Adaptive Voice-Quality Control Based on One-to-Many Eigenvoice Conversion

Kumi Ohta[†], *Tomoki Toda*, *Yamato Ohtani*[‡], *Hiroshi Saruwatari*, *Kiyohiro Shikano*

Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan

tomoki@is.naist.jp

Abstract

This paper presents adaptive voice-quality control methods based on one-to-many eigenvoice conversion. To intuitively control the converted voice quality by manipulating a small number of control parameters, a multiple regression Gaussian mixture model (MR-GMM) has been proposed. The MR-GMM also allows us to estimate the optimum control parameters if target speech samples are available. However, its adaptation performance is limited because the number of control parameters is too small to widely model voice quality of various target speakers. To improve the adaptation performance while keeping capability of voice-quality control, this paper proposes an extended MR-GMM (EMR-GMM) with additional adaptive parameters to extend a subspace modeling target voice quality. Experimental results demonstrate that the EMR-GMM yields significant improvements of the adaptation performance while allowing us to intuitively control the converted voice quality.

Index Terms: voice-quality control, voice conversion, eigenvoices, unsupervised adaptation

1. Introduction

There are still several situations where we face difficulties in our speech communication, although numerous helpful technologies have been developed: e.g., even if linguistic information is conveyed by speech-to-speech translation, the user's own voice quality conveying information, such as personality, is lost; or even if people having physical problems in speech production can speak in mechanical voices, they lose their own original voices. Some of these situations can be addressed if voice quality is controlled beyond our physical constraints. A voice conversion (VC) technique used to convert voice quality while keeping linguistic information unchanged has the potential of making our speech communication more convenient.

Statistical approaches to VC have been studied since the late 1980s [1]. One of the most promising approaches is the use of a probabilistic model, such as a Gaussian mixture model (GMM), for converting speech parameters [2, 3]. In this approach, a GMM of the joint probability density of source and target speech parameters is trained in advance with parallel data consisting of dozens of utterance pairs of the source and target voices. This training process using the parallel data results in many limitations in the use of VC. It is indispensable in some applications used to flexibly convert the user's voice quality into the target voice quality if only a small amount of target speech data is available or even if any of them are not available.

To make the training process more flexible, there have been proposed several adaptive VC methods (e.g., [4]). As one of them, we have proposed one-to-many eigenvoice conversion (EVC) [5] to make it possible to flexibly convert the source speaker's voice into an arbitrary speaker's voice. A one-to-many eigenvoice GMM (EV-GMM) is trained in advance with multiple parallel data sets consisting of utterance pairs of the source speaker and many prestored target speakers. The eigenvoice technique [6] is capable of modifying the probability density function by changing a small number of global weighting parameters. This feature allows us 1) to rapidly adapt the one-to-many EV-GMM to arbitrary target speakers using only a small amount of their speech data without any linguistic restrictions and 2) to efficiently generate converted speech with various voice characteristics by manipulating the weighting parameters. However, it is difficult to intuitively control the converted voice quality because manipulating only a single weighting parameter causes simultaneous changes in various voice quality factors, such as gender, age, and so on.

Inspired by the use of multiple regression hidden Markov models (HMMs) for intuitively controlling the voice quality of synthetic speech in HMM-based text-to-speech synthesis [7], we have proposed a voice-quality control method with a multiple regression GMM (MR-GMM) in the one-to-many EVC framework [8]. The structure of the MR-GMM is the same as that of the EV-GMM, but representative vectors spanning a subspace to model the voice qualities of various target speakers are intentionally optimized so that each of them captures the voice characteristics described by each of the primitive word pairs expressing voice quality factors, such as male/female for gender or elder/younger for age. To prevent the degradation of controllability caused by the use of many control parameters, a small number of important primitive word pairs that effectively describe the voice qualities of various speakers have to be selected. However, a decrease in the number of representative vectors reduces the adaptation performance when target speech samples are available because the subspace spanned by such a small number of representative vectors does not sufficiently cover the voice qualities of various speakers.

In this paper, we propose an extended MR-GMM (EMR-GMM) to improve the adaptation performance of the MR-GMM while keeping the controllability of voice quality sufficiently high. Both representative vectors capturing the voice characteristics described by the primitive word pairs and additional representative vectors are used for extending the subspace so as to model the voice characteristics not well described by the primitive word pairs. Experimental results demonstrate that the EMR-GMM yields significant improvements of the adaptation performance while keeping the capability of manual control of the converted voice quality.

2. Voice-Quality Control and Adaptation with MR-GMM

2.1. MR-GMM

Let $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ and $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$ be $2D$-dimensional source and target acoustic feature vectors consist-

[†] Presently, with Brother Industries, Ltd., Japan.

[‡] Presently, with TOSHIBA CORPORATION, Japan.

ing of D-dimensional static and dynamic feature vectors at frame t, where $^\top$ denotes the transposition of the vector. The joint probability density of the source and target feature vectors is modeled by the MR-GMM as

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}^{(MR)}, \boldsymbol{w}_c) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{X}_t, \boldsymbol{Y}_t; \boldsymbol{\mu}_{m,\boldsymbol{w}_c}^{(X,Y)}, \boldsymbol{\Sigma}_m^{(X,Y)}), \quad (1)$$

where

$$\boldsymbol{\mu}_{m,\boldsymbol{w}_c}^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_{m,\boldsymbol{w}_c}^{(Y)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{B}_m^{(Y)}\boldsymbol{w}_c + \boldsymbol{b}_m^{(Y)}(0) \end{bmatrix} \quad (2)$$

$$\boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \quad (3)$$

and $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ shows the normal distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The total number of mixture components is M. The weight of the m^{th} mixture component is α_m. The target mean vector of the m^{th} mixture component is modeled by a linear combination of a bias vector $\boldsymbol{b}_m^{(Y)}(0)$ and representative vectors $\boldsymbol{B}_m^{(Y)} = [\boldsymbol{b}_m^{(Y)}(1), \cdots, \boldsymbol{b}_m^{(Y)}(J)]$ with a J-dimensional voice-quality (VQ) control vector $\boldsymbol{w}_c = [w_c(1), \cdots, w_c(J)]^\top$, each component of which captures a weighting factor for primitive word pairs expressing voice quality. The parameter set $\boldsymbol{\lambda}^{(MR)}$ consists of mixture-dependent parameters tied over every target speaker, i.e., the mixture-component weights, the source mean vectors, the bias and representative vectors, and the covariance matrices.

Note that the structure of the MR-GMM is the same as that of the EV-GMM, but the definitions of the weighting parameters for the representative vectors are different.

2.2. Training of MR-GMM

Multiple parallel data sets consisting of the source speaker's voices and many prestored target speakers' voices are used for training the MR-GMM. First, perceptual scores for the primitive word pairs (e.g., -1: elder, 0: neutral, and 1: younger for an elder/younger pair) are manually assigned into each prestored target speaker. These perceptual scores are used as components of the VQ control vector. Then, the MR-GMM parameter set is optimized using all of the multiple parallel data sets with the target-speaker-dependent VQ control vectors as

$$\hat{\boldsymbol{\lambda}}^{(MR)} = \arg\max_{\boldsymbol{\lambda}^{(MR)}} \prod_{s=1}^{S} \prod_{t=1}^{T_s} P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}|\boldsymbol{\lambda}^{(MR)}, \boldsymbol{w}_c^{(s)}), \quad (4)$$

where $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t^{(s)}$ are the feature vectors of the source and s^{th} target speakers at frame t, respectively. The VQ control vector of the s^{th} prestored target speaker is $\boldsymbol{w}_c^{(s)}$. This optimization can be performed with the EM algorithm to iteratively maximize the following auxiliary function,

$$\mathcal{Q}(\boldsymbol{\lambda}^{(MR)}; \hat{\boldsymbol{\lambda}}^{(MR)}) = \sum_{s=1}^{S} \sum_{t=1}^{T_s} \sum_{m=1}^{M} \gamma_{m,t}^{(s)} \log P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, m|\hat{\boldsymbol{\lambda}}^{(MR)}, \boldsymbol{w}_c^{(s)}) \quad (5)$$

$$\gamma_{m,t}^{(s)} = P(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \boldsymbol{\lambda}^{(MR)}, \boldsymbol{w}_c^{(s)}). \quad (6)$$

2.3. Manual Control of Voice Quality

To intuitively control the converted voice quality, the VQ control vector is manually determined using scores for the individual primitive word pairs. The target mean vectors of the joint probability density function of the source and target voices are defined on the basis of the determined VQ control vector.

In spectral conversion, the maximum-likelihood trajectory-based conversion method considering global variance (GV) [3] is employed. The GV probability density function given by the normal distribution also depends on the target voices. Therefore, its mean vector is modeled by multiple-linear regression with the VQ control vector and its covariance matrix is tied over every target speakers, such as that in the MR-GMM.

In F_0 conversion, the source F_0 is converted as

$$\log \hat{F}_{0,t}^{(y)} = \frac{\sigma^{(y)}}{\sigma^{(x)}} \left(\log F_{0,t}^{(x)} - \mu^{(x)} \right) + \mu^{(y)}, \quad (7)$$

where $F_{0,t}^{(x)}$ and $\hat{F}_{0,t}^{(y)}$ are the source and converted F_0 values at frame t, respectively. The mean and standard deviation of log-scaled F_0 are shown by $\mu^{(x)}$ and $\sigma^{(x)}$ for the source voice and by $\mu^{(y)}$ and $\sigma^{(y)}$ for the target voice, respectively. The target-dependent parameters $\mu^{(y)}$ and $\sigma^{(y)}$ are also modeled by multiple-linear regression with the VQ control vector.

2.4. Unsupervised Adaptation of Voice Quality

The MR-GMM is adapted to arbitrary target speakers by estimating of the optimum VQ control vector from the given speech samples. This process is performed in a completely unsupervised manner without parallel data and linguistic information.

The VQ control vector is estimated so that the likelihood of the marginal distribution of the adapted MR-GMM for a time sequence of the given target feature vectors $\boldsymbol{Y}_1', \cdots, \boldsymbol{Y}_T'$ is maximized as

$$\hat{\boldsymbol{w}}_c = \arg\max_{\boldsymbol{w}_c} \prod_{t=1}^{T} \int P(\boldsymbol{X}_t, \boldsymbol{Y}_t'|\boldsymbol{\lambda}^{(MR)}, \boldsymbol{w}_c) d\boldsymbol{X}_t. \quad (8)$$

The following auxiliary function is iteratively maximized,

$$\mathcal{Q}(\boldsymbol{w}_c; \hat{\boldsymbol{w}}_c) = \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{m,t}' \log P(\boldsymbol{Y}_t', m|\boldsymbol{\lambda}^{(MR)}, \hat{\boldsymbol{w}}_c) \quad (9)$$

$$\gamma_{m,t}' = P(m|\boldsymbol{Y}_t', \boldsymbol{\lambda}^{(MR)}, \boldsymbol{w}_c). \quad (10)$$

The ML estimate of the weight vector is given by

$$\hat{\boldsymbol{w}}_c = \left[\sum_{m=1}^{M} \overline{\gamma}_m' \boldsymbol{B}_m^{(Y)\top} \boldsymbol{\Sigma}_m^{(YY)-1} \boldsymbol{B}_m^{(Y)} \right]^{-1} \cdot \left[\sum_{m=1}^{M} \boldsymbol{B}_m^{(Y)\top} \boldsymbol{\Sigma}_m^{(YY)-1} \overline{\boldsymbol{Y}}_m' \right], \quad (11)$$

where

$$\overline{\gamma}_m' = \sum_{t=1}^{T} \gamma_{m,t}' \quad (12)$$

$$\overline{\boldsymbol{Y}}_m' = \sum_{t=1}^{T} \gamma_{m,t}' \left(\boldsymbol{Y}_t' - \boldsymbol{b}_m^{(Y)}(0) \right). \quad (13)$$

In this paper, we directly calculate the mean vector of the GV probability density function and the mean and standard deviation of the target log-scaled F_0 from the given target speech samples instead of estimating the VQ control vectors for those parameters.

3. Voice-Quality Control and Adaptation with EMR-GMM

In the manual control of voice quality, it is better to reduce the number of primitive word pairs as much as possible to simplify the manual control process. However, the reduction in the number of representative vectors degrades the adaptation performance of the MR-GMM because the subspace spanned by those vectors does not cover the voice qualities of various target speakers sufficiently. To improve the adaptation performance of the MR-GMM while keeping its capability of voice-quality control, the subspace modeling the voice qualities of various target speakers is extended by the use of additional representative vectors.

The joint probability density function of the source and target feature vectors $P(\boldsymbol{X}_t, \boldsymbol{Y}_t|\boldsymbol{\lambda}^{(EMR)}, \boldsymbol{w}_c, \boldsymbol{w}_h)$ is modeled by the EMR-GMM, the target mean vector of the m^{th} mixture component of which is given by

$$\mu_m^{(Y)}(\boldsymbol{w}_c, \boldsymbol{w}_h) = \boldsymbol{B}_m^{(Y)}\boldsymbol{w}_c + \boldsymbol{H}_m^{(Y)}\boldsymbol{w}_h + \boldsymbol{b}_m^{(Y)}(0), \quad (14)$$

where $\boldsymbol{H}_m^{(Y)} = [\boldsymbol{h}_m^{(Y)}(1), \cdots, \boldsymbol{h}_m^{(Y)}(K)]$ is the set of additional representative vectors and $\boldsymbol{w}_h = [w_h(1), \cdots, w_h(K)]^\top$ is the K-dimensional adaptive weight vector. The parameter set $\boldsymbol{\lambda}^{(EMR)}$ consists of the original MR-GMM parameters and the additional representative vectors. The voice characteristics well described by the primitive word pairs are modeled on the subspace spanned by the original representative vectors and the residual voice characteristics are modeled on the subspace spanned by the additional representative vectors.

In training, both the EMR-GMM parameter set $\boldsymbol{\lambda}^{(EMR)}$ and the set of adaptive weight vectors for individual prestored target speakers $\boldsymbol{w}_h^{(1:S)} = \left\{\boldsymbol{w}_h^{(1)}, \cdots, \boldsymbol{w}_h^{(S)}\right\}$ are optimized using all of the multiple parallel data sets with the manually determined target-speaker-dependent VQ control vectors as

$$\left\{\hat{\boldsymbol{\lambda}}^{(EMR)}, \hat{\boldsymbol{w}}_h^{(1:S)}\right\} = \arg\max_{\left\{\boldsymbol{\lambda}^{(EMR)}, \boldsymbol{w}_h^{(1:S)}\right\}} \prod_{s=1}^{S}\prod_{t=1}^{T_s} P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}|\boldsymbol{\lambda}^{(EMR)}, \boldsymbol{w}_c^{(s)}, \boldsymbol{w}_h^{(s)}). \quad (15)$$

This optimization is performed with EM algorithm maximizing the following auxiliary function:

$$\mathcal{Q}\left(\left\{\boldsymbol{\lambda}^{(EMR)}, \boldsymbol{w}_h^{(1:S)}\right\}; \left\{\hat{\boldsymbol{\lambda}}^{(EMR)}, \hat{\boldsymbol{w}}_h^{(1:S)}\right\}\right) = \sum_{s=1}^{S}\sum_{t=1}^{T_s}\sum_{m=1}^{M}\gamma_{m,t}^{(s)} \log P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}|\hat{\boldsymbol{\lambda}}^{(EMR)}, \boldsymbol{w}_c^{(s)}, \hat{\boldsymbol{w}}_h^{(s)}) \quad (16)$$

$$\gamma_{m,t}^{(s)} = P(m|\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)}, \boldsymbol{\lambda}^{(EMR)}, \boldsymbol{w}_c^{(s)}, \boldsymbol{w}_h^{(s)}). \quad (17)$$

It is straightforward to apply an adaptive training method for the EV-GMM to EMR-GMM training.

In the manual control of voice quality, only the VQ control vector is manipulated in the same manner as the MR-GMM.

In the unsupervised adaptation, both the VQ control vector and the adaptive weight vector are estimated so that the likelihood of the marginal distribution of the adapted EMR-GMM for a time sequence of the given target feature vectors is maximized as

$$\{\hat{\boldsymbol{w}}_c, \hat{\boldsymbol{w}}_h\} = \arg\max_{\{\boldsymbol{w}_c, \boldsymbol{w}_h\}} \prod_{t=1}^{T}\int P(\boldsymbol{X}_t, \boldsymbol{Y}_t'|\boldsymbol{\lambda}^{(EMR)}, \boldsymbol{w}_c^{(s)}, \boldsymbol{w}_h^{(s)})d\boldsymbol{X}_t. \quad (18)$$

The following auxiliary function is iteratively maximized,

$$\mathcal{Q}(\{\boldsymbol{w}_c, \boldsymbol{w}_h\}; \{\hat{\boldsymbol{w}}_c, \hat{\boldsymbol{w}}_h\}) = \sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_{m,t}' \log P(\boldsymbol{Y}_t', m|\boldsymbol{\lambda}^{(EMR)}, \hat{\boldsymbol{w}}_c^{(s)}, \hat{\boldsymbol{w}}_h^{(s)}) \quad (19)$$

$$\gamma_{m,t}' = P(m|\boldsymbol{Y}_t', \boldsymbol{\lambda}^{(EMR)}, \boldsymbol{w}_c^{(s)}, \boldsymbol{w}_h^{(s)}). \quad (20)$$

The ML estimate of the joint vector $\boldsymbol{w} = [\boldsymbol{w}_c^\top, \boldsymbol{w}_h^\top]^\top$ is given by

$$\hat{\boldsymbol{w}} = \left[\sum_{m=1}^{M}\overline{\gamma}_m' \left[\boldsymbol{B}_m^{(Y)}, \boldsymbol{H}_m^{(Y)}\right]^\top \boldsymbol{\Sigma}_m^{(YY)-1}\left[\boldsymbol{B}_m^{(Y)}, \boldsymbol{H}_m^{(Y)}\right]\right]^{-1} \cdot \left[\sum_{m=1}^{M}\left[\boldsymbol{B}_m^{(Y)}, \boldsymbol{H}_m^{(Y)}\right]^\top \boldsymbol{\Sigma}_m^{(YY)-1}\overline{\boldsymbol{Y}}_m'\right]. \quad (21)$$

4. Experimental Evaluations

4.1. Experimental Conditions

We used 40 speakers consisting of 10 male and 10 female speakers from the Japanese Newspaper Article Sentences (JNAS) database [9] and 10 senior male and 10 senior female speakers from the Senior-Japanese Newspaper Article Sentences (S-JNAS) database [10] to train the MR-GMM and the EMR-GMM. Each speaker uttered one of the phonetically balanced 50 sentence sets. We used a female speaker as the source speaker, who uttered the same sentence sets as the prestored speakers. The number of extended representative vectors of the EMR-GMM was set to 15. The number of mixture components was set to 128 in both the MR-GMM and the EMR-GMM.

To develop the VQ control vector for each prestored target speaker, we used a 7-scaled score (-3: very, -2: quite, -1: somewhat, 0: neutral, 1: somewhat, 2: quite, 3: very) for 2 Japanese primitive word pairs, elder/younger and heavy/light. One Japanese female subject manually assigned these scores to each of the prestored target speakers by listening to natural speech samples of various sentences. Scores for each primitive word pair over different prestored target speakers were normalized into the Z-score (zero mean and unit variance), and the normalized scores were used as components of the VQ control vector.

The 1^{st} through 24^{th} mel-cepstral coefficients were used as spectral parameters. STRAIGHT [11] was employed as the analysis-synthesis method. The sampling frequency was set to 16 kHz. The frame shift was set to 5 ms.

4.2. Evaluation of Voice-Quality Control

We conducted the subjective evaluation of voice-quality control. The number of listeners was 5. The fifty test sentences not included in the training data were used for such evaluation. For each sentence, we synthesized 5 samples of the converted speech by varying only one component of the 2-dimensional VQ control vector from -2 to 2 in 5 steps while setting the other component to zero. The converted speech samples when setting every component of the VQ control vector to zero were also synthesized as reference speech. Each listener compared the voice quality of the converted speech with that of the reference speech using a 5-scaled score (-2: very, -1: somewhat, 0: no difference, 1: somewhat, 2: very) for the primitive word pair corresponding to the varied component. In the EMR-GMM, every component of the adaptive weight vector for the additional representative vectors was set to zero.

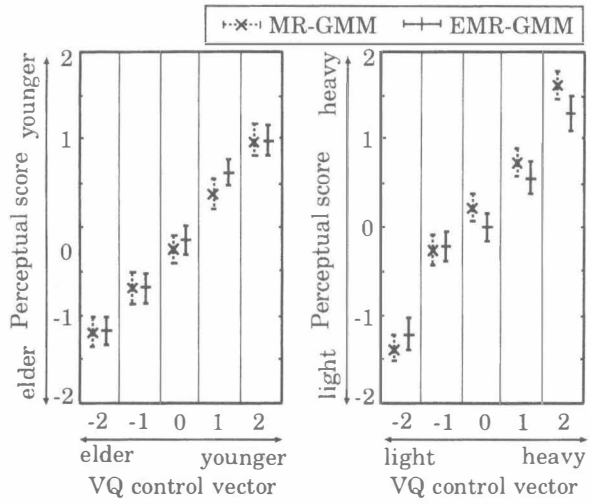

Figure 1: *Results of perceptual evaluation of voice-quality control (left: elder/younger and right: light/heavy). Mean values and their 95% confidence intervals are shown.*

Experimental results are shown in **Figure 1**. In voice-quality control based on elder/younger and light/heavy pairs, perceptual scores relatively well correlate to the setting of the VQ control vector. These tendencies are observed in both the MR-GMM and the EMR-GMM. Therefore, both the MR-GMM and the EMR-GMM allow us to manually control voice quality by manipulating the VQ control vector.

4.3. Evaluations of Voice-Quality Adaptation

We also conducted objective and subjective evaluations of unsupervised adaptation. We used 10 target speakers consisting of 5 male and 5 female speakers from JNAS, who were not included in the group of prestored target speakers. For each speaker, 2 sentences were used for adaptation and 51 sentences were used for evaluation from the 53 phonetically balanced sentence set not used in the training process. In the objective evaluation, the mel-cepstral distortion between the converted and target spectra was used as an evaluation metric. In the subjective evaluation, we conducted a preference test of speech quality and an XAB test of speaker individuality. In the preference test, sample pairs of the converted speech with the MR-GMM and with the EMR-GMM were presented to listeners in random order and the listeners were asked which sample sounded better. In the XAB test, those sample pairs were presented to listeners in random order after the target speaker's speech was presented to them as reference. Then, the listeners were asked which sample sounded more similar to the reference target. The number of listeners was 7. Each listener evaluated 160 sample pairs in each test.

Results of the objective and subjective evaluations are shown in **Table 1** and **Figure 2**, respectively. The EMR-GMM yields higher spectral conversion accuracy and more significant improvements in both speech quality and speaker individuality compared with the MR-GMM because the EMR-GMM is capable of sufficiently modeling the voice qualities of various target speakers.

5. Conclusions

In this paper, adaptive voice-quality control methods based on a one-to-many eigenvoice conversion framework have been described. To intuitively control the converted voice quality, a multiple regression GMM (MR-GMM) has been proposed. Furthermore, to improve the unsupervised adaptation performance of the MR-GMM when target speech samples are available, we have proposed an extended MR-GMM (EMR-GMM). Experimental results showed that the EMR-GMM is effective for improving the adaptation performance of the MR-GMM while keeping the controllability of voice quality sufficiently high.

Table 1: *Mel-cepstral distortion [dB] between the converted and target spectra in voice-quality adaptation. All the components of the VQ control vector and the adaptive weight vector are set to zero before adaptation. Mel-cepstral distortion between the source and the target spectra is 8.28 [dB].*

	before adaptation	after adaptation
MR-GMM	5.73	5.54
EMR-GMM	5.77	5.20

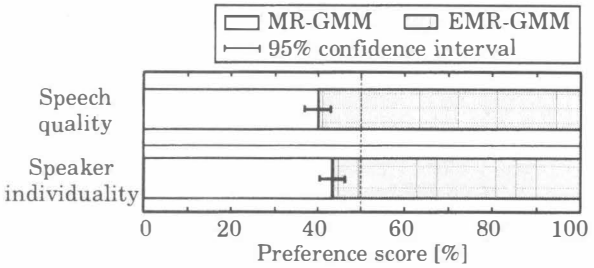

Figure 2: *Results of preference test of speech quality and XAB test of speaker individuality in voice-quality adaptation.*

Acknowledgements: The authors are grateful to Professor Hideki Kawahara of Wakayama University, Japan, for permission to use the STRAIGHT analysis-synthesis method. This research was supported in part by MIC SCOPE and MEXT Grant-in-Aid for Young Scientists (A).

6. References

[1] H. Kuwabara and Y. Sagisaka. Acoustic characteristics of speaker individuality: control and conversion. *Speech Communication*, Vol. 16, No. 2, pp. 165–173, 1995.

[2] Y. Stylianou, O. Cappé, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Trans. SAP*, Vol. 6, No. 2, pp. 131–142, 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, 2007.

[4] A. Mouchtaris, J.V. der Spiegel, and P. Mueller. Nonparallel training for voice conversion based on a parameter adaptation approach. *IEEE Trans. ASLP*, Vol. 14, No. 3, pp. 952–963, 2006.

[5] T. Toda, Y. Ohtani, and K. Shikano. One-to-many and many-to-one voice conversion based on eigenvoices. *Proc. ICASSP*, pp. 1249–1252, Hawaii, USA, Apr. 2007

[6] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Trans. SAP*, Vol. 8, No. 6, pp. 695–707, 2000.

[7] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi. A style control technique for HMM-based expressive speech synthesis. *IEICE Trans.*, Vol. E90-D, No. 9, pp. 1406–1413, 2007.

[8] K. Ohta, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Regression approaches to voice quality control based on one-to-many eigenvoice conversion. *Proc. SSW6*, pp. 101–106, Bonn, Germany, Aug. 2007.

[9] JNAS: Japanese Newspaper Article Sentences. *http://research.nii.ac.jp/src/eng/list/detail.html#JNAS*

[10] S-JNAS: Senior-Japanese Newspaper Article Sentences. *http://research.nii.ac.jp/src/eng/list/detail.html#S-JNAS*

[11] H. Kawahara, I. Masuda-Katsuse, and A.de 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, 1999.