PAPER Special Section on Advances in Modeling for Real-world Speech Information Processing and its Application

Voice Timbre Control Based on Perceived Age in Singing Voice Conversion

Kazuhiro KOBAYASHI[†a)], ***Student Member***, **Tomoki TODA**[†], **Hironori DOI**[†], ***Members***, **Tomoyasu NAKANO**[††], ***Nonmember***, **Masataka GOTO**[††], ***Member***, **Graham NEUBIG**[†], ***Nonmember***, **Sakriani SAKTI**[†], ***and*** **Satoshi NAKAMURA**[†], ***Members***

SUMMARY The perceived 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 an investigation of acoustic features that have an effect on the perceived age, and a novel voice timbre control technique based on the perceived 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 this limitation through the use of statistical voice conversion. This technique makes it possible to convert singing voice timbre 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 first perform an investigation of the factors that play a part in the listener's perception of the singer's age at first. Then, we applied a multiple-regression Gaussian mixture models (MR-GMM) to SVC for the purpose of controlling voice timbre based on the perceived age and we propose SVC based on the modified MR-GMM for manipulating the perceived age while maintaining singer's individuality. The experimental results show that 1) the perceived age of singing voices corresponds relatively well to the actual age of the singer, 2) prosodic features have a larger effect on the perceived age than spectral features, 3) the individuality of a singer is influenced more heavily by segmental features than prosodic features 4) the proposed voice timbre control method makes it possible to change the singer's perceived age while not having an adverse effect on the perceived individuality.

key words: *singing voice, voice conversion, perceived 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 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 singing voice characteristics, statistical voice conversion (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 singing VC (SVC) 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 SVC system, eigenvoice conversion (EVC) techniques [6] have been applied to SVC [7]. In an SVC 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 to arbitrary source and target singers by automatically estimating a few adaptation 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 adaptation 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 adaptation parameter.

In the area of statistical parametric speech synthesis [9], there have been several attempts at developing techniques for manually controlling voice characteristics of synthetic speech by manipulating intuitively controllable parameters corresponding to specific physical traits, such as gender and age. Nose et al. proposed a method for con-

Manuscript received September 28, 2013.
Manuscript revised January 18, 2014.

[†]The authors are with Nara Institute of Science and Technology (NAIST), Ikoma-shi, 630–0192 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.E97.D.1419

trolling speaking styles [10] and emotional expressions [11] in synthetic speech with multiple regression hidden Markov models (HMM). Tachibana et al. extended this method to control voice characteristics of synthetic speech using a voice characteristics control vector assigned to expressive word pairs describing voice characteristics, such as "warm – cold" and "smooth – non-smooth" [12]. A similar method has also been proposed in statistical VC [13] with multiple regression GMM (MR-GMM). Although these methods have only been applied to voice characteristics 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 perceived 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. The age has several good properties; e.g., the age is a measurement on the ratio scale unlike measurements on a nominal scale, such as gender; the age is more understandable than other expressive word pairs because it is observable; the age is widely distributed over people. The perceived age is also expected to have some of these good properties and to be conveniently used as a control factor to continuously and intuitively modify singing voice characteristics. There are several researches related to the age or the perceived age of normal speech. It has been reported that there is a correlation between the actual age and the perceived age [14]. As an investigation of an impact of aging on speech acoustics, it has been found that aperiodicity of excitation signals tends to increase with aging [15] and the perceived age of normal speech is varied by manipulating its F_0 variations, duration, and aperiodicity [16]. A method to classify speech of elderly people and non-elderly people using spectral and prosodic features has also been developed [17]. On the other hand, the perceived age of singing voices has not yet been studied deeply. Therefore, it is not obvious that these findings are also found in singing voices.

As fully understanding the acoustic features that contribute to the perceived age of singing voices is essential to the development of VC techniques to modify a singer's perceived age, in this paper we first perform an investigation of the acoustic features that play a part in the listener's perception of the singer's age at first. We conduct several types of perceptual evaluation to investigate 1) how well the perceived age of singing voices corresponds to the actual age of the singer, 2) whether or not singing VC processing causes adverse effects on the perceived age of singing voices, 3) which spectral or prosodic features have a larger effect on the perceived age, and 4) which spectral or prosodic features have a individuality of a singer. Then, we propose a novel voice timbre conversion method that converts the singer's perceived age while maintaining individuality in SVC.

2. Statistical Singing Voice Conversion (SVC)

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 [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. $\boldsymbol{\lambda}$ 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 [18]. 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) [3] is also considered in conversion.

3. Investigation of Acoustic Features Affecting Perceived Age

In the traditional SVC [5], [7], only the spectral features such as mel-cepstrum are converted. It is also straightforward to convert the aperiodic components (ACs) [19], which capture noise strength on each frequency band of the excitation signal, as in the traditional VC for natural voices [20]. If the perceived age of singing voices is captured well by these acoustic features, it will make it possible to develop a

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

Features	Analysis/synthesis (w/ ACs)	Analysis/synthesis (w/o ACs)	Intra-singer SVC (source)	Intra-singer SVC (target)	SVC
Power, F_0, duration	Source singer	Source singer	Source singer	Target singer	Source singer
Mel-cepstrum	Source singer	Source singer	Converted to source singer	Converted to target singer	Converted to target singer
Aperiodic components	Source singer	Removed	Converted to source singer	Converted to target singer	Converted to target singer

real-time SVC system capable of controlling the perceived age of singing voices by combining SVC with MR-GMM (described in Sect. 5.1) and real-time statistical VC techniques [21], [22]. On the other hand, if the perceived age of singing voices is not captured at all 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 characteristics control framework of HMM-based speech synthesis [10], [12] can be used in the SVC system to control the perceived age of singing voices, although it is not straightforward to develop a real-time SVC system in this framework. In this section, we compare the perceived age of natural singing voices with that of several types of synthesized singing voices by modifying acoustic features as shown in Table 1 for the purpose of investigating acoustic features affecting the perceived age in singing voices to clarify which types of techniques can be implemented for the SVC system.

3.1 Effects of Analysis/Synthesis

In the analysis/synthesis framework, a voice is first converted into parameters of a source-filter model, then simply re-synthesized into a waveform using these parameters without change. We define this re-synthesized singing voice as analysis/synthesis (w/ ACs). 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 perceived age of singing voices. We use STRAIGHT [23] as a widely used high-quality analysis/synthesis method to extract acoustic features consisting of spectral envelope, F_0, and ACs. The spectral envelope is further parameterized with mel-cepstrum.

3.2 Effects of Aperiodic Components

As mentioned above, previous research [15] has shown that ACs tend to change with aging in normal speech. We investigate the effects of ACs on the perceived 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 [23] instead of STRAIGHT mixed excitation [20] is used to generate voiced excitation signals. We define this re-synthesized singing voice as analysis/synthesis (w/o ACs).

3.3 Effects of Conversion Errors

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 acoustic 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 perceived 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 (source/target) in this paper.

To achieve intra-singer SVC (source/target) 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. It is impossible to train such a GMM by simply using the source feature vector of the source singer $\boldsymbol{X}_t$ as the target feature vector $\boldsymbol{Y}_t$ because this duplication causes the rank deficiency of the covariance matrix. Namely, the following conditions need to hold; $\boldsymbol{X}_t$ is different from $\boldsymbol{X}'_t$; they depend on each other; and both are identically distributed. This GMM can be trained using a parallel data set consisting of the song pairs sung by the source singer but the source singer needs to sing the same songs twice to develop such a parallel data set. As a more convenient way to develop the GMM for intra-singer SVC (source/target), we use the framework of many-to-many EVC. The 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 as follows:

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

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

Using this GMM, intra-singer SVC (source/target) is performed in the same manner as described in Sect. 2. The converted singing voice sample essentially has the same singing voice characteristics as those before the conversion although they suffer from conversion errors. We define this converted singing voice as intra-singer SVC (source/target).

3.4 Effects of Prosodic and Segmental Features

To investigate which acoustic features have a larger effect on the perceived age of singing voices, segmental features or prosodic features, we use the SVC for converting only segmental features, such as mel-cepstrum and ACs, 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 for Investigation of Acoustic Features

4.1 Experimental Conditions

In our experiments, we first investigated the correspondence between the perceived age and the actual age of the singer. We used the AIST humming database [24] consisting of singing voices of 25 songs with Japanese lyrics sung by Japanese male and female amateur singers in their 20s, 30s, 40s, and 50s. The total number of singers in the database was 75. The length of each song was approximately 20 seconds. For evaluation, only one Japanese song (No. 39) was used. Eight Japanese male subjects in their 20s were asked to guess the age of each singer by listening to his/her singing voices.

In the second experiment, we investigated the acoustic features that affect the perceived age of singing voices. We did so by comparing the perceived age of natural singing voices with that of each type of synthesized singing voice as shown in Table 1. Eight Japanese male subjects in his 20s assigned the perceived age to each synthesized singing voice. 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 as evaluation singers. The singers were also separated into two groups, A and B, so that one group always included 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.

In the third experiment, we investigated which acoustic features more affected the singer's individuality of singing voices. We divided the 16 evaluation singers into four groups, M1, M2, F1 and F2, so that each group included four male or female singers from all age groups. The subjects were also randomly separated into four groups. Converted singing voices with SVC were created in every combination of source and target singer pairs in each group (i.e., 12 combinations) as evaluation samples. Converted singing voices with intra-singer SVC (source/target) were also created for individual singers (four male or female singers) in each group as reference samples. The subjects were asked to separate the evaluation samples into four classes corresponding to the reference samples on the basis of similarity of singer's individuality. The subjects were allowed to listen to the evaluation and reference samples as many times as they wanted. We gave instructions to the subjects to evaluate the singer's individuality considering a possibility of changes of singing voice characteristics caused by aging.

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 ACs 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 (source/target) and SVC, we used 18 songs including the evaluation song (No. 39). In the intra-singer SVC (source/target), GMMs for converting the mel-cepstrum and ACs 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 ACs 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 Comparison between Perceived Age and Actual Age

Figure 1 indicates the correlation between the perceived age of natural singing voices and the actual age of the singer. Each point indicates the perceived age of each singer averaged over all subjects. The standard deviation of the perceived age in each singer over all subjects is 6.17. The correlation coefficient between the perceived age and the actual age in this figure is 0.81. These results show quite high correlation between the perceived age and the actual age.

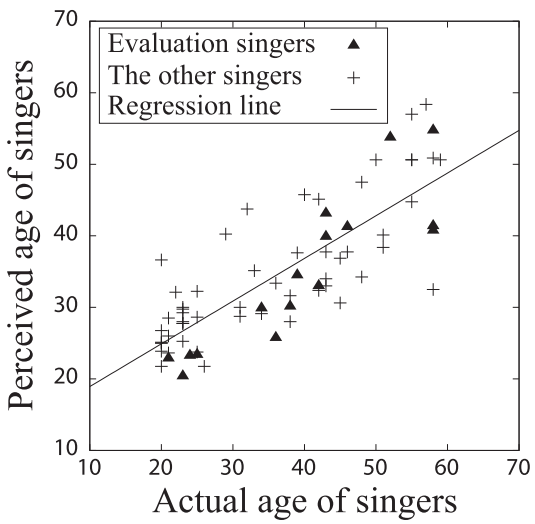

Fig. 1 Correlation between singer's actual age and perceived age.

4.3 Acoustic Features Affecting Perceived Age

Table 2 indicates average values and standard deviations of differences between the perceived age of natural singing voices and each type of intra-singer synthesized singing voice: analysis/synthesis (w/ ACs), analysis/synthesis (w/o ACs) and the intra-singer SVC (source/target). The table also indicates correlation coefficients between the perceived age of natural and synthesized voices. From the results, we can see that in analysis/synthesis (w/ ACs), the perceived age difference is small and the correlation coefficient is very high. Therefore, distortion caused by analysis/synthesis processing does not affect the perceived age. It can be observed from analysis/synthesis (w/o ACs) that this result does not change even if not using ACs. Therefore, ACs do not affect the perceived age of singing voices. On the other hand, intra-singer SVC (source/target) causes slightly larger differences between natural singing voices and the synthesized singing voices. Therefore, some acoustic cues to the perceived age are removed through the statistical conversion processing. Nevertheless, the perceived age differences are relatively small, and therefore, it is likely that important acoustic cues to the perceived age are still kept in the converted acoustic features.

Figures 2 and 3 indicate a comparison between the perceived age of singing voices generated by SVC and intra-singer SVC (source/target). In each figure, the vertical axis indicates the perceived age of converted singing voices by SVC (prosodic features: source singer, segmental features: target singer). The horizontal axis in Fig. 2 indicates the perceived age of singing voices generated by intra-singer SVC (source) and that in Fig. 3 indicates the perceived age of singing voices generated by intra-singer SVC (target). Therefore, if the prosodic features more strongly affect the perceived age than the segmental features, a higher correlation will be observed in Fig. 2. If the segmental features more strongly affect the perceived 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 perceived 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 perceived age than the segmental features.

4.4 Acoustic Features Affecting Singer Individuality

In this experiment, we investigated which prosodic and segmental features have a larger impact on singer's individuality. Table 3 indicates the ratios judged by subjects based on similarity of between the converted singing voice from the source singer into the target singer with SVC and the

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

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

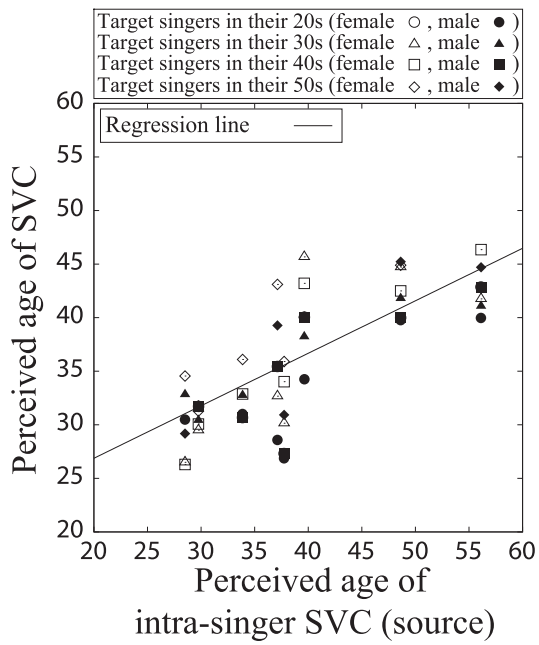

Fig. 2 Correlation of perceived age between singing voices generated by the intra-singer SVC (source) and the SVC.

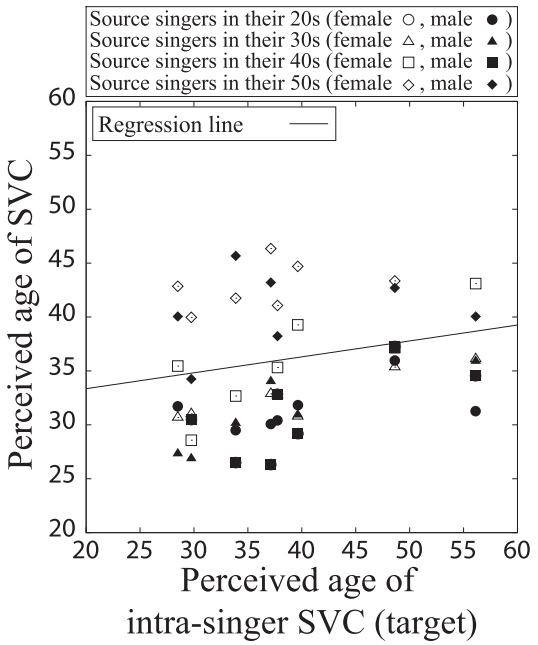

Fig. 3 Correlation of perceived age between singing voices generated by the intra-singer SVC (target) and the SVC.

Table 3 Evaluation of singer identification in SVC.

Acoustic features	Ratio
Prosodic features	52.08
Segmental features	35.42
Disagreement	12.50

source, target, or other singers' reference singing voices that were generated by intra-singer SVC (source/target). If the prosodic features more strongly have the individuality of singer than segmental features, then singing voice converted with SVC is classified into intra-singer SVC (source). On the other hand, if the segmental features more strongly have the individuality of singer than prosodic features, then the singing voice converted with SVC is classified into intra-singer SVC (target). If the singing voice converted with SVC is classified to the other singers' reference singing voices, it was counted as a disagreement sample. This table demonstrates that individuality of a singer is distinguished from prosodic features rather than segmental features. This result has a similar tendency on Figs. 2 and 3. Namely, there is a correlation between singer's individuality and perceived age. These results suggest that if it is necessary to make large changes in the perceived age, then prosodic features are the most suitable acoustic features. However, it will also cause changes of singer's individuality. In contrast, if it is required to change only the perceived age while remaining singer's individuality, segmental features are more appropriate features although a range of changes of the perceived age is limited.

5. Voice Timbre Control Based on Perceived age

In the last evaluation, we indicated that segmental features are suitable to control the perceived age to retain singer individuality. In this section, we develop a perceived age controllable SVC technique for a specific singer. MR-GMM is applied to SVC to convert segmental features by manipulating the perceived age. Moreover, we propose a modified MR-GMM to maintain the singer's individuality. In this section, we apply MR-GMM to SVC. Then, we modify MR-GMM to maintain the singer's individuality.

5.1 SVC with Multiple Regression GMM (MR-GMM)

SVC with MR-GMM also consists of a training process and a conversion process. The MR-GMM is trained using multiple parallel data sets consisting of the source singer's singing voices and many pre-stored target singers' singing voices. The joint probability density of $2D$-dimensional joint static and dynamic feature vectors modeled by the MR-GMM is given by

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{\lambda}^{(MR)}, w^{(s)}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t^{(s)} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}(s) \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}\right), \tag{5}$$

where $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ and $\boldsymbol{Y}_t^{(s)} = [\boldsymbol{Y}_t^{(s)\top}, \Delta\boldsymbol{Y}_t^{(s)\top}]^\top$ show static and delta feature vectors of the source and s-th pre-stored target singer. The mean vector of the s-th pre-stored singer is given by

$$\boldsymbol{\mu}_m^{(Y)}(s) = \boldsymbol{b}_m^{(Y)} w^{(s)} + \overline{\boldsymbol{\mu}}_m^{(Y)}, \tag{6}$$

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 perceived age score, which is manually assigned for each pre-stored target singer.

In the conversion process, the perceived age score is manually set to a desired value. Then, the converted feature vector is determined in the same manner as described in Sect. 2.

5.2 MR-GMM Implementation Based on Many-to-Many SVC

In this paper, to make it easier to develop the MR-GMMs for individual source singers (i.e., users), we apply the framework of many-to-many SVC [7] to SVC based on MR-GMM. The joint probability density of many-to-many MR-GMM follows:

$$\begin{aligned} &P\left(\boldsymbol{Y}_t^{(i)}, \boldsymbol{Y}_t^{(o)} | \boldsymbol{\lambda}^{(MR)}, w^{(i)}, w^{(o)}\right) \\ &= \sum_{m=1}^{M} P\left(m | \boldsymbol{\lambda}^{(MR)}\right) \int P\left(\boldsymbol{Y}_t^{(i)} | \boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(MR)}, w^{(i)}\right) \\ &\qquad P\left(\boldsymbol{Y}_t^{(o)} | \boldsymbol{X}_t, m, \boldsymbol{\lambda}^{(MR)}, w^{(o)}\right) P\left(\boldsymbol{X}_t | m, \boldsymbol{\lambda}^{(MR)}\right) \mathrm{d}\boldsymbol{X}_t \\ &= \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix} \boldsymbol{Y}_t^{(i)} \\ \boldsymbol{Y}_t^{(o)} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_m^{(Y)}(i) \\ \boldsymbol{\mu}_m^{(Y)}(o) \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(YY)} & \boldsymbol{\Sigma}_m^{(YXY)} \\ \boldsymbol{\Sigma}_m^{(YXY)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}\right), \end{aligned} \tag{7}$$

$$\boldsymbol{\Sigma}_m^{(YXY)} = \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(XX)-1} \boldsymbol{\Sigma}_m^{(XY)}, \tag{8}$$

where $w^{(i)}$ and $w^{(o)}$ indicate the perceived age score of the source singer and that of the target singers, respectively. Source and target mean vectors are given by Eq. (6).

It is possible to use Eq. (6) to describe the input mean vectors $\boldsymbol{\mu}_m^{(Y)}(i)$ based on the perceived age score of the input singer. However, accuracy of acoustic modeling by the MR-GMM tends to decrease since the acoustic characteristics of the input singer are not always modeled well on a subspace spanned by the basis vector. Namely, we suppose that it is possible to prepare a parallel data set of each user and the reference singer. This condition is still practical in the development of the user-dependent SVC system. Using a parallel data of the input singer's singing voice and the reference singer's singing voice, the input mean vector of the MR-GMM is updated in the sense of a maximum likelihood criterion. Consequently, the input mean vector is given by

$$\boldsymbol{\mu}_m^{(Y)}(i) = \hat{\boldsymbol{\mu}}_m^{(Y)}, \tag{9}$$

where $\hat{\boldsymbol{\mu}}_m^{(Y)}$ is its maximum likelihood estimate. Note that it is also possible to train all parameters of the MR-GMM using the parallel data sets of the user and all pre-stored

target singers without using the many-to-many SVC framework. However, the training method presented here is still useful to effectively reduce computational cost to develop the MR-GMM because it is necessary to update only input mean vectors as shown in Eq. (9). Moreover, there is a possibility to reduce the amount of singing voice data of the user used for training or implement an unsupervised training approach without the parallel data set based on model adaptation techniques.

5.3 Modified MR-GMM to Retain Singer Individuality

In SVC with many-to-many MR-GMM, it is possible to convert voice timbre of the input singer into desired voice timbre corresponding to an output perceived age score. However, the output mean vector given by Eq. (6) only expresses average voice characteristics of several pre-stored target singers. Therefore, a converted singing voice doesn't express voice timbre of the input singer.

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

$$\begin{aligned}\boldsymbol{\mu}_m^{(Y)}(o) &= \boldsymbol{b}_m^{(Y)} w^{(o)} + \overline{\boldsymbol{\mu}}_m^{(Y)} \\ &= \boldsymbol{b}_m^{(Y)} (w^{(i)} + \Delta w) + \overline{\boldsymbol{\mu}}_m^{(Y)} \\ &= \boldsymbol{b}_m^{(Y)} w^{(i)} + \overline{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{b}_m^{(Y)} \Delta w \\ &\simeq \hat{\boldsymbol{\mu}}_m^{(Y)} + \boldsymbol{b}_m^{(Y)} \Delta w \end{aligned} \tag{10}$$

where the perceived age score of the output singing voice $w^{(o)}$ is represented by that of the input singing voice $w^{(i)}$ and a difference perceived age score Δw between them. In the modified representative form, the output mean vector is represented by the input mean vector $\hat{\boldsymbol{\mu}}_m^{(Y)}$ and the additional vector corresponding to a difference perceived age score Δw. As the input 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 input singer's singing voice are well preserved in this modified representative form.

6. Experimental Evaluation of Perceived Age Control

6.1 Experimental Conditions

In the first experiment, we evaluated the variation of perceived age achieved by the modified MR-GMM. 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 perceived age score in Eq. (10) 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 the conventional and modified MR-GMMs. Subjects and evaluation singers were separated into two groups in the same manner as the first experiment. We changed the perceived age score in Eq. (10) into −60, −30, 30 and 60 in the modified MR-GMM. In the conventional MR-GMM, the perceived age score in Eq. (6) was varied ±30, 60 from the perceived age of each evaluation singer, which was determined by listening to samples of the intra-singer SVC (source/target) in the previous experiment. A pair of songs generated by the modified and conventional MR-GMM of the same singer and variation of the perceived age scores was presented to subjects after presenting the intra-singer SVC (source) 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). Subjects and evaluation singers were the same as in the first experiment. The perceived age score was the same as for the second evaluation. Subjects rated the naturalness of the converted singing voices using a 5-point scale: "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad.

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 singing voices was 25 in each singer. We used parallel data sets of the reference singer and 16 evaluation singers to update the input mean vectors by Eq. (9) for each evaluation singer. The perceived age score for each singer was determined as an average score over 25 singing voices of the singer rated by one male subject in his 20s. The number of mixture components of the MR-GMM was 128 for spectral envelope and 32 for ACs. The other experimental conditions were the same as Sect. 4.1.

6.2 Experimental Results

Figure 4 indicates the varieties of perceived age in the modified MR-GMM. To change the perceived age score from −60 to 60, the perceived age of the singer was almost linearly varied. Especially, we can see the same tendency as observed in the investigation of segmental features shown in Fig. 3. The result in Fig. 3 indicates that the change of observed perceived age from 20 to 60 years old in the hori-

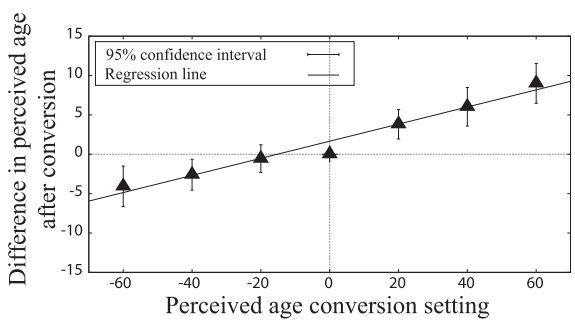

Fig. 4 Setting and actual differential in perceived age after conversion.

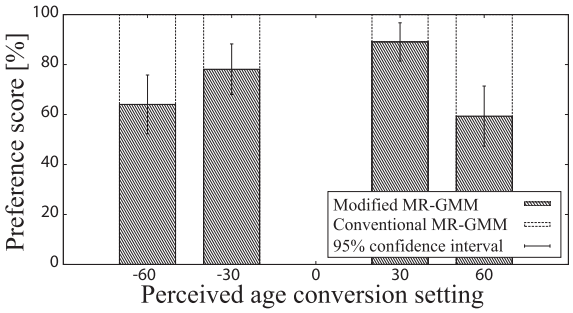

Fig. 5 Comparing singer individuality of conventional MR-GMM and Proposed MR-GMM converted singing voice.

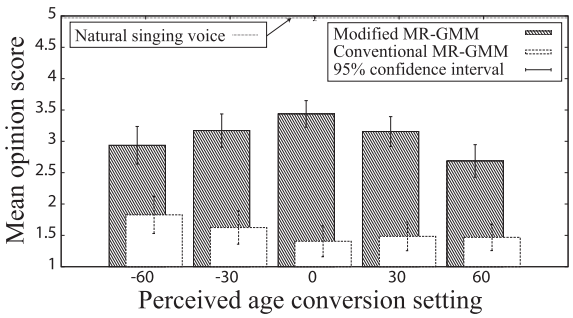

Fig. 6 Mean opinion score of conventional MR-GMM and modified MR-GMM.

zontal line is about 5 years. This means that modified MR-GMM can appropriately control the perceived age of singing voices.

Figure 5 indicates the result of the XAB test for the singer individuality. We can see that as we make larger changes in the perceived age, the preference score of the modified MR-GMM tends to decrease. However the modified MR-GMM has a higher preference score than the conventional MR-GMM for each setting.

Figure 6 indicates the results of MOS test for the naturalness. This figure has the same tendency as displayed in Fig. 5. The modified MR-GMM has a higher MOS than the conventional MR-GMM for each setting. The bias vectors of the modified MR-GMM ($\hat{\boldsymbol{\mu}}_m^{(Y)}$ in Eq. (10)) model singing voice characteristics of a single singer (i.e., the source singer). On the other hand, those of the conventional MR-GMM ($\overline{\boldsymbol{\mu}}_m^{(Y)}$ in Eq. (10)) model voice characteristics of multiple pre-stored target singers. Therefore, over-smoothing effects of the conventional MR-GMM tend to be larger than those of the modified MR-GMM. Consequently, the naturalness of the singing voices is also improved by using the modified MR-GMM.

These results suggest that 1) the modified MR-GMM enables to control the perceived age of singing voices relatively well, 2) the modified MR-GMM enables to retain the singer individuality better than the conventional MR-GMM during the perceived age control, and 3) the modified MR-GMM also generates better quality of converted singing voices compared with the conventional MR-GMM.

7. Conclusions

In order to develop voice timbre control based on the perceived age, we have investigated acoustic features that affect the perceived age. To factorize the effect of several acoustic features on the perceived age of singing voices, several types of synthetic singing voices were constructed and evaluated. We have also proposed a method for controlling the perceived age that maintains the singer's individuality. A conventional voice timbre control technique based on multiple-regression Gaussian mixture model (MR-GMM) was not able to control the singer's perceived age while maintaining the singer's individuality. To address this problem, we have proposed the modified MR-GMM. The experimental results have demonstrated that 1) Analysis/synthesis process, aperiodic components, and voice conversion errors have only small effects on the perceived age of singing voices. 2) the prosodic features more strongly affect the perceived age than the segmental features, 3) the conversion of the prosodic features also causes a larger change of the singer individuality compared with the conversion of the segmental features, and 4) the proposed method makes it possible to retain singer's individuality better than the conventional MR-GMM during the perceived age control. In future work, we plan to investigate the effect of dealing with variations of the perceived age caused by different listeners. Moreover, it is worthwhile to investigate a method to factorize the prosodic features into components related to the perceived age and singer's individuality to make it possible to more widely manipulate the perceived age while retaining singer's individuality.

Acknowledgments

This work was supported in part by JSPS KAKENHI Grant Number 22680016 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," in Entertainment Computing–ICEC 2009, pp.185–190, Springer, 2009.
[2] 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.
[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, April 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 Commun., 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," IEICE Trans. Inf. & Syst., vol.E90-D, no.9, pp.1406–1413, Sep. 2007.
[11] T. Nose and T. Kobayashi, "An intuitive style control technique in HMM-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model," Speech Commun., vol.55, no.2, pp.347–357, 2013.
[12] 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.
[13] 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.
[14] W. Ryan and K. Burk, "Perceptual and acoustic correlates of aging in the speech of males," J. Communication Disorders, vol.7, no.2, pp.181–192, 1974.
[15] H. Kasuya, H. Yoshida, S. Ebihara, and H. Mori, "Longitudinal changes of selected voice source parameters," Proc. INTERSPEECH, pp.2570–2573, Sept. 2010.
[16] J.D. Harnsberger, W.S. Brown Jr., R. Shrivastav, and H. Rothman, "Noise and tremor in the perception of vocal aging in males," J. Voice, vol.24, no.5, pp.523–530, 2010.
[17] 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.
[18] 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.
[19] 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.
[20] 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.
[21] 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.
[22] T. Toda, T. Muramatsu, and H. Banno, "Implementation of computationally efficient real-time voice conversion," Proc. INTERSPEECH, Sept. 2012.
[23] 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 Commun., vol.27, no.3-4, pp.187–207, April 1999.
[24] 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.

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 master's course at the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST) in Japan. He is a student member of ISCA, and ASJ.

Tomoki Toda was born in Aichi, Japan on January 18, 1977. He earned his B.E. degree from Nagoya University, Aichi, Japan, in 1999 and his M.E. and D.E. degrees from the Graduate School of Information Science, NAIST, Nara, Japan, in 2001 and 2003, respectively. He was a Research Fellow of JSPS in the Graduate School of Engineering, Nagoya Institute of Technology, Aichi, Japan, from 2003 to 2005. He was an Assistant Professor of the Graduate School of Information Science, NAIST from 2005 to 2011, where he is currently an Associate Professor. He has also been a Visting Researcher at the NICT, Kyoto, Japan, since May 2006. From March 2001 to March 2003, he was an Intern Researcher at the ATR Spoken Language Communication Research Laboratories, Kyoto, Japan, and then he was a Visiting Researcher at the ATR until March 2006. He was also a Visiting Researcher at the Language Technologies Institute, CMU, Pittsburgh, USA, from October 2003 to September 2004 and at the Department of Engineering, University of Cambridge, Cambridge, UK, from March to August 2008. His research interests include statistical approaches to speech processing such as voice transformation, speech synthesis, speech analysis, speech production, and speech recognition. He received the 18th TELECOM System Technology Award for Students and the 23rd TELECOM System Technology Award from the TAF, the 2007 ISS Best Paper Award and the 2010 ISS Young Researcher's Award in Speech Field from the IEICE, the 10th Ericsson Young Scientist Award from Nippon Ericsson K.K., the 4th Itakura Prize Innovative Young Researcher Award and the 26th Awaya Prize Young Researcher Award from the ASJ, the 2009 Young Author Best Paper Award from the IEEE SPS, the Best Paper Award (Short Paper in Regular Session Category) from APSIPA ASC 2012, the 2012 Kiyasu Special Industrial Achievement Award from the IPSJ, and the 2013 Best Paper Award (Speech Communication Journal) from EURASIP-ISCA. He was a member of the Speech and Language Technical Committee of the IEEE SPS from 2007 to 2009. He is a member of IEEE, ISCA, IPSJ, and ASJ.

Hironori Doi graduated from the Department of Information and Image Science, Faculty of Engineering, Chiba University, Japan in 2008. He received his M.E. and D.E. dgrees from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan, in 2009 and 2012, respectively. He was a Research Fellow of JSPS in NAIST from 2011 to 2012. He currently works for DWANGO Co., Ltd., Japan. He mainly studies singing voice conversion and speaking-aid systems.

Tomoyasu Nakano received the Ph.D. degree in informatics from University of Tsukuba, Tsukuba, Japan in 2008. He is currently 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). He is a member of the IPSJ and the Acoustical Society of Japan (ASJ).

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).

Graham Neubig received his B.E. from University of Illinois, Urbana-Champaign, U.S.A, in 2005, and his M.E. and Ph.D. in informatics from Kyoto University, Kyoto, Japan in 2010 and 2012 respectively. He is currently an assistant professor at the Nara Institute of Science and Technology, Nara, Japan. His research interests include speech and natural language processing, with a focus on machine learning approaches for applications such as machine translation, speech recognition, and spoken dialog.

Sakriani Sakti received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received "DAAD-Siemens Program Asia 21st Century" Award to study in Communication Technology, University of Ulm, Germany, and received her M.Sc. degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003–2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006–2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005–2008) with Dialog Systems Group University of Ulm, Germany, and received her Ph.D. degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003–2007), A-STAR and U-STAR (2006–2011). She also served as a visiting professor of Computer Science Department, University of Indonesia (UI) in 2009–2011. Currently, she is an assistant professor of the Augmented Human Communication Lab, NAIST, Japan. She is a member of JNS, SFN, ASJ, ISCA, IEICE and IEEE. Her research interests include statistical pattern recognition, speech recognition, spoken language translation, cognitive communication, and graphical modeling framework.

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.