


RESEARCH

Open Access



Non-parallel dictionary learning for voice conversion using non-negative Tucker decomposition

Yuki Takashima^{1*} , Toru Nakashika², Tetsuya Takiguchi¹ and Yasuo Ariki¹

Abstract

Voice conversion (VC) is a technique of exclusively converting speaker-specific information in the source speech while preserving the associated phonemic information. Non-negative matrix factorization (NMF)-based VC has been widely researched because of the natural-sounding voice it achieves when compared with conventional Gaussian mixture model-based VC. In conventional NMF-VC, models are trained using parallel data which results in the speech data requiring elaborate pre-processing to generate parallel data. NMF-VC also tends to be an extensive model as this method has several parallel exemplars for the dictionary matrix, leading to a high computational cost. In this study, an innovative parallel dictionary-learning method using non-negative Tucker decomposition (NTD) is proposed. The proposed method uses tensor decomposition and decomposes an input observation into a set of mode matrices and one core tensor. The proposed NTD-based dictionary-learning method estimates the dictionary matrix for NMF-VC without using parallel data. The experimental results show that the proposed method outperforms other methods in both parallel and non-parallel settings.

Keywords: Voice conversion, Non-negative Tucker decomposition, Non-negative matrix factorization, Non-parallel training

1 Introduction

Voice conversion (VC) is a technique used to convert speaker-specific information in the speech of a source speaker into that of a target speaker while retaining linguistic information. Lately, VC techniques have been garnering particular attention [1], and various statistical approaches to VC have been studied [2, 3] as these techniques can be applied to numerous tasks [4–8]. Of these approaches, the Gaussian mixture model (GMM)-based mapping method [9] is the most prevalent, and a number of enhancements have been proposed [10–12]. Other VC methods, such as approaches based on non-negative matrix factorization (NMF) [13–15], neural networks [16], deep learning [17, 18], restricted Boltzmann machines [19–21], variational autoencoders [22], and a generative adversarial network [23], have also been

proposed. Notably, in recent years, the NMF has outperformed GMM in parallel data conditions. Exemplar-based NMF-VC retains the high naturality of the converted speech, and many of its variants have been proposed [24, 25]. Although more recent deep learning methods require significantly large training data, NMF-VC requires comparatively less training data. Therefore, this study focuses on NMF-VC.

NMF [26] is one of the most popular sparse representation methods. The goal of NMF is to decompose the input observation into two matrices: the basis matrix and weight matrix. In this study, the basis matrix is referred to as the “dictionary,” and the weight matrix as the “activity.” The NMF-based method can be classified into two approaches: the dictionary-learning approach [14] and exemplar-based approach [27]. In the dictionary-learning approach, the dictionary and activity are estimated simultaneously during the training, and the estimated dictionary is used in conversion. However, in the exemplar-based approach, the training data is straight-away used as exemplars in the conversion step. By using

*Correspondence: takashima@stu.kobe-u.ac.jp

¹Graduate School of System Informatics, Kobe University, 1-1 Rokkodai, Nada-ku, Kobe 657-8501, Japan

Full list of author information is available at the end of the article

the learned dictionary instead of the exemplars, the VC is executed with lower computation times.

However, both the NMF-based approaches require parallel data (aligned speech data from the source and the target speakers, so that each frame of the source speaker's data corresponds to that of the target speaker's data) for training the models, which leads to several problems. First, the data are limited to predefined statements (both speakers must utter the same statements). Second, the training data (the parallel data) are not the original speech data anymore, as the speech data are stretched and modified along the time axis when aligned, and there is no certainty that each frame is aligned perfectly. As the dictionary is assembled from parallel data, the error of alignment in the parallel data might adversely affect VC performance. Several other approaches have been proposed that do not use (or minimally use) parallel data of the source and the target speakers [28–30]. For example, in [28], the spectral relationships between two arbitrary speakers (reference speakers) is modeled using GMMs and the source speaker's speech is converted using the matrix that projects the feature space of the source speaker into that of the target speaker through that of the reference speakers. In this study, the conventional NMF-based VC method is expanded into a non-parallel VC method. A previous study [30] proposed using the phone segmentation results from automatic speech recognition to construct a sub-dictionary for each phone for an exemplar-based NMF voice conversion. This particular technique was applied to the non-parallel VC.

To tackle the non-parallel approach, a non-negative Tucker decomposition (NTD) [31–33]-based dictionary-learning method is proposed. The NTD is a non-negative extension of the Tucker decomposition that decomposes the input observation into a set of matrices and one core tensor. Tucker decomposition is generally introduced to deal with a high-order tensor. In recent studies, Tucker decomposition has been widely applied in visual question-answering systems [34] and speech recognition [35]. As spectral features are used for input observation, a set of matrices consists of two mode matrices for frequency and time and a core tensor corresponding to a core matrix. It is assumed that these matrices correspond to the frequency basis matrix, the phonemic information, and a codebook between the frequency basis and each phone, respectively. In the proposed approach, the activity matrix in NMF is decomposed into the codebook and the phonemic information. When learning the dictionaries, while the activity matrix is shared between speakers using parallel data in the conventional NMF-VC, in the proposed method, the codebook is shared between speakers, and the phonemic information is dependent on a speaker. Hence, the time-varying phonemic information can be captured for each speaker. During the conversion, only the phonemic

information matrix is estimated as the activity matrix. As the proposed method can have time-dependent factors for each speaker, there is no necessity for parallel data. To the best of authors' knowledge, NTD-based VC has not been attempted, except [36] where Tucker decomposition was used to represent the speaker space and the conversion mechanism was based on GMM. The present VC is based on NMF, and this approach is fundamentally different from those presented previously [36].

Several methods have been proposed for tensor decomposition [37–39]. In [37], NMF is applied to variational Bayesian matrix factorization, where each observed entry is assumed to be a beta distribution. Shi et al. [38] proposed tensor decomposition with variance maximization for feature extraction. In [39], pairwise similarity information is incorporated into Tucker tensor decomposition. While these methods have useful properties, it is difficult to adapt them directly to VC. NTD can be readily integrated with NMF-based VC, because NMF is the second-order case of the Tucker decomposition with the non-negative constraint.

The rest of this paper is organized as follows. In Section 2, a conventional NMF-based VC is described. Section 3 includes the description of the proposed method. Section 4 details the evaluation of the experimental data, and Section 5 details the Experiments on VCC 2018. Finally, in Section 6, the conclusions are presented.

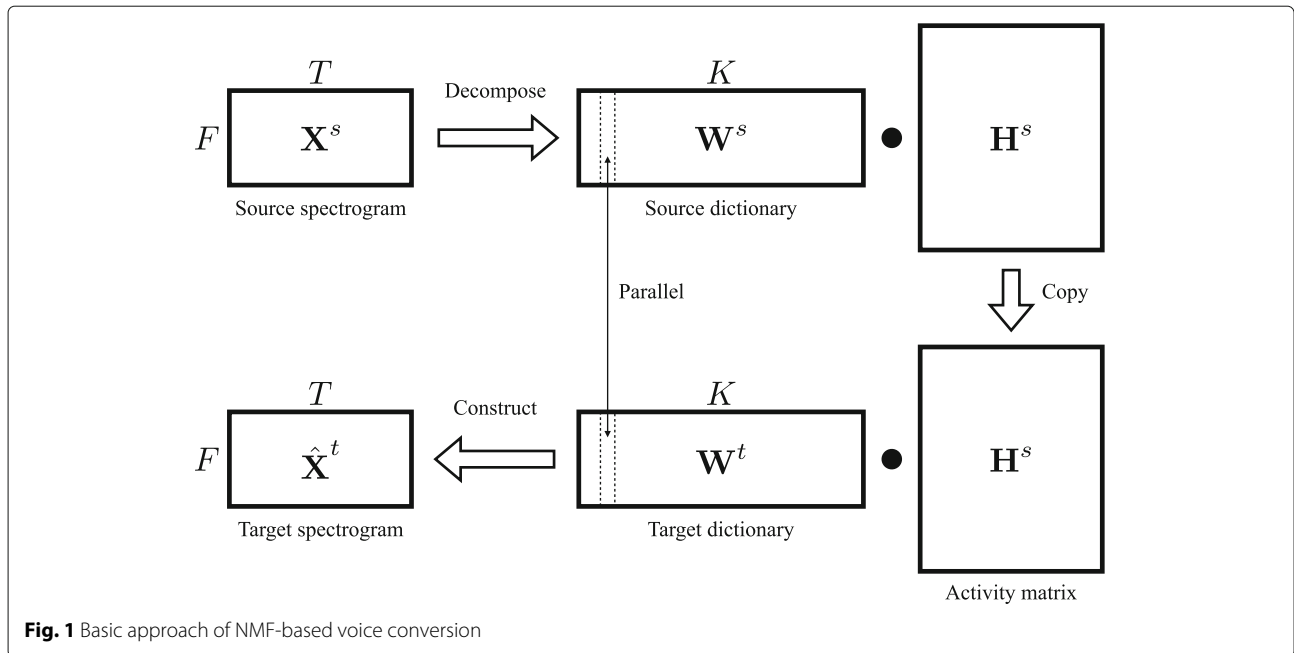
2 NMF-based voice conversion

NMF is a matrix decomposition method under non-negative constraints. The basic idea behind decomposing a matrix $\mathbf{X} \in \mathbb{R}^{F \times T}$ is to find two matrices $\mathbf{W} \in \mathbb{R}^{F \times K}$ and $\mathbf{H} \in \mathbb{R}^{K \times T}$ that minimize the distance between \mathbf{X} and \mathbf{WH} under non-negative constraints. F and T represent the number of dimensions and frames. In NMF, \mathbf{W} is called a basis matrix and contains K bases in columns. \mathbf{H} is called an activity matrix and indicates the activity of each basis along the time index.

VC approaches using NMF are divided into two categories: supervised and unsupervised approaches. The supervised approach, known as the exemplar-based VC, estimates only the activity from observation and the dictionary must be provided. However, the unsupervised approach, i.e., the dictionary-learning VC, estimates both the dictionary and the activity from observation. The proposed method is based on the latter, i.e., the dictionary-learning approach.

2.1 Dictionary learning using nMF

Figure 1 shows the basic approach of the dictionary-learning NMF-based VC [14], where F , T , and K represent the number of dimensions, frames, and bases, respectively. This VC method needs two dictionaries that are phonemically parallel. $\mathbf{W}^s \in \mathbb{R}^{F \times K}$ represents a source



dictionary, and $\mathbf{W}^t \in \mathbb{R}^{F \times K}$ represents a target dictionary. In exemplar-based VC, these two dictionaries consist of the same words or sentences and are aligned with dynamic time warping (DTW), which is comparable with the conventional GMM-based VC. In dictionary-learning VC, these two dictionaries are estimated simultaneously and as a result have the same number of bases.

For the training source speaker data $\mathbf{X}^s \in \mathbb{R}^{F \times T}$ and the training target speaker data $\mathbf{X}^t \in \mathbb{R}^{F \times T}$, two dictionaries \mathbf{W}^s , \mathbf{W}^t , and the activity $\mathbf{H} \in \mathbb{R}^{K \times T}$ are simultaneously estimated. The cost function of this joint NMF is defined as follows:

$$d_{KL}(\mathbf{X}^s, \mathbf{W}^s \mathbf{H}) + d_{KL}(\mathbf{X}^t, \mathbf{W}^t \mathbf{H}) + \lambda \|\mathbf{H}\|_1$$

$$s.t. \mathbf{W}^s, \mathbf{W}^t, \mathbf{H} \geq 0, \quad (1)$$

where \mathbf{X}^s and \mathbf{X}^t represent parallel data. In Eq. (1), $d_{KL}(\mathbf{A}, \mathbf{B})$ denotes the Kullback-Leibler divergence between the two matrices \mathbf{A} and \mathbf{B} , and the last term is the sparsity constraint with the L1-norm regularization term that causes the activity matrix to be sparse. λ represents the weight of the sparsity constraint. This function is minimized by iteratively updating parameters, as is done in the traditional NMF.

This method assumes that when the source and the target spectra (which are from the same words but spoken by different speakers) are expressed with sparse representations of the source dictionary and the target dictionary, respectively, the obtained activity matrices are approximately equivalent to each other. In the conversion process, for the input source spectrogram \mathbf{X}^s , only the activity \mathbf{H}^s is estimated while fixing the source dictionary \mathbf{W}^s .

The estimated source activity \mathbf{H}^s is multiplied with the target dictionary \mathbf{W}^t , and the target spectrogram $\hat{\mathbf{X}}^t$ is constructed as follows:

$$\hat{\mathbf{X}}^t = \mathbf{W}^t \mathbf{H}^s. \quad (2)$$

2.2 Problems

NMF-based VC has several problems. First, if the source and target utterances are aligned using DTW in advance, the estimated parameters are affected by the quality of the alignment. And a mismatch of alignment appears to persist. Aihara et al. [24] have shown that this mismatch degrades the performance of exemplar-based VC. Second, it appears that the activity matrix contains other information along with the phonetic information. Aihara et al. [25, 27] assumed that the activity matrix contains the phonetic information and speaker information, and accordingly proposed certain frameworks to overcome this effect, thereby improving the performance of NMF-based VC. In this study, an alternative approach is proposed. The activity matrix is decomposed into the speaker-shared matrix and the speaker-dependent phonetic information matrix. This decomposition makes parallel data unnecessary. Moreover, during the conversion, estimating only the phonetic information matrix as the activity matrix is expected to improve the accuracy of activity estimation.

3 Methods

3.1 NTD

Given a non-negative N-way tensor, NTD [40] decomposes the input tensor into a core tensor and a set of mode

matrices that are restricted to have only non-negative elements. In this study, as the spectral features are used as the input observation, a core tensor is represented as a matrix, and there are two mode matrices. Under these conditions, NTD is simply defined as follows:

$$\mathbf{X} \approx \mathbf{UGV}^T \text{ s.t. } \mathbf{U}, \mathbf{G}, \mathbf{V} \geq 0, \quad (3)$$

where $\mathbf{X} \in \mathbb{R}^{F \times T}$, $\mathbf{U} \in \mathbb{R}^{F \times M}$, $\mathbf{V} \in \mathbb{R}^{T \times L}$, and $\mathbf{G} \in \mathbb{R}^{M \times L}$ represent an input spectrogram, a mode matrix along the frequency axis, a mode matrix along the time axis, and a core matrix, respectively. F , T , M , and L indicate the number of frequency bins, frames, frequency basis, and time basis, respectively. The cost function of NTD is defined as follows:

$$d_{KL}(\mathbf{X}, \mathbf{UGV}^T). \quad (4)$$

NTD provides a general form of the non-negative tensor factorization including a special case of NMF; updating algorithms have been proposed in [40]. These updating algorithms are based on that NMF.

3.2 Dictionary learning using nTD

This section describes the method of estimating a parallel dictionary between the source and target speakers by NTD. The objective function is represented as follows:

$$\begin{aligned} \alpha d_{KL}(\mathbf{X}^s, \mathbf{U}^s \mathbf{G} \mathbf{V}^{sT}) + \beta d_{KL}(\mathbf{X}^t, \mathbf{U}^t \mathbf{G} \mathbf{V}^{tT}) \\ + \lambda \|\mathbf{V}^{sT}\|_1 + \lambda \|\mathbf{V}^{tT}\|_1 \\ \text{s.t. } \mathbf{U}^s, \mathbf{U}^t, \mathbf{G}, \mathbf{V}^s, \mathbf{V}^t \geq 0, \end{aligned} \quad (5)$$

where $\mathbf{X}^s \in \mathbb{R}^{F \times T_s}$, $\mathbf{X}^t \in \mathbb{R}^{F \times T_t}$, $\mathbf{U}^s \in \mathbb{R}^{F \times M}$, $\mathbf{U}^t \in \mathbb{R}^{F \times M}$, $\mathbf{V}^s \in \mathbb{R}^{T_s \times L}$, $\mathbf{V}^t \in \mathbb{R}^{T_t \times L}$, and $\mathbf{G} \in \mathbb{R}^{M \times L}$ represent the source and target spectrograms, the source and target frequency basis matrices, the source and target time basis

matrices, and a core matrix, respectively. α and β represent the weight of each term. F , T_s , T_t , M , and L indicate the number of frequency bins, source and target frames, frequency basis, and time basis, respectively. This function is minimized by iteratively updating the following equations in the same manner as the NTD:

$$\mathbf{U}^s \leftarrow \mathbf{U}^s \cdot \left(\tilde{\mathbf{H}}^s (\mathbf{X}^s ./ \mathbf{U}^s \tilde{\mathbf{H}}^s)^T ./ \tilde{\mathbf{H}}^s \mathbf{1}^{(T_s \times F)} \right)^T \quad (6)$$

$$\mathbf{U}^t \leftarrow \mathbf{U}^t \cdot \left(\tilde{\mathbf{H}}^t (\mathbf{X}^t ./ \mathbf{U}^t \tilde{\mathbf{H}}^t)^T ./ \tilde{\mathbf{H}}^t \mathbf{1}^{(T_t \times F)} \right)^T \quad (7)$$

$$\begin{aligned} \mathbf{V}^s \leftarrow \mathbf{V}^s \cdot \left(\left(\mathbf{X}^s ./ \tilde{\mathbf{W}}^s \mathbf{V}^{sT} \right)^T \tilde{\mathbf{W}}^s \right) \\ ./ \left(\mathbf{1}^{(T_s \times F)} \tilde{\mathbf{W}}^s + \lambda \mathbf{1}^{(T_s \times L)} \right) \end{aligned} \quad (8)$$

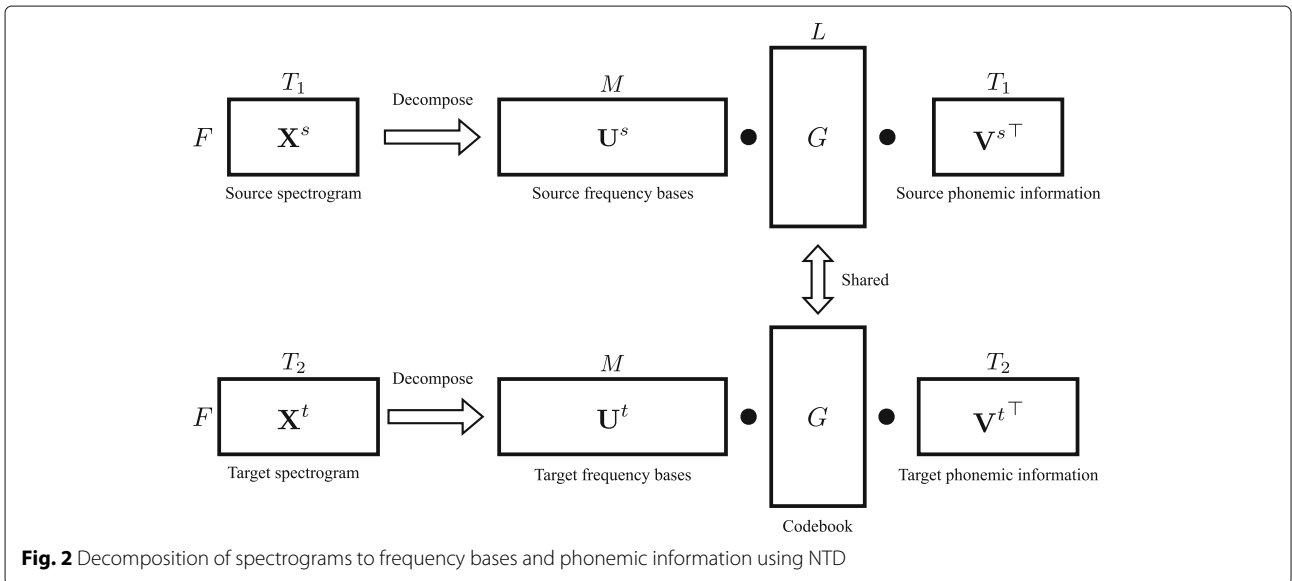
$$\begin{aligned} \mathbf{V}^t \leftarrow \mathbf{V}^t \cdot \left(\left(\mathbf{X}^t ./ \tilde{\mathbf{W}}^t \mathbf{V}^{tT} \right)^T \tilde{\mathbf{W}}^t \right) \\ ./ \left(\mathbf{1}^{(T_t \times F)} \tilde{\mathbf{W}}^t + \lambda \mathbf{1}^{(T_t \times L)} \right) \end{aligned} \quad (9)$$

$$\begin{aligned} \mathbf{G} \leftarrow \mathbf{G} \cdot \left(\mathbf{U}^{sT} (\mathbf{X}^s ./ \tilde{\mathbf{X}}^s) \mathbf{V}^s + \mathbf{U}^{tT} (\mathbf{X}^t ./ \tilde{\mathbf{X}}^t) \mathbf{V}^t \right) \\ ./ \left(\mathbf{U}^{sT} \mathbf{1}^{(F \times T_s)} \mathbf{V}^s + \mathbf{U}^{tT} \mathbf{1}^{(F \times T_t)} \mathbf{V}^t \right) \end{aligned} \quad (10)$$

$$\begin{aligned} \tilde{\mathbf{H}}^s = \mathbf{G} \mathbf{V}^{sT}, \tilde{\mathbf{H}}^t = \mathbf{G} \mathbf{V}^{tT}, \tilde{\mathbf{W}}^s = \mathbf{U}^s \mathbf{G}, \tilde{\mathbf{W}}^t = \mathbf{U}^t \mathbf{G}, \\ \tilde{\mathbf{X}}^s = \mathbf{U}^s \mathbf{G} \mathbf{V}^{sT}, \tilde{\mathbf{X}}^t = \mathbf{U}^t \mathbf{G} \mathbf{V}^{tT}, \end{aligned}$$

where \cdot and $./$ denote element-wise multiplication and division, respectively. In this framework, only a core matrix \mathbf{G} is shared, and time-varying matrices \mathbf{V}^s and \mathbf{V}^t are dependent on each speaker, as shown in Fig. 2. Therefore, there is no necessity for parallel data.

After each matrix in the model is estimated, the source and target parallel dictionaries are calculated as $\mathbf{U}^s \mathbf{G}$ and $\mathbf{U}^t \mathbf{G}$, respectively. During conversion, for the given source spectrogram \mathbf{X}^s , only \mathbf{V}^s is estimated as $\mathbf{X}^s = \mathbf{U}^s \mathbf{G} \mathbf{V}^{sT}$.



Then, the target spectrogram $\hat{\mathbf{X}}^t$ is obtained as $\hat{\mathbf{X}}^t = \mathbf{U}^t \mathbf{G} \mathbf{V}^s \mathbf{T}$.

It is assumed that \mathbf{U}^s and \mathbf{U}^t represent the frequency basis matrices, and \mathbf{V}^s and \mathbf{V}^t represent the phonemic information. As the core matrix is not dependent on either the frequency or the time, this matrix represents the codebook between the frequency bases and the phones. Based on this assumption, the core matrix makes a correspondence between frequency bases and phones. Specifically, there are L phones, and a spectrum of each phone is constructed using M frequency bases. Although the information contained in the activity matrix is not only the phonemic information, in conventional NMF-based approaches, the activity matrix is assumed to contain only the phonemic information. Therefore, the estimated activity is degraded. In contrast, the proposed NTD-based approach specifically decomposes the activity matrix into the speaker-shared information and the speaker-dependent phonemic information. Therefore, it is expected that the performance of the activity estimation will be improved during conversion.

4 Experimental evaluation

4.1 Conditions

The proposed VC technique was evaluated in a speaker-conversion task using clean speech data by comparing its results with the conventional GMM-based method [10], the conventional NMF-based dictionary-learning method [14], and an adaptive restricted Boltzmann machine (ARBM)-based method [20] that does not use parallel data. For the evaluation, voice samples of speech data stored in the ATR Japanese speech database [41] of three males and three females were used. The sampling rate was 16 kHz. A total of 45 sentences were used for training, and another 50 sentences were used for testing. Parallel data aligned using dynamic programming matching (DPM) was used to train the GMM-based and NMF-based methods. The proposed method and the ARBM-based method do not require parallel data. As training data, the same utterances were used for the source and the target speaker in the parallel setting, and completely different utterances for each speaker were used in the non-parallel setting.

Parameter initialization has a significant impact on the conversion performance. In this study, \mathbf{V}^s and \mathbf{V}^t are initialized randomly. Table 1 shows the initialization algorithm for \mathbf{U}^s , \mathbf{U}^t , and \mathbf{G} . In the parallel setting, the initialization is based on the NMF framework using parallel data calculated by the source and target training data. In the non-parallel setting, the initialization is based on the NMF and NTD frameworks. This initialization method uses an adaptive matrix [42]. Finally, initialized parameters are optimized by Eqs. (6) to (10).

Table 1 Algorithm for initializing parameters

Initializing in the parallel setting

- Set source and target parallel data \mathbf{X}_s and \mathbf{X}_t
- Optimize \mathbf{W}_s , \mathbf{W}_t , and \mathbf{H} minimizing $d_{KL}(\mathbf{X}_s, \mathbf{W}_s \mathbf{H}) + d_{KL}(\mathbf{X}_t, \mathbf{W}_t \mathbf{H})$
- Optimize \mathbf{U}_s , \mathbf{U}_t , and \mathbf{G} minimizing $d_{KL}(\mathbf{W}_s, \mathbf{U}_s \mathbf{G}) + d_{KL}(\mathbf{W}_t, \mathbf{U}_t \mathbf{G})$

Initializing in the non-parallel setting

- Set source training data \mathbf{X}_s
- Optimize \mathbf{W}_s and \mathbf{H}_s while minimizing $d_{KL}(\mathbf{X}_s, \mathbf{W}_s \mathbf{H}_s)$
- Set target training data \mathbf{X}_t
- Optimize \mathbf{A} and \mathbf{H}_t while minimizing $d_{KL}(\mathbf{X}_t, \mathbf{A} \mathbf{W}_s \mathbf{H}_t)$ while fixing \mathbf{W}_s
- Optimize \mathbf{U}_s , \mathbf{U}_t , and \mathbf{G} while minimizing $d_{KL}(\mathbf{W}_s, \mathbf{U}_s \mathbf{G}) + d_{KL}(\mathbf{A} \mathbf{W}_s, \mathbf{U}_t \mathbf{G})$

In the conventional NMF-based method and the proposed method, a 513-dimensional WORLD spectrum [43] is used for spectral features. The hyperparameters α and β are used to control the length of the training data for the source and the target speaker, respectively. These parameters were set as follows:

$$\alpha = \min(T_s, T_t) / T_s \quad (11)$$

$$\beta = \min(T_s, T_t) / T_t \quad (12)$$

where T_s and T_t represent the number of frames of source and target training data, respectively. The sparse constraint λ was set to 0.2. The parameters are updated until the convergence condition $|F_t - F_{t-1}| < \epsilon |F_T|$ is fulfilled, where $|F_t|$ indicates a value of an objective function at an iteration t . ϵ was set to $\exp(-9)$. The GMM experiments were implemented using sprocket [44]. In the conventional NMF-based dictionary-learning method, the number of bases is 1000. In the ARBM-based method, a 32-dimensional Mel-cepstrum that was calculated from the 513-dimensional WORLD spectra was used as an input vector. Softmax constraints were set to hidden units.

In this study, a conventional linear regression based on the mean and standard deviation [10] was used to convert F0 information. Other information, such as aperiodic components, was synthesized without any conversion.

The proposed method was evaluated both objectively and subjectively. Mel-cepstral distortion (MCD) [dB] was used as a measure of the objective evaluations, defined as follows:

$$\text{MCD} = (10 / \ln 10) \sqrt{2 \sum_{d=1}^{24} (mc_d^{\text{conv}} - mc_d^{\text{tar}})^2} \quad (13)$$

where mc_d^{conv} and mc_d^{tar} represent the d th dimension of the converted and target Mel-cepstral coefficients, respectively.

The subjective evaluation was based on “speech quality” and “similarity to the target speaker (individuality).” In the subjective evaluation, 25 sentences were evaluated by 10

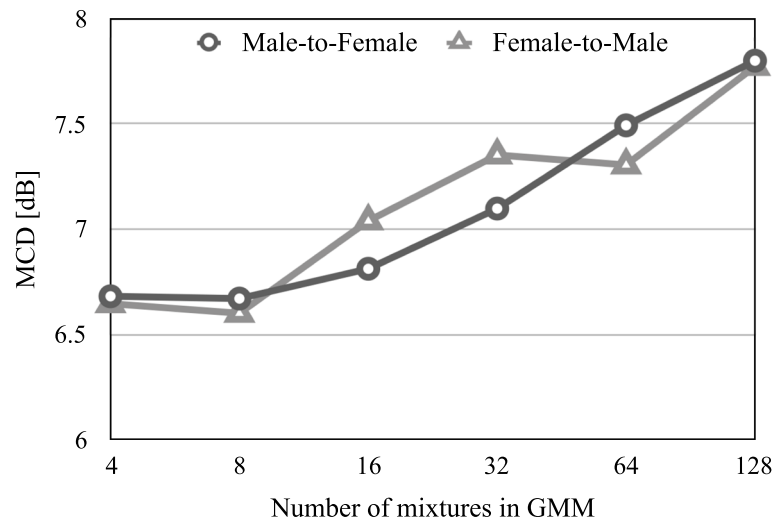


Fig. 3 Average MCD [dB] of the conventional GMM-based VC when varying the number of mixtures

native Japanese speakers. To evaluate the speech quality, a mean opinion score (MOS) test was performed. The opinion score was set to a 5-point scale (5, excellent; 4, good; 3, fair; 2, poor; 1, bad). For the similarity evaluation, an XAB test was conducted, in which each participant listened to the voice of the target speaker and then to the voice converted using the two methods. The participant was then asked to judge which sample sounded most similar to the target speaker's voice.

4.2 Parameters

The performance of each method was evaluated using different parameters. One male speaker and one female

speaker were selected and male-to-female conversion and female-to-male conversion was evaluated.

First, the performance of the conventional GMM-based VC was evaluated using different number of mixtures. The results obtained when using 4, 8, 16, 32, 64, and 128 mixtures are shown in Fig. 3. A lower value indicates a better result. As shown in Fig. 3, the optimal numbers were close to 8. Therefore, eight mixtures were used in the subsequent experiments.

Next, the performance of the conventional ARBM-based VC was evaluated using a different number of hidden units. The results are shown in Fig. 4 when using 2, 4, 8, 16, 32, and 48 hidden units. As shown in Fig. 4,

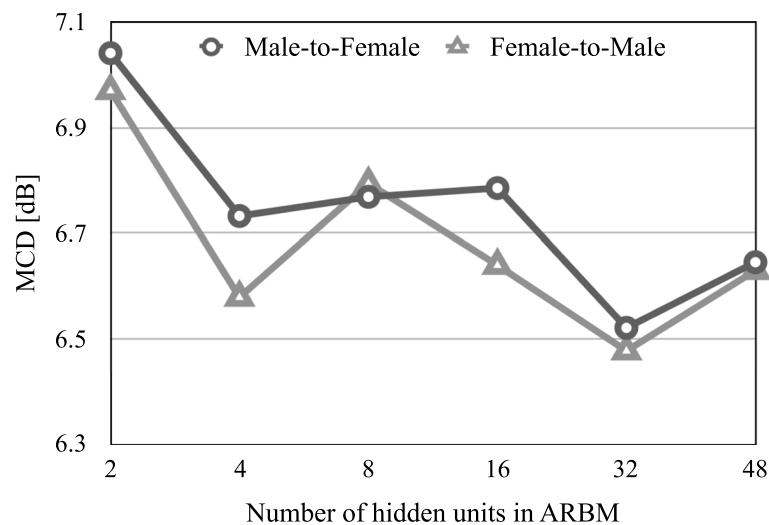


Fig. 4 Average MCD [dB] of the conventional ARBM-based VC when varying the number of hidden units

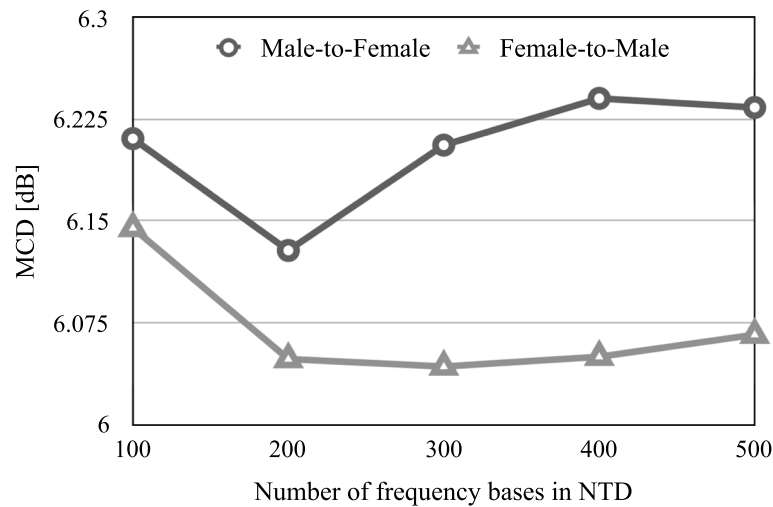


Fig. 5 Average MCD [dB] of the conventional NTD-based VC when varying the number of frequency bases

the optimal number was around 32. Therefore, 32 hidden units were used in the later experiments.

Finally, the performance of the proposed method was evaluated using a different number of frequency bases. The results are shown in Fig. 5 when the numbers of frequency bases M were 100, 200, 300, 400, and 500. The optimal number was around 200. Therefore, 200 was used as the number of frequency bases in the subsequent experiments. In the experiments, to control the number of dictionary bases during conversion, the number of time bases L was fixed to 1000.

4.3 Results

In this section, the proposed method is compared with conventional GMM, NMF, and ARBM-based methods.

Initially, the proposed method is compared with the parallel method in a parallel setting. Table 2 shows the average MCD values for male-to-female conversion, female-to-male conversion, male-to-male conversion, and female-to-female conversion. In this table, “ M_i ” and “ F_j ” indicate the i th male speaker and j th female speaker, respectively, and $\text{src} \rightarrow \text{tar}$ denotes the src -to- tar conversion. The rightmost column in the table indicates the mean value for each method with a 95% confidence interval. Here, a lower value indicates a better result. In these experiments, the

models were trained using parallel utterances. The GMM and NMF frameworks require parallel data. For these, parallel utterances were used to calculate the parallel data. Table 2 clearly demonstrates that the proposed NTD-based dictionary learning is not affected by the alignment error in DTW, and hence yields 10.1% and 1.8% relative improvements when compared with the conventional GMM-based method and the conventional NMF-based dictionary learning, respectively. Moreover, it confirms that the proposed method achieved a significantly better score than both the comparative methods, when using a p value test of 0.05.

Next, the method was compared with the non-parallel method in a non-parallel setting. Table 3 shows the average MCD values for male-to-female conversion, female-to-male conversion, male-to-male conversion, and female-to-female conversion. These results show that the proposed method has a comparable performance to the conventional non-parallel method, ARBM. However, the proposed method achieved a notably worse score than the ARBM-based method, when using a p value test of 0.05. This difference is explained in the next section.

Figure 6 shows the results of the MOS test on speech quality. The error bar shows a 95% confidence interval. Here, a higher value indicates a better

Table 2 Average MCD [dB] using parallel utterances

	Male-to-female			Female-to-male			Male-to-male		Female-to-female		Mean
	$M_1 \rightarrow F_1$	$M_2 \rightarrow F_2$	$M_3 \rightarrow F_3$	$F_1 \rightarrow M_1$	$F_2 \rightarrow M_2$	$F_3 \rightarrow M_3$	$M_1 \rightarrow M_3$	$M_3 \rightarrow M_2$	$F_1 \rightarrow F_3$	$F_3 \rightarrow F_2$	
GMM	6.67	7.35	6.76	6.60	6.97	7.04	6.41	7.36	6.42	7.21	6.88 ± 0.04
NMF	6.24	6.62	6.32	6.14	6.34	6.23	6.20	6.68	6.11	6.08	6.30 ± 0.03
NTD	6.12	6.50	6.31	6.04	6.23	6.08	6.05	6.66	5.99	5.86	6.19 ± 0.03

Table 3 Average MCD [dB] using non-parallel data

	Male-to-female			Female-to-male			Male-to-male		Female-to-female		Mean
	M1→F1	M2→F2	M3→F3	F1→M1	F2→M2	F3→M3	M1→M3	M3→M2	F1→F3	F3→F2	
ARBM	6.52	6.69	6.27	6.48	6.76	6.62	6.98	7.04	6.37	6.21	6.59 ± 0.03
NTD	6.75	7.30	6.56	6.75	7.04	6.99	6.85	7.04	6.64	6.03	6.67 ± 0.04

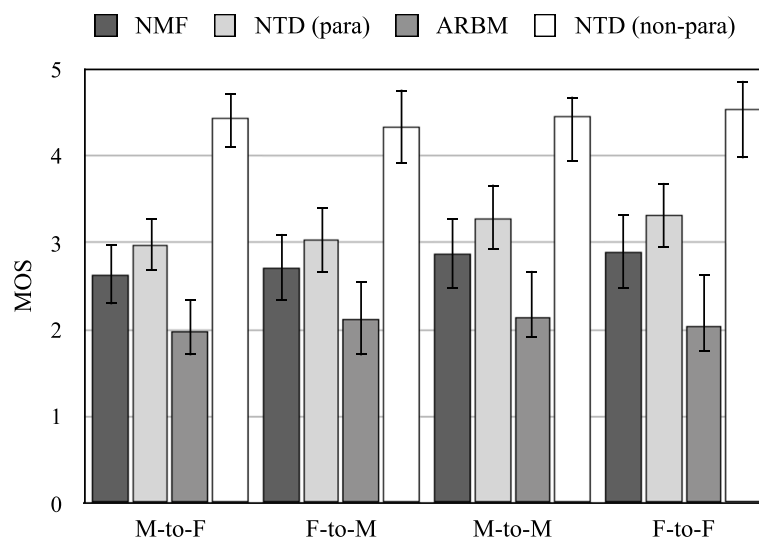
result. M-to-F, F-to-M, M-to-M, and F-to-F denote male-to-female conversion, female-to-male conversion, male-to-male conversion, and female-to-female conversion, respectively. “NTD (para)” and “NTD (non-para)” denote the proposed method with parallel utterances training and non-parallel utterances training, respectively. The proposed method achieved a significantly better score than the conventional methods. Specifically, NTD with the non-parallel setting showed the best results across all conversions.

Figures 7 and 8 show the results of the XAB test. The error bar shows a 95% confidence interval. For this test, a higher value indicates a better result. In Fig. 7, the results of the proposed method and conventional NMF-based dictionary-learning method are compared. In the male-to-female and female-to-female conversions, the proposed method achieved a better score than NMF-based dictionary learning. In the male-to-male and female-to-male conversions, the proposed method achieved a lower score than NMF-based dictionary learning. However, the difference between the two methods is not statistically significant, because $p > 0.3$ in the p value test. The proposed NTD-based dictionary learning without calculating parallel data showed comparable performance to the conventional NMF-based dictionary learning, which requires parallel data. In Fig. 8, the results of the proposed method

and the ARBM-based VC are compared. In conversions to male, the proposed method achieved a better score than ARBM-based VC. In conversions to female, the proposed method achieved a lower score than ARBM-based VC. In only the male-to-female conversion, the difference was significant — $p < 0.05$. However, in other conversions, the difference was not statistically significant. These tests show that the proposed non-parallel VC approach effectively converts the individuality of the source speaker’s voice to the target speaker’s voice while preserving high speech quality.

4.4 Discussion

In the objective evaluations, the proposed method achieved a better MCD value than the conventional VC, which uses parallel data. This is due to the fact that the proposed method is not affected by the mismatch of DPM. Moreover, the proposed NTD-based method yielded better performance, although the number of learned parameters decreased by approximately 60% of the conventional NMF-based one. This result indicates that the proposed dictionary learning has better spectral representation while keeping the number of bases of dictionaries constant during conversion. In addition, the average difference in MCD between the proposed method and the ARBM-based method was approximately 0.08 dB. This

**Fig. 6** MOS of speech quality

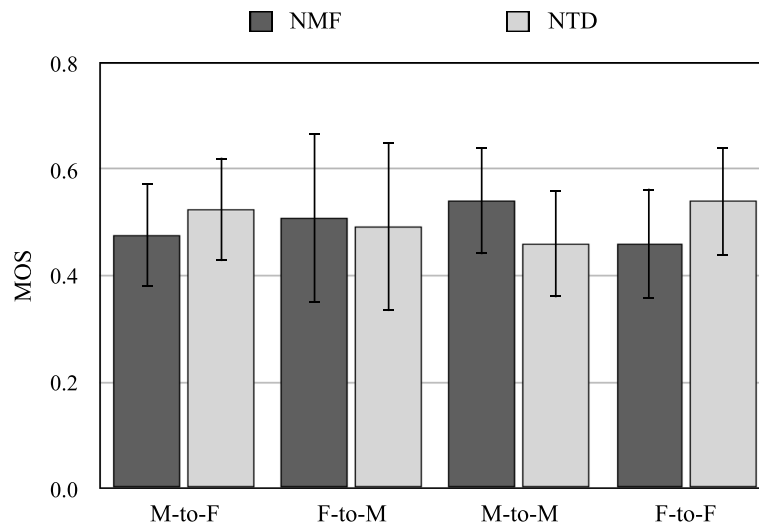


Fig. 7 XAB test between NMF and NTD

difference is relatively small. It is assumed that MCD is superior to the ARBM-based method, as it uses Mel-cepstrum as an input feature, whereas NTD-based methods use a WORLD spectrum. In the speech quality test, the proposed method using non-parallel training data achieved a better MOS score than that using parallel utterances. This is due to the model's ability to learn diverse phonemic information by using non-parallel data when compared with parallel utterances. For example, n sentences are used for each speaker as training data. In the instances using parallel utterances, which consist of the same context for both speakers, the frequency base matrices \mathbf{U}^s and \mathbf{U}^t and the codebook \mathbf{G} are learned from n context patterns. However, in the non-parallel setting, where

a different context was used for the source and target speakers, the frequency base matrices and the codebook were learned from n and $2n$ context patterns, respectively. A codebook was effectively learned while improving the generalization ability. Therefore, the method using non-parallel data outperformed that using parallel utterances.

5 Experiments on voice conversion challenge 2018

The proposed method was also evaluated on the Voice Conversion Challenge (VCC) 2018 [45], which includes both parallel and non-parallel recordings from native English speakers from the USA. VCC 2018 consists of a total of 12 speakers. Each speaker has sets of 81 and 35

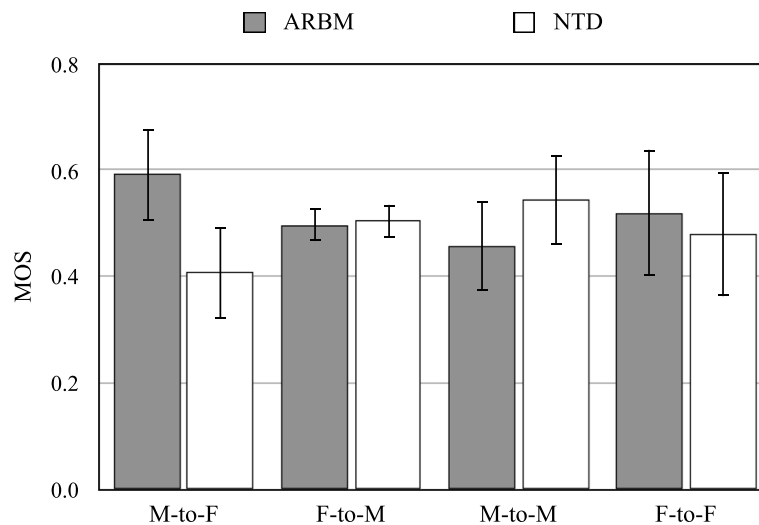


Fig. 8 XAB test between ARBM and NTD

sentences for training and evaluation, respectively. The recordings were down sampled to 16 kHz. Systems were conducted for the 16 combinations of source-target pairs.

The results of this objective evaluation are shown in Table 4. Our proposed method did not outperform the GMM-based VC in the parallel setting, while the NTD-based method achieved 3.89% relative improvement compared with the ARBM-based method in the non-parallel setting. These results demonstrate that our method is especially effective in non-parallel settings.

6 Conclusion

An innovative dictionary-learning method of NMF-based voice conversion was proposed. It makes NMF-VC possible for non-parallel training. While exemplar-based VC retains the naturality of the converted speech to a high degree, the source and target dictionaries expand significantly. Although dictionary-learning VC achieves compact dictionary representation, the parallel dictionaries of the source and target speakers are difficult to learn. These conventional NMF-VC methods require parallel utterances by the source and target speakers to construct the source and target dictionaries. In this study, a method parallel dictionary learning for NMF-VC based on NTD was proposed that does not require parallel data during training. NTD decomposes an input observation into a set of mode matrices and one core tensor. In the proposed framework, it is assumed that NTD decomposes the spectrogram into the frequency basis matrix, phonemic information matrix, and codebook matrix. Recently, several studies have been conducted for NMF-VC, and the scope of possible applications is widening. It is assumed that the proposed method assists these applications with non-parallel training. It was confirmed that the proposed method achieved an almost identical MCD to the conventional NMF-based dictionary learning that uses parallel data. Furthermore, the performance of the proposed method was comparable to that of the conventional ARBM-based method in a non-parallel setting.

In future work, we plan to apply the method to assistive technology for speakers with articulation disorders. The speech of such speakers is considerably different from that of the speech of unimpaired persons, and it is difficult to align correctly. The proposed method does not require the same texts of speech data for the source and target speakers or the framewise matching between acoustic

features of both speakers. Furthermore, the NTD-based dictionary learning is a natural expansion of the NMF-based method, and it can read parallel and non-parallel data to learn the dictionary. Therefore, we also aim to investigate a semi-supervised dictionary-learning method that improves the performance of a model trained with a small set of parallel data using a large set of non-parallel data.

In the real world, background noise deteriorates conversion performance. However, the proposed model has not been designed with noise robustness in mind. In order to retain the quality of converted voices in a noisy environment, noise robustness is required. In our previous study [46], a noise-robust NMF-based VC was proposed, where the performance was improved by 25% compared with the GMM-based method. As the currently proposed method is based on NMF-based VC, it will be easy to apply the noise-robust conversion. The evaluation of our proposed method for a noisy environment will be a topic for our future work.

Abbreviations

ARBM: Adaptive restricted Boltzmann machine; DPM: Dynamic programming matching; DTW: Dynamic time warping; GMM: Gaussian mixture model; MCD: Mel-Cepstral Distortion; MOS: Mean opinion score; NMF: Non-negative matrix factorization; NTD: Non-negative Tucker decomposition; VC: Voice conversion; VCC: Voice conversion challenge

Acknowledgements

Not applicable.

Authors' contributions

YT performed the experiments and wrote the paper. YT, TN, and TT reviewed and edited the manuscript. All of the authors discussed the final results. All of the authors read and approved the final manuscript.

Funding

This work was supported in part by JSPS KAKENHI (no. JP17J04380) and PRESTO, JST (no. PMJPR15D2).

Availability of data and materials

All data used in this study are included in the ATR Japanese speech database [41].

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Graduate School of System Informatics, Kobe University, 1-1 Rokkodai, Nada-ku, Kobe 657-8501, Japan. ²Graduate School of Informatics and Engineering, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan.

Received: 19 December 2018 Accepted: 14 August 2019

Published online: 11 September 2019

References

1. T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, J. Yamagishi, in *Proc. Interspeech*. The voice conversion challenge 2016 (ISCA, San Francisco, 2016), pp. 1632–1636

Table 4 Average MCD [dB] on VCC 2018

	GMM	ARBM	NTD
Parallel	6.55		7.03
Non-parallel		7.97	7.70

2. R. Gray, Vector quantization. *IEEE Assp. Mag.* **1**(2), 4–29 (1984)
3. H. Valbret, E. Moulines, J.-P. Tubach, Voice transformation using PSOLA technique. *Speech Comm.* **11**(2–3), 175–187 (1992)
4. A. Kain, M. W. Macon, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Spectral voice conversion for text-to-speech synthesis (IEEE, Seattle, 1998), pp. 285–288
5. C. Veaux, X. Rodet, in *Proc. Interspeech.* Intonation conversion from neutral to expressive speech (ISCA, Florence, 2011), pp. 2765–2768
6. K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. *Speech Comm.* **54**(1), 134–146 (2012)
7. L. Deng, A. Acero, L. Jiang, J. Droppo, X. Huang, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* High-performance robust speech recognition using stereo training data (IEEE, Salt Lake City, 2001), pp. 301–304
8. A. Kunikoshi, Y. Qiao, N. Minematsu, K. Hirose, in *Proc. Interspeech.* Speech generation from hand gestures based on space mapping (ISCA, Brighton, 2009), pp. 308–311
9. Y. Stylianou, O. Cappé, E. Moulines, Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech Audio Process.* **6**(2), 131–142 (1998)
10. T. Toda, A. W. Black, K. Tokuda, Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Trans. Audio Speech Lang. Process.* **15**(8), 2222–2235 (2007)
11. E. Helander, T. Virtanen, J. Nurminen, M. Gabbouj, Voice conversion using partial least squares regression. *IEEE Trans. Audio Speech Lang. Process.* **18**(5), 912–921 (2010)
12. D. Saito, H. Doi, N. Minematsu, K. Hirose, in *Proc. Interspeech.* Application of matrix variate gaussian mixture model to statistical voice conversion (ISCA, Singapore, 2014), pp. 2504–2508
13. R. Takashima, T. Takiguchi, Y. Ariki, in *IEEE Workshop on Spoken Language Technology.* Exemplar-based voice conversion in noisy environment (IEEE, Miami, 2012), pp. 313–317
14. R. Takashima, R. Aihara, T. Takiguchi, Y. Ariki, in *Speech Synthesis Workshop.* Noise-robust voice conversion based on spectral mapping on sparse space (ISCA, Barcelona, 2013), pp. 71–75
15. Z. Wu, T. Virtanen, E. Chng, H. Li, Exemplar-based sparse representation with residual compensation for voice conversion. *IEEE/ACM Trans. Audio Speech Lang. Process.* **22**(10), 1506–1521 (2014)
16. S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, K. Prahallad, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Voice conversion using artificial neural networks (IEEE, Taipei, 2009), pp. 3893–3896
17. T. Nakashika, R. Takashima, T. Takiguchi, Y. Ariki, in *Proc. Interspeech.* Voice conversion in high-order eigen space using deep belief nets (ISCA, Lyon, 2013), pp. 369–372
18. T. Nakashika, T. Takiguchi, Y. Ariki, Voice conversion using rnn pre-trained by recurrent temporal restricted Boltzmann machines. *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**(3), 580–587 (2015)
19. L.-H. Chen, Z.-H. Ling, Y. Song, L.-R. Dai, in *Proc. Interspeech.* Joint spectral distribution modeling using restricted Boltzmann machines for voice conversion (ISCA, Lyon, 2013), pp. 3052–3056
20. T. Nakashika, T. Takiguchi, Y. Minami, Non-parallel training in voice conversion using an adaptive restricted Boltzmann machine. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(11), 2032–2045 (2016)
21. Z. Wu, E. Chng, H. Li, in *ChinaSIP.* Conditional restricted Boltzmann machine for voice conversion (IEEE, Beijing, 2013), pp. 104–108
22. Y. Saito, Y. Ijima, K. Nishida, S. Takamichi, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Non-Parallel Voice Conversion Using Variational Autoencoders Conditioned by Phonetic Posteriorgrams and D-Vectors (IEEE, Calgary, 2018), pp. 5274–5278
23. F. Fang, J. Yamagishi, I. Echizen, J. Lorenzo-Trueba, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* High-Quality Nonparallel Voice Conversion Based on Cycle-Consistent Adversarial Network (IEEE, Calgary, 2018), pp. 5279–5283
24. R. Aihara, T. Nakashika, T. Takiguchi, Y. Ariki, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary (IEEE, Florence, 2014), pp. 7894–7898
25. R. Aihara, T. Takiguchi, Y. Ariki, in *Proc. Interspeech.* Parallel dictionary learning for voice conversion using discriminative graph-embedded non-negative matrix factorization (ISCA, San Francisco, 2016), pp. 292–296
26. D. D. Lee, H. S. Seung, in *NIPS.* Algorithms for non-negative matrix factorization (MIT Press, Denver, 2000), pp. 556–562
27. R. Aihara, T. Takiguchi, Y. Ariki, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* Activity-mapping non-negative matrix factorization for exemplar-based voice conversion (IEEE, South Brisbane, 2015), pp. 4899–4903
28. A. Mouchtaris, J. V. der Spiegel, P. Mueller, Nonparallel training for voice conversion based on a parameter adaptation approach. *IEEE Trans. Audio Speech Lang. Process.* **14**(3), 952–963 (2006)
29. T. Hashimoto, H. Uchida, D. Saito, N. Minematsu, in *Proc. Interspeech.* Parallel-data-free many-to-many voice conversion based on dnn integrated with eigenspace using a non-parallel speech corpus (ISCA, Stockholm, 2017), pp. 1278–1282
30. B. Sisman, H. Li, K. C. Tan, in *ASRU.* Sparse representation of phonetic features for voice conversion with and without parallel data (IEEE, Okinawa, 2017), pp. 677–684
31. L. D. Lathauwer, B. D. Moor, J. Vandewalle, A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* **21**(4), 1253–1278 (2000)
32. P. M. Kroonenberg, J. De Leeuw, Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika.* **45**, 69–97 (1980)
33. L. R. Tucker, Some mathematical notes on three-mode factor analysis. *Psychometrika.* **31**, 279–311 (1966)
34. H. Ben younes, R. Cadène, M. Cord, N. Thome, in *ICCV.* Mutan: Multimodal tucker fusion for visual question answering (IEEE Computer Society, Venice, 2017), pp. 2631–2639
35. J.-T. Chien, C. Shen, in *Proc. Interspeech.* Deep neural factorization for speech recognition, (2017), pp. 3682–3686
36. D. Saito, K. Yamamoto, N. Minematsu, K. Hirose, in *Proc. Interspeech.* One-to-many voice conversion based on tensor representation of speaker space (ISCA, Florence, 2011), pp. 653–656
37. Z. Ma, A. E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, J. Guo, Variational bayesian matrix factorization for bounded support data. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(4), 876–889 (2015)
38. Q. Shi, Y.-M. Cheung, Q. Zhao, H. Lu, Feature extraction for incomplete data via low-rank tensor decomposition with feature regularization. *IEEE Trans. Neural Netw. Learn. Syst.* **30**(6), 1803–1817 (2019)
39. B. Jiang, C. Ding, J. Tang, B. Luo, Image representation and learning with graph-laplacian Tucker tensor decomposition. *IEEE Trans. Cybernet.* **49**(4), 1417–1426 (2019)
40. Y. Kim, S. Choi, in *Computer Vision and Pattern Recognition.* Nonnegative tucker decomposition (IEEE Computer Society, Minneapolis, 2007), pp. 1–8
41. A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, K. Shikano, ATR Japanese speech database as a tool of speech recognition and synthesis. *Speech Commun.* **9**(4), 357–363 (1990)
42. R. Aihara, T. Fujii, T. Nakashika, T. Takiguchi, Y. Ariki, Small-parallel exemplar-based voice conversion in noisy environments using affine non-negative matrix factorization. *EURASIP J. Audio Speech Music Process.* **2015**, 32 (2015)
43. M. Morise, F. Yokomori, K. Ozawa, World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans.* **99-D**(7), 1877–1884 (2016)
44. K. Kobayashi, T. Toda, in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop.* sprocket: Open-source voice conversion software (ISCA, Les Sables d’Olonne, 2018), pp. 203–210
45. J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, Z.-H. Ling, in *Odyssey.* The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods (ISCA, Les Sables d’Olonne, 2018), pp. 195–202
46. R. Aihara, R. Takashima, T. Takiguchi, Y. Ariki, Noise-robust voice conversion based on sparse spectral mapping using non-negative matrix factorization. *IEICE Trans. Inf. Syst.* **97-D**(6), 1411–1418 (2014)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.