PAPER

Tensor Factor Analysis for Arbitrary Speaker Conversion

Daisuke SAITO[†a)], **Nobuaki MINEMATSU**[†], ***Members***, ***and*** **Keikichi HIROSE**[†*], ***Fellow***

SUMMARY This paper describes a novel approach to flexible control of speaker characteristics using tensor representation of multiple Gaussian mixture models (GMM). 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 GMM (EV-GMM) was proposed. In the EVC, 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 factor 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 dimension of the mean vector and the Gaussian component. The speaker space is derived by the tensor factor 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. In addition, in this paper, effects of speaker adaptive training before factorization are also investigated. Experimental results of one-to-many voice conversion demonstrate the effectiveness of the proposed approach.

key words: *voice conversion, Gaussian mixture models, eigenvoice, tensor factor analysis, Tucker decomposition*

1. Introduction

Voice conversion (VC), or speaker conversion is a technique to partly transform an input utterance of a speaker to another utterance that sounds like another speaker while its linguistic content is 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 kinds of 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 approaches, GMM-based approaches have been widely used in particular because GMMs have good properties of flexibility and solid theoretical background.

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. Saito *et al.* proposed a voice conversion framework based on noisy channel model to effectively integratethe the speaker GMM with the joint density GMM [9]. Non-parallel data can be utilized through the speaker GMM. 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 [11]. 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 pivot speaker henceforth, and many pre-stored speakers. Based on joint density models of the pivot and the pre-stored speakers, the speaker GMMs of the pre-stored speakers can be extracted.

From the help of the feature space of the pivot speaker, Components of Gaussian in GMM are aligned. Hence a speaker space can be 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 [12], an arbitrary speaker is represented as a vector of this speaker space. Hence the joint density GMM of the pivot and the target speaker is flexibly developed by estimating a small number of weight parameters for the bases of the space. Inspired by speaker recognition studies, Wu *et al.* also proposed a voice conversion method utilizing mixture of factor analyzers, where the factor analysis of GMM supervectors was embedded in mean vectors of the joint density GMM [13]. This work is free from a pivot speaker to construct the speaker space, and aims for mitigating the problem of the sparse parallel data for the joint density model.

However, the representation of GMM supervector itself has an inherent problem that multiple factors of acoustic variations are included in the same space. Namely, Gaus-

Manuscript received June 14, 2019.
Manuscript revised February 7, 2020.
Manuscript publicized March 13, 2020.
[†]The authors are with The University of Tokyo, Tokyo, 113–8656 Japan.
[*]Presently, with the National Institute of Informatics.
a) E-mail: dsk_saito@gavo.t.u-tokyo.ac.jp
DOI: 10.1587/transinf.2019EDP7166

sian 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, for more tractable treatment of the VC framework, we propose a method to construct the speaker space based on tensor factor analysis. In our approach, an arbitrary speaker is not represented as a supervector, but a matrix whose row and column respectively correspond to the dimension of the mean vector and the component of GMM. Based on this representation, the data set of the pre-stored speakers was expressed as a third-order tensor, and the tensor factor analysis is introduced to obtain the speaker space. Since the tensor analysis can treat multiple factors of variations properly [14], it will be expected to improve the performance of VC. The approach was first proposed in [15]. This paper provides detailed investigation for tensor factor analysis in the proposed framework, and it provides generalized viewpoints for eigenvoice-based approaches. In addition, it also introduces a new strategy for speaker adaptive training which keeps orthogonalities of multiple factors. 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 mixture of factor analyzers [13], or non-parallel training for many-to-many EVC [17]. Although this paper mainly focuses on GMM-based voice conversion, the proposed factor analysis itself can be utilized for neural-network-based approaches [18].

The remainder of this paper is organized as follows. Section 2 describes the basic EVC approach. Then, our proposed approach using the tensor factor analysis to construct the speaker space is described in Sect. 3. Section 4 describes a new strategy for speaker adaptive training to keep both the performance and flexibility. In Sect. 5, experimental evaluations are described. Finally, Sect. 6 concludes the paper.

2. Eigenvoice Conversion (EVC)

2.1 Eigenvoice GMM (EV-GMM)

In this section, one-to-many EVC [19] is described. Let $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ and $\boldsymbol{Y}_t^{(s)} = [\boldsymbol{y}_t^{(s)\top}, \Delta\boldsymbol{y}_t^{(s)\top}]^\top$ be D-dimensional vectors of the pivot speaker and the s-th target speaker, respectively. They consist of $D/2$-dimensional static and dynamic features. The notation $(\cdot)^\top$ denotes transpose of a vector. The joint probability density of the pivot 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 $\boldsymbol{\Sigma}$. $\boldsymbol{\lambda}^{(EV)}$ denotes model parameters of EV-GMM; α_m, $\boldsymbol{\mu}_m^{(X)}$, $\boldsymbol{B}_m$, $\boldsymbol{b}_m^{(0)}$, and $\boldsymbol{\Sigma}_m^{(Z)}$. 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 $K \leq S - 1$. 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 $\boldsymbol{B} = [\boldsymbol{B}_1^\top, \boldsymbol{B}_2^\top, \ldots, \boldsymbol{B}_m^\top]^\top \in \mathcal{R}^{DM\times K}$ and the bias supervector $\boldsymbol{b} = \left[\boldsymbol{b}_1^{(0)^\top}, \boldsymbol{b}_2^{(0)^\top}, \ldots, \boldsymbol{b}_m^{(0)^\top}\right]^\top \in \mathcal{R}^{DM\times 1}$.

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. Let $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$ be the joint vector of the pivot and the target speakers, and $\boldsymbol{Y}_t$ denotes a vector from a pre-stored speaker. The probability of TI-GMM is as follows:

$$P(\boldsymbol{Z}_t|\boldsymbol{\lambda}^{(TI)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(Z)}, \boldsymbol{\Sigma}_m^{(Z)}), \quad (3)$$

$$\boldsymbol{\mu}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}, \quad (4)$$

where $\boldsymbol{\lambda}^{(TI)}$ denotes model parameters of TI-GMM; α_m, $\boldsymbol{\mu}_m^{(Z)}$, and $\boldsymbol{\Sigma}_m^{(Z)}$. In the case of one-to-many conversion, $\boldsymbol{\mu}_m^{(Y)}$ is inferred from multiple target speakers.

Next, each target dependent GMM $\boldsymbol{\lambda}^{(s)}$ (TD-GMM) is trained by updating only the target mean vectors ($\boldsymbol{\mu}_m^{(s)}$) using each of the corresponding parallel data set. $\boldsymbol{\mu}_m^{(Y)}$ are used for initial values for $\boldsymbol{\mu}_m^{(s)}$. Because this process is achieved before PCA-based factorization, $\boldsymbol{w}^{(s)}$ is not included in the parameter set $\boldsymbol{\lambda}^{(s)}$. Note that α_m, $\boldsymbol{\mu}_m^{(X)}$, and $\boldsymbol{\Sigma}_m^{(Z)}$ are not updated. $\boldsymbol{\mu}_m^{(s)}$ is updated by EM algorithm as follows:

$$\hat{\boldsymbol{\mu}}_m^{(s)} = \overline{\boldsymbol{Y}}_m^{(s)} - \boldsymbol{\Sigma}_m^{(YX)}\boldsymbol{\Sigma}_m^{(XX)-1}\left(\overline{\boldsymbol{X}}_m^{(s)} - \boldsymbol{\mu}_m^{(X)}\right), \quad (5)$$

$$\overline{\boldsymbol{X}}_m^{(s)} = \frac{1}{\overline{\gamma}_m^{(s)}} \sum_{t_s=1}^{T_s} \gamma_{m,t_s}^{(s)} \boldsymbol{X}_{t_s}^{(s)}, \overline{\boldsymbol{Y}}_m^{(s)} = \frac{1}{\overline{\gamma}_m^{(s)}} \sum_{t_s=1}^{T_s} \gamma_{m,t_s}^{(s)} \boldsymbol{Y}_{t_s}^{(s)} \quad (6)$$

$$\gamma_{m,t_s}^{(s)} = P\left(m|\boldsymbol{Z}_{t_s}^{(s)}, \boldsymbol{\lambda}^{(s)}\right), \overline{\gamma}_m^{(s)} = \sum_{t_s=1}^{T_s} \gamma_{m,t_s}^{(s)}. \quad (7)$$

In Eq. (5), the second term corresponds to the effect that $\boldsymbol{\mu}_m^{(X)}$ is fixed among all the TD-GMMs. 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 TD-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 [10]. Let $\boldsymbol{Y}^{(tar)}$ be a sequence of the target features. $\boldsymbol{w}$ is estimated as follows:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \int P(\boldsymbol{X}, \boldsymbol{Y}^{(tar)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}) d\boldsymbol{X}. \tag{8}$$

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

$$\hat{\boldsymbol{w}} = \left\{ \sum_{m=1}^{M} \overline{\gamma}_m^{(tar)} \boldsymbol{B}_m^\top \boldsymbol{\Sigma}_m^{(YY)^{-1}} \boldsymbol{B}_m \right\}^{-1} \sum_{m=1}^{M} \boldsymbol{B}_m^\top \boldsymbol{\Sigma}_m^{(YY)^{-1}} \overline{\boldsymbol{Y}}_m^{(tar)}, \tag{9}$$

$$\overline{\gamma}_m^{(tar)} = \sum_{t=1}^{T} \gamma_{m,t}, \ \overline{\boldsymbol{Y}}_m^{(tar)} = \sum_{t=1}^{T} \gamma_{m,t} (\boldsymbol{Y}_t^{(tar)} - \boldsymbol{b}_m^{(0)}), \tag{10}$$

$$\gamma_{m,t} = P(m | \boldsymbol{Y}_t^{(tar)}, \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}). \tag{11}$$

Equation (9) 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 Eq. (11). After adaptation, the step of parameter generation is the same as [20].

3. Tensor Factor Analysis for the 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 [21]. 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 Fig. 1. Using this flattening operation, the product of a tensor and a matrix is 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 mode-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:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top = \mathcal{S} \times_1 \boldsymbol{U} \times_2 \boldsymbol{V}. \tag{12}$$

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

$$\mathcal{A} = \mathcal{S} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times_3 \boldsymbol{U}_3. \tag{13}$$

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 Eq. (13) is called high-order SVD, or Tucker decomposition [21], [22]. 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*.

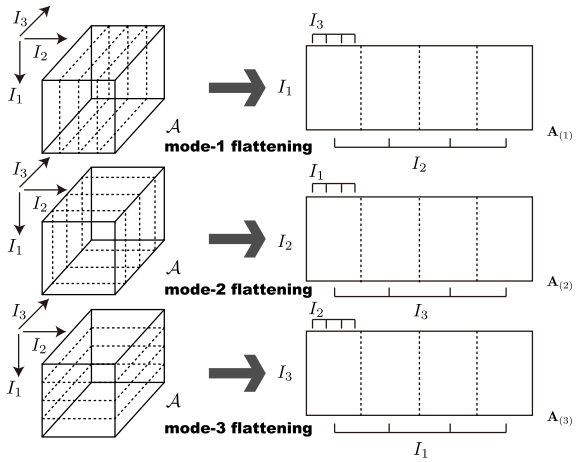


Fig. 1 Flattening of the $(I_1 \times I_2 \times I_3)$-tensor $\mathcal{A}$ to the flattened matrices $\boldsymbol{A}_{(1)}$, $\boldsymbol{A}_{(2)}$ and $\boldsymbol{A}_{(3)}$.

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 $D \times M$ matrix [23], where D is the dimension of the feature, and M is the number of mixtures. First, the bias matrix $\boldsymbol{b}' = \left[\boldsymbol{b}_1^{(0)}, \boldsymbol{b}_2^{(0)}, \ldots, \boldsymbol{b}_m^{(0)}\right]$ 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 $\mathcal{M} \in \mathcal{R}^{D \times M \times S}$. Then, $\mathcal{M}$ can be represented as follows:

$$\mathcal{M} = \mathcal{G}^{D \times M \times S} \times_1 \boldsymbol{U}^{(D)} \times_2 \boldsymbol{U}^{(M)} \times_3 \boldsymbol{U}^{(S)}, \tag{14}$$

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

$$\boldsymbol{\mu}^{(n)} = \mathcal{G} \times_1 \boldsymbol{U}^{(D)} \times_2 \boldsymbol{U}^{(M)} \times_3 \boldsymbol{u}_s, \tag{15}$$

For efficient representation, truncated matrices and tensor are considered, namely $\mathcal{G} \in \mathcal{R}^{K_D \times K_M \times K_S}$, $\boldsymbol{U}^{(D)} \in \mathcal{R}^{D \times K_D}$ ($K_D \leq D$), $\boldsymbol{U}^{(M)} \in \mathcal{R}^{M \times K_M}$ ($K_M \leq M$), and $\boldsymbol{u}_s \in \mathcal{R}^{1 \times K_s}$. We call this factorization as tensor factor analysis (TFA) henceforth. There are several candidates for TFA. In the previous paper [15], $\boldsymbol{U}^{(M)}$ becomes the bases, and the others become weights as similar to [23]. In this paper, the other factorizations are comprehensively investigated. Table 1 shows several kinds of groupings investigated in the paper. Henceforth, for readability, several subscripts are omitted and

Table 1 Kinds of tensor factor analysis.

method	Base	Weight	Footprint	# of parameters for adaptation
Eigenvoice (EV)	$\mathcal{G} \times_1 \boldsymbol{U}^{(D)} \times_2 \boldsymbol{U}^{(M)}$	$\boldsymbol{u}_s$	DMK_s	K_s
truncated EV	$\mathcal{G} \times_1 \boldsymbol{U}^{(D)} \times_2 \boldsymbol{U}^{(M)}$	$\boldsymbol{u}_s$	$K_D K_M K_s$	K_s
TFA-mix	$\boldsymbol{U}^{(M)}$	$\mathcal{G} \times_1 \boldsymbol{U}^{(D)} \times_3 \boldsymbol{u}_s$	MK_M	DK_M
TFA-feat	$\boldsymbol{U}^{(D)}$	$\mathcal{G} \times_2 \boldsymbol{U}^{(M)} \times_3 \boldsymbol{u}_s$	DK_D	MK_D
TFA-bilinear	$\boldsymbol{U}^{(D)}$,$\boldsymbol{U}^{(M)}$	$\mathcal{G} \times_3 \boldsymbol{u}_s$	$DK_D + MK_M$	$K_D K_M$

some variables are repeatedly used.

3.2.1 Eigenvoice (EV)

When only $\boldsymbol{u}_s$ is regarded as a weight, bases of the factorization are $\mathcal{G} \times_1 \boldsymbol{U}^{(D)} \times_2 \boldsymbol{U}^{(M)} \in \mathcal{R}^{D \times M \times K_s}$. When $D = K_D$ and $M = K_M$, this factorization is identical to Eigenvoice (EV). That is to say, Eigenvoice can be viewed as a special case of the proposed factorization. In this factorization, the number of parameters which should be estimated for a new speaker is quite small, while the footprint of the stored model is large.

3.2.2 Truncated EV

In order to reduce the footprint of EV-based adaptation, the number of base vectors has been reduced. On the other hand, in the case of TFA, different modes can be independently truncated, i.e. the feature space and GMM components. When the EV bases are truncated in advance by TFA, the footprint of EV can be reduced to $K_D K_M K_S$ from DMK_S while the number of parameters for adaptation is kept.

3.2.3 TFA-Mix

When the truncation of mixture information is focused on, $\boldsymbol{U}^{(M)}$ is selected as the bases. Taking the bias matrix $\boldsymbol{b}'$ into account, we obtain the matrix $\boldsymbol{\mu}$ for a new speaker as

$$\boldsymbol{\mu} = \boldsymbol{W}\boldsymbol{U}^{(M)\top} + \boldsymbol{b}', \tag{16}$$

where $\boldsymbol{W} \in \mathcal{R}^{D \times K_M}$ is a weight matrix. Hence, in this factorization, parameters to be estimated become a $D \times K_M$ matrix, while the footprint of the model is MK_M.

In [23], 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:

$$\mathrm{vec}(\boldsymbol{W}) = \left[\sum_{m=1}^{M} \overline{\gamma}_m^{(tar)} \boldsymbol{U}_m^\top \boldsymbol{U}_m \otimes \boldsymbol{\Sigma}_m^{(YY)^{-1}} \right]^{-1} \mathrm{vec}(\boldsymbol{C}), \tag{17}$$

$$\boldsymbol{C} = \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{\Sigma}_m^{(YY)^{-1}} (\boldsymbol{Y}_t^{(tar)} - \boldsymbol{b}_m^{(0)}) \boldsymbol{U}_m, \tag{18}$$

$$\boldsymbol{U}_m = \boldsymbol{U}^{(M)}(m, :) \in \mathcal{R}^{1 \times K_M}, \tag{19}$$

where vec() is the vec-operator that stacks the columns of a matrix into a vector. Compared with Eq. (9), Eq. (17) has a similar form, but it estimates $D \times K_M$ parameters rather than K (or K_s) parameters in Eq. (9). This means that our proposed method might be more flexible to adapt for the data. We verify it by the experiments.

3.2.4 TFA-Feat

When the dimensionality reduction for acoustic features is focused on, $\boldsymbol{U}^{(D)}$ is selected as the bases. The matrix $\boldsymbol{\mu}$ for a new speaker is

$$\boldsymbol{\mu} = \boldsymbol{U}^{(D)}\boldsymbol{W} + \boldsymbol{b}', \tag{20}$$

where $\boldsymbol{W} \in \mathcal{R}^{K_D \times M}$ is a weight matrix. Similarly to Eq. (17), the updating equations based on maximum likelihood criterion are derived as

$$\boldsymbol{W}(:, m) = \left[\overline{\gamma}_m^{(tar)} \boldsymbol{U}^{(D)\top} \boldsymbol{\Sigma}_m^{(YY)^{-1}} \boldsymbol{U}^{(D)} \right]^{-1} \boldsymbol{c}_m, \tag{21}$$

$$\boldsymbol{c}_m = \boldsymbol{U}^{(D)\top} \boldsymbol{\Sigma}_m^{(YY)^{-1}} \overline{\boldsymbol{Y}}_m^{(tar)}. \tag{22}$$

In the case of dimensionality reduction for acoustic features, weight column vectors $\boldsymbol{w}_m = \boldsymbol{W}(:, m)$ are independently inferred. That is to say, this inference is sensitive to the occupation count $\overline{\gamma}_m^{(tar)}$ for the focused component m. Although this property is not always effective, in particular, in the case that the amount of adaptation data is limitted, this truncation affects the footprint to be small.

3.2.5 TFA-Bilinear

Taking into both the effects from GMM components and acoustic features, we obtain a bilinear form for representing a new speaker as follows:

$$\boldsymbol{\mu} = \boldsymbol{U}^{(D)}\boldsymbol{W}\boldsymbol{U}^{(M)\top} + \boldsymbol{b}', \tag{23}$$

where $\boldsymbol{W} \in \mathcal{R}^{K_D \times K_M}$ is a weight matrix. Compared with Eigenvoice, although both the effects from GMM components and acoustic features are included in Eq. (23), they can be separately controlled by row and column of the weight matrix. The updating equations based on maximum likelihood criterion are derived as

$$\mathrm{vec}(\boldsymbol{W}) = \boldsymbol{E}^{-1} \mathrm{vec}(\boldsymbol{C}), \tag{24}$$

$$\boldsymbol{C} = \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{m,t} \boldsymbol{U}^{(D)\top} \boldsymbol{\Sigma}_m^{(YY)^{-1}} (\boldsymbol{Y}_t^{(tar)} - \boldsymbol{b}_m^{(0)}) \boldsymbol{U}_m, \tag{25}$$

$$\boldsymbol{E} = \sum_{m=1}^{M} \overline{\gamma}_m^{(tar)} \boldsymbol{U}_m^\top \boldsymbol{U}_m \otimes \boldsymbol{U}^{(D)\top} \boldsymbol{\Sigma}_m^{(YY)^{-1}} \boldsymbol{U}^{(D)}. \tag{26}$$

4. Speaker Adaptive Training before Factorization

This section describes a new strategy of speaker adaptive training (SAT) for arbitrary speaker conversion. SAT was introduced for training a canonical speaker-independent model [24]. The effectiveness of SAT in arbitrary speaker conversion was shown in [16] and [25]. In SAT, shared parameters in the canonical model are estimated by maximizing likelihood of all the models for individual pre-stored speakers. Usually, mean vectors are factorized, and factorized parameters are regarded as the shared parameters.

One of the largest effects of SAT is compacting variance. In the process of SAT for the joint density GMM, the shared covariance matrix are calculated as the mean of the covariance matrices, each of which corresponds to covariance of a pair of speakers. Hence, shrinkage of the covariance is expected. On the other hand, the shared parameters obtained from factorization is not essential. On the contrary, orthogonalities in the factorization are lost in the process of SAT. Factorization based on orthogonal bases is convenient to control the complexity of the models. To receive both the merits of compact variance and orthogonal factorization, we propose a SAT process before factorization of mean parameters.

The proposed process replaces the construction of TD-GMMs in Sect. 2.2. Compared with Sect. 2.2, α_m, $\boldsymbol{\mu}_m^{(X)}$, and $\boldsymbol{\Sigma}_m^{(Z)}$ is not fixed, but shared. Shared parameters in the canonical model are estimated by maximizing likelihood of all the models for individual pre-stored speakers:

$$\hat{\boldsymbol{\lambda}}(1\dots S) = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}} \prod_{s=1}^{S} \prod_{t_s=1}^{T_s} P(\boldsymbol{Z}_{t_s}^{(s)}|\boldsymbol{\lambda}(1\dots S)), \tag{27}$$

where $\boldsymbol{Z}_{t_s}^{(s)} = [\boldsymbol{X}_{t_s}^{\top}, \boldsymbol{Y}_{t_s}^{(s)^{\top}}]^{\top}$, and $\boldsymbol{\lambda}(1\dots S)$ denotes a set of all the TD-GMMs. In SAT, the shared parameters of the canonical model are estimated in a maximum likelihood manner. To realize it, the following auxiliary function is derived:

$$Q\left(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}\right) = \sum_{s=1}^{S} \sum_{m=1}^{M} \overline{\gamma}_m^{(s)} \log P\left(\boldsymbol{Z}^{(s)}, m|\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{W}}_s)\right), \tag{28}$$

$$\gamma_{m,t_s}^{(s)} = P\left(m|\boldsymbol{Z}_{t_s}^{(s)},\ \boldsymbol{\lambda}(\boldsymbol{W}_s)\right), \overline{\gamma}_m^{(s)} = \sum_{t_s=1}^{T_s} \gamma_{m,t_s}^{(s)}. \tag{29}$$

As mentioned in [16], simultaneous update for all parameters based on Eq. (28) is difficult because of their interdependency on each other. Hence, the following update scheme is adopted. (1) Using the current shared parameters and Eq. (29), $\gamma_{m,t_s}^{(s)}$ and $\overline{\gamma}_m^{(s)}$ are calculated. (2) Using $\gamma_{m,t_s}^{(s)}$, $\overline{\gamma}_m^{(s)}$ and the current shared parameters, each target dependent mean vector $\boldsymbol{\mu}_m^{(s)}$ of the pre-stored speakers is updated. (3) Using the results of the previous steps, the shared weight parameters $\hat{\alpha}_m$ and $\boldsymbol{\mu}_m^{(X)}$ for GMM 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, Eq. (5) is used to update $\boldsymbol{\mu}_m^{(s)}$. In Steps 3 and 4, the shared parameters are updated as follows:

$$\boldsymbol{\mu}_m^{(X)} = \frac{1}{\sum_{s=1}^{S} \overline{\gamma}_m^{(s)}} \sum_{s=1}^{S} \sum_{t_s=1}^{T_s} \gamma_{m,t_s}^{(s)} \boldsymbol{X}_{t_s} \tag{30}$$

$$\hat{\alpha}_m = \frac{\sum_{s=1}^{S} \overline{\gamma}_m^{(s)}}{\sum_{m=1}^{M} \sum_{s=1}^{S} \overline{\gamma}_m^{(s)}}, \tag{31}$$

$$\boldsymbol{\Sigma}_m^{(ZZ)} = \frac{1}{\sum_{s=1}^{S} \overline{\gamma}_m^{(s)}} \sum_{s=1}^{S} \boldsymbol{\sigma}_{m,s}, \tag{32}$$

$$\boldsymbol{\sigma}_{m,s} = \overline{\boldsymbol{V}}_m^{(s)} + \overline{\gamma}_m^{(s)} \hat{\boldsymbol{\mu}}_m^{(s)} \hat{\boldsymbol{\mu}}_m^{(s)\top} - \left(\hat{\boldsymbol{\mu}}_m^{(s)} \overline{\boldsymbol{Z}}_m^{(s)\top} + \overline{\boldsymbol{Z}}_m^{(s)\top} \hat{\boldsymbol{\mu}}_m^{(s)}\right), \tag{33}$$

$$\overline{\boldsymbol{V}}_m^{(s)} = \sum_{t_s=1}^{T} \gamma_{m,t_s}^{(s)} \boldsymbol{Z}_{t_s}^{(s)} \boldsymbol{Z}_{t_s}^{(s)^{\top}}, \tag{34}$$

$$\hat{\boldsymbol{\mu}}_m^{(s)} = \begin{bmatrix} \hat{\boldsymbol{\mu}}_m^{(X)} \\ \hat{\boldsymbol{\mu}}_m^{(s)} \end{bmatrix}, \tag{35}$$

Compared with the update equations in [16], Eqs. (31) to (34) have the same forms. That is to say, updating the shared parameters is carried out in the same manner as [16]. In Eqs. (32), (33) and (34), 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 the proposed SAT protocol also affects for compacting variations as well as SAT for EVC. On the other hand, construction of the mean vectors is based on Eq. (5). Hence, any factorization can be carried out in the same manner as that in Sect. 2.2. Finally orthogonal properties of bases in the factorization are preserved theoretically.

5. Experimental Evaluation

5.1 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 pivot speaker from ATR Japanese speech database B-set [26], and 256 pre-stored speakers including 127 male and 129 female speakers [27]. 50 sentences were uttered from each pre-stored speaker, which were included in one of nine subsets. The pivot speaker uttered all of the nine subsets and an additional subsets used for evaluation. In the evaluation, we selected new 10 speakers of 5 male and 5 female speakers, which were not included in the pre-stored speakers. We used 1 to 32 utterances for adaptation, and other 21 utterances for evaluation.

We used 24-dimensional mel-cepstrum vectors for spectrum representation. Finally D=48 because both static and dynamic features were included. These were derived by STRAIGHT analysis [28]. The number of mixture components (M) was fixed to 128. Aperiodic components, which

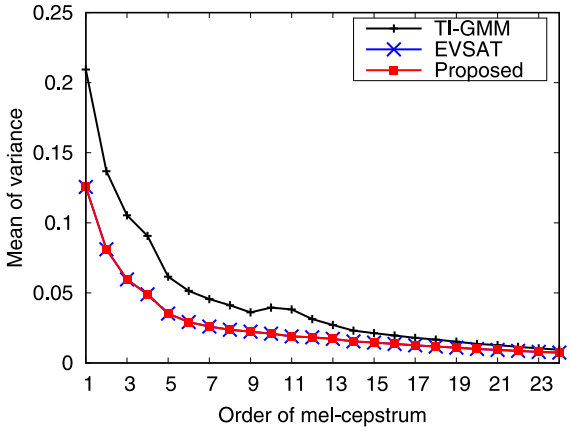


Fig. 2 Mean of variances for target features; TI-GMM, EVSAT, and the proposed protocol.

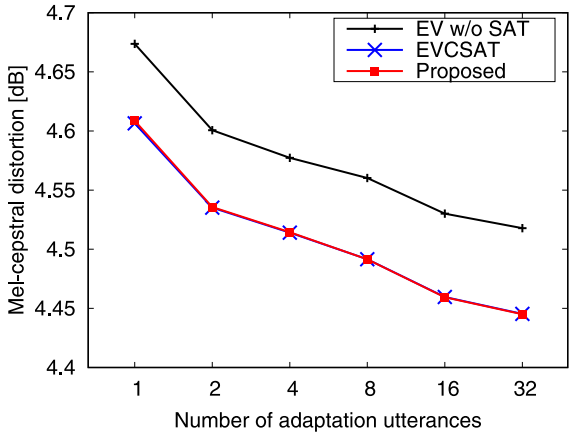


Fig. 3 Results of conversion by three kinds of adaptive training.

are features to construct STRAIGHT mixed excitation, were not converted in this study, and they were parameterized by 5-dimensional banded aperiodicity components (bap). 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. Global variance models were constructed for each test speaker [20]. They were used only for subjective evaluations.

We investigated the effectiveness of the proposed tensor factor analysis in Table 1 on one-to-many VC. As objective evaluation, the conversion performance was evaluated by using mel-cepstral distortion between the converted vectors and the vectors of the targets.

5.2 Effectiveness of the Proposed Protocol for SAT

In this section, the effectiveness of the proposed protocol for SAT was verified. Eigenvoice was used for factorization. The number of bases was fixed to $K = 255$. We compared diagonal components of the target covariance matrices $\mathbf{\Sigma}_m^{(YY)}$ of TI-GMM, EVC-based model after SAT [16], and the constructed model after the proposed SAT. Figure 2 shows the mean of variances for target features in individual Gaussian components of these methods. Values of diagonal components of TI-GMM are relatively larger than those of EVSAT, and the proposed protocol. Compared the proposed protocol with EV-based SAT, there is no significant differences between values of diagonal components of them. This result shows the effects of the proposed method. In addition, since the proposed approach is carried out before factorization, orthogonality of the bases is guaranteed.

Figure 3 shows the results of conversion by three kinds of adaptive training. Compared the proposed protocol with EV-based SAT, achieved mel-cepstral distortions is not significantly different. It can be said that the proposed protocol effectively captures the essence of speaker adaptive training, i.e. shrikage of variances of Gaussian components. Henceforth, the speaker space constructed by the proposed SAT protocol was investigated.

5.3 Objective Evaluations

5.3.1 Eigenvoice vs. Truncated EV

By using TFA, the footprint of the Eigenvoice framework can be reduced effectively. Figure 4 shows the average mel-cepstral distortion as a function of data compression rate in the truncated EV method. In the truncated EV method, the hyperparameters K_D, K_M were varied before adaptation, and the EV bases were reconstructed by tensor product $\mathcal{G} \times_1 \boldsymbol{U}^{(D)} \times_2 \boldsymbol{U}^{(M)}$. In Fig. 4, each point corresponds to each condition of truncation (K_D, K_M, K_S). The data compression rate is defined by $K_D K_M K_S / DMS$. From Fig. 4 (a), if the compression rate is ignored, Eigenvoice has achieved best performance, when the number of adaptation utterances is $N = 2$. If the low data compression rate is required, the original Eigenvoice could not achive the better performance without degradation. On the other hand, the proposed framework of truncated EV can flexibly control the data compression rate while keeping the conversion performance. In addition, from Fig. 4 (b), when the number of adaptation utterances is larger, the proposed framework balances the conversion performance and the data compression.

5.3.2 Feature Space Truncation in TFA-Feat

Figure 5 shows the result of TFA-feat method. The mel-cepstral distortion is shown as a function of the number of column vectors in $U^{(D)}(= K_D)$. From Fig. 5, even when bases matrix $U^{(D)}$ is tructaed to $K_D = 25$, the performance of conversion was maintained. This effect might be caused by concatenation of static and dynamic features. In GMM-based VC, dynamic features are used to capture interframe correlations. On the other hand, they would not strongly depend on speakers. The results of truncation in TFA-feat would reflect these properties.

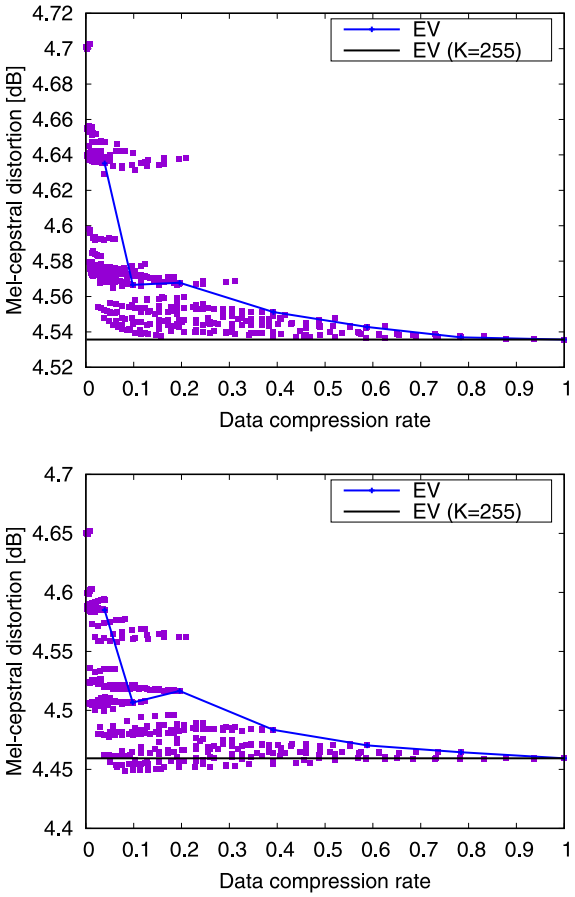


Fig. 4 Data compression effects of truncated EV.

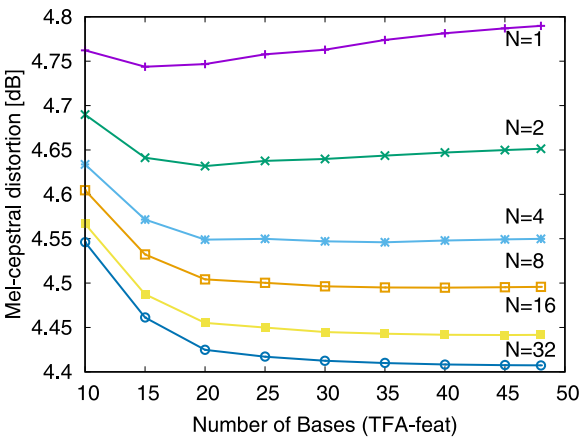


Fig. 5 Effect of feature space truncation in TFA-feat. The mel cepstral distortion is shown as a function of the number of column vectors in $U^{(D)}$.

5.3.3 Mixture Truncation in TFA-Mix

Figure 6 shows the results of TFA-mix method. The mel-cepstral distortion is shown as a function of the number of column vectors in $U^{(M)}(= K_M)$. From Fig. 6, reducing the bases matrix $U^{(M)}$ improves the conversion performance. Although slight reduction of the bases makes worse results

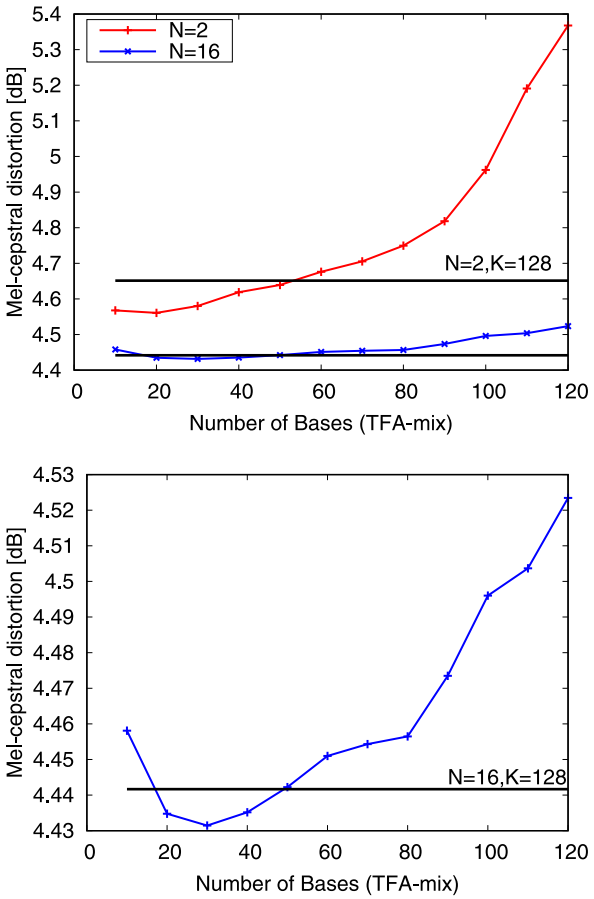


Fig. 6 Effect of feature space truncation in TFA-mix. The mel cepstral distortion is shown as a function of the number of column vectors in $U^{(M)}$.

of conversion, the proposed TFA-mix would effectively capture the essence of the speaker space.

5.3.4 Bilinear Adaptation

Figure 7 shows the results of TFA-bilinear method. The mel-cepstral distortions are depicted as heatmaps in Fig. 7. The indices in the figure correspond to the number of bases for mixtures (K_M) and that for features (K_D), respectively. From Fig. 7, TFA-bilinear involves both properties of TFA-mix and TFA-feat. Degradation caused by slight reduction of $U^{(M)}$ is observed around $K_M \in [100, 120]$. Truncation of $U^{(D)}$ in $N = 16$ is more sensitive than that in $N = 2$ when $K_D \leq 25$. Finally, optimal parameters (K_D, K_M) were (30,20) for $N = 2$ and (48,20) for $N = 16$, respectively.

5.3.5 Comprehensive Comparison

Figure 8 shows the result of average mel-cepstral for the test data as a function of the number of adaptation utterances. For each case, the optimal numbers of the bases parameters (K_*) were selected. Table 2 shows the optimal numbers for each of the methods. Compared with Eigenvoice, the performances of the TFA approaches are better when the number of adaptation utterances is larger than 8. This means that

Table 2 The optimal numbers of the bases parameters for each method.

# of utterances	1	2	4	8	16	32
Eigenvoice (K)	200	255	255	255	255	255
Truncated EV [$K_D = 25, K_M = 20$] (K_s)	100	200	255	255	255	255
TFA-mix (K_M)	10	20	20	30	30	128
TFA-feat (K_D)	15	20	35	40	45	48
TFA-bilinear (K_D, K_M)	(20,20)	(30,20)	(48,20)	(48,20)	(48,20)	(48,128)

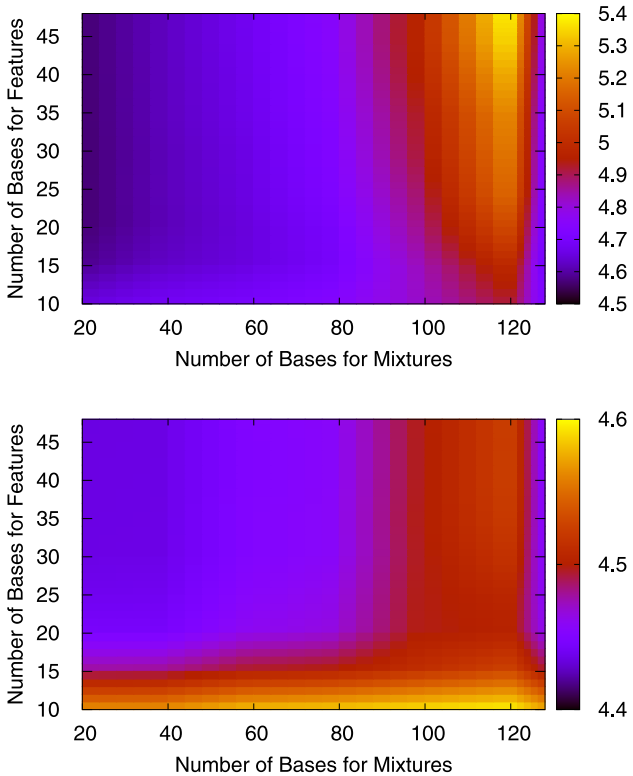


Fig. 7 Results of objective evaluations for TFA-bilinear.

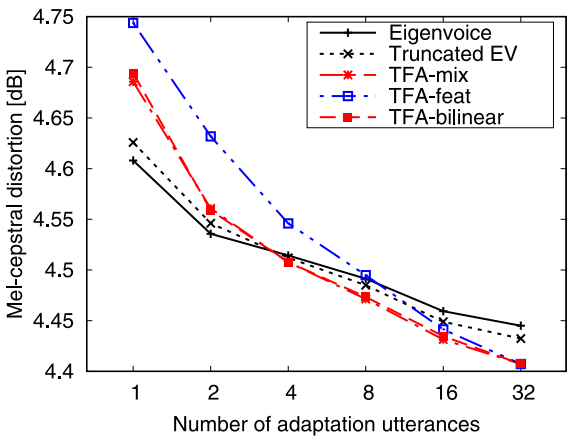


Fig. 8 Results of objective evaluations by average mel-cepstral distortion (MCD).

the TFA-based approaches correctly are scaled out for the increase of the number of adaptation data. Comparing TFA-mix with TFA-bilinear, we can observe that they achieved the similar performances to each other. On the other hand, TFA-feat is much more sensitive to the variation of the number of adaptation utterances. This would mean that the base matrix $U^{(M)}$ is more important than the base matrix $U^{(D)}$ because it captures the correlation of Gaussian components in GMM. Compared with Eigenvoice, truncated EV also achieved the similar performance to that of original Eigenvoice. Note that the data compression rate of the truncated EV in this comparison is about 3.2% to 8.1%. It can be said that the proposed scheme balances the conversion performance and the data compression. When the number of adaptation utterances is small such as 1 or 2, Eigenvoice achieved better performance. This is caused by the large footprint as the prior knowledge ($DMK_s \sim 1.5 \times 10^6$). On the other hand, in spite of the small footprint of the models ($MK_M \sim 1200$, $DK_D \sim 700$), TFA approaches make good results. This means that, some important knowledge could be captured by $\boldsymbol{U}^{(M)}$ and $\boldsymbol{U}^{(D)}$. It might be said that our proposed approach effectively captures the essence of the speaker space with the small footprint.

5.4 Subjective Evaluations

5.4.1 Overview

Listening tests were carried out to evaluate the naturalness of converted speech. 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 more natural. Each paired test was conducted with at least 25 subjects, which they were collected by a crowdsource system. The number of sample pairs evaluated by each subject was 10 in each test. As the samples for the subjective evaluation, the converted utterances of two target speakers (one male and one female) were selected. They were selected based on the results of objective evaluations.

5.4.2 Eigenvoice vs. Truncated EV

Figure 9 shows the results of comparison between Eigenvoice and truncated EV. From Fig. 9, it can be observed that truncated EV achieved comparable performances to that achieved by Eigenvoice. In addition, the truncated EV reduced the footprint to 8.1% of the original Eigenvoice. It can be said that TFA effectively captures the essence of the model.

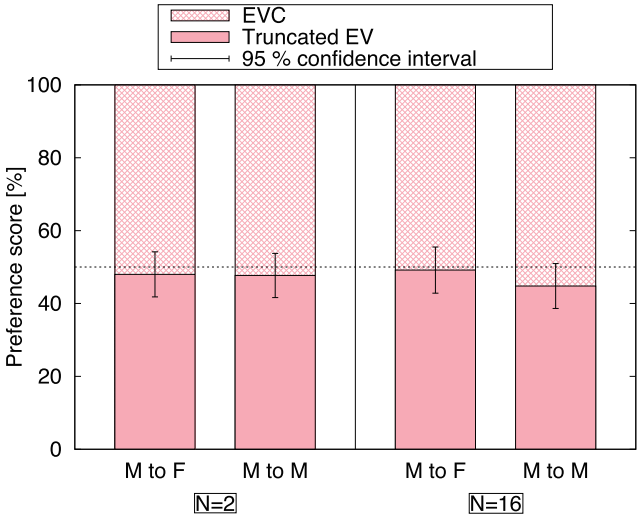


Fig. 9 Results of subjective evaluations between Eigenvoice and truncated EV. The number N means the number of adaptation utterances.

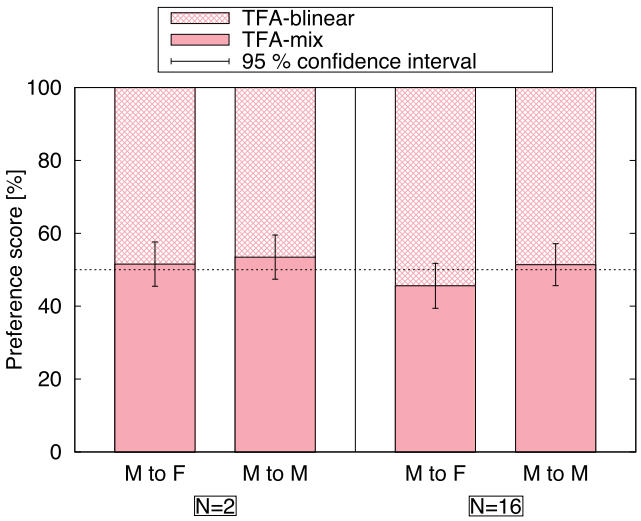


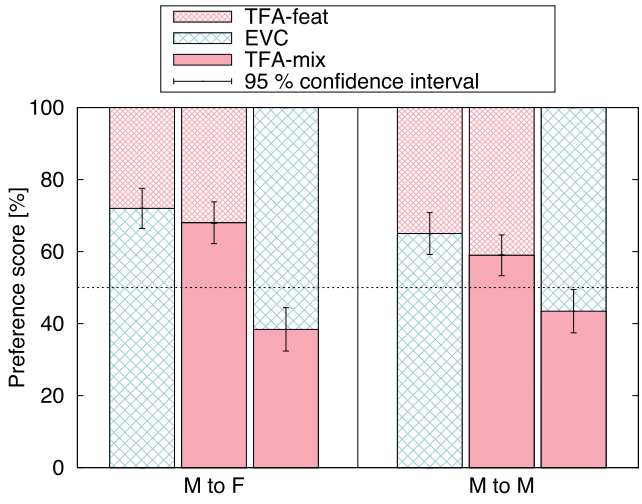
Fig. 10 Results of subjective evaluations between TFA-mix and TFA-bilinear. The number N means the number of adaptation utterances.

5.4.3 TFA-Mix vs. TFA-Bilinear

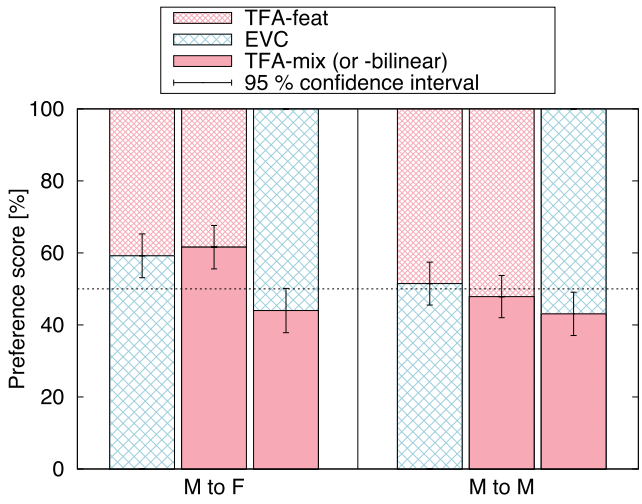
Figure 10 shows the results of comparison between TFA-mix and TFA-bilinear. Eigenvoice and truncated EV. From Fig. 10, TFA-mix and TFA-bilinear were comparable to each other. According to the results, the base matrix $\boldsymbol{U}^{(M)}$ would be basically a dominant factor for adaptation. Although the achieved performance by TFA-bilinear was comparable to that by TFA-mix, TFA-blinear has possibility to reduce the number of adaptation parameters from DK_M in TFA-mix to $K_D K_M$ by considering the base matrix $\boldsymbol{U}^{(D)}$.

5.4.4 Comprehensive Comparison

Following the previous subjective evaluations, we conducted a comprehensive comparison among TFA approaches. Although the observed differences are small in Figs. 9 and 10, one method were selected from each flavor, i.e. an EV flavor and a TFA-mix flavor. Based on Fig. 9, original EVC was selected as an Eigenvoice-flavored method. From Fig. 10, TFA-mix was selected as a TFA-mix-flavored method except for the condition of '$N = 16$ and M to F'. TFA-bilinear was selected for that condition. In addition to them, TFA-feat was added as the third competitor.



(a) $N = 2$



(b) $N = 16$

Fig. 11 Results of subjective evaluations.

Figure 11 shows the results. When using two adaptation utterances (Fig. 11 (a)), "EVC" achieved the best performances both in male-to-male and male-to-female conversion. This would be caused by the effect of the larger footprint of EVC. Under this condition, "TFA-mix" outperformed "TFA-feat". This might be caused by the limited amount of occupation counts. When using 16 adaptation utterances (Fig. 11 (b)), performances of TFA-mix and TFA-feat were totally improved from $N = 2$. In male-to-male conversion, TFA-feat achieved a comparable performance to that of EVC, while the footprint of the methods is quite

smaller. Both the objective and subjective evaluations suggest that our proposed method works effectively with the much smaller footprint than that of original EVC.

6. 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 non-parallel training, or mixture of factor analyzers should be verified. The utilization of the core tensor $\mathcal{G}$ without increase of footprints is another further work. We also plan to apply the proposed scheme to neural-network-based approaches.

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. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," IEEE Trans. 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. 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] D. Saito, S. Watanabe, A. Nakamura, and N. Minematsu, "Statistical voice converison based on noisy channel model," IEEE Trans. Speech and Audio Processing, vol.20, no.6, pp.1784–1794, 2012.
[10] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on Gaussian mixture model," Proc. INTERSPEECH, pp.2446–2449, 2006.
[11] R. Kuhn, J-C. 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.
[12] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," IEEE Trans. Audio, Speech, and Language Processing, vol.16, no.5, pp.980–988, 2008.
[13] Z. Wu, T. Kinnunnen, E.S. Chng, and H. Li, "Mixture of factor analyzers using priors non-parallel speech for voice converison," IEEE Signal Processing letters, vol.19, no.12, pp.914–917, 2012.
[14] M.A.O. Vasilescu and D. Terzopoulos, "Mutilinear analysis of image ensembles: TensorFaces," Proc. ECCV, vol.2350, pp.447–460, 2002.
[15] 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.
[16] 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.
[17] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Non-parallel training for many-to-many eigenvoice conversion," Proc. ICASSP, pp.4822–4825, 2010.
[18] T. Hashimoto, D. Saito, and N. Minematsu, "Many-to-Many and Completely Parallel-Data-Free Voice Conversion Based on Eigenspace DNN," IEEE/ACM Transaction on Audio, Speech, and Language Processing, vol.27, no.2, pp.332–341, 2019.
[19] 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.
[20] T. Toda, A.W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," IEEE Trans. Audio, Speech, and Language Processing, vol.15, no.8, pp.2222–2235, 2007.
[21] 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.
[22] L.R. Tucker, "Some mathematical notes on three-mode factor analysis," Psychometrika, vol.31, no.3, pp.279–311, 1966.
[23] Y. Jeong, "Speaker adaptation based on the multilinear decomposition of training speaker models," Proc. ICASSP, pp.4870–4873, 2010.
[24] T. Anastasakos, J. McDonough, R. Schwarts, and J. Makhoul, "A compact model for speaker adaptive training," Proc. ICSLP, vol.2, pp.1137–1140, 1996.
[25] D. Saito, N. Minematsu, and K. Hirose, "Effects of speaker adaptive training on tensor-based arbitrary speaker conversion," Proc. INTERSPEECH, pp.98–101, 2012.
[26] 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, no.4, pp.357–363, 1990.
[27] "JNAS: Japanese newspaper article sentences," http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html
[28] 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.

Daisuke Saito received the B.E., M.S., and Dr. Eng. degrees from the University of Tokyo, Tokyo, Japan, in 2006, 2008, and 2011, respectively. From 2010 to 2011, he was a Research Fellow (DC2) of the Japan Society for the Promotion of Science. He is currently a lecturer (senior assistant professor) in the Graduate School of Engineering, University of Tokyo. He is interested in various areas of speech engineering, including voice conversion, speech synthesis, acoustic analysis, speaker recognition, and speech recognition. Dr. Saito is a member of the International Speech Communication Association (ISCA), the Acoustical Society of Japan (ASJ), the Information Processing Society of Japan (IPSJ), the Institute of Electronics, Information and Communication Engineers (IEICE), and the Institute of Image Information and Television Engineers (ITE). He received the ISCA Award for the best student paper of INTERSPEECH 2011, the Awaya Award from the ASJ in 2012, and the Itakura Award from ASJ in 2014.

Nobuaki Minematsu was born in Hyogo, Japan in 1966. He received the doctor in Engineering from the University of Tokyo in 1995. From 1995 to 2000, he was a research associate at Toyohashi University of Technology. From 2000, he was an associate professor of the University of Tokyo and since 2012, he has been a full professor there. He has a wide interest in speech communication covering the areas of speech science and speech engineering, especially he has expert knowledge on Computer-Aided Language Learning (CALL). He received paper awards from RISP, JSAI, ICIST, O-COCOSDA, IEICE in 2005, 2007, 2011, 2014, and 2016 and received an encouragement award from PSJ in 2014. He was a distinguished lecturer of APSIPA from 2015 to 2016. He is a member of IEEE, ISCA, SLaTE, IPA, APSIPA, IEICE, IPSJ, ASJ, PSJ, etc.

Keikichi Hirose received his B.E. degree in electrical engineering in 1972 and his Ph.D. degree in electronic engineering in 1977 from the University of Tokyo, Tokyo, Japan. In 1977, he joined the University of Tokyo as a Lecturer in the Department of Electrical Engineering and in 1994 became a Professor in the Dept. of Electronic Engineering. From 1996, he was a Professor at the Graduate School of Engineering, Department of Information and Communication Engineering, University of Tokyo. In 1999, he moved to the University's Graduate School of Frontier Sciences (Department of Frontier Informatics), and again moved to Graduate School of Information Science and Technology (Department of Information and Communication Engineering) in 2004. From March 1987 to January 1988, he was a Visiting Scientist of the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, U.S.A. He is a professor emeritus of the University of Tokyo. His research interests widely cover spoken language information processing. He led a project "Realization of advanced spoken language information processing from prosodic feature" as part of Scientific Research on Priority Areas, Grant-in-Aid on Scientific Research, Ministry of Education, Culture, Sports, Science and Technology, Japan. He is a member of the Institute of Electrical and Electronics Engineers, the Acoustical Society of America, the International Speech Communication Association, the Institute of Electronics, Information and Communication Engineers (Fellow), the Acoustical Society of Japan, and other professional organizations. He received a 2007 paper award from the Research Institute of Signal Processing.