha_m \left(\boldsymbol{\mu}_m^{(22)}(l)-\hat{\boldsymbol{\mu}}_m^{(i)}(l)\right)^{\top}\Sigma_m^{(YY)^{-1}}\left(\boldsymbol{\mu}_m^{(22)}(l)-\hat{\boldsymbol{\mu}}_m^{(i)}(l)\right), \quad (25)$$

where L denotes the number of evaluation singers. $\hat{\boldsymbol{\mu}}_m^{(i)}(l)$ denotes the adapted singer-dependent MR-GMM for the evaluation singer l using his/her i phrases in the unsupervised adaptation or i parallel phrases in the supervised adaptation. Note that the mean vectors of the singer-dependent MR-GMM used as a target $\boldsymbol{\mu}_m^{(22)}(l)$ in this distance calculation is equivalent to those determined using the supervised ML adaptation using 22 parallel phrases.

In the second experiment, we evaluated the conversion accuracy using the mel-cepstrum distortion as an evaluation metric in the different settings of the hyper-parameter τ. The mel-cepstrum distortion was calculated as follows:

$$\text{Mel-CD}(i)\ [dB] = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{24}\left(mc_d^{(22)} - \hat{mc}_d^{(i)}\right)^2}, \quad (26)$$

where $mc_d^{(22)}$ denotes the mel-cepstrum coefficients analyzed from the converted singing voice generated with the singer-dependent MR-GMM developed with the supervised ML adaptation using 22 parallel phrases, and $\hat{mc}_d^{(i)}$ denotes those developed with the unsupervised MAP adaptation using i phrases. The setting of the hyper-parameter τ is varied from 0, 1, 3, 6, 12, to 24. Note that the setting of $\tau = 0$ corresponds to the unsupervised ML adaptation.

In the third experiment, we evaluated the perceived age controllability. The number of adaptation phrases was set to 1 and 6. The 16 evaluation singers were divided into four groups. 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 in one group of the evaluation singers by listening to it in random order. The number of evaluation samples in each subject was 48.

In the final experiment, we evaluated the quality of the converted singing voice using a opinion test. The number of subjects and evaluation singers were the same as in the second experiment. The subjects evaluated quality of the converted singing voices in the same manner as described in 4.2.1.

4.3.2 Experimental Results

Figure 8 indicates the Mahalanobis distances as a function of the number of adaptation phrases. The distance when using 1 parallel phrase in the ML adaptation is very large. On the other hand, the distance using 1 phrase in the MAP adaptation is significantly lower than it. In the ML adaptation, it is necessary to use 6 or more parallel phrases to reduce the distance as small as in the MAP adaptation.

Figure 9 shows the mel-cepstrum distortion as a function of the number of adaptation phrases in each hyper-parameter setting. We can see that the unsupervised adaptation using either ML or MAP is effective. The unsupervised ML adaptation ($\tau = 0$) causes significantly large degradation when using only one adaptation phrase. On the other hand, such a degradation is effectively alleviated by using

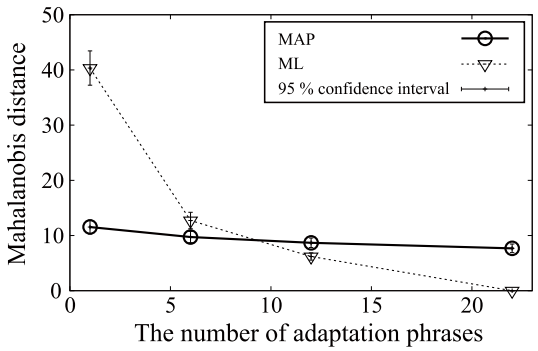

Fig. 8 Mahalanobis distance as a function of the number of adaptation phrases.

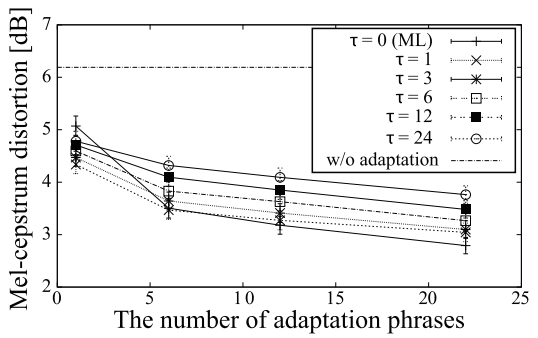

Fig. 9 Mel-cepstrum distortion as a function of the number of adaptation phrases and hyper-parameters settings.

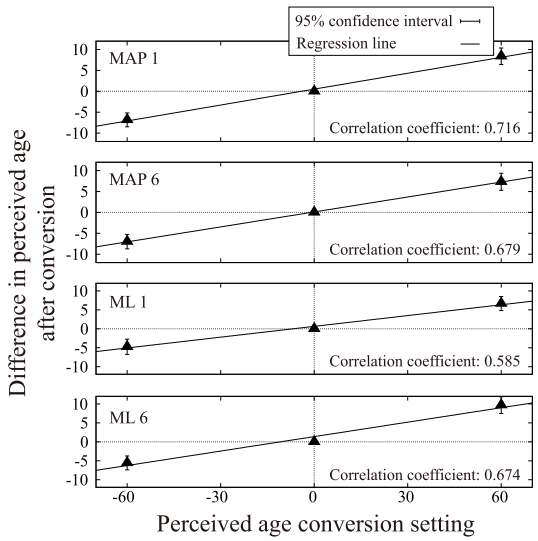

Fig. 10 Experimental result on perceived age controllability of the adapted MR-GMMs.

the proposed MAP adaptation. We can also see that performance of the proposed MAP adaptation is affected by the hyper-parameter setting, and relatively good performance is achieved by setting the hyper parameter to a small value.

Figure 10 shows the experimental result on the perceived age controllability. We can see that the MAP adaptation using only 1 phrase has higher controllability compared to the ML adaptation in 1 parallel phrase and its controllability is similar to that of the MAP adaptation using 6 phrases and that of the ML adaptation using 6 parallel phrases. This tendency is consistent with that observed in the previous objective evaluation shown in Fig. 8. Moreover, comparing to the result described in Fig. 5, we can see that the proposed MAP adaptation method using only 1 phrase achieves similar controllability to the conventional method using 22 parallel phrases.

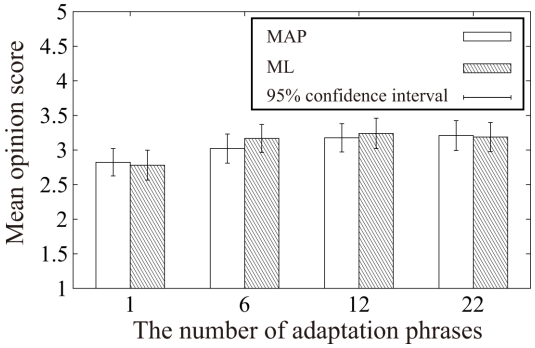

Fig. 11 Mean opinion score of speech quality depending on the number of adaptation phrases.

Figure 11 indicates the results of the opinion test on the speech quality. We can see that there is no significantly large quality difference between the MAP adaptation and the ML adaptation. We can also see that the quality of the converted singing voice tends to degrade if using only 1 phrase. This quality degradation is alleviated by increasing the number of adaptation phrases to 6 and the resulting quality reaches to that of the conventional method using 22 parallel phrases.

These results suggest that 1) the MAP adaptation outperforms the ML adaptation when a few phrases are available, and 2) the MAP adaptation by using only a small number of arbitrary phrases (e.g., 6 phrases) achieves almost the same controllability and quality of the converted singing voice as in the conventional method that needs a larger number of parallel phrases (e.g., 22 phrases).

5. 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 the gender-dependent modeling technique, the direct waveform modification technique with spectral differential, and the unsupervised adaptation technique based on maximum a posteriori (MAP) estimation. The experimental results have demonstrated that 1) the proposed methods can expand a range of the controllable perceived age wider, 2) the proposed methods can significantly improve quality of the converted singing voice, and 3) the proposed methods needs only a small number of arbitrary phrases from each user to develop the voice timbre control system for him/her. In future work, we will investigate

singer individuality of prosodic features and develop a technique to control the prosodic features by manipulating the perceived age while retaining the singer individuality.

Acknowledgments

This work was supported in part by JSPS KAKENHI Grant Number 26280060, Grant-in-Aid for JSPS Research Fellow Number 16J10726, and by the JST OngaCREST project.

References

[1] M. Morise, M. Onishi, H. Kawahara, and H. Katayose, "v.morish' 09: A morphing-based singing design interface for vocal melodies," Proc. ICEC, vol.5709, pp.185–190, 2009.
[2] 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, April 2009.
[3] F. Villavicencio and J. Bonada, "Applying voice conversion to concatenative singing-voice synthesis," Proc. INTERSPEECH, pp.2162–2165, Sept. 2010.
[4] Y. Kawakami, H. Banno, and F. Itakura, "GMM voice conversion of singing voice using vocal tract area function," IEICE Technical Report. SP2010-81, Nov. 2010.
[5] 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.
[6] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," IEEE Trans. SAP, vol.6, no.2, pp.131–142, March 1998.
[7] 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.
[8] 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.
[9] 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.
[10] W. Endres, W. Bambach, and G. Flösser, "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.842–1848, 1971.
[11] 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.
[12] H. Dudley, "Remaking speech," JASA, vol.11, no.2, pp.169–177, 1939.
[13] H. Zen, K. Tokuda, and A.W. Black, "Statistical parametric speech synthesis," Speech Communication, vol.51, no.11, pp.1039–1064, Nov. 2009.
[14] T. Merritt, T. Raitio, and S. King, "Investigating source and filter contributions, and their interaction, to statistical parametric speech synthesis," Proc. INTERSPEECH, pp.1509–1513, Sept. 2014.
[15] 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, April 1999.
[16] M. Morise, "An attempt to develop a singing synthesizer by collaborative creation," Proc. SMAC, pp.287–292, Aug. 2013.
[17] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," IEEE Trans. SAP, vol.9, no.1, pp.21–29, 2001.
[18] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," IEEE J-STSP, vol.8, no.2, pp.184–194, 2014.
[19] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, "Voice conversion with smoothed GMM and MAP adaptation," Proc. INTESPEECH, 2003, pp.1–4, Sept. 2003.
[20] C.H. Lee and C.H. Wu, "Map-based adaptation for speech conversion using adaptation data selection and non-parallel training," Proc. INTERSPEECH, 2006, pp.17–21, Sept. 2006.
[21] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," IEEE Trans. ASLP, vol.18, no.5, pp.944–953, 2010.
[22] K. Kobayashi, T. Toda, T. Nakano, M. Goto, G. Neubig, S. Sakti, and S. Nakamura, "Gender-dependent spectrum differential models for perceived age control based on direct waveform modification in singing voice conversion," Proc. APSIPA ASC, pp.1–4, Dec. 2014.
[23] 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.
[24] 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, pp.2514–2418, Sept. 2014.
[25] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion based on direct waveform modification with global variance," Proc. INTERSPEECH, pp.2754–2758, Sept. 2015.
[26] 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.
[27] 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, pp.13–15, Sept. 2001.
[28] 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.

Kazuhiro Kobayashi graduated from the Department of Electrical and Electronic Engineering, Faculty of Engineering Science, Kansai University in Japan in 2012. He is currently in the Ph.D. course at the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST) in Japan. He recieved a few awards including best presentation award in the Acoustical Society of Japan (ASJ). He is a student member of IEEE, ISCA, and ASJ.

Tomoki Toda received his B.E. degree from Nagoya University, Japan, in 1999 and his M.E. and D.E. degrees from Nara Institute of Science and Technology (NAIST), Japan, in 2001 and 2003, respectively. He was a Research Fellow of the Japan Society for the Promotion of Science from 2003 to 2005. He was then an Assistant Professor (2005-2011) and an Associate Professor (2011-2015) at NAIST. From 2015, he has been a Professor in the Information Technology Center at Nagoya University. His research interests include statistical approaches to speech processing. He received more than 10 paper/achievement awards including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (Speech Communication Journal).

Tomoyasu Nakano received the Ph.D. degree in Informatics from University of Tsukuba, Tsukuba, Japan in 2008. He is currently working as a Senior Researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. His research interests include singing information processing, human-computer interaction, and music information retrieval. He has received several awards including the IPSJ Yamashita SIG Research Award from the Information Processing Society of Japan (IPSJ) and the Best Paper Award from the Sound and Music Computing Conference 2013. He is a member of the IPSJ and the Acoustical Society of Japan.

Masataka Goto received the Doctor of Engineering degree from Waseda University in 1998. He is currently a Prime Senior Researcher and the Leader of the Media Interaction Group at the National Institute of Advanced Industrial Science and Technology (AIST), Japan. Over the past 21 years, he has published more than 190 papers in refereed journals and international conferences and has received 38 awards, including several best paper awards, best presentation awards, and the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology (Young Scientists' Prize). He has served as a committee member of over 90 scientific societies and conferences, including the General Chair of the 10th and 15th International Society for Music Information Retrieval Conferences (ISMIR 2009 and 2014). In 2011, as the Research Director he began a 5-year research project (OngaCREST Project) on music technologies, a project funded by the Japan Science and Technology Agency (CREST, JST).

Satoshi Nakamura received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was a director of ATR Spoken Language Communication Research Laboratories in 2000-2008, and a vice president of ATR in 2007-2008. He was a director general of Keihanna Research Laboratories, National Institute of Information and Communications Technology, Japan in 2009-2010. He is currently a professor and a director of Augmented Human Communication laboratory, Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in modeling and systems of spoken dialog system, speech-to-speech translation. He is one of the leaders of speech-to-speech translation research projects including C-STAR, IWSLT and A-STAR. He headed the world first network-based commercial speech-to-speech translation service for 3-G mobile phones in 2007 and VoiceTra project for iPhone in 2010. He received LREC Antonio Zampoli Award, the Commendation for Science and Technology by the Ministry of Science and Technology in Japan. He is an elected board member of ISCA, International Speech Communication Association, and an elected member of IEEE SPS, speech and language TC.