Effects of Speaker Adaptive Training on Tensor-based Arbitrary Speaker Conversion

Daisuke Saito[1], *Nobuaki Minematsu*[2], *Keikichi Hirose*[1]

[1]Graduate School of Information Science and Technology, The University of Tokyo, Japan
[2]Graduate School of Engineering, The University of Tokyo, Japan
dsaito@hil.t.u-tokyo.ac.jp, {mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract

This paper introduces speaker adaptive training techniques to tensor-based arbitrary speaker conversion. 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), which is based on an eigenvoice Gaussian mixture model (EV-GMM), was proposed. Although the EVC can effectively construct the conversion model for arbitrary target speakers using only a few utterances, increase of the utterances used to construct the conversion model does not always improve the conversion performance. This is because the EV-GMM method has an inherent problem in representation of GMM supervectors. We previously proposed tensor-based speaker space as a solution for this problem, and realized more flexible control of speaker characteristics. In this paper, to aim larger improvement of the performance of VC, speaker adaptive training and tensor-based speaker representation are integrated. The proposed method can construct the flexible and precise conversion model, and experimental results of one-to-many voice conversion demonstrate the effectiveness of the proposed approach.

Index Terms: voice conversion, Gaussian mixture model, eigenvoice, Tucker decomposition, speaker adaptive training

1. Introduction

Voice conversion (VC), or speaker conversion 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 techniques can apply to various applications besides speech synthesis [2, 3]. Among several statistical approaches to construct the conversion model, GMM-based approaches are widely used because of their flexibility [2, 4].

When constructing the conversion model, however, a parallel training corpus, which are a set of utterance pairs of the same sentences spoken by a source and a target speakers, are required. This requirement limits the applicability of the conversion model to the specific speaker pair. Hence, flexible control of speaker characteristics with little need of a parallel corpus is an important objective of VC. For this purpose, several adaptation techniques using voices of other speakers have been proposed [5, 6]. These approaches are inspired by speaker adaptation techniques in speech recognition studies. Among these, eigenvoice conversion (EVC) [6], which uses the eigenvoice technique proposed in speech recognition [7], is implemented by constructing a speaker space. Based on training with multiple pre-stored parallel data sets, a speaker space is constructed utilizing GMM supervector, in a similar manner to speaker recognition studies [8]. Then, adaptation to an arbitrary speaker becomes the problem to locate that speaker in the constructed speaker space. Hence, precise construction of the speaker space is important for improvement of the performance of voice conversion. However, the representation of GMM supervector has an inherent problem that multiple factors of acoustic variations are included in the same space. Hence, scalability of adaptation performance of EVC is limited caused by the problem.

We have recently proposed a new representation of speaker space based on tensor analysis for arbitrary speaker conversion [9]. In our approach, an arbitrary speaker is not represented as a supervector, but as a matrix whose row and column respectively correspond to the component of GMM and the dimension of the mean vector. Using this representation, we can express the data set of the pre-stored speakers as a third-order tensor, and introduce the tensor analysis to obtain the speaker space. Based on this speaker space, Tensor-based Arbitrary Speaker Conversion (TASC) has been realized and the effectiveness of TASC, compared with EVC, was shown by the one-to-many VC task [9].

Because our approach is a new method of representing a speaker space, it can be flexibly integrated with other effective techniques which are independent of speaker space. In this paper, we introduce speaker adaptive training for TASC. Speaker adaptive training was introduced for training a canonical speaker-independent model [10], and its effectiveness in arbitrary speaker conversion was shown in [11]. This paper investigates the effects of speaker adaptive training when it is applied to tensor-based flexible speaker representation.

2. Eigenvoice conversion (EVC)

2.1. Eigenvoice GMM (EV-GMM)

In this section, one-to-many EVC [6] is briefly 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:

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

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

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 as 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 J representative vectors $\boldsymbol{B}_m = \left[\boldsymbol{b}_m^{(1)}, \boldsymbol{b}_m^{(2)}, \ldots, \boldsymbol{b}_m^{(J)}\right]$, where $J < S$. In EV-GMM, the speaker individuality of the target is controlled with the J-dimensional vector $\boldsymbol{w}^{(s)}$. Namely, a speaker space is constructed by J bases of supervectors $\boldsymbol{B} = [\boldsymbol{B}_1^\top, \boldsymbol{B}_2^\top, \ldots, \boldsymbol{B}_M^\top]^\top \in \mathcal{R}^{2DM \times J}$ and the bias supervector $\boldsymbol{b} = \left[\boldsymbol{b}_1^{(0)^\top}, \boldsymbol{b}_2^{(0)^\top}, \ldots, \boldsymbol{b}_M^{(0)^\top}\right]^\top \in \mathcal{R}^{2DM \times 1}$. Construction of the speaker space is realized by principal component analysis (PCA). 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.

2.2. 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 [6]. Let $\boldsymbol{Y}^{(tar)}$ be a sequence of the target features. $\boldsymbol{w}$ is estimated as follows:

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

3. Tensor-based speaker space

In this section, construction of the speaker space based on the tensor analysis is described [9]. In the EVC approach, GMM supervector is a representation of speaker. However, the representation of GMM supervector has an inherent problem that multiple factors of acoustic variations are embedded 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. To solve this problem, 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 Tucker decompostion of the set of the matrices, instead of PCA of the set of the supervectors. Tucker decompostion can be viewed as expansion of SVD [12], and treat multiple factors of variations properly.

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 [13], 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 $\mathcal{M} \in \mathcal{R}^{M \times D' \times S}$. By Tucker decompostion of $\mathcal{M}$, $\boldsymbol{U}^{(M)} \in \mathcal{R}^{M \times M}$, $\boldsymbol{U}^{(D')} \in \mathcal{R}^{D' \times D'}$, and $\boldsymbol{U}^{(S)} \in \mathcal{R}^{S \times S}$ are extracted as bases matrices. Focusing on relation of mixture components, we adopt $\boldsymbol{U}^{(M)}$ for the bases, as similar to [13]. Finally, using the truncated bases, consequently, we obtain the matrix for a new speaker as

$$\boldsymbol{\mu}^{(new)} = \boldsymbol{U}^{(M)} \boldsymbol{W}_{(new)}^\top + \boldsymbol{b}', \qquad (4)$$

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

For adaptation data $\boldsymbol{Y}^{(tar)}$, we derive the following updating equations based on maximum likelihood criterion [9]:

$$\operatorname{vec}(\boldsymbol{W}) = \left(\sum_{m=1}^{M} \overline{\gamma}_m^{(tar)} \boldsymbol{U}_m^\top \boldsymbol{U}_m \otimes \boldsymbol{\Sigma}_m^{(YY)^{-1}} \right)^{-1} \operatorname{vec}(\boldsymbol{C}), \qquad (5)$$

$$\boldsymbol{C} = \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{\Sigma}_m^{(YY)^{-1}} (\boldsymbol{Y}_t^{(tar)} - \boldsymbol{b}_m^{(0)}) \boldsymbol{U}_m, \qquad (6)$$

$$\boldsymbol{U}_m = \boldsymbol{U}^{(M)}(m, :) \in \mathcal{R}^{1 \times K}, \qquad (7)$$

$$\gamma_{m,t} = P(m | \boldsymbol{Y}_t^{(tar)}, \boldsymbol{\lambda}, \boldsymbol{W}), \overline{\gamma}_m^{(tar)} = \sum_{t=1}^{T} \gamma_{m,t}. \qquad (8)$$

where vec() is the vec-operator that stacks the columns of a matrix into a vector.

4. Speaker adaptive training for TASC

This section describes speaker adaptive training (SAT) for tensor-based arbitrary speaker conversion. Applying SAT to tensor-based speaker representation is expected to construct a canonical model and to realize more flexible and precise voice conversion.

Similarly to SAT for EVC [11], shared parameters in the canonical model for TASC are estimated by maximizing likelihood of all the models for individual pre-stored speakers:

$$\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\mathcal{W}}}_1^S) = \underset{\lambda, \mathcal{W}_1^S}{\operatorname{argmax}} \prod_{s=1}^{S} \prod_{t_s=1}^{T_s} P(\boldsymbol{Z}_{t_s}^{(s)} | \boldsymbol{\lambda}(\boldsymbol{W}_s)), \qquad (9)$$

where $\boldsymbol{Z}_{t_s}^{(s)} = [\boldsymbol{X}_{t_s}^\top, \boldsymbol{Y}_{t_s}^{(s)^\top}]^\top$, and $\boldsymbol{\lambda}(\boldsymbol{W}_s)$ denotes the adapted model to the s-th pre-stored speaker with the weight matrix $\boldsymbol{W}_s$. $\boldsymbol{\mathcal{W}}_1^S$ denotes a tensor representing a set of the weight matrices of S pre-stored speakers $(\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_S)$. In SAT, the shared parameters of the canonical model and $\boldsymbol{\mathcal{W}}_1^S$ are estimated in a maximum likelihood manner. To realize it, the following auxiliary function is derived:

$$Q\left(\boldsymbol{\lambda}(\boldsymbol{\mathcal{W}}_1^S), \hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\mathcal{W}}}_1^S)\right) = \sum_{s=1}^{S} \sum_{m=1}^{M} \overline{\gamma}_m^{(s)} \log P\left(\boldsymbol{Z}^{(s)}, m | \hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{W}}_s)\right), (10)$$

$$\gamma_{m,t_s}^{(s)} = P\left(m | \boldsymbol{Z}_{t_s}^{(s)}, \ \boldsymbol{\lambda}(\boldsymbol{W}_s)\right), \overline{\gamma}_m^{(s)} = \sum_{t_s=1}^{T_s} \gamma_{m,t_s}^{(s)}. \qquad (11)$$

As mentioned in [11], simultaneous update for all parameters based on Equation 11 is difficult because of their interdependency on each other. Hence, the following update scheme is adopted. (1) Using the current shared parameters and Equation 11, $\gamma_{m,t_s}^{(s)}$ and $\overline{\gamma}_m^{(s)}$ are calculated. (2) Using $\gamma_{m,t_s}^{(s)}$ and $\overline{\gamma}_m^{(s)}$ and the current shared parameters, each weight matrix $\hat{\boldsymbol{W}}_s$ of the pre-stored speakers is updated. (3) Using the results of the previous steps, the shared weight parameters $\hat{\alpha}_m$ for GMM and the bases matrices $\hat{\boldsymbol{U}}^{(M)}$ are updated. (4) The covariance matrices $\hat{\boldsymbol{\Sigma}}_m^{(ZZ)}$ are updated using the updated parameters in the previous steps. (5) Step 1 to 4 are repeated until the number of repetition equals to the preset value. Note that each step in the update scheme can monotonically increase the likelihood of the adapted models for individual pre-stored speakers.

In Step 2, the updated weight matrix $\hat{\boldsymbol{W}}_s$ for the s-th pre-stored speaker is written as

$$\mathrm{vec}(\hat{\boldsymbol{W}}_s)=\left[\sum_{m=1}^{M}\overline{\gamma}_m^{(s)}\boldsymbol{U}_m^{\top}\boldsymbol{U}_m\otimes\boldsymbol{P}_m^{(YY)}\right]^{-1}\mathrm{vec}\left[\sum_{m=1}^{M}\boldsymbol{C}'_m\boldsymbol{U}_m\right],\quad(12)$$

$$\boldsymbol{C}'_m=\boldsymbol{P}_m^{(YX)}(\overline{\boldsymbol{X}}^{(s)}-\overline{\gamma}_m^{(s)}\boldsymbol{\mu}_m^{(X)})+\boldsymbol{P}_m^{(YY)}(\overline{\boldsymbol{Y}}^{(s)}-\overline{\gamma}_m^{(s)}\boldsymbol{b}_m^{(0)}),\quad(13)$$

$$\overline{\boldsymbol{Z}}_m^{(s)}=\begin{bmatrix}\overline{\boldsymbol{X}}_m^{(s)}\\ \overline{\boldsymbol{Y}}_m^{(s)}\end{bmatrix}=\begin{bmatrix}\sum_{t_s=1}^{T_s}\gamma_{i,t_s}^{(s)}\boldsymbol{X}_{t_s}^{(s)}\\ \sum_{t_s=1}^{T_s}\gamma_{i,t_s}^{(s)}\boldsymbol{Y}_{t_s}^{(s)}\end{bmatrix},\quad(14)$$

$$\boldsymbol{\Sigma}_m^{(ZZ)-1}=\begin{bmatrix}\boldsymbol{P}_m^{(XX)}\boldsymbol{P}_m^{(XY)}\\ \boldsymbol{P}_m^{(YX)}\boldsymbol{P}_m^{(YY)}\end{bmatrix}\equiv\boldsymbol{P}_m^{(ZZ)}\quad(15)$$

Compared with Equations 5 and 6, Equations 12 and 13 have similar forms, but they include the effects of the vectors of the reference speaker $\boldsymbol{X}$. In Steps 3 and 4, the shared parameters are updated as follows:

$$\hat{\alpha}_m=\frac{\sum_{s=1}^{S}\overline{\gamma}_m^{(s)}}{\sum_{m=1}^{M}\sum_{s=1}^{S}\overline{\gamma}_m^{(s)}},\quad(16)$$

$$\hat{\boldsymbol{u}}_m=\left(\sum_{s=1}^{S}\overline{\gamma}_m^{(s)}\hat{\boldsymbol{E}}_s^{\top}\boldsymbol{P}_m^{(ZZ)}\hat{\boldsymbol{E}}_s\right)^{-1}\left(\sum_{s=1}^{S}\hat{\boldsymbol{E}}_s\boldsymbol{P}_m^{(ZZ)}\overline{\boldsymbol{Z}}_m^{(s)}\right),\quad(17)$$

$$\boldsymbol{\Sigma}_m^{(ZZ)}=\frac{1}{\sum_{s=1}^{S}\overline{\gamma}_m^{(s)}}\sum_{s=1}^{S}\Big\{\overline{\boldsymbol{V}}_m^{(s)}+\overline{\gamma}_m^{(s)}\hat{\boldsymbol{\mu}}^{(s)}\hat{\boldsymbol{\mu}}^{(s)\top} - \left(\hat{\boldsymbol{\mu}}_m^{(s)}\overline{\boldsymbol{Z}}_m^{(s)\top}+\overline{\boldsymbol{Z}}_m^{(s)\top}\hat{\boldsymbol{\mu}}_m^{(s)}\right)\Big\},\quad(18)$$

$$\overline{\boldsymbol{V}}_m^{(s)}=\sum_{t_s=1}^{T}\gamma_{m,t_s}^{(s)}\boldsymbol{Z}_{t_s}^{(s)}\boldsymbol{Z}_{t_s}^{(s)\top},\quad(19)$$

$$\hat{\boldsymbol{\mu}}_m^{(s)}=\hat{\boldsymbol{E}}_s\hat{\boldsymbol{u}}_m=\begin{bmatrix}\hat{\boldsymbol{\mu}}_m^{(X)}\\ \hat{\boldsymbol{W}}_s\hat{\boldsymbol{U}}_m^{\top}+\hat{\boldsymbol{b}}_m^{(0)}\end{bmatrix},\quad(20)$$

$$\hat{\boldsymbol{u}}_m=\left[\hat{\boldsymbol{\mu}}_m^{(X)\top},\hat{\boldsymbol{b}}_m^{(0)\top},\hat{\boldsymbol{U}}_m\right]^{\top}\in\mathcal{R}^{(2D'+K)\times1},\quad(21)$$

$$\hat{\boldsymbol{E}}_s=\begin{bmatrix}\boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{I} & \hat{\boldsymbol{W}}_s\end{bmatrix}\in\mathcal{R}^{2D'\times(2D'+K)},\quad(22)$$

and $\boldsymbol{I}$ denotes an identity matrix whose size is D'. Compared with the update equations in [11], Equations 16 to 19 have the same forms. That is to say, updating the shared parameters is carried out in the same manner as [11]. In Equations 18 and 19, the shared covariance matrix is calculated as the mean of the covariance matrices, each of which corresponds to covariances of a pair of speakers. Hence, it is expected that SAT for TASC also affects for compacting variations as well as SAT for EVC. On the other hand, construction of the mean vectors by the bases and the weight parameters (Equations 20 to 22) is different from that in [11]. In the proposed method, we need to calculate $(2D'+K)\times(2D'+K)$-sized inverse matrix for updating $\hat{\boldsymbol{u}}_m$. In [11], the update of $\hat{\boldsymbol{v}}_m$ that corresponds to Equation 17 in the proposed method requires calculating $\{D'\cdot(J+2)\}\times\{D'\cdot(J+2)\}$-sized inverse matrix. Although the assumption that covariance matrices $\boldsymbol{\Sigma}_m^{(XX)}$, $\boldsymbol{\Sigma}_m^{(XY)}$ and $\boldsymbol{\Sigma}_m^{(YY)}$ are diagonal reduces computational cost to D' times calculation of $(J+2)\times(J+2)$-sized inverse matrices, it is more computationally expensive than the proposed method.

5. Experimental evaluation

5.1. Experimental conditions

To evaluate the performance of our proposed method and the effects of SAT on both EVC and TASC, 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 [14], and 273 pre-stored speakers including 137 male and 136 female speakers[1]. 50 sentences were uttered by 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.

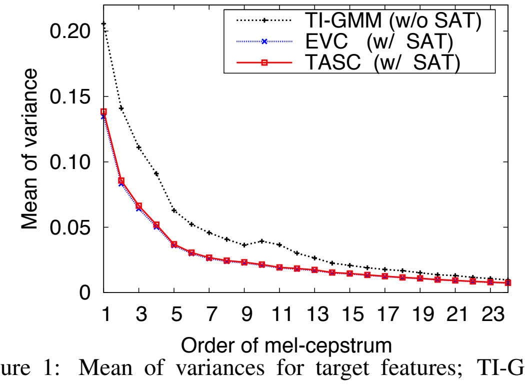

Figure 1: Mean of variances for target features; TI-GMM, EVC-SAT and TASC-SAT.

We used 24-dimensional mel-cepstrum vectors for spectrum representation (D=24, D'=48). These were derived by STRAIGHT analysis [15]. The number of mixture components (M) was fixed to 128.

In SAT for EVC, the number of representative vectors was fixed to 272 (J). In SAT for TASC, the size of representative matrices was fixed to 80 (K). Both values were determined from the results of adaptation using models without SAT. The number of iterations for SAT was fixed to 7.

We evaluated four methods for arbitrary speaker conversion; the tensor-based one-to-many VC with and without SAT (TASC w/ SAT and TASC w/o SAT), the one-to-many EVC with and without SAT (EVC w/ SAT and EVC w/o SAT). In addition, traditional VC with the parallel training (Traditional) was also compared with them [16]. Note that traditional VC requires th parallel data for the target speakers. In both EVC and TASC, the number of representative vectors ($J'\leq J$) and the size of representative matrix ($K'\leq K$) were varied.

5.2. Effects of SAT on compacting variance

We compared diagonal components of the target covariance matrices $\boldsymbol{\Sigma}_m^{(YY)}$ of TI-GMM, EVC-based model after SAT, and tensor-based model after SAT. Figure 1 shows the mean of variances for target features in individual Gaussian components of these models. Values of diagonal components of TI-GMM are relatively larger than those of EVC-based model after SAT and tensor-based model after SAT. This result shows the effects of SAT that they compact variations of the models trained by many speakers. Compared tensor-based model with EVC-based model, there is no significant difference between values of diagonal components of them.

5.3. 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

[1] http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html

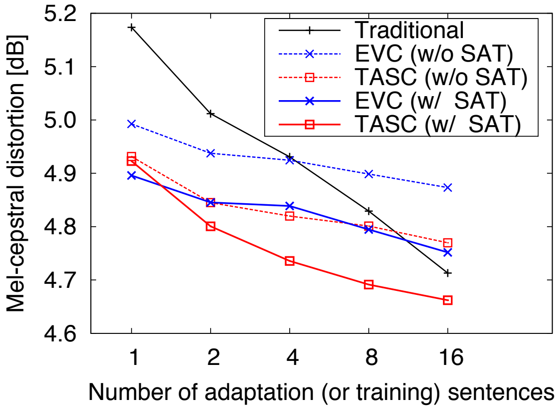

Figure 2: Results of objective evaluations by mel-cepstral distortion (MCD). The optimal numbers of J' and K' were selected in each condition in the sense of MCD.

Table 1: The optimal numbers of the representative parameters for each method.

# of sentences	1	2	4	8	16
EVC w/o SAT (J)			272		
EVC w/ SAT (J')			272		
TASC w/o SAT (K)	20	20	40	80	80
TASC w/ SAT (K')	10	20	30	40	40

of mixture components is selected. Compared with the conversion methods without SAT, the performance of the methods using SAT is better in both of EVC and TASC. This means that SAT works effectively to compact variances, and that conversion performances were improved by constructing the canonical model which works as a speaker-dependent models more likely. Compared with EVC, the performance of TASC is better both with and without SAT. This means that our proposed representation of the speaker space works well rather than supervector representation of the speaker space. TASC with SAT outperformed "Traditional" even when the number of training or adaptation utterances is 16. This means that combination of tensor-based representation and SAT can effectively capture the information in the adaptation data.

Table 1 shows the optimal numbers of the representative parameters in each method. In the cases of EVC, $J = 272$ and $J' = 272$ are optimal even the number of sentences is varied. This means the flexibility of EVC is limited because of the high-dimensional ($D'M$) representative vectors. On the other hand, in the cases of TASC, the optimal numbers of the representative parameters are varied depending on the number of adaptation utterances. Since the number of mixtures is 128, the size of the representative matrix was effectively reduced. Interestingly, after SAT, the optimal size of the representative matrix is slightly reduced. It might be caused by the effects of SAT on compacting variations.

6. Conclusions

We have proposed speaker adaptive training (SAT) for the tensor-based speaker space, to improve the performance of speaker conversion for arbitrary target speakers. In the tensor-based speaker representation that we have previously proposed, SAT works effectively to construct a canonical and precise model for voice conversion as well as in the case of EVC. For further works, the effectiveness of SAT on tensor-based arbitrary speaker conversion should be investigated in large-scale subjective evaluations. In addition, the optimization of the representative parameters in the proposed framework also should be investigated. Baysian treatment of arbitrary speaker conversion including representation of speaker space, the structure optimization, and adaptation based on tensor representation is another further direction.

7. Acknowledgment

This work was supported by KAKENHI Grant-in-Aid for Research Activity Start-up (11025460).

8. 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] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," Proc. INTERSPEECH, pp. 653–656, 2011.

[10] T. Anastasakos, J. McDonough, R. Schwarts and J. Makhoul, "A compact model for speaker adaptive training," Proc. ICSLP, vol. 2, pp. 1137–1140, 1996.

[11] 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.

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

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

[14] 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.

[15] 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.

[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.