One-to-Many Voice Conversion Based on Tensor Representation of Speaker Space.

Conference Paper · January 2011

Source: DBLP

CITATIONS
30

READS
105

4 authors, including:

Daisuke Saito
The University of Tokyo
52 PUBLICATIONS **603** CITATIONS
SEE PROFILE

Keikichi Hirose
The University of Tokyo
430 PUBLICATIONS **2,626** CITATIONS
SEE PROFILE

One-to-Many Voice Conversion Based on Tensor Representation of Speaker Space

Daisuke Saito, Keisuke Yamamoto, Nobuaki Minematsu, Keikichi Hirose

Graduate School of Information Science and Technology, The University of Tokyo, Japan
{dsk_saito,yama,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract

This paper describes a novel approach to flexible control of speaker characteristics using tensor representation of speaker space. In voice conversion studies, realization of conversion from/to an arbitrary speaker's voice is one of the important objectives. For this purpose, eigenvoice conversion (EVC) based on an eigenvoice Gaussian mixture model (EV-GMM) was proposed. In the EVC, similarly to speaker recognition approaches, a speaker space is constructed based on GMM supervectors which are high-dimensional vectors derived by concatenating the mean vectors of each of the speaker GMMs. In the speaker space, each speaker is represented by a small number of weight parameters of eigen-supervectors. In this paper, we revisit construction of the speaker space by introducing the tensor analysis of training data set. In our approach, each speaker is represented as a matrix of which the row and the column respectively correspond to the Gaussian component and the dimension of the mean vector, and the speaker space is derived by the tensor analysis of the set of the matrices. Our approach can solve an inherent problem of supervector representation, and it improves the performance of voice conversion. Experimental results of one-to-many voice conversion demonstrate the effectiveness of the proposed approach.

Index Terms: voice conversion, Gaussian mixture model, eigenvoice, tensor analysis, Tucker decomposition

1. Introduction

Voice conversion (VC) is a technique to transform an input utterance of a speaker to another utterance that sounds like another speaker with its linguistic content preserved [1]. VC can be regarded as a framework of modification between two feature spaces, not limited to speaker spaces. Hence VC techniques can apply to various applications, including the modification of speaker identity in Text-to-Speech (TTS) systems [2], speech enhancement [3], hand motion to speech conversion [4], and so on. Statistical approaches have often been used for implementing the conversion from source features to target ones [1, 2, 5, 6]. Among these, GMM-based approaches are widely used in particular because of their flexibility.

To construct the conversion model, however, these methods require a training corpus, which contains plenty of utterances with the same linguistic content from both the source and target speakers. In addition, application of the conversion model is limited to this specific pair of speakers. Namely, flexible control of speaker characteristics for VC framework is an important objective. For this purpose, it is effective to utilize voices of other speakers as prior knowledge. There have been several proposed approaches which do not require a large parallel corpus but use other non-parallel data. Mouchtaris *et al.* proposed an unsupervised training method based on maximum likelihood constrained adaptation of the GMM trained with an existing parallel data set of a different speaker pair [7]. Lee *et al.* proposed another approach based on maximum a posteriori (MAP) adaptation [8]. They are inspired by speaker adaptation techniques in speech recognition studies. To use prior knowledge from many other speakers more effectively, Toda *et al.* proposed eigenvoice conversion (EVC) based on the eigenvoice technique in speech recognition [10]. In the EVC, eigenvoice GMM (EV-GMM) is trained with multiple parallel data sets consisting of utterance pairs of a single speaker, which is called the reference speaker henceforth, and many pre-stored speakers. Based on joint density models of the reference and the pre-stored speakers, the speaker GMMs of the pre-stored speakers can be extracted, and a speaker space is constructed based on GMM supervectors which are high-dimensional vectors derived by concatenating all the mean vectors of each of the speaker GMMs. Similarly to speaker recognition studies [11], an arbitrary speaker is represented as a vector of this speaker space. Hence the joint density GMM of the reference and the target speaker is flexibly developed by estimating a small number of weight parameters for the bases of the space.

However, the representation of GMM supervector has an inherent problem that multiple factors of acoustic variations are included in the same space. Namely, Gaussian component of GMM and the dimension of the mean vector are treated interdependently, and the speaker space becomes a high-dimensional vector space. In this paper, to represent the speaker space for the VC framework as more tractable, we propose one-to-many voice conversion based on tensor representation of the speaker space. In our approach, an arbitrary speaker is not represented as a supervector, but a matrix whose row and column respectively correspond to the component of GMM and the dimension of the mean vector. Using this representation, we express the data set of the pre-stored speakers as a third-order tensor, and introduce the tensor analysis to obtain the speaker space. Since the tensor analysis can treat multiple factors of variations properly [12], it will be expected to improve the performance of VC. Although we tackle the task of one-to-many VC in this paper, our proposed method can also apply to many-to-one VC, or tasks of speaker recognition. Because our approach mainly focuses on the representation of the speaker space, there still exists the flexibility to integrate our method with other effective methods such as speaker adaptive training for EVC [13] or non-parallel training for many-to-many EVC [14].

The remainder of this paper is organized as follows. Section 2 describes the basic EVC approach. Then, our proposed approach using the tensor representation of the speaker space is described in Section 3. In Section 4, experimental evaluations are described. Finally, Section 5 concludes the paper.

28 – 31 August 2011, Florence, Italy

2. Eigenvoice conversion (EVC)

2.1. Eigenvoice GMM (EV-GMM)

In this section, one-to-many EVC [15] is described. Let $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \boldsymbol{\Delta x}_t^\top]^\top$ and $\boldsymbol{Y}_t^{(s)} = [\boldsymbol{y}_t^{(s)\top}, \boldsymbol{\Delta y}_t^{(s)\top}]^\top$ be $2D$-dimensional vectors of the source speaker and the s-th target speaker, respectively. They consist of D-dimensional static and dynamic features. The notation $(\cdot)^\top$ denotes transposition of a vector. The joint probability density of the source and the target vectors is modeled by an EV-GMM as follows:

$$P(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}^{(s)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^{(s)\top}]^\top; \boldsymbol{\mu}_m^{(Z)}(\boldsymbol{w}^{(s)}), \boldsymbol{\Sigma}_m^{(Z)}), \quad (1)$$

$$\boldsymbol{\mu}_m^{(Z)}(\boldsymbol{w}^{(s)}) = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{B}_m \boldsymbol{w}^{(s)} + \boldsymbol{b}_m^{(0)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}, \quad (2)$$

where $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The weight of the m-th component is denoted by α_m, and the number of mixture components is M. In EV-GMM, when we use the S pre-stored speakers, the target mean vector $\boldsymbol{\mu}_m^{(Y)}$ is represented as a linear combination of the bias vector $\boldsymbol{b}_m^{(0)}$ and the K representative vectors $\boldsymbol{B}_m = \left[\boldsymbol{b}_m^{(1)}, \boldsymbol{b}_m^{(2)}, \ldots, \boldsymbol{b}_m^{(K)}\right]$, where $K < S$. In EV-GMM, the speaker individuality of the target is controlled with the K-dimensional vector $\boldsymbol{w}^{(s)}$. Namely, a speaker space is constructed by K bases of supervectors $\boldsymbol{B} = [\boldsymbol{B}_1^\top, \boldsymbol{B}_2^\top, \ldots, \boldsymbol{B}_m^\top]^\top \in \mathcal{R}^{2DM \times K}$ and the bias supervector $\boldsymbol{b} = \left[\boldsymbol{b}_1^{(0)\top}, \boldsymbol{b}_2^{(0)\top}, \ldots, \boldsymbol{b}_m^{(0)\top}\right]^\top \in \mathcal{R}^{2DM \times 1}$.

2.2. Construction of the speaker space for EVC

When we employ EV-GMM based on principal component analysis (PCA), to construct the speaker space for EVC, first, a target independent joint density GMM (TI-GMM) is trained using all of the multiple parallel data sets simultaneously. Next, each target dependent GMM is trained by updating only the target mean vectors of TI-GMM using each of the corresponding parallel data set. As a feature vector of the speaker space, a supervector for each pre-stored target speaker is constructed by concatenating the mean vectors of the target dependent GMM. The bias vector $\boldsymbol{b}$ and representative vectors $\boldsymbol{B}$ are determined with PCA for all the supervectors of the target speakers.

2.3. Adaptation of EV-GMM

The EV-GMM is adapted for arbitrary speakers by estimating the weight vector $\boldsymbol{w}$ for given their speech samples based on maximum likelihood criterion [9]. Let $\boldsymbol{Y}^{(tar)}$ be a sequence of the target features. $\boldsymbol{w}$ is estimated as follows:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \int P(\boldsymbol{X}, \boldsymbol{Y}^{(tar)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) d\boldsymbol{X}. \quad (3)$$

Using EM-algorithm for the estimation, we can derive the following updating equations for $\hat{\boldsymbol{w}}$:

$$\hat{\boldsymbol{w}} = \left\{ \sum_{m=1}^{M} \overline{\gamma}_m^{(tar)} \boldsymbol{B}_m^\top \boldsymbol{\Sigma}_m^{(YY)^{-1}} \boldsymbol{B}_m \right\}^{-1} \sum_{m=1}^{M} \boldsymbol{B}_m^\top \boldsymbol{\Sigma}_m^{(YY)^{-1}} \overline{\boldsymbol{Y}}_m^{(tar)}, \quad (4)$$

$$\overline{\gamma}_m^{(tar)} = \sum_{t=1}^{T} \gamma_{m,t}, \quad \overline{\boldsymbol{Y}}_m^{(tar)} = \sum_{t=1}^{T} \gamma_{m,t} (\boldsymbol{Y}_t^{(tar)} - \boldsymbol{b}_m^{(0)}), \quad (5)$$

$$\gamma_{m,t} = P(m | \boldsymbol{Y}_t^{(tar)}, \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}). \quad (6)$$

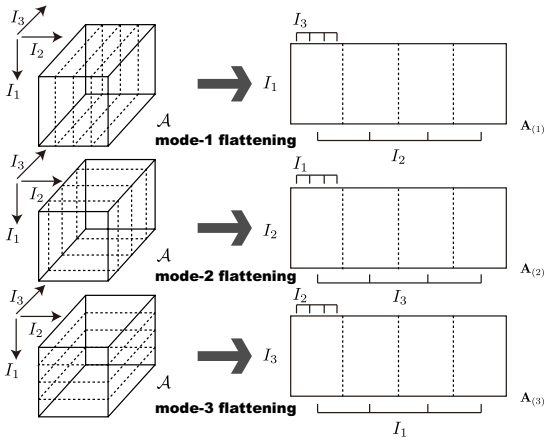

Figure 1: Flattening of the $(I_1 \times I_2 \times I_3)$-tensor $\mathcal{A}$ to the flattened matrices $\boldsymbol{A}_{(1)}$, $\boldsymbol{A}_{(2)}$ and $\boldsymbol{A}_{(3)}$.

Equation 4 approximately means the calculation of the projection weights of the target for each basis of the speaker space. TI-GMM is used for the initialization for Equation 6. After adaptation, the step of parameter generation is the same as [16].

3. Tensor representation of speaker space

3.1. Multilinear algebra

In this section, construction of the speaker space based on the tensor analysis is described. First, we introduce some of the multilinear algebra related to our approach [17]. Tensor is a multidimensional array which generalizes matrix representation. Each dimension in tensor is called "mode." Let $\mathcal{A} \in \mathcal{R}^{I_1 \times I_2 \times I_3}$ be a third-order tensor. Generally, a high-order tensor can be expressed as a matrix using a mode-n flattening, which slices a tensor $\mathcal{A}$ along the mode-n axis and splices the sliced matrices to one matrix $\boldsymbol{A}_{(n)}$ as shown in Figure 1. Using this flattening operation, the product of a tensor and a matrix can be defined. The expression $\mathcal{A} = \mathcal{G} \times_n \boldsymbol{B}$ denotes the mode-n product of a tensor $\mathcal{G}$ with a matrix $\boldsymbol{B}$, and it is performed by using the model-n flattened matrices as $\boldsymbol{A}_{(n)} = \boldsymbol{B} \cdot \boldsymbol{G}_{(n)}$.

One of the most important operations of matrix algebra is Singular Value Decomposition (SVD). Since a matrix can be viewed as a second-order tensor, SVD of matrix $\boldsymbol{A}$ can be represented as the following mode-n products:

$$\boldsymbol{A} = \boldsymbol{U} \boldsymbol{S} \boldsymbol{V}^\top = \mathcal{S} \times_1 \boldsymbol{U} \times_2 \boldsymbol{V}. \quad (7)$$

Expanding SVD in the case of second-order tensors to that of high-order ones, we can derive the following decomposition:

$$\mathcal{A} = \mathcal{S} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3. \quad (8)$$

When $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, and $\boldsymbol{U}_3$ are orthogonal and the tensor $\mathcal{S}$ is dense, i.e. not diagonal as the case of second-order, the decomposition of Equation 8 is called high-order SVD, or Tucker decomposition [17, 18]. Since PCA can be regarded as SVD of a data *matrix*, the construction of the space can also be expanded by Tucker decomposition when we introduce a data *tensor*.

3.2. Proposed construction of the speaker space

To construct the speaker space based on Tucker decomposition, each speaker in the pre-stored data sets is expressed as an $M \times D'$ matrix [19], where M is the number of mixtures, and $D' = 2D$. First, the bias matrix $\boldsymbol{b}' = \left[\boldsymbol{b}_1^{(0)}, \boldsymbol{b}_2^{(0)}, \ldots, \boldsymbol{b}_m^{(0)}\right]^\top$ is

subtracted from each speaker matrix in advance. When we have the S pre-stored speakers, the training data sets are represented as the tensor $\mathcal{M} \in \mathcal{R}^{M \times D' \times S}$. Then, $\mathcal{M}$ can be represented as follows:

$$\mathcal{M} = \mathcal{G}^{M \times D' \times S} \times_1 \boldsymbol{U}^{(M)} \times_2 \boldsymbol{U}^{(D')} \times_3 \boldsymbol{U}^{(S)}, \quad (9)$$

where $\boldsymbol{U}^{(M)} \in \mathcal{R}^{M \times M}$, $\boldsymbol{U}^{(D')} \in \mathcal{R}^{D' \times D'}$, and $\boldsymbol{U}^{(S)} \in \mathcal{R}^{S \times S}$. These matrices separately capture the effects from GMM components, dimensions of the mean vector, and speaker indices, respectively, and the tensor $\mathcal{G}$ puts them together. Fixing the index of the third mode, we obtain the matrix representing the speaker n as

$$\boldsymbol{\mu}^{(n)} = \mathcal{G} \times_1 \boldsymbol{U}^{(M)} \times_2 \boldsymbol{U}^{(D')} \times_3 \boldsymbol{U}^{(S)}(n,:). \quad (10)$$

Although there are several candidates for bases of the speaker space, in this paper, as similar to [19], Equation 10 is grouped as follows:

$$\boldsymbol{\mu}^{(n)} = \boldsymbol{U}^{(M)} \left\{ \mathcal{G} \times_2 \boldsymbol{U}^{(D')} \times_3 \boldsymbol{U}^{(S)}(n,:) \right\}^\top = \boldsymbol{U}^{(M)} \boldsymbol{W}_n^\top, \quad (11)$$

where $\boldsymbol{U}^{(M)}$ becomes the bases, and $\boldsymbol{W}_n \in \mathcal{R}^{D' \times M}$ is a weight matrix. Using the truncated bases, consequently, we obtain the matrix for a new speaker as

$$\boldsymbol{\mu}^{(new)} = \boldsymbol{U}^{(M)} \boldsymbol{W}_{(new)}^\top + \boldsymbol{b}', \quad (12)$$

where $\boldsymbol{U}^{(M)} \in \mathcal{R}^{M \times K} (K \leq S)$ and $\boldsymbol{W}_{(new)} \in \mathcal{R}^{D' \times K}$ are a representative matrix and a weight one, respectively. Hence, in our proposed method, parameters to be estimated become a $D' \times K$ matrix, while they become a K-dimensional vector in the conventional EVC.

In [19], the equation for adaptation is derived based on minimum mean square error. On the other hand, in this paper, for adaptation data $\boldsymbol{Y}^{(tar)}$, we derive the following updating equations based on maximum likelihood criterion:

$$\mathrm{vec}(\boldsymbol{W}) = \left[\sum_{m=1}^{M} \overline{\gamma}_m^{(tar)} \boldsymbol{U}_m^\top \boldsymbol{U}_m \otimes \boldsymbol{\Sigma}_m^{(YY)^{-1}} \right]^{-1} \mathrm{vec}(\boldsymbol{C}), \quad (13)$$

$$\boldsymbol{C} = \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{\Sigma}_m^{(YY)^{-1}} (\boldsymbol{Y}_t^{(tar)} - \boldsymbol{b}_m^{(0)}) \boldsymbol{U}_m, \quad (14)$$

$$\boldsymbol{U}_m = \boldsymbol{U}^{(M)}(m,:) \in \mathcal{R}^{1 \times K}, \quad (15)$$

where vec() is the vec-operator that stacks the columns of a matrix into a vector. Compared with Equation 4, Equation 13 has a similar form, but it estimates $D' \times K$ parameters rather than K parameters in Equation 4. This means that our proposed method might be more flexible to adapt for the data. We verify it by the experiments.

4. Experimental evaluation

4.1. Experimental conditions

To evaluate the performance of our proposed method, one-to-many voice conversion experiments were carried out. We used one male speaker as the reference speaker from ATR Japanese speech database B-set [20], and 273 pre-stored speakers including 137 male and 136 female speakers [21]. 50 sentences were uttered from each speaker. In the evaluation, we selected new 6 speakers of 3 male and 3 female speakers. We used 1 to 16 utterances for adaptation, and other 21 utterances for evaluation.

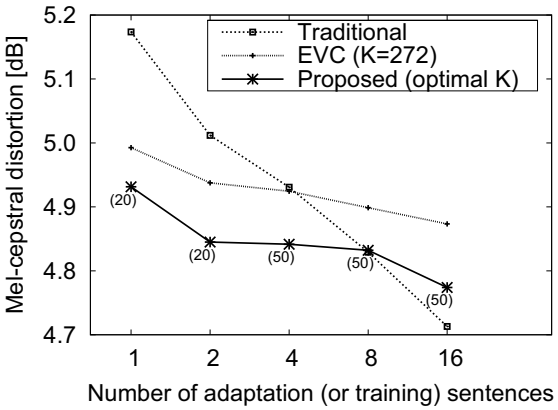

Figure 2: Results of objective evaluations by mel-cepstral distortion (MCD). The numbers in parentheses mean the optimal number of K in each condition in the sense of the MCD.

We used 24-dimensional mel-cepstrum vectors for spectrum representation (D=24). These are derived by STRAIGHT analysis [22]. The number of mixture components (M) was fixed to 128. Aperiodic components, which are features to construct STRAIGHT mixed excitation, are not converted in this study, and they are fixed to -30 dB at all frequencies. Prosodic features, the power coefficient and the fundamental frequency were converted in a simple manner that only considers the mean and the standard deviation of the parameters.

We compared the proposed one-to-many VC algorithms (Proposed) and the one-to-many EVC (EVC) with traditional VC with the parallel training (Traditional) [16]. We note that speaker adaptive training is not applied in both the proposed and the EVC method.

4.2. Objective evaluations

We evaluated the conversion performance using mel-cepstral distortion between the converted vectors and the vectors of the targets. Figure 2 shows the result of average mel-cepstral for the test data as a function of the number of adaptation, or training sentences. In "Traditional," for each case, the optimal number of mixture components is selected. Both the proposed method and the EVC significantly outperform "Traditional" when using a small amount of adaptation data less than 8. This means that prior knowledge from the pre-stored data set is effectively utilized for improvement of the performance. Compared with "EVC", the performance of the proposed method is better. This means that our proposed representation of the speaker space works well rather than supervector representation of the speaker space. In this experiment, the size of representative matrix in the proposed method (K) was optimally determined in each number of adaptation sentences in the sense of the mel-cepstral distortion. When the number of adaptation sentences was small, K is also small. $K = 272$ was optimal in EVC, i.e. all of the representative vectors are used. On the other hand, in our proposed method, the size of the representative matrix was effectively reduced, since the full size of the representative matrix is $K = 127$ when $M = 128$. It might be said that our proposed approach effectively captures the essence of the speaker space.

4.3. Subjective evaluations

A listening test was carried out to evaluate the naturalness of converted speech and conversion accuracy for speaker individ-

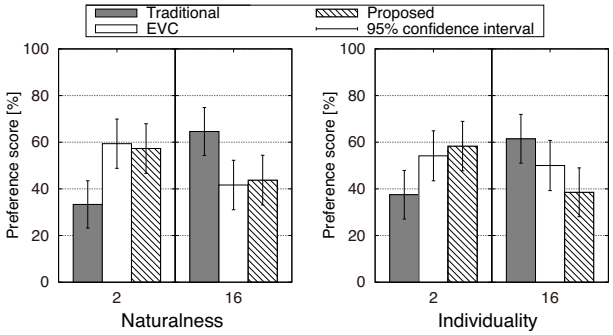

Figure 3: Results of subjective evaluations. The number in x axis is the number of adaptation (or training) sentences.

uality. The test was conducted with 8 subjects. To evaluate the naturalness, a paired comparison was carried out. In this test, pairs of two different types of the converted samples were presented to subjects, and then each subject judged which sample sounded better. To evaluate conversion accuracy, an RAB test was performed, where pairs of two different types of the samples were presented after presenting the reference sample of the target speech. The number of sample pairs evaluated by each subject was 36 in each test.

Figure 3 shows the results. When using two adaptation data, both the "Proposed" and "EVC" outperform "Traditional." When using 16 adaptation data, "Traditional" has the best performance in both naturalness and speaker individuality. Compared with the EVC method, the performance of the proposed method is comparable or slightly better to that of the EVC except in speaker individuality when using 16 adaptation data.

Both the objective and subjective evaluations suggest that our proposed method works effectively. For further improvements, we need to investigate the properties of our proposed method in detail, e.g. other candidates for grouping in Equation 10.

5. Conclusions

We have proposed a new method for speaker adaptation in voice conversion which represents the pre-stored data set as the tensor representation. In our approach, each speaker is represented as a matrix whose row and vector respectively correspond to the Gaussian component and the dimension of the mean vector. The treatment of the data set as the tensor representation enables the conversion framework to model the speaker characteristics more flexibly. For further improvements of the conversion performance, first, integration of our method with other effective methods such as speaker adaptive training or non-parallel training should be verified. We are also planning to investigate other grouping methods in Equation 10. The optimization of the size of the representative matrix, i.e. the optimization of K, is another further work.

6. Acknowledgment

The first author is partly supported by Japan Society of Promotion for Science, and this work was supported by KAKENHI Grant-in-Aid for JSPS Fellows (22-8861).

7. References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," Proc. ICASSP, pp. 655–658, 1988.

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," Proc. ICASSP, vol. 1, pp. 285–288, 1998.

[3] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, "High-performance robust speech recognition using stereo training data," Proc. ICASSP, pp. 301–304, 2001.

[4] A. Kunikoshi, Y. Qiao, N. Minematsu, and K. Hirose, "Speech generation from hand gestures based on space mapping," Proc. INTERSPEECH, pp. 308–311, 2009.

[5] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," Proc. ICASSP, pp. 3893–3896, 2009.

[6] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," IEEE Trans. on Speech and Audio Processing, vol. 6, no. 2, pp. 131–142, 1998.

[7] A. Mouchtaris, J. V. der Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," IEEE Trans. on Audio, Speech, and Language Processing, vol. 14, no. 3, pp. 952–963, 2006.

[8] C. H. Lee and C. H. Wu, "Map-based adaptation for speech conversion using adaptation data selection and non-parallel training," Proc. INTERSPEECH, pp. 2254–2257, 2006.

[9] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," Proc. INTERSPEECH, pp. 2446–2449, 2006.

[10] R. Kuhn, J-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in Eigenvoice space," IEEE Trans. on Speech and Audio Processing, vol. 8, no. 6, pp. 695–707, 2000.

[11] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," IEEE Trans. on Audio, Speech, and Language Processing, vol. 16, no. 5, pp. 980–988, 2008.

[12] M. A. O. Vasilescu and D. Terzopoulos, "Mutilinear analysis of image ensembles: TensorFaces," Proc. ECCV, pp. 447–460, 2002.

[13] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Speaker adaptive training for one-to-many eigenvoice conversion based on Gaussian mixture model," Proc. INTERSPEECH, pp. 1981–1984, 2007.

[14] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Non-parallel training for many-to-many eigenvoice conversion," Proc. ICASSP, pp. 4822–4825, 2010.

[15] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," Proc. ICASSP, vol. IV, pp. 693–696, 2007.

[16] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Trans. on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2222–2235, 2007.

[17] L. De Lathauwer, B. De Moor and J. Vandewalle, "A multilinear singular value decomposition," SIAM Journal on Matrix Analysis and Applications, vol. 21, No. 4, pp. 1253–1278, 2000.

[18] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," Psychometrika, vol. 31, no. 3, pp. 279–311, 1966.

[19] Y. Jeong, "Speaker adaptation based on the multilinear decomposition of training speaker models," Proc. ICASSP, pp. 4870–4873, 2010.

[20] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K.Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Communication, vol.9, pp.357–363, 1990.

[21] "Jnas: Japanese newspaper article sentences," http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html

[22] 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, pp.187–207, 1999.