An Investigation of How to Design Control Parameters for Statistical Voice Timbre Control

Kazutaka Kubo* and Kazuhiro Kobayashi† and Tomoki Toda† and
Graham Neubig‡ and Sakriani Sakti* and Satoshi Nakamura*
* Nara Institute of Science and Technology (NAIST), Japan
† Information Technology Center, Nagoya University, Japan
E-mail: kobayashi.kazuhiro@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp
‡ Language Technologies Institute, Carnegie Mellon University, USA

***Abstract*—Multiple-regression Gaussian mixture models (MR-GMM) allow for control of voice timbre along several axes each described by a voice timbre expression word. To create these axes, perceptual scores corresponding to multiple voice timbre expression words are manually assigned to individual pre-stored target speakers as the voice timbre control parameters, and then acoustic basis vectors corresponding to the individual control parameters are learned. The voice timbre expression words are usually selected from various words using factor analysis so that the voice timbre control parameters are independent of each other. However, the resulting basis vectors are not often orthogonal to each other, and they practically cause difficulties in intuitively controlling the converted voice timbre. Towards the development of the MR-GMM capable of intuitively controlling converted voice timbre, we investigate how to design the voice timbre control parameters so that not only the voice timbre control parameters but also the corresponding acoustic basis vectors are independent of each other. Experimental results demonstrate that 1) a method for annotation of the voice timbre control parameters using the converted voices rather than natural voices is effective, and 2) the independences of the voice timbre control parameters and acoustic basis vectors is helpful for improving the converted voice timbre controllability of the MR-GMM.**

I. Introduction

Varieties of voice characteristics, such as voice timbre and fundamental frequency (F_0) patterns, produced by individual speakers are always restricted by their own physical constraints imposed by the speech production mechanism. Voice conversion (VC) is a potential technique for us to produce speech sounds beyond our own physical constraints [1].

As one of the most popular statistical VC methods that converts voice timbre of a source speaker into that of a target speaker, a conversion method using a Gaussian mixture models (GMM) have been proposed [2], [3]. This technique converts acoustic features of the source speaker into those of the target speaker based on a previously trained GMM using parallel utterances of the source and target speakers. To make it possible to convert an arbitrary source speaker into an arbitrary target speaker, many-to-many conversion based on eigenvoice GMM (EV-GMM) has been proposed [4]. The EV-GMM is trained in advance using multiple parallel data sets including utterance pairs of a single reference speaker and many other pre-stored target speakers. This technique models various source/target speakers' acoustic features using orthogonal basis vectors (i.e., eigenvectors). Therefore, it is straightforward to manually change converted voice timbre by manipulating interpolation weights for the basis vectors. However, it is essentially difficult to intuitively control them because a voice timbre component modeled with each basis vector does not have any specific meaning.

In order to manually control voice timbre of the source speaker based on intuitively understandable parameters, statistical VC with multiple-regression GMM (MR-GMM) [5] has been proposed. In the training of the MR-GMM, the model learns basis vectors corresponding to individual voice timbre control parameters, which are perceived scores on specific voice timbre expression word pairs, using the multiple parallel data sets along with the voice timbre control parameters manually assigned to each pre-stored target speaker. The MR-GMM makes it possible to intuitively control converted voice timbre thanks to the voice timbre control parameters having specific meanings represented by the corresponding voice timbre expression word pairs. However, it is still difficult to achieve sufficiently high controllability because the resulting basis vectors are not usually orthogonal to each other and the use of them causes difficulties in simultaneously manipulating the corresponding voice timbre control parameters.

In this paper, towards improvements of controllability of converted voice timbre using the MR-GMM with multiple voice timbre control parameters, we carefully investigate how to design these parameters. We assume that it is effective to design them so that 1) the voice timbre control parameters are independent of each other in a perceptual space, and 2) the corresponding basis vectors are orthogonal to each other in an acoustic space. The experimental results demonstrate that 1) an annotation method of the voice timbre control parameters using the converted voices rather than natural voices is effective, and 2) the voice timbre controllability of the MR-GMM is significantly improved by making the voice timbre control parameters independent of each other and also basis vectors orthogonal to each other.

II. Voice timbre control based on Modified MR-GMM

In this paper, we utilize a modified representation of the MR-GMM (Modified MR-GMM) [6] as an enhanced version of the traditional MR-GMM [5]. In the training process of the Modified MR-GMM, a canonical MR-GMM consisting of basis vectors corresponding to individual voice timbre

control parameters defined by the perceived scores is previously trained using multiple parallel data sets consisting of a reference speaker's voices and many pre-stored target speakers' voices in the same manner as the traditional MR-GMM. Then, a part of the canonical MR-GMM parameters is modified so that voice timbre of a specific target speaker is well converted while retaining speaker individuality.

In the conversion process, the voice timbre control parameter differentials are manually set to desired values. Then, the voice timbre of the source speaker is converted into the desired one according to the given voice timbre control parameter differentials. The maximum likelihood estimation of speech parameter trajectories [3] is used in conversion.

III. Design of voice timbre control parameters

Several voice timbre expression words were carefully selected from various candidate words using factor analysis through a large-sized perceptual test [7]. In our previous work [5], these selected words were used to assign the voice timbre control parameters to the individual pre-stored target speakers through listening to their natural voices. The assigned voice timbre control parameters are usually less correlated to each other. On the other hand, the resulting basis vectors of the MR-GMM trained with these parameters are not usually orthogonal to each other, causing difficulties in intuitively controlling the converted voice timbre in practice. We carefully investigate how to address this issue, focusing on how to design the voice timbre control parameters.

A. Investigation of how to define voice timbre control parameters

Previous work on the voice timbre control of singing voices revealed that a perceived age can be effectively used as a voice timbre control parameter [6]. Our informal experimental results also revealed that the same tendency was observed in the voice timbre control of normal speech. In this paper, to simplify the problem, we use the perceived age as the first voice timbre control parameter and investigate only how to define the second voice timbre control parameter.

We hypothesize that the development of the orthogonal basis vectors corresponding to uncorrelated voice timbre control parameters is helpful for improving controllability of the converted voice timbre. However, in spite of the use of the uncorrelated voice timbre control parameters, the second basis vectors often fail to be orthogonal with the first basis vectors as mentioned above. One possible reason is that acoustic space modeled by the MR-GMM is differ from natural voice because the MR-GMM can model only voice characteristics on a subspace spanned by the basis vectors. These differences are expected to strongly affect the voice timbre control parameters as the acoustic cues not modeled by the MR-GMM are also evaluated in their assignment.

We investigate the effectiveness of using converted voices rather than natural voices in the assignment of the voice timbre control parameters. After the development of the first basis vector for the perceived age, we generate the converted voices of the pre-stored target speakers with the Modified MR-GMM by setting the target perceived age to be constant over all speakers. In other words, the perceived ages of the individual pre-stored target speakers are normalized into the constant value. Because the acoustic variations modeled on the subspace spanned by the basis vector are removed from these converted voices, the use of them for assigning another voice timbre control parameter to the individual pre-stored target speakers is expected to be effective for making the second basis vectors orthogonal to the first one.

B. Voice timbre normalization

We perform perceived age normalization of all pre-stored target speakers to generate the converted voices to be used for the assignment of the second voice timbre control parameter. To also remove the effects of prosodic features not handled in the MR-GMM-based voice timbre control on the converted voices, voices of a single reference speaker are used as the source voices to be converted. In the perceived age normalization process, the perceived age score differential needs to be determined for each pre-stored target speaker.

1) Normalization based on perceptual score differential: The perceived age normalization is performed in the perceived age space previously developed by evaluating natural voices, which is used for the development of the first basis vector. The perceived age score differential $\Delta w(s)$ of the s-the pre-stored target speaker is determined as follows:

$$\Delta w(s) = w^{(T)} - w^{(O)}(s), \tag{1}$$

where $w^{(O)}(s)$ and $w^{(T)}$ are the perceived age score of the s-th pre-stored target speaker and the normalization target age, respectively.

2) Normalization based on acoustic feature distance: Perceived age normalization is performed in the acoustic space by minimizing the Mahalanobis distance between the target mean vectors and the normalized mean vectors on the subspace spanned by the first basis vector. The perceived age differential $\Delta w(s)$ is determined as follows:

$$\Delta w(s) = \underset{\Delta w(s)}{\operatorname{argmin}} (\boldsymbol{\mu}_m^{(T)} - \boldsymbol{\mu}_m^{(Y)}(s) - \boldsymbol{b}_m^{(Y)} \Delta w(s))^\top {\boldsymbol{\Sigma}_m^{(YY)}}^{-1} (\boldsymbol{\mu}_m^{(T)} - \boldsymbol{\mu}_m^{(Y)}(s) - \boldsymbol{b}_m^{(Y)} \Delta w(s)), \tag{2}$$

$$\boldsymbol{\mu}_m^{(T)} = \boldsymbol{b}_m^{(Y)} w^{(T)} + \overline{\boldsymbol{\mu}}_m^{(Y)}, \tag{3}$$

where $\boldsymbol{b}_m^{(T)}$ is the first basis vector corresponding to the perceived age and $\boldsymbol{\mu}_m^{(Y)}(s)$ is the mean vector of the speaker-dependent model of the s-th pre-stored target speaker. $\overline{\boldsymbol{\mu}}_m^{(Y)}$ is the bias vector. $\boldsymbol{\Sigma}_m^{(YY)}$ is the covariance matrix of the MR-GMM. m is the mixture component index.

3) Refinement considering perceptual scores assigned to normalized voice: We also investigate the effectiveness of refining the perceived age differential considering perceived ages newly annotated using the perceived age normalized voices. After generating the perceived age normalized voices using the normalization method described in III-B2, we additionally assign perceived ages to the pre-stored target speakers by listening to the normalized voices. Although it is ideal that these ages are identical to the normalization target age over all pre-stored target speakers, they are actually different from it due to insufficient performance of the MR-GMM-based voice timbre control. To improve the perceived age normalization

performance, we refine the perceived age differential as follows:

$$\Delta w' = \Delta w \frac{w^{(T)} - w^{(O)}(s)}{\hat{w}(s) - w^{(O)}(s)}, \quad (4)$$

where $\Delta w'$ is the refined perceived age differential and the $\hat{w}(s)$ is the newly assigned perceived age using the perceived age normalized voices.

IV. Experimental evaluations

In this evaluation, we performed the following evaluations: 1) comparison of the accuracy of the perceived age normalization techniques, 2) simultaneous voice timbre control using two voice timbre expression words.

A. Experimental conditions

We used JNAS [8], consisting of Japanese speech utterances spoken by about 300 Japanese male and female speakers in their 20s, 30s, 40s, 50s, and 60s. The sampling frequency was set to 16 kHz. The 1st through 24th mel-cepstral coefficients parameterised from spectral envelope extracted by STRAIGHT analysis [9] 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 [10]. The frame shift was 5 ms.

For the training of the MR-GMM, we used parallel data sets of a reference speaker and 277 pre-stored target speakers including 137 male and 140 female speakers. The duration of each utterance was approximately 8 seconds. The number of training utterances for each pre-stored target speaker was about 50. The number of mixture components of each MR-GMM was 128 for the spectral features and 64 for the aperiodic components. In this paper, because we focused on only how to design the voice timbre control parameters, the number of subjects was set to 1 in order to remove the effect of perceptual differences among multiple subjects.

B. Evaluation of accuracy of perceived age normalization techniques

We evaluated normalization accuracy of the several perceived age normalization techniques. The following converted voices were evaluated.

- w/o norm: converted voices w/o the perceived age normalization (i.e., perceived age score differential Δw was set to 0),
- w/ PS: converted voices w/ the perceived age normalization on the perceptual age space described in Sect. III-B1,
- w/ AS: converted voices w/ the perceived age normalization on the acoustic space described in Sect. III-B2.
- w/ ref: converted voices w/ the perceived age normalization using the refinement process described in Sect. III-B3.

We performed perceptual evaluation twice to assign the perceived age score to these converted voices. In the first evaluation, the converted voices of w/o norm, w/ PS, and w/ AS were evaluated. Using the results of w/ AS, the perceived age score differentials were refined and the converted voices of w/ ref were generated. Then, we performed the second evaluation using the converted voices of w/o norm, w/ AS, and w/ ref. We used 30 utterances spoken by randomly selected 30 evaluation speakers. We used original perceived age scores of the evaluation speakers as the reference, which were assigned in our preliminary experiment. The normalization target age for all evaluation speakers was set to 45, which was derived by averaging the original perceived age scores for all pre-stored target speakers. To reduce the noise of annotation, the subject annotated the perceived age score to each utterance five times, and then, an average value over these five scores was calculated for each speaker.

TABLE I
Correlation coefficients between target perceived age and perceived age scores of various converted voices

Method	Correlation coefficients	
	First	Second
w/o norm	0.86	0.80
w/ PS	0.81	N/A
w/ AS	0.78	0.67
w/ ref	N/A	0.34

Table I describes correlation coefficients calculated between the averaged perceived scores and the original (reference) ones over the evaluation speakers. If the perceived age normalization effectively works, the correlation coefficient should be close to zero. We can see that the effect of the perceived age normalization by w/ PS and w/ AS is limited but w/ ref is capable of effectively normalizing the perceived age of the converted voices. These results suggest that 1) the perceived age control method based on the MR-GMM is capable of effectively normalizing the perceived age of the converted voices but 2) it is still necessary to manually tune the manipulated parameters (i.e., the perceived age differential) to achieve sufficient performance of the perceived age normalization.

C. Evaluation of voice timbre control based on simultaneous manipulation of multiple voice timbre control parameters

We evaluated controllability of the voice timbre control simultaneously using two voice timbre control parameters. We used the perceived age score as the first voice timbre control parameter. To define the second voice timbre control parameter, we conducted a perceptual evaluation in a similar manner as done in the previous work [7]. We assigned the perceptual scores on several voice timbre expression word pairs to the individual pre-stored target speakers by listening to their voices. Unlike the previous work, we used the perceived age normalized voices (by w/ ref) in this evaluation to more carefully evaluate the remaining voice timbre variations after the perceived age normalization, which were also well modeled by the MR-GMM. We used 11 voice timbre expression word pairs in the evaluation and apply factor analysis to the corresponding 11 perceptual scores over all pre-stored target speakers. As a result, we found that "carrying voice (less carrying - more carrying)" was the most effective voice timbre expression word pairs to model the remaining voice timbre variations. Therefore, we decided to use the carrying voice score as the second voice timbre control parameter.

In order to evaluate the effects of how to design the voice

TABLE II
CORRELATION COEFFICIENTS BETWEEN TWO VOICE TIMBRE CONTROL PARAMETERS AND ANGLES BETWEEN CORRESPONDING TWO BASIS VECTORS.

Method	Correlation coefficients	Angle
Method 1	-0.29	57.1°
Method 2	-0.33	80.2°
Method 3	-0.21	87.8°

timbre control parameters on the voice timbre controllability of the resulting MR-GMM, we assigned the perceived age score and the carrying voice score to the individual pre-stored target speakers using the following three methods:

- Method 1) the natural voice samples of the pre-stored target speakers were used in the annotation.
- Method 2) the converted voice samples from the reference speaker's voices into the pre-stored target speakers' voices were used in the annotation.
- Method 3) the perceived age scores were annotated with the method 2 but the carrying voice scores were annotated using the perceived age normalized voice samples generated by converting the reference speaker's voices into the pre-stored target speakers' voices using the perceived age normalization method with the refinement (w/ ref).

In the method 2, the voice timbre control parameters were designed considering the acoustic variations possibly modeled by the MR-GMM. Furthermore, in the method 3, the second voice timbre control parameter was more carefully designed compared to the method 2 considering only the residual acoustic variations not modeled on a subspace spanned by the first basis vector of the MR-GMM. An annotation scale for the carrying voice was set to 7 (-3 through 3). Using each of the resulting three types of the voice timbre control parameters, the MR-GMM was trained. Then, the converted voice samples were generated using the resulting three MR-GMMs by setting a pair of the perceived age score differential and the carrying voice score differential to (20, -10), (20, 0), (20, 10), (0, -10), (0, 10), (-20, -10), (-20, 0), and (-20, 10). Each converted voice sample was presented to the subject only once in random order, and then, the subject assigned the perceived age score and the carrying voice score to each sample.

Figure 1 and Table II describe the subjective and objective experimental results for the simultaneous voice timbre control using perceived age and carrying voice. In Fig. 1, we can see that the method 3 makes it possible to control voice timbre more widely compared with the other two methods. Moreover, in Tab. II, we can see that the method 3 is capable of keeping the two voice timbre control parameters less correlated to each other and also making the two corresponding basis vectors very close to orthogonal. These results suggest that the voice timbre controllability is significantly improved by carefully designing multiple voice timbre control parameters considering not only 1) correlation between different voice timbre control parameters but also 2) acoustic variations modeled by the MR-GMM to develop a subspace spanned by the basis vectors holding independence on both the perceptual score space and the acoustic feature space.

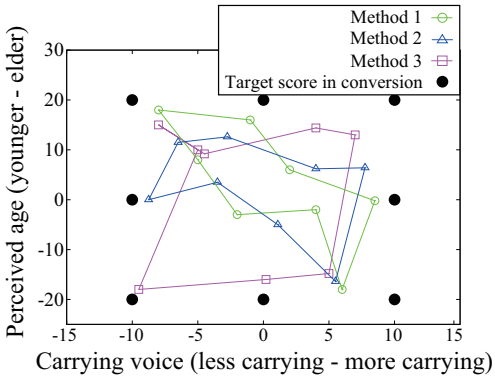

Fig. 1. Result of subjective evaluation on voice timbre control simultaneously manipulating two voice timbre control parameters.

V. CONCLUSIONS

In this paper, in order to improve controllability of the voice timbre control based on the MR-GMM with multiple voice timbre control parameters, we have investigated how to design these parameters to hold independence between the different parameters on both a perceptual space and an acoustic space. The experimental results have demonstrated that the proposed approach to carefully annotating the voice timbre control parameters to each speaker's voice using the converted voices is capable of improving the voice timbre controllability of the MR-GMM. In future work, we plan to develop a model training framework to reduce the annotation efforts while preserving high controllability.

VI. ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI Grant Number 17H01763, Grant-in-Aid for JSPS Research Fellow Number 16J10726, and by JST, PRESTO Grant Number JPMJPR1657.

REFERENCES

[1] T. Toda, "Augmented speech production based on real-time statistical voice conversion," *Proc. GlobalSIP*, pp. 755–759, Dec. 2014.
[2] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. SAP*, vol. 6, no. 2, pp. 131–142, Mar. 1998.
[3] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. ASLP*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.
[4] 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.
[5] 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.
[6] 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. and Syst.*, vol. E97-D, no. 6, pp. 1419–1428, 2014.
[7] H. Kido and H. Kasuya, "Everyday expressions associated with voice quality normal utterance extraction by perceptual evaluation," *The Acoustical Society of Japan (Japanese edition)*, pp. 337–344, May 2001.
[8] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, and K. Shikano, "Japanese Newspaper Article Sentences," *Journal of the Acoustical Society of Japan*, , no. 3, pp. 199–206, 1995.
[9] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, Apr. 1999.
[10] 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.