SSW6

universitätbonn

Regression Approaches to Voice Quality Control Based on One-to-Many Eigenvoice Conversion

Kumi Ohta, Yamato Ohtani, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology, Japan

{kumi-o, yamato-o, tomoki, sawatari, shikano}@is.naist.jp

Abstract

This paper proposes techniques for flexibly controlling voice quality of converted speech from a particular source speaker based on one-to-many eigenvoice conversion (EVC). EVC realizes a voice quality control based on the manipulation of a small number of parameters, i.e., weights for eigenvectors, of an eigenvoice Gaussian mixture model (EV-GMM), which is trained with multiple parallel data sets consisting of a single source speaker and many pre-stored target speakers. However, it is difficult to control intuitively the desired voice quality with those parameters because each eigenvector doesn't usually represent a specific physical meaning. In order to cope with this problem, we propose regression approaches to the EVC-based voice quality controller. The tractable voice quality control of the converted speech is achieved with a low-dimensional voice quality control vector capturing specific voice characteristics. We conducted experimental verifications of each of the proposed approaches.

1. Introduction

Voice conversion (VC) is a technique for converting non-linguistic features such as speaker individuality while keeping the linguistic features. One of the most typical VC applications is speaker conversion that converts a certain speaker's voice into another speaker's voice [1]. This technique realizes a voice quality controller which converts one user's voice quality into another voice, which is very useful not only as an amusement device but also as a speech enhancement device for a speaking aid system recovering a disabled person's voice or as a hearing aid system to make speech sounds more intelligible.

Speech morphing [2] [3] is one of the techniques for constructing a voice quality controller. Input speech is usually converted by manipulating acoustic features such as fundamental frequency (F0) and spectral envelope in a simple manner, e.g., linear spectral warping. One advantage of this method is that it is easily used without training for a specific conversion model. On the other hand, this system allows very limited voice quality control of the converted speech. Since, in this case, the resulting voice quality strongly depends on the user's own voice quality, it is indeed difficult to convert any arbitrary voice into any desired speaker's voice.

As a technique for realizing a specific speaker's voice, a statistical approach to VC has been studied [1]. This framework trains a conversion model between a source speaker and a target speaker in advance, using parallel data consisting of utterance pairs of those two speakers [4]. A Gaussian mixture model (GMM) is often used as the conversion model [5]. The resulting model allows the determination of target speech parameters given the source parameters based on minimum mean square error (MMSE) estimation [5] or maximum likelihood estimation (MLE) [6] without any linguistic restrictions. Thus an arbitrary sentence uttered by the source speaker is rendered as an utterance by the target speaker. Because this framework needs training samples of the desired target speaker's voices, it is very difficult to construct a voice quality controller with the flexibility to vary voice quality of the converted speech.

As a novel VC framework, eigenvoice conversion (EVC) has been proposed [7] [8]. The eigenvoice is a popular speaker adaptation technique in the speech recognition area [9] [10]. It has also been applied to HMM-based TTS [11]. EVC realizes the conversion from a particular source speaker's voice into arbitrary speakers' voices (one-to-many EVC) or that from arbitrary speakers' voices into a particular target speaker's voice (many-to-one EVC). In one-to-many EVC, the eigenvoice Gaussian mixture model (EV-GMM) is trained in advance, using multiple parallel data sets consisting of utterance-pairs of the source speaker and multiple pre-stored target speakers. The voice quality of the converted speech is controlled by a small number of free parameters for eigenvectors capturing dominant voice characteristics extracted from pre-stored target speakers, which are called eigenvoices. Therefore, this framework allows us to control manually the voice quality of the converted speech. However, it is difficult to control intuitively the desired voice quality because each eigenvoice doesn't usually represent a specific physical meaning.

Recently, a multiple regression approach has been proposed for intuitively controlling voice quality of synthetic speech in the HMM/HSMM-based TTS [12] [13]. HMM/HSMM parameters are controlled with a low-dimensional vector called the **voice quality control vector**. Each component of the voice quality control vector captures specific characteristics of voice quality described by expression words such as sex, age and brightness. This paper proposes multiple regression approaches to EVC for constructing the voice quality controller that allows us to intuitively control the voice quality of the converted speech. We conducted experimental verifications for showing the advantages and disadvantages of each of them.

This paper is organized as follows. In **Section 2**, the framework of EVC is described. In **Section 3**, the proposed methods for constructing the EVC-based voice quality controller are described. In **Section 4**, experimental verifications are described. Finally, we summarize this paper in **Section 5**.

2. One-to-Many Eigenvoice Conversion (EVC) [7] [8]

The framework of EVC is shown in **Figure 1**.

2.1. Eigenvoice GMM (EV-GMM)

We use $2D$-dimensional acoustic features $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top,\ \Delta\boldsymbol{x}_t^\top\right]^\top$ (source speaker's) and $\boldsymbol{Y}_t^{(s)} = \left[\boldsymbol{y}_t{}^{(s)\top},\ \Delta\boldsymbol{y}_t{}^{(s)\top}\right]^\top$ (the

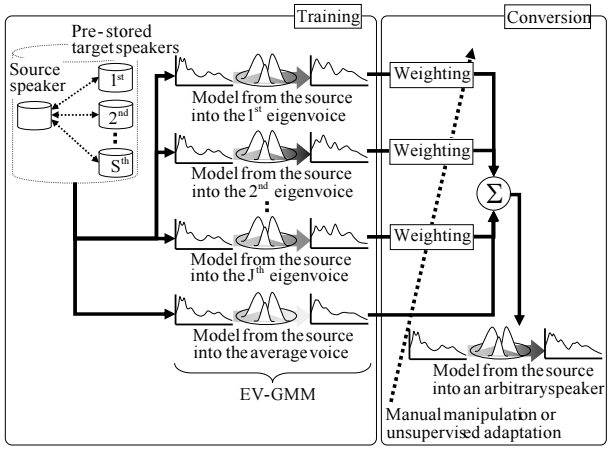

Figure 1: *Framework of EVC.*

$s^{\rm th}$ pre-stored target speaker's) consisting of D-dimensional static and dynamic features at frame t, where $\top$ denotes transposition of the vector. Using a parallel training data set consisting of time-aligned source and target features $\boldsymbol{Z}_t^{(s)} = \left[\boldsymbol{X}_t^\top,\ \boldsymbol{Y}_t{}^{(s)\top}\right]^\top$ determined by Dynamic Time Warping (DTW), the EV-GMM $\boldsymbol{\lambda}^{(EV)}$ on joint probability density $P\left(\boldsymbol{Z}_t^{(s)}|\boldsymbol{\lambda}^{(EV)}\right)$ is trained in advance. The joint probability density is written as

$$
\begin{aligned}
P\left(\boldsymbol{Z}_t^{(s)}|\boldsymbol{\lambda}^{(EV)}\right) &= \sum_{i=1}^{M} \alpha_i \mathcal{N}\left(\boldsymbol{Z}_t; \boldsymbol{\mu}_i^{(Z)},\ \boldsymbol{\Sigma}_i^{(ZZ)}\right), \\
\boldsymbol{\mu}_i^{(Z)} &= \begin{bmatrix} \boldsymbol{\mu}_i^{(X)} \\ \boldsymbol{B}_i^{(Y)}\boldsymbol{w} + \boldsymbol{b}_i^{(Y)}(0) \end{bmatrix}, \\
\boldsymbol{\Sigma}_i^{(ZZ)} &= \begin{bmatrix} \boldsymbol{\Sigma}_i^{(XX)} & \boldsymbol{\Sigma}_i^{(XY)} \\ \boldsymbol{\Sigma}_i^{(YX)} & \boldsymbol{\Sigma}_i^{(YY)} \end{bmatrix},
\end{aligned} \tag{1}
$$

where $\mathcal{N}\left(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$ shows the normal distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The $i^{\rm th}$ mixture weight is α_i. The total number of mixtures is M. In the EV-GMM, the target mean vector for the $i^{\rm th}$ mixture is represented as a linear combination of a bias vector $\boldsymbol{b}_i^{(Y)}(0)$ and eigenvectors $\boldsymbol{B}_i^{(Y)} = \left[\boldsymbol{b}_i^{(Y)}(1),\ \boldsymbol{b}_i^{(Y)}(2),\ \cdots,\ \boldsymbol{b}_i^{(Y)}(J)\right]$. The number of eigenvectors is J. The target speaker individuality is controlled with only the J-dimensional weight vector $\boldsymbol{w} = \left[w(1),\ w(2),\ \cdots,\ w(J)\right]^\top$ for eigenvectors. Consequently, the EV-GMM has a parameter set $\boldsymbol{\lambda}^{(EV)}$ consisting of the single weight vector and parameters for individual mixtures such as the mixture weights, the source mean vectors, the bias and eigenvectors, and the covariance matrices. This paper employs diagonal covariance matrices for individual blocks, $\boldsymbol{\Sigma}_i^{(XX)}$, $\boldsymbol{\Sigma}_i^{(XY)}$, $\boldsymbol{\Sigma}_i^{(YX)}$, and $\boldsymbol{\Sigma}_i^{(YY)}$.

2.2. Training of EV-GMM

In order to train the EV-GMM, we use multiple parallel data sets. Each of them consists of utterance-pairs of the source speaker and one of the multiple pre-stored target speakers.

Firstly, we train a target independent GMM $\boldsymbol{\lambda}^{(0)}$ simultaneously, using all of the multiple parallel data sets as follows:

$$
\boldsymbol{\lambda}^{(0)} = \arg\max \prod_{s=1}^{S} \prod_{t=1}^{T_s} P\left(\boldsymbol{Z}_t^{(s)}|\boldsymbol{\lambda}\right), \tag{2}
$$

The number of feature vectors for the $s^{\rm th}$ speaker is T_s. The number of pre-stored target speakers is S. Secondly, we train each target dependent GMM $\boldsymbol{\lambda}^{(s)}$ by updating only target mean vectors $\boldsymbol{\mu}_i^{(Y)}$ of the target independent GMM $\boldsymbol{\lambda}^{(0)}$ using each of multiple parallel data sets as follows:

$$
\boldsymbol{\lambda}^{(s)} = \arg\max \prod_{t=1}^{T_s} P\left(\boldsymbol{Z}_t^{(s)}|\boldsymbol{\lambda}\right). \tag{3}
$$

Lastly, we determine the bias vector $\boldsymbol{b}_i^{(Y)}(0)$ and the eigenvectors $\boldsymbol{B}_i^{(Y)}$. We prepare a $(2D \times M)$-dimensional supervector $\boldsymbol{\mu}^{(Y)}(s) = \left[\boldsymbol{\mu}_1^{(Y)}(s)^\top,\ \boldsymbol{\mu}_2^{(Y)}(s)^\top,\ \cdots,\ \boldsymbol{\mu}_M^{(Y)}(s)^\top\right]^\top$ for each pre-stored target speaker by concatenating the target mean vectors $\boldsymbol{\mu}_i^{(Y)}(s)$ of the target dependent GMM $\boldsymbol{\lambda}^{(s)}$. We extract the eigenvectors with principal component analysis (PCA) for the supervectors. Consequently, the supervector is written as

$$
\begin{aligned}
\boldsymbol{\mu}^{(Y)}(s) &\simeq \boldsymbol{B}^{(Y)}\boldsymbol{w}^{(s)} + \boldsymbol{b}^{(Y)}, \\
\boldsymbol{B}^{(Y)} &= \left[\boldsymbol{B}_1{}^{(Y)\top}, \boldsymbol{B}_2{}^{(Y)\top},\ \cdots,\ \boldsymbol{B}_M{}^{(Y)\top}\right]^\top, \\
\boldsymbol{b}^{(Y)} &= \left[\boldsymbol{b}_1^{(Y)}(0)^\top, \boldsymbol{b}_2^{(Y)}(0)^\top,\ \cdots,\ \boldsymbol{b}_M^{(Y)}(0)^\top\right]^\top,
\end{aligned} \tag{4}
$$

$$
\boldsymbol{b}_i^{(Y)}(0)^\top = \frac{1}{S}\sum_{s=1}^{S} \boldsymbol{\mu}_i^{(Y)}(s), \tag{5}
$$

where $\boldsymbol{w}^{(s)}$ consists of the principal components for the $s^{\rm th}$ pre-stored target speaker. We construct the EV-GMM $\boldsymbol{\lambda}^{(EV)}$ from the resulting bias and eigenvectors and the tied parameters, i.e., the mixture weights, the source mean vectors, and the covariance matrices of the target independent GMM. Now, various supervectors, i.e., the target mean vectors are created by varying only J $(< S \ll 2D \times M)$ free parameters of $\boldsymbol{w}$.

2.3. Problems in EV-GMM

The EV-GMM allows the control of voice quality of the converted speech by manually changing the weight vector. However, individual eigenvectors only capture dominant voice characteristics among pre-stored target speakers, which don't represent a specific physical meaning such as a masculine voice, a feminine voice, a hoarse voice, or a clear voice. Therefore, it is difficult to intuitively control the desired voice quality.

3. EVC-Based Voice Quality Controller

We propose regression approaches to the EVC-based voice controller for realizing the control of target mean vectors $\boldsymbol{\mu}_i^{(Y)}$ with the K-dimensional voice quality control vector $\boldsymbol{w}_e$ as follows:

$$
\boldsymbol{\mu}_i^{(Y)} = \hat{\boldsymbol{B}}_i^{(Y)}\boldsymbol{w}_e + \hat{\boldsymbol{b}}_i^{(Y)}(0), \tag{6}
$$

where

$$
\hat{\boldsymbol{B}}_i^{(Y)} = \left[\hat{\boldsymbol{b}}_i^{(Y)}(1),\ \hat{\boldsymbol{b}}_i^{(Y)}(2),\ \cdots,\ \hat{\boldsymbol{b}}_i^{(Y)}(K)\right].
$$

First, appropriate components of the voice quality control vector are manually assigned to each pre-stored target speaker. And then, the regression parameters $\hat{\boldsymbol{B}}_i^{(Y)}$ and $\hat{\boldsymbol{b}}_i^{(Y)}(0)$ are estimated by the following three methods: **A**) least squares estimation (LSE) of a regression matrix converting the voice quality control vector into principal components, **B**) LSE of a regression matrix converting the voice quality control vector into target mean vectors, and **C**) MLE of all parameters of EV-GMM under the condition of Eq. (6). **Figures 2**, **3**, and **4** show these proposed methods **A**, **B**, and **C**, respectively.

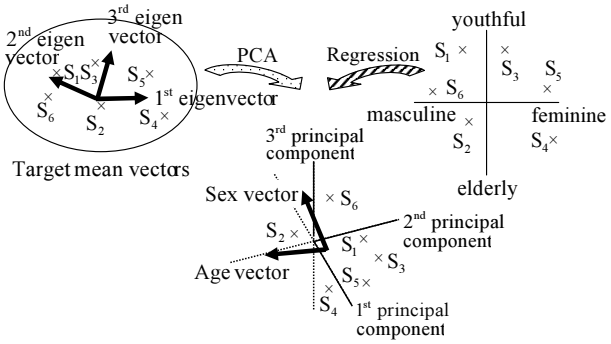

Figure 2: *Proposed method A.*

3.1. Proposed Method A: Regression of Principal Components on Voice Quality Control Vector

The target mean vectors of each pre-stored target speaker are efficiently represented as principal components by using eigenvectors. The proposed method A performs a regression of principal components on the voice quality control vector.

Principal components for the s^{th} target speaker $\boldsymbol{p}^{(s)}$ modeled by the following linear equation,

$$\begin{aligned}\boldsymbol{p}^{(s)} &\simeq \boldsymbol{R}\boldsymbol{w}_e^{(s)} + \boldsymbol{r} \\ &= \left[\begin{array}{cc} \boldsymbol{r} & \boldsymbol{R} \end{array}\right] \left[\begin{array}{c} 1 \\ \boldsymbol{w}_e^{(s)} \end{array}\right] \\ &= \boldsymbol{R}'\boldsymbol{w}_e^{(s)\prime}, \end{aligned} \tag{7}$$

where $\boldsymbol{R}$ is a regression matrix and $\boldsymbol{r}$ is a bias vector. $\boldsymbol{w}_e^{(s)}$ is the voice quality control vector for the s^{th} target speaker. In order to estimate the matrix $\boldsymbol{R}'$, we minimize the following error function:

$$\varepsilon_A^2 = \sum_{s=1}^{S} \left(\boldsymbol{p}^{(s)} - \boldsymbol{R}'\boldsymbol{w}_e^{(s)\prime}\right)^\top \left(\boldsymbol{p}^{(s)} - \boldsymbol{R}'\boldsymbol{w}_e^{(s)\prime}\right). \tag{8}$$

The LS estimate of $\boldsymbol{R}'$ is given by

$$\hat{\boldsymbol{R}}' = \boldsymbol{P}\boldsymbol{W}_e{}'^\top \left(\boldsymbol{W}_e'\boldsymbol{W}_e{}'^\top\right)^{-1}, \tag{9}$$

where

$$\begin{aligned}\boldsymbol{P} &= \left[\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)}, \ \cdots, \ \boldsymbol{p}^{(S)}\right], \\ \boldsymbol{W}_e' &= \left[\boldsymbol{w}_e{}^{(1)\prime}, \boldsymbol{w}_e{}^{(2)\prime}, \ \cdots, \ \boldsymbol{w}_e{}^{(S)\prime}\right].\end{aligned}$$

Therefore, using the obtained regression matrix and the bias vector, the regression parameters in Eq. (6) are written as

$$\begin{aligned}\hat{\boldsymbol{B}}_i^{(Y)} &= \boldsymbol{B}_i^{(Y)}\hat{\boldsymbol{R}}, \\ \hat{\boldsymbol{b}}_i^{(Y)}(0) &= \boldsymbol{B}_i^{(Y)}\hat{\boldsymbol{r}} + \boldsymbol{b}_i^{(Y)}(0).\end{aligned} \tag{10}$$

3.2. Proposed Method B: Regression of Target Mean Vectors on Voice Quality Control Vector

Voice characteristicsto be controlled might not be properly represented as a linear combination of eigenvectors. If so, it is necessary to change the eigenvectors themselves. The proposed method B performs a regression of the target mean vectors on the voice quality control vector.

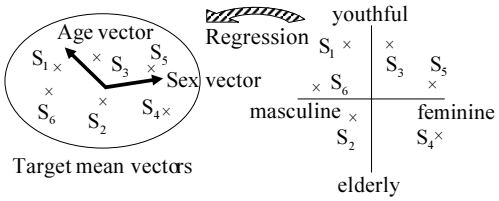

Figure 3: *Proposed method B.*

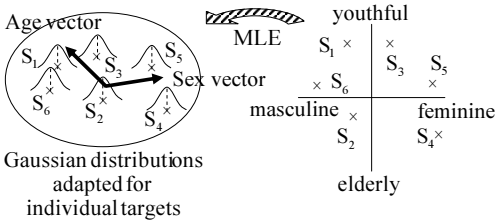

Figure 4: *Proposed method C.*

The target mean vector for the s^{th} target speaker $\boldsymbol{\mu}^{(Y)}(s)$ is modeled by

$$\begin{aligned}\boldsymbol{\mu}^{(Y)}(s) &\simeq \boldsymbol{B}^{(Y)}\boldsymbol{w}_e^{(s)} + \boldsymbol{b}^{(Y)}(0) \\ &= \left[\begin{array}{cc} \boldsymbol{b}^{(Y)} & \boldsymbol{B}^{(Y)} \end{array}\right] \left[\begin{array}{c} 1 \\ \boldsymbol{w}_e^{(s)} \end{array}\right] \\ &= \boldsymbol{B}^{(Y)\prime}\boldsymbol{w}_e^{(s)\prime}, \end{aligned} \tag{11}$$

In order to estimate the matrix $\boldsymbol{B}^{(Y)\prime}$, we minimize the following error function:

$$\varepsilon_B^2 = \sum_{s=1}^{S} \left(\boldsymbol{\mu}^{(Y)}(s) - \boldsymbol{B}^{(Y)\prime}\boldsymbol{w}_e^{(s)\prime}\right)^\top \left(\boldsymbol{\mu}^{(Y)}(s) - \boldsymbol{B}^{(Y)\prime}\boldsymbol{w}_e^{(s)\prime}\right). \tag{12}$$

The LS estimate of $\boldsymbol{B}^{(Y)\prime}$ is given by

$$\hat{\boldsymbol{B}}^{(Y)\prime} = \boldsymbol{\mu}^{(Y)}\boldsymbol{W}_e{}'^\top \left(\boldsymbol{W}_e'\boldsymbol{W}_e{}'^\top\right)^{-1}, \tag{13}$$

where

$$\boldsymbol{\mu}^{(Y)} = \left[\boldsymbol{\mu}^{(Y)}(1), \boldsymbol{\mu}^{(Y)}(2), \ \cdots, \ \boldsymbol{\mu}^{(Y)}(S)\right].$$

3.3. Proposed method C: MLE of EV-GMM Parameters

The desired voice quality might not always be realized by the methods mentioned above because voice quality of the converted speech is affected not only by the target mean vectors but also the other EV-GMM parameters. In order to realize more precise voice quality control, the proposed method C optimizes all of the EV-GMM parameters in the sense of ML under the condition that the weight vector is set to $\boldsymbol{w}_e^{(s)}$. This process is considered to be speaker adaptive training (SAT) [14] [15]. Most parameters of the EV-GMM in the previous methods are affected by acoustic variations of the pre-stored target speakers because they are from the target independent GMM. SAT reduces those variations by training the EV-GMM while considering the adaptation process.

The EV-GMM is trained by maximizing the likelihood of the adapted models for individual pre-stored target speakers as follows:

$$\hat{\boldsymbol{\lambda}}^{(EV)} = \arg\max_{\boldsymbol{\lambda}} \prod_{s=1}^{S}\prod_{t=1}^{T_s} P\left(\boldsymbol{Z}_t^{(s)} | \, \boldsymbol{\lambda}^{(EV)}, \, \boldsymbol{w}_e^{(s)}\right), \tag{14}$$

where the voice quality control vector $\boldsymbol{w}_e^{(s)}$ is employed in the adapted model for the s^{th} pre-stored target speaker. In order to estimate the EV-GMM parameters including the regression parameters, we maximize the following auxiliary function with the EM algorithm,

$$Q\left(\boldsymbol{\lambda}^{(EV)},\, \hat{\boldsymbol{\lambda}}^{(EV)}\right) = \sum_{s=1}^{S}\sum_{i=1}^{M} \bar{\gamma}_i^{(s)} \log P\left(\boldsymbol{Z}_t^{(s)},\, m_i|\, \hat{\boldsymbol{\lambda}}^{(EV)},\, \boldsymbol{w}_e^{(s)}\right), \quad (15)$$

where

$$\bar{\gamma}_i^{(s)} = \sum_{t=1}^{T_s} P\left(m_i|\, \boldsymbol{Z}_t^{(s)},\, \boldsymbol{\lambda}^{(EV)},\, \boldsymbol{w}_e^{(s)}\right). \quad (16)$$

Because it is difficult to estimate all parameters simultaneously, we estimate them at the following order,

$$\begin{aligned} &Q\left(\boldsymbol{\lambda}^{(EV)},\, \boldsymbol{\lambda}^{(EV)}\right) \\ &\leq Q\left(\boldsymbol{\lambda}^{(EV)},\, (\hat{\boldsymbol{B}}_i^{(Y)},\, \boldsymbol{b}_i^{(Y)}(0),\, \boldsymbol{\mu}_i^{(X)},\, \alpha_i,\, \boldsymbol{\Sigma}_i^{(zz)})\right) \\ &\leq Q\left(\boldsymbol{\lambda}^{(EV)},\, (\hat{\boldsymbol{B}}_i^{(Y)},\, \hat{\boldsymbol{b}}_i^{(Y)}(0),\, \hat{\boldsymbol{\mu}}_i^{(X)},\, \hat{\alpha}_i,\, \hat{\boldsymbol{\Sigma}}_i^{(zz)})\right), \end{aligned}$$

ML estimates of those parameters are written as

$$\begin{aligned} \hat{\boldsymbol{v}}_i &= \left(\sum_{s=1}^{S} \bar{\gamma}_i^{(s)} \boldsymbol{W}_s^\top \boldsymbol{\Sigma}_i^{(ZZ)^{-1}} \boldsymbol{W}_s\right)^{-1} \\ &\quad \times \left(\sum_{s=1}^{S} \boldsymbol{W}_s^\top \boldsymbol{\Sigma}_i^{(ZZ)^{-1}} \bar{\boldsymbol{Z}}_i^{(s)}\right), \end{aligned} \quad (17)$$

$$\hat{\alpha}_i = \frac{\sum_{s=1}^{S} \bar{\gamma}_i^{(s)}}{\sum_{i=1}^{M}\sum_{s=1}^{S} \bar{\gamma}_i^{(s)}}, \quad (18)$$

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_i^{(ZZ)} &= \frac{1}{\sum_{s=1}^{S} \bar{\gamma}_i^{(s)}} \sum_{s=1}^{S} \Big\{ \bar{\boldsymbol{V}}_i^{(s)} + \bar{\gamma}_i^{(s)} \hat{\boldsymbol{\mu}}_i^{(s)} \hat{\boldsymbol{\mu}}_i^{(s)^\top} \\ &\quad - \left(\hat{\boldsymbol{\mu}}_i^{(s)} \bar{\boldsymbol{Z}}_i^{(s)^\top} + \bar{\boldsymbol{Z}}_i^{(s)} \hat{\boldsymbol{\mu}}_i^{(s)^\top}\right) \Big\}, \end{aligned} \quad (19)$$

where

$$\bar{\boldsymbol{Z}}_i^{(s)} = \begin{bmatrix} \bar{\boldsymbol{X}}_i^{(s)} \\ \bar{\boldsymbol{Y}}_i^{(s)} \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{T_s} P\left(m_i|\, \boldsymbol{Z}_t^{(s)},\, \boldsymbol{\lambda}^{(EV)},\, \boldsymbol{w}_e^{(s)}\right) \boldsymbol{X}_t^{(s)} \\ \sum_{t=1}^{T_s} P\left(m_i|\, \boldsymbol{Z}_t^{(s)},\, \boldsymbol{\lambda}^{(EV)},\, \boldsymbol{w}_e^{(s)}\right) \boldsymbol{Y}_t^{(s)} \end{bmatrix},$$

$$\boldsymbol{\Sigma}_i^{(ZZ)^{-1}} = \begin{bmatrix} \boldsymbol{P}_i^{(XX)} & \boldsymbol{P}_i^{(XY)} \\ \boldsymbol{P}_i^{(YX)} & \boldsymbol{P}_i^{(YY)} \end{bmatrix},$$

$$\bar{\boldsymbol{V}}_i^{(s)} = \sum_{t=1}^{T_s} P\left(m_i|\, \boldsymbol{Z}_t^{(s)},\, \boldsymbol{\lambda}^{(EV)},\, \boldsymbol{w}_e^{(s)}\right) \boldsymbol{Z}_t^{(s)} \boldsymbol{Z}_t^{(s)^\top},$$

$$\hat{\boldsymbol{\mu}}_i^{(s)} = \boldsymbol{W}_s \hat{\boldsymbol{v}}_i = \begin{bmatrix} \hat{\boldsymbol{\mu}}_i^{(X)} \\ \hat{\boldsymbol{B}}_i^{(Y)} \boldsymbol{w}_e^{(s)} + \hat{\boldsymbol{b}}_i^{(Y)}(0) \end{bmatrix},$$

$$\hat{\boldsymbol{v}}_i = \left[\hat{\boldsymbol{\mu}}_i^{(X)^\top},\, \hat{\boldsymbol{b}}_i^{(Y)}(0)^\top,\, \hat{\boldsymbol{b}}_i^{(Y)}(1)^\top,\, \cdots,\, \hat{\boldsymbol{b}}_i^{(Y)}(K)^\top\right]^\top,$$

$$\boldsymbol{W}_s = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} & w_e^{(s)}(1)\boldsymbol{I} & w_e^{(s)}(2)\boldsymbol{I} & \cdots & w_e^{(s)}(K)\boldsymbol{I} \end{bmatrix},$$

and the matrix $\boldsymbol{I}$ is the $D \times D$ unit matrix. This paper employs the target independent GMM $\boldsymbol{\lambda}^{(0)}$ in Eq. (2) for calculating occupancies $\bar{\gamma}_i^{(s)}$ at the first E-step.

4. Experimental Verifications

4.1. Experimental Conditions

We used 30 speakers, 15 male and 15 female, as the pre-stored speakers. These speakers were included in the Japanese Newspaper Article Sentences (JNAS) database [16]. Each of them uttered a set of phonetically balanced 50 sentences. We used a female speaker not included in JNAS as the source speaker, who uttered the same sentence sets as uttered by the pre-stored speakers.

As the voice quality control vector, we used a 7-scaled categorical score (-3: very, -2: quite, -1: somewhat, 0: no preference, 1: somewhat, 2: quite, 3: very) for 5 Japanese word pairs expressing voice quality (masculine/feminine, hoarse/clear, elderly/youthful, thin/deep, and lax/tense), which were in the major expression word pairs extracted by Kido et al. [17, 18]. One Japanese female subject assigned these scores to each of the pre-stored target speakers by listening to natural speech samples of various sentences uttered by each of them. Scores for each word pair were normalized into the Z-score (zero mean and unit variance).

The STRAIGHT analysis method [19] was employed for the spectral extraction. The first through 24^{th} mel-cepstral coefficients into which the extracted STRAIGHT spectrum were converted were used as the spectral parameter. The shift length was set to 5 ms. Sampling frequency was 16 kHz.

First, we trained the EV-GMM as described in **Section 2.2**. And then, its parameters were further updated with each proposed method. In the proposed method A, all 29 eigenvectors were employed with no loss of information. The number of mixtures of the EV-GMM was set to 128.

4.2. Objective Verification

In order to validate whether the resulting EV-GMM appropriately models a correspondence between the voice quality control vector and voice quality of the converted speech, we calculated the Euclidian distance between the manually assigned scores for each of pre-stored target speakers and the estimated ones, so that voice quality of the converted speech was similar to that of each pre-stored target speaker. We also calculated the correlation coefficient between those two kinds of scores. Since it was difficult to determine manually the best score settings, they were approximately determined with maximum likelihood eigen-decomposition(MLED) [9] for the target adaptation data in the same manner as described in [7]. We used 2 sentences for each pre-stored target speaker as the adaptation data in the score determination.

Figure 5 and **Figure 6** shows the Euclidean distances and the correlation coefficients between the manually assigned scores and the estimated scores. Moreover, an example of the assigned and the estimated scores on sex and hoarseness is shown in **Figure 7**. As a reference, each figure also shows the results of the reassigned scores, which were assigned by the same subject a second time on a different day.

We can see that the proposed method A doesn't work at all. These results show that the relationship between the voice

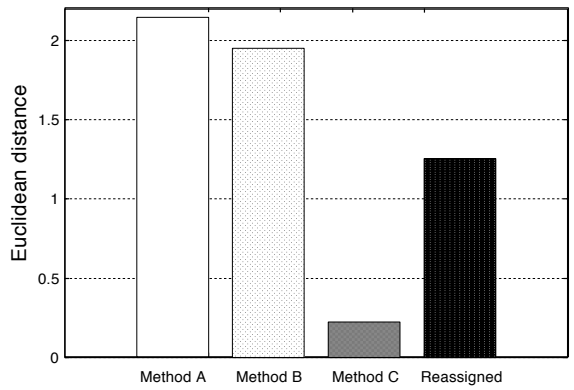

Figure 5: *Euclidean distances between the manually assigned scores and the estimated scores. We show averaged distance over 30 pre-stored target speakers.*

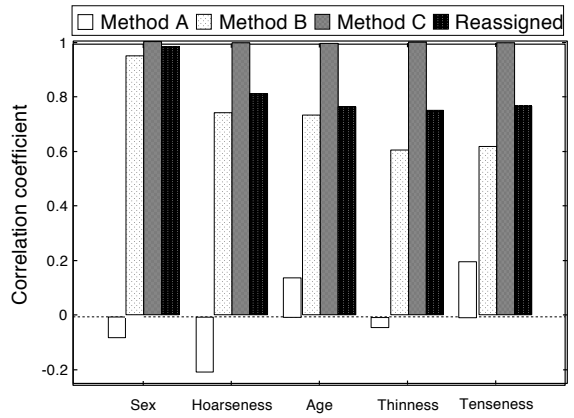

Figure 6: *Correlation coefficients between the manually assigned scores and the estimated scores.*

quality control vector and principal components is difficult to model as a linear conversion.

The proposed method B yields slightly lower distance and much better correlation coefficients than the proposed method A. It is necessary to estimate the regression matrix by directly modeling the relationship between the target mean vectors and the voice quality control vector instead of by using eigenvectors for designing the desired voice quality control. Although the proposed method B works much better than the proposed method A, the score distance is still large. Moreover, the estimated scores are quite different from the assigned ones for several speakers, as shown in **Figure 7**. These degradations are caused by a fact that the training criterion (LS) doesn't correspond to the conversion criterion (ML) and the trained parameters of the EV-GMM are limited to only regression parameters.

The proposed method C causes the best results. This is reasonable because the training criterion corresponds to the conversion criterion and every parameter of the EV-GMM is optimized so that the assigned scores capture the voice quality of the pre-stored target speakers as accurately as possible. However, we can observe from the results of the reassigned scores that even human judgment is not so consistent in scoring. These results imply that it is not always necessary to realize such strict score-consistency as found in the proposed method C.

4.3. Subjective Verification

To compare proposed method B with C, we conducted a preference test on the speech quality of the converted voices. Average voices were used as stimuli. Average voices were converted voices from the source speaker's voices when setting every component of the voice quality control vector to zero. Because the resulting bias vectors were almost the same in those methods, average voices produced by individual EV-GMMs had

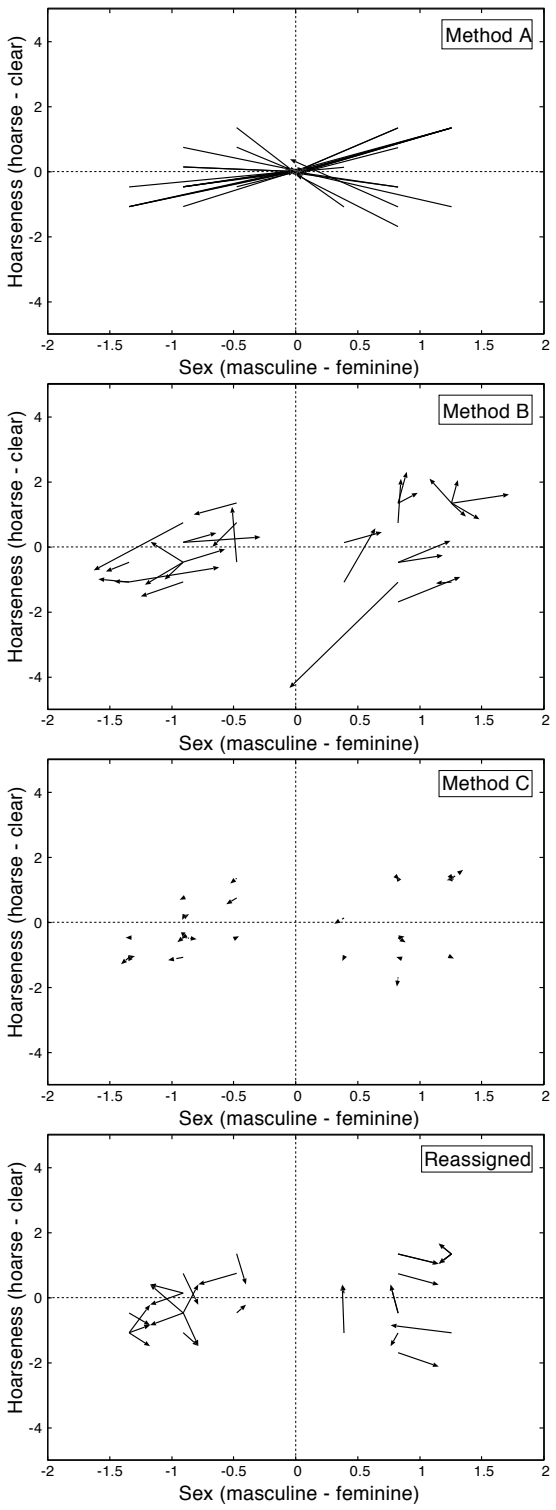

Figure 7: *An example of the manually assigned scores and the estimated scores for sex and hoarseness. The starting point of each arrow shows the manually assigned score, and its ending point shows the estimated score for each of pre-stored target speakers.*

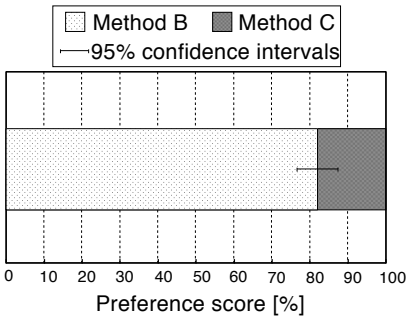

Figure 8: *Result of subjective evaluation.*

very similar speaker individuality. As for F0 conversion, a simple linear conversion based on mean values and standard deviations of log-scaled F0 was employed for converting the source to the average voice. In the preference test in those two proposed methods, we randomly presented a pair of the average voices produced by the EV-GMMs. The subjects were asked which sample sounded more natural. The 50 utterances not included in the training data were evaluated. The number of subjects was 5.

The result of the preference test is shown in **Figure 8**. It is observed that method B outperforms method C. The converted speech in method C sometimes has unstable sound quality. As mentioned above, method C causes the EV-GMM modeling the correspondence between the voice quality control vector and the pre-stored target voice quality as precisely as possible. It is possible that such strict modeling causes large projection errors on the high-dimensional acoustic space, especially if the low-dimensional space represented by the voice quality control vectors covers only a very limited sub-space. Those errors directly affect the estimation of the EV-GMM parameters. We have to cope with this problem in order to realize both high-quality and high-controllability of the EVC-based voice quality controller. It is also possible that the trained parameters in method C converge to local optima due to using inappropriate initial model, i.e., the target independent GMM in this paper.

5. Conclusions

We proposed regression approaches to the voice quality controller based on one-to-many eigenvoice conversion (EVC). First, the voice quality control vector was defined and proper component values of the vector were manually assigned to each of the pre-stored target speakers. And then, the eigenvoice Gaussian mixture model (EV-GMM) was trained so that voice characteristics of the pre-stored target speakers were properly represented by the voice quality control vectors. We conducted experimental verifications for showing advantages and disadvantages of each of the proposed methods.

6. Acknowledgements

This research was supported in part by MEXT's (the Japanese Ministry of Education, Culture, Sports, Science and Technology) e-Society project.

7. 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] M. Abe, "Speech morphing by gradually changing spectrum parameter and fundamental frequency," in *Proc. ICSLP 96*, Oct. 1996, Vol. 4, pp. 2235-2238.

[3] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino, and H. Kawahara, "Implementation of realtime STRAIGHT speech manipulation system: Report on its implementation," *Acoust. Sci. & Tech.*, Vol. 28, No. 3, pp. 140-146, 2007.

[4] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. Jpn. (E)*, Vol. 11, No. 2, pp. 71-76, 1990.

[5] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 2, pp. 131-142, 1998.

[6] T. Toda, A.W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Proc. ICASSP 2005*, Mar. 2005, Vol. 1, pp. 9-12.

[7] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," in *Proc. INTERSPEECH2006-ICSLP*, Sep. 2006, pp. 2446-2449.

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

[9] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech and Audio Processing*, Vol. 8, No. 6, pp. 695-707, 2000.

[10] P. Kenny, G. Boulianne, and P. Dumouchel, "Maximum likelihood estimation of eigenvoices and residual variances for large vocabulary speech recognition tasks," in *Proc. ICSLP 2002*, Sep. 2002, pp. 57-60.

[11] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *Proc. ICSLP 2002*, Sep. 2002, pp. 1269-1272.

[12] K. Miyanaga, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based speech synthesis," in *Proc. INTERSPEECH2004-ICSLP*, Oct. 2004, pp. 1437-1440.

[13] M. Tachibana, T. Nose, J. Yamagishi, and T. Kobayashi, "A technique for controlling voice quality of synthetic speech using multiple regression HSMM," in *Proc. INTERSPEECH2006-ICSLP*, Sep. 2006, pp. 2438-2441.

[14] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP 96*, Oct. 1996, Vol. 2, pp. 1137-1140.

[15] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Speaker adaptive training for voice conversion based on eigenvoice," *IEICE Tech. Rep.*, SP2006-40, pp. 31-36, 2006 [in Japanese].

[16] JNAS: Japanese Newspaper Article Sentences. http://www.mibel.cs.tsukuba.ac.jp/jnas/instruct.html

[17] H. Kido, and H. Kasuya, "Extraction of everyday expression associated with voice quality of normal utterance," *J. Acoust. Soc. Jpn.*, Vol. 55, No. 6, pp. 405-411, 1999 [in Japanese].

[18] H. Kido, and H. Kasuya, "Everyday expressions associated with voice quality of normal utterance —Extraction by perceptual evaluation—," *J. Acoust. Soc. Jpn.*, Vol. 57, No. 5, pp. 337-344, 2001 [in Japanese].

[19] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, Vol. 27, No. 3-4, pp. 187-207, 1999.