EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Speaker-adaptive-trainable Boltzmann machine and its application to non-parallel voice conversion

Toru Nakashika* and Yasuhiro Minami

Abstract

In this paper, we present a voice conversion (VC) method that does not use any parallel data while training the model. Voice conversion is a technique where only speaker-specific information in the source speech is converted while keeping the phonological information unchanged. Most of the existing VC methods rely on parallel data—pairs of speech data from the source and target speakers uttering the same sentences. However, the use of parallel data in training causes several problems: (1) the data used for the training is limited to the pre-defined sentences, (2) the trained model is only applied to the speaker pair used in the training, and (3) a mismatch in alignment may occur. Although it is generally preferable in VC to not use parallel data, a non-parallel approach is considered difficult to learn. In our approach, we realize the non-parallel training based on speaker-adaptive training (SAT). Speech signals are represented using a probabilistic model based on the Boltzmann machine that defines phonological information and speaker-related information explicitly. Speaker-independent (SI) and speaker-dependent (SD) parameters are simultaneously trained using SAT. In the conversion stage, a given speech signal is decomposed into phonological and speaker-related information, the speaker-related information is replaced with that of the desired speaker, and then voice-converted speech is obtained by combining the two. Our experimental results showed that our approach outperformed the conventional non-parallel approach regarding objective and subjective criteria.

Keywords: Voice conversion, Boltzmann machine, Unsupervised training, Energy-based model, Speaker adaptation, Non-parallel training, SAT

1 Introduction

In recent years, voice conversion (VC), which is a technique used to change speaker-specific information in the speech of a source speaker into that of a target speaker while retaining linguistic information, has been garnering much attention since the VC techniques can be applied to various tasks [1–5]. Most of the existing approaches rely on statistical models [6, 7], and the approach based on the Gaussian mixture model (GMM) [8–11] is one of the mainstream methods used nowadays. Other statistical models, such as non-negative matrix factorization (NMF) [12, 13], neural networks (NNs) [14], restricted Boltzmann machines (RBMs) [15, 16], and deep learning [17, 18], are also used in VC. However, almost all of the existing VC methods 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) for training the models, which leads to several problems. First, the data is limited to pre-defined articles (both speakers must utter the same articles). Second, the trained model is only applied to the speaker pair used in the training, and it is difficult to reuse the model on the conversion of another speaker pair. Third, the training data (the parallel data) is not the original speech data anymore because the speech data is stretched and modified in the time axis when aligned. Furthermore, it is not guaranteed that each frame is aligned perfectly, and the mismatching may cause some errors in training.

Several approaches that do not use *parallel data from the source to the target speakers*[1] have been also proposed [19–23]. In [19], for example, the authors model the spectral relationships between two arbitrary speakers (reference speakers) using GMMs and convert the source

*Correspondence: nakashika@uec.ac.jp
Graduate School of Information Systems, University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, 182-8585 Tokyo, Japan

speaker's speech using the matrix that projects the feature space of the source speaker into that of the target speaker through that of the reference speakers. As a result, parallel data from the source and target speakers is not required. In [21, 22], codebooks (eigenvoice) are obtained using the parallel data of the reference speakers, and a many-to-many VC is achieved by mapping the source speaker's speech into an eigenvoice and the eigenvoice into the target speaker's speech. The multistep VC [24] is also proposed to reduce the training cost of estimating the mapping functions for each speaker pair.

In this paper, we propose a totally-parallel-data-free[2] VC method using an energy-based probabilistic model and speaker adaptive training (SAT). The idea is simple and intuitive. A speech signal of an arbitrary speaker is composed of neutral speech (the speech with the averaged voice calculated from a collection of speech samples from multiple speakers) that directly links to the phonological information belonging to no one, accompanied with the speaker-specific information. In this assumption, VC is achieved by three steps: decomposing a speech signal into neutral speech and speaker-specific information, replacing the speaker-specific information with that of the desired speaker, and composing a speech signal using the neutral speech and the replaced speaker information. The proposed model, called a speaker adaptive trainable Boltzmann machine (SATBM), is designed to help such a decomposition. The above VC steps can be viewed as a simplified version of the combination of automatic speech recognition (ASR) and text-to-speech (TTS) systems: text estimation from the input speech using the ASR system, followed by speech generation of the target speaker from the text using the TTS system. Although the VC can be realized by this approach, our VC scheme has several advantages. First, in our approach, we can reduce (or omit) the cost of training two different systems. Second, the combination approach requires a large amount of training data of the target speaker in TTS, while our approach does not. Third, the latent phonological features in our approach can be optimized for the VC. Fourth, ideally, the voice-converted speech can be generated in real time by our approach due to the frame-wise conversion.

We attempted the non-parallel training using another probabilistic model named the adaptive restricted Boltzmann machine (ARBM) [25] in our previous work. The architecture is different from the proposed model in this paper, which makes some differences, e.g., while an ARBM is based on the model-space transformation, a SATBM is based on the constrained model-space transformation. In the following sections, we will discuss this in more detail.

2 Formulation

In general, it is known that the differences of speech signals in terms of speakers can be represented as multiplication in the cepstrum-based domain. After the general form, we define an acoustic feature vector[3] $\hat{\boldsymbol{x}}_{rt} = \left[\hat{x}_{rt}^1, \cdots, \hat{x}_{rt}^D\right]^\top \in \mathbb{R}^D$ (D is the number of dimensions) of a speaker r at time t as follows:

$$\hat{\boldsymbol{x}}_{rt} = \mathbf{A}_r \boldsymbol{x}_t + \boldsymbol{b}_r, \tag{1}$$

where $\boldsymbol{x}_t = \left[x_t^1, \cdots, x_t^D\right]^\top \in \mathbb{R}^D$, $\mathbf{A}_r \in \mathbb{R}^{D\times D}$, and $\boldsymbol{b}_r \in \mathbb{R}^D$ denote the speaker-normalized acoustic feature vector (acoustic features of the neutral speaker or the averaged speaker), a speaker adaptation matrix, and a bias vector of the speaker r, respectively. Note that $\mathbf{A}_r$ is a global matrix for all the phonemes or kinds of speech sounds unlike in MLLR or similar techniques. Here, we assume that $\boldsymbol{x}_t$ is normally distributed with time-varying (phoneme-dependent) mean $\boldsymbol{\mu}_t \in \mathbb{R}^D$ and time-invariant diagonal variance $\Sigma = \text{diag}\,(\boldsymbol{\sigma}^2), \boldsymbol{\sigma}^2 = \left[\sigma_1^2, \cdots, \sigma_D^2\right]^\top \in \mathbb{R}^D$ given a latent phonological vector $\boldsymbol{h}_t = \left[h_t^1, \cdots, h_t^H\right]^\top \in \mathbb{B}^H$ ($\mathbb{B}$ is a binary space and H is the number of dimensions of the latent vector). At this time, $\hat{\boldsymbol{x}}_{rt}$ is also normally distributed; that is,

$$\begin{aligned} \hat{\boldsymbol{x}}_{rt}|\boldsymbol{h}_t &\sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{rt}, \hat{\Sigma}_r), \\ \hat{\boldsymbol{\mu}}_{rt} &= \mathbf{A}_r \boldsymbol{\mu}_t + \boldsymbol{b}_r \\ \hat{\Sigma}_r &= \mathbf{A}_r \Sigma \mathbf{A}_r^\top. \end{aligned} \tag{2}$$

The dependence on $\boldsymbol{h}_t$ in Eq. (2) is explained as follows. The speech of the neutral speaker at a certain time is supposed to be determined by the latent, phonological information that must exist behind but is not observable. For example, if the phoneme /e/ is intended at time t, then the neutral speech at t should correspond acoustically to the phoneme /e/. Therefore, we assume that the mean vector of the neutral speaker $\boldsymbol{\mu}_t$ is determined using a latent phonological vector $\boldsymbol{h}$ as

$$\boldsymbol{\mu}_t = \mathbf{W}\boldsymbol{h}_t + \boldsymbol{b}, \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{D\times H}$ and $\boldsymbol{b} \in \mathbb{R}^D$ are a matrix and a bias vector, respectively, that project the phonological space into the acoustic space. Furthermore, $\boldsymbol{b}$ is also regarded as a bias vector to realize the speaker characteristics of the neutral speaker. Incidentally, the conditional probability $p(\hat{\boldsymbol{x}}_{rt}|\boldsymbol{h}_t)$ given $\boldsymbol{h}_t$ in Eq. (2) can be calculated as follows:

$$\begin{aligned} p(\hat{\boldsymbol{x}}_{rt}|\boldsymbol{h}_t) &= \mathcal{N}(\hat{\boldsymbol{\mu}}_{rt}, \hat{\Sigma}_r) \\ &\propto e^{-\frac{1}{2}(\hat{\boldsymbol{x}}_{rt}-\hat{\boldsymbol{\mu}}_{rt})^\top \hat{\Sigma}_r^{-1}(\hat{\boldsymbol{x}}_{rt}-\hat{\boldsymbol{\mu}}_{rt})} \\ &\propto e^{-\left\{\frac{1}{2}(\hat{\boldsymbol{x}}_{rt}-\hat{\boldsymbol{b}}_r)^\top \hat{\Sigma}_r^{-1}(\hat{\boldsymbol{x}}_{rt}-\hat{\boldsymbol{b}}_r) - \hat{\boldsymbol{x}}_{rt}^\top \hat{\Sigma}_r^{-1}\hat{\mathbf{W}}_r \boldsymbol{h}_t\right\}}, \end{aligned} \tag{4}$$

where we introduce $\hat{\boldsymbol{b}}_r = \mathbf{A}_r\boldsymbol{b} + \boldsymbol{b}_r$ and $\hat{\mathbf{W}}_r = \mathbf{A}_r\mathbf{W}$.

On the other hand, in this paper, we assume that the phonological information can be determined by the acoustic features as well. It means h_t^j is Bernoulli-distributed and its parameter $\pi_t^j \in \boldsymbol{\pi}_t$ ($j = 1, \cdots, H$) that represents the probability $p\left(h_t^j = 1\right)$ is assumed to be a function of $\boldsymbol{x}_t$. In this formulation, it is beneficial in terms of reducing the number of parameters to use the already-defined parameters. We define $\boldsymbol{\pi}_t$ as follows:

$$\boldsymbol{\pi}_t = \boldsymbol{\phi}\left(\mathbf{W}^\top \Sigma^{-1} \boldsymbol{x}_t + \boldsymbol{c}\right), \tag{5}$$

where $\boldsymbol{\phi}(\cdot)$ denotes an element-wise sigmoid function and $\boldsymbol{c} \in \mathbb{R}^H$ is a bias term for the phonological information that is independent of time. Considering that $\boldsymbol{x}_t = \mathbf{A}_r^{-1}(\hat{\boldsymbol{x}}_{rt} - \boldsymbol{b}_r)$ and $\Sigma^{-1} = \mathbf{A}_r^\top \hat{\Sigma}_r^{-1} \mathbf{A}_r$, the conditional probability, $p(\boldsymbol{h}_t|\hat{\boldsymbol{x}}_{rt})$ forms incidentally as follows:

$$\begin{aligned} p(\boldsymbol{h}_t|\hat{\boldsymbol{x}}_{rt}) &= \mathcal{B}(\boldsymbol{\pi}_t) \\ &\propto e^{\left(\mathbf{W}^\top \Sigma^{-1} \mathbf{A}_r^{-1}(\hat{\boldsymbol{x}}_{rt} - \boldsymbol{b}_r) + \boldsymbol{c}\right)^\top \boldsymbol{h}_t} \\ &= e^{-\left(-\hat{\boldsymbol{x}}_{rt}^\top \hat{\Sigma}_r^{-1} \hat{\mathbf{W}}_r \boldsymbol{h} - \hat{\boldsymbol{c}}_r^\top \boldsymbol{h}\right)}, \end{aligned} \tag{6}$$

where we use the replacement of $\hat{\boldsymbol{c}}_r = \boldsymbol{c} - \hat{\mathbf{W}}_r^\top \hat{\Sigma}_r^{-1} \boldsymbol{b}_r$.

Now, we consider the joint probability of $\hat{\boldsymbol{x}}_{rt}$ and $\boldsymbol{h}_t$. From Eqs. (4) and (6), we notice that the same term $-\hat{\boldsymbol{x}}_{rt}^\top \hat{\Sigma}_r^{-1} \hat{\mathbf{W}}_r \boldsymbol{h}$ appears in the exponential. Consequently, the following joint probability satisfies Eqs. (4) and (6):

$$\begin{aligned} p(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t) =& \frac{1}{Z} e^{-E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)} \\ E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t) =& \frac{1}{2}(\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r)^\top \hat{\Sigma}_r^{-1} (\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r) \\ & - \hat{\boldsymbol{x}}_{rt}^\top \hat{\Sigma}_r^{-1} \hat{\mathbf{W}}_r \boldsymbol{h}_t - \hat{\boldsymbol{c}}_r^\top \boldsymbol{h}_t, \end{aligned} \tag{7}$$

where $Z = \int^D \sum_{\boldsymbol{h}_t} e^{-E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)} d^D \hat{\boldsymbol{x}}_{rt}$ is a normalization term. Furthermore, substituting Eq. (1) for Eq. (7) forms

$$\begin{aligned} p(\boldsymbol{x}_t, \boldsymbol{h}_t) &= \tfrac{1}{Z} e^{-E(\boldsymbol{x}_t, \boldsymbol{h}_t)} \\ E(\boldsymbol{x}_t, \boldsymbol{h}_t) &= \tfrac{\|\boldsymbol{x}_t - \boldsymbol{b}\|_2^2}{2\boldsymbol{\sigma}^2} - \left(\tfrac{\boldsymbol{x}_t}{\boldsymbol{\sigma}^2}\right)^\top \mathbf{W} \boldsymbol{h}_t - \boldsymbol{c}^\top \boldsymbol{h}_t, \end{aligned} \tag{8}$$

which is nothing more than the definition of a Gaussian-Bernoulli restricted Boltzmann machine (GB-RBM) [26]. In other words, the model defined in Eq. (7) implies that it adapts the neutral speech to that of a speaker r when using a GB-RBM with the visible units of acoustic features of the neutral speaker (or the averaged speaker) and the hidden units of latent phonological features, as shown in Fig. 1a. In another viewpoint, it can be regarded as a sort of semi-RBM [27] since there are shared connections $\hat{\mathbf{W}}_r$ between $\hat{\boldsymbol{x}}_{rt}$ and $\boldsymbol{h}_t$ and connections $\hat{\Sigma}_r^{-1}$ among $\hat{\boldsymbol{x}}_{rt}$ but no connections among $\hat{\boldsymbol{h}}_t$ (Fig. 1b). The difference is that the model in Eq. (7) assumes the existence of the neutral speaker and defines additional parameters that enable speaker-adaptive training. In this paper, we call the probabilistic model defined in Eq. (7) speaker-adaptive-trainable Boltzmann machine (SATBM). In our previous work [25], we have proposed another probabilistic model named adaptive restricted Boltzmann machine (ARBM), that is an extension of an RBM where only the connection weights between the visible and hidden units are speaker-adaptive. The ARBM is based on a model-space transformation, whereas the SATBM is based on both a model-space transformation and a feature-space transformation (i.e., constrained model-space transformation), as Eqs. (1) and (2) indicate. Specifically, in the SATBM, the speaker-dependent parameters (means and covariance matrix) of Gaussian visible units are represented as $\mathbf{A}_r \mathbf{W} \boldsymbol{h}_t + \mathbf{A}_r \boldsymbol{b} + \boldsymbol{b}_r$ of means and $\mathbf{A}_r \Sigma \mathbf{A}_r^\top$ of a covariance matrix. On the other hand, in ARBM, the speaker-dependent parameters of Gaussian visible units are represented as $\mathbf{A}_r \mathbf{W} \boldsymbol{h}_t + \boldsymbol{b} + \boldsymbol{b}_r$ of means and Σ of a covariance matrix. This indicates that the speaker-dependent Gaussian parameters in the SATBM are more strongly influenced by the speaker and changed to adapt to the speaker more than those in the ARBM. In another perspective, the SATBM directly models the correlations between the dimensions in the observed features while the ARBM does not. The observed features take different values every time a specific speaker pronounces the same phoneme, and the extent of the variation depends on the speaker. The SATBM also represents such characteristics of each speaker. For this reason, we expect the SATBM would be superior in acoustic modeling to the ARBM.

2.1 One-hot activation of $\boldsymbol{h}_t$

We can further add constraints $\sum_{j=1}^H h_t^t = 1$ to our model resulting in a one-hot vector $\boldsymbol{h}_t$, which indicates that only a certain phonological component is activated. In the real speech, only one phoneme, such as /a/ and /e/, should be activated in the background at a certain frame. Therefore, this modification may give better representation for speech. The use of a one-hot representation is inspired by such a phonological reason.

Such constraints give the following conditional probability that h_t^j is activated (that is, π_t^j) as

$$\begin{aligned} \pi_t^j &= p\left(h_t^j = 1|\hat{\boldsymbol{x}}_{rt}\right) \\ &= \frac{e^{\hat{\boldsymbol{w}}_r^{j\top} \hat{\Sigma}_r^{-1} \hat{\boldsymbol{x}}_{rt} + \hat{c}_r^j}}{\sum_{j'} e^{\hat{\boldsymbol{w}}_r^{j'\top} \hat{\Sigma}_r^{-1} \hat{\boldsymbol{x}}_{rt} + \hat{c}_r^{j'}}} \\ &= \psi\left(\hat{\boldsymbol{w}}_r^{j\top} \hat{\Sigma}_r^{-1} \hat{\boldsymbol{x}}_{rt} + \hat{c}_r^j\right), \end{aligned} \tag{9}$$

where $\hat{\boldsymbol{w}}_r^j$ and $\hat{c}_r^j$ indicate the jth column vector in $\hat{\mathbf{W}}_r$ and the jth element in $\hat{\boldsymbol{c}}_r$, respectively, and $\psi(\cdot)$ denotes a softmax function (we also define $\boldsymbol{\psi}(\cdot)$ as an element-wise softmax function for convenience). Equation 9 is used when we sample $\boldsymbol{h}_t$ instead of Eq. (5), as discussed in the following sections.

3 Parameter estimation based on SAT

In this section, we describe the method of the parameter estimation in the previously defined model, a SATBM, based on SAT [28]. For convenience, we use symbols $\Theta^{SD} = \{\mathbf{A}_r, \boldsymbol{b}_r\}_{r=1}^R$ for SD parameters and $\Theta^{SI} = \{\mathbf{W}, \boldsymbol{\sigma}^2, \boldsymbol{b}, \boldsymbol{c}\}$ for SI parameters. Given a collection of the

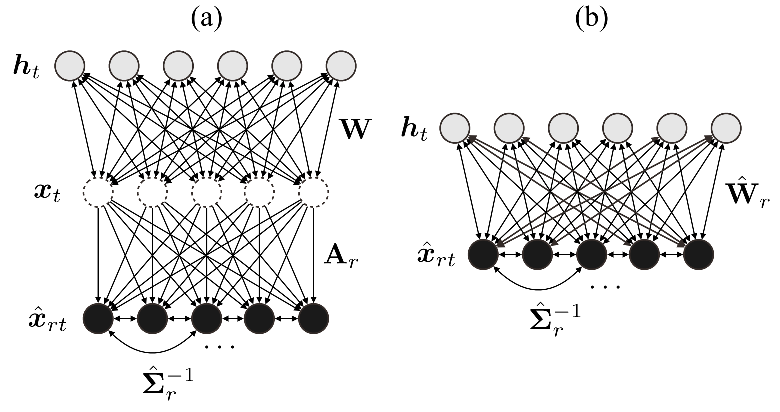
(a)
(b)
$\boldsymbol{h}_t$
$\mathbf{W}$
$\boldsymbol{x}_t$
$\mathbf{A}_r$
$\hat{\boldsymbol{x}}_{rt}$
$\hat{\boldsymbol{\Sigma}}_r^{-1}$
$\boldsymbol{h}_t$
$\hat{\mathbf{W}}_r$
$\hat{\boldsymbol{x}}_{rt}$
$\hat{\boldsymbol{\Sigma}}_r^{-1}$

Fig. 1 a Proposed model: speaker-adaptive-trainable Boltzmann machine (SATBM) and **b** its simplified representation, which can be seen as a type of semi-RBM

speech data, $\boldsymbol{X} = \{\boldsymbol{X}_r\}_{r=1}^{R}$, $\boldsymbol{X}_r = \{\hat{\boldsymbol{x}}_{rt}\}_{t=1}^{T_r}$ that is composed of R speakers, these parameters are simultaneously estimated so as to maximize the likelihood as

$$(\hat{\Theta}^{SD}, \hat{\Theta}^{SI}) \triangleq \underset{(\Theta^{SD},\Theta^{SI})}{\operatorname{argmax}} \prod_{r=1}^{R} \prod_{t=1}^{T_r} p(\hat{\boldsymbol{x}}_{rt}). \tag{10}$$

According to the SAT paradigm, the SD parameters Θ^{SD} undertake the speaker-induced variation, and the SI parameters Θ^{SI} capture the remaining information, i.e., phonetically relevant variation. Unlike the conventional SAT+MLLR (maximum likelihood linear regression), the SATBM explicitly models the relationships between the speaker-normalized acoustic features and the phonological information, which implies the possibility that the model represents the speech data better than SAT+MLLR.

The parameters are iteratively updated based on gradient descent. The partial differential of the log-likelihood $l = \log \prod_r \prod_t p(\hat{\boldsymbol{x}}_{rt}) = \sum_r \sum_t \log \sum_{\boldsymbol{h}} p(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t))$ in terms of a parameter $\theta \in \{\Theta^{SD}, \Theta^{SI}\}$ is derived as follows:

$$\frac{\partial l}{\partial \theta} = \sum_r \left(\left\langle \frac{\partial E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)}{\partial \theta} \right\rangle_{\text{data}} - \left\langle \frac{\partial E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)}{\partial \theta} \right\rangle_{\text{model}} \right),$$

where $\langle \cdot \rangle_{\text{data}}$ and $\langle \cdot \rangle_{\text{model}}$ denote expectations of the empirical data and the inner model, respectively. It is generally difficult to compute the expectations of the inner model; however, we can still use contrastive divergence (CD) [29] and efficiently approximate them with the expectations of the reconstructed data. We can analytically calculate the partial gradients $\frac{\partial E(\hat{\boldsymbol{x}}_{rt}, \boldsymbol{h}_t)}{\partial \theta}$ for each parameter as follows:

$$\begin{aligned}
\frac{\partial E(\hat{\boldsymbol{x}}_{rt},\boldsymbol{h}_t)}{\partial \mathbf{A}_r} &= -\frac{1}{2}\left(\mathbf{A}_r^{-1}\mathbf{C}_{rt}\hat{\Sigma}_r^{-1} + \hat{\Sigma}_r^{-1}\mathbf{D}_{rt}\mathbf{A}_r^{-\top}\right) \\
\frac{\partial E(\hat{\boldsymbol{x}}_{rt},\boldsymbol{h}_t)}{\partial \boldsymbol{b}_r} &= -\hat{\Sigma}_r^{-1}(\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r - \hat{\mathbf{W}}_r\boldsymbol{h}_t) \\
\frac{\partial E(\hat{\boldsymbol{x}}_{rt},\boldsymbol{h}_t)}{\partial \mathbf{W}} &= -\mathbf{A}_r^{\top}\hat{\Sigma}_r^{-1}(\hat{\boldsymbol{x}}_{rt} - \boldsymbol{b}_r)\boldsymbol{h}_t^{\top} \\
\frac{\partial E(\hat{\boldsymbol{x}}_{rt},\boldsymbol{h}_t)}{\partial \boldsymbol{\sigma}^2} &= -\frac{1}{2}\mathrm{diag}\left(\mathbf{A}_r^{\top}\hat{\Sigma}_r^{-1}\mathbf{E}_{rt}\hat{\Sigma}_r^{-1}\mathbf{A}_r\right) \\
\frac{\partial E(\hat{\boldsymbol{x}}_{rt},\boldsymbol{h}_t)}{\partial \boldsymbol{b}} &= -\mathbf{A}_r^{\top}\hat{\Sigma}_r^{-1}(\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r) \\
\frac{\partial E(\hat{\boldsymbol{x}}_{rt},\boldsymbol{h}_t)}{\partial \boldsymbol{c}} &= -\boldsymbol{h}_t,
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{C}_{rt} &= (\hat{\boldsymbol{x}}_{rt} - \boldsymbol{b}_r)(\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r - 2\hat{\mathbf{W}}_r\boldsymbol{h}_t)^{\top} \\
\mathbf{D}_{rt} &= (\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r)(\hat{\boldsymbol{x}}_{rt} - \boldsymbol{b}_r)^{\top} \\
\mathbf{E}_{rt} &= (\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r)(\hat{\boldsymbol{x}}_{rt} - \hat{\boldsymbol{b}}_r)^{\top} - 2(\hat{\boldsymbol{x}}_{rt} - \boldsymbol{b}_r)(\hat{\mathbf{W}}_r\boldsymbol{h}_t)^{\top}.
\end{aligned}$$

4 Application to VC

To use the proposed model for the VC, we follow three stages of training, adaptation, and conversion. In the training stage, speaker-independent parameters $\hat{\Theta}^{SI}$ are obtained as in Eq. (10) using R reference speakers' speech (we discard the speaker-dependent parameters $\hat{\Theta}^{SD}$). In the adaptation stage, new speaker-dependent parameters $\Theta_i^{SD} = \{\mathbf{A}_i, \boldsymbol{b}_i\}$ and $\Theta_o^{SD} = \{\mathbf{A}_o, \boldsymbol{b}_o\}$ are estimated using adaptation data of the source and the target speakers $\{\hat{\boldsymbol{x}}_{it}\}_{t=1}^{Ti}$, $\{\hat{\boldsymbol{x}}_{ot}\}_{t=1}^{To}$ while keeping $\hat{\Theta}^{SI}$ fixed. That is,

$$\hat{\Theta}_r^{SD} \triangleq \underset{\Theta_r^{SD}}{\operatorname{argmax}} \prod_{t=1}^{T_r} p\left(\hat{\boldsymbol{x}}_{rt}; \Theta_r^{SD}, \hat{\Theta}^{SI}\right), \; r \in \{i, o\}. \tag{11}$$

To convert the frame-wise acoustic feature vector of the source speaker $\boldsymbol{x}_{it}$ into that of the target speaker $\boldsymbol{x}_{ot}$, we take an ML-based approach. In this approach, $\boldsymbol{x}_{ot}$ is computed so as to maximize the probability given $\boldsymbol{x}_{it}$, formulated as

$$\begin{aligned}\boldsymbol{x}_{ot} &\triangleq \underset{\boldsymbol{x}_{ot}}{\operatorname{argmax}}\, p(\boldsymbol{x}_{ot}|\boldsymbol{x}_{it})\\ &= \underset{\boldsymbol{x}_{ot}}{\operatorname{argmax}} \textstyle\sum_{\boldsymbol{h}_t} p(\boldsymbol{h}_t|\boldsymbol{x}_{it})p(\boldsymbol{x}_{ot}|\boldsymbol{h}_t)\\ &\simeq \underset{\boldsymbol{x}_{ot}}{\operatorname{argmax}}\, p(\hat{\boldsymbol{h}}_t|\boldsymbol{x}_{it})p(\boldsymbol{x}_{ot}|\hat{\boldsymbol{h}}_t)\\ &= \underset{\boldsymbol{x}_{ot}}{\operatorname{argmax}}\, p(\boldsymbol{x}_{ot}|\hat{\boldsymbol{h}}_t)\\ &= \mathbf{A}_o\mathbf{W}\boldsymbol{\psi}\left(\mathbf{W}^{\top}\Sigma^{-1}\mathbf{A}_i^{-1}(\boldsymbol{x}_{it}-\boldsymbol{b}_i)+\boldsymbol{c}\right)+\mathbf{A}_o\boldsymbol{b}+\boldsymbol{b}_o,\end{aligned} \tag{12}$$

where we give $\hat{\boldsymbol{h}}_t \triangleq \mathbb{E}[p(\boldsymbol{h}_t|\boldsymbol{x}_{it})]$. It is worth noting that the conversion function is based on the non-linear transformation.

5 Experimental evaluation

5.1 System configuration

In our VC experiments, we evaluated the performance of our model, a SATBM, using ASJ Continuous Speech Corpus for Research (ASJ-JIPDEC[4]). In the training stage where the SI parameters are estimated, we randomly selected and used the speech data of five sentences (approx. 160 k frames) uttered by 56 speakers (26 males and 30 females) from set A in the corpus. For adaptation and evaluation, a male (identified as "ECL0001") and female ("ECL1003") speakers that were not included in the training were used as source and target speakers, respectively, unless otherwise stated. We also evaluated the proposed SATBM using the other speaker pairs, which will be discussed in Section 5.4. The amount of adaptation data was five sentences for each person. As an acoustic feature vector, we used 32-dimensional mel-cepstral features that were calculated from the 513-dimensional WORLD [30] spectra without dynamic features. In the training of the system, we used up to 64 softmax hidden units, a learning rate of 0.01, a momentum of 0.9, and a batch-size of $R \times 100 (= 5600)$ and set the number of iterations as 200. For the evaluation of the proposed method, we used parallel data (of 10 different sentences from the training and adaptation data) of the source and the target speakers, which was created using dynamic programming. But again, note that all speech data used for the training and the adaptation was NOT parallel.

Mel-cepstral distortion (MCD) is generally used for objective evaluation in the VC. However, we used the mel-cepstral distortion improvement ratio (MDIR) instead in this paper because it does not make sense to view the distance between the spectral features in the mel scale of the source and the target speakers when we want to recognize the differences in speaker identities and because the scale of the MCD varies in the evaluation data. For more discussion about the differences between MDIR and MCD, see Section 5.4. The MDIR is defined as follows:

$$\text{MDIR [dB]} = \frac{10\sqrt{2}}{\ln 10}(\|\boldsymbol{m}_o - \boldsymbol{m}_i\|_2 - \|\boldsymbol{m}_o - \boldsymbol{m}_c\|_2)$$

where $\boldsymbol{m}_i$, $\boldsymbol{m}_o$, and $\boldsymbol{m}_c$ are mel-cepstral features at a frame of the source speaker's speech, target speaker's speech, and converted speech, respectively. The higher the value of MDIR is, the better the performance of the VC is. The MDIR was calculated for each frame from the parallel data of 10 sentences and averaged.

5.1.1 *Methods to be compared*

It is difficult and unfair to evaluate the proposed method because most of the existing VC approaches use parallel data in training and our method does not. Nevertheless, we can still compare the proposed method with our earlier model, ARBM [25]. In addition, a linear-transform-based approach, which has not been proposed, is interesting to compare with. This approach is simple: the vector $\boldsymbol{x}_{ot}$ is calculated as

$$\boldsymbol{x}_{ot} \triangleq \mathbf{A}_o\mathbf{A}_i^{-1}(\boldsymbol{x}_{it} - \boldsymbol{b}_i) + \boldsymbol{b}_o, \tag{13}$$

which was derived from the equation $\boldsymbol{x}_t = \mathbf{A}_i^{-1}(\boldsymbol{x}_{it} - \boldsymbol{b}_i) = \mathbf{A}_o^{-1}(\boldsymbol{x}_{ot} - \boldsymbol{b}_o)$ starting with Eq. (1). However, it is under the assumption that the *true* feature space of the neutral speaker was obtained. The parameters $\mathbf{A}_r$ and $\boldsymbol{b}_r$ are estimated in SAT using the gradient decent, the same as our proposed method. So, the difference between the linear-transform approach and the proposed model is whether the latent phonological features are modeled or not.

For reference, we also compared our proposed model with a popular GMM-based VC with 8, 16, 32, and 64 mixtures using the parallel data of five sentences.

5.1.2 *Optimal number of hidden units*

From the pre-experiment, we see the effects of changing the number of softmax hidden units. Figure 2 shows the performance when changing the number of hidden units between 8, 16, 32, and 64. From Fig. 2, the optimal number was 16, and the performance degraded as the number of hidden units increased more than 16. This is

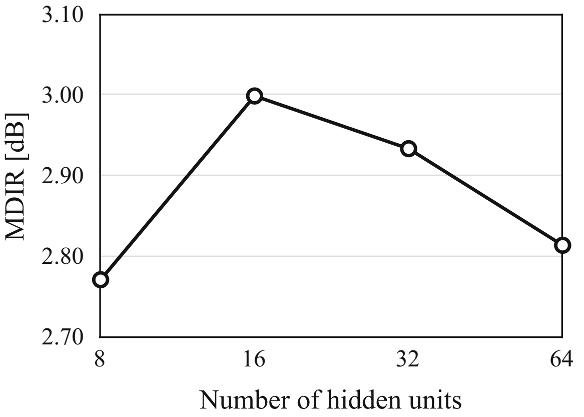

Fig. 2 VC performance when changing the number of hidden units

since the hidden layer becomes speaker-dependent with a large number of units, and hence, it cannot convert the voice properly. To verify this, we examined how the hidden units were speaker-dependent by analyzing the distribution of the hidden units with the cases of $H = 16$ (the optimal number of units) and $H = 64$ (too large a number of units). Figures 3 and 4 show examples of the expected values of the hidden units for the cases of $H = 16$ and $H = 64$, respectively, comparing the distributions obtained from the source and target speakers. From Figs. 3 and 4, we observed that more hidden units when $H = 64$ were speaker-dependent than those when $H = 16$. To measure objectively how close the distributions are to each other, we further calculated the Euclidean distances and the cosine similarities between the two distributions obtained from the source and target speakers' speech, as shown in Table 1. Obviously, Table 1 shows that the two hidden activations of $H = 16$ are close to each other, more than those of $H = 64$. A similar discussion can be seen in our previous work using the ARBM [25].

In speech recognition in Japanese, 43 phonemes are often used [31], which consists of seven short vowels, five long vowels, 28 consonants, and three special symbols. These kinds of Japanese phonemes were defined by the Acoustical Society of Japan (ASJ) committee. Comparing these artificial numbers with the optimal number of hidden units as $H = 16$, we could state that this is reasonable because using the static short-term acoustic features does not represent the consonants and long vowels sufficiently and because the natural speech should contain some allophones.

Considering the above, we will use 16 softmax hidden units in the following experiments unless otherwise stated.

5.2 Objective comparison

The VC performance of the linear-transform-based approach, the ARBM, and the proposed model (SATBM) is compared using objective criteria, as shown in Fig. 5. Each method was evaluated by varying the number of diagonals of the adaptation matrices used as 1 (that means diagonal matrices), 3 (tridiagonal), 5 (pentadiagonal), 7 (heptadiagonal), and 9 (nonadiagonal). In this experiment, we used 16 softmax hidden units for the SATBM and the ARBM. As shown in Fig. 5, the proposed model performed best with any type of the adaptation matrix. The performance of the SATBM and the ARBM was improved when adding the diagonals up to seven. On the other hand, the linear approach barely improved with the number of diagonals. Interestingly, the proposed model achieved a high MDIR even when the diagonal adaptation matrices were used, unlike the ARBM and linear approaches. In the linear approach, the diagonal matrix could not capture the correlations between the dimensions of the mel cepstrum, which makes it impossible to match the vocal tracts among the speakers. Meanwhile, the SATBM could make the source speech resemble the target voice more or less even when a diagonal adaptation matrix was used, due to modeling the latent phonological information. The ARBM models the latent phonological information as well; however, the speaker-dependent Gaussian parameters of the ARBM are not sufficient and failed to represent the speech correctly with the diagonal adaptation matrices.

We also found that the SATBM and the ARBM degraded with nine diagonals and more. This is due to the over-fitting caused by the large number of parameters. In some literature, such as [32–34], it is known that warping cepstral-based features between different speakers is achieved by linear transformation with an adaptation matrix, and a few diagonal elements (such

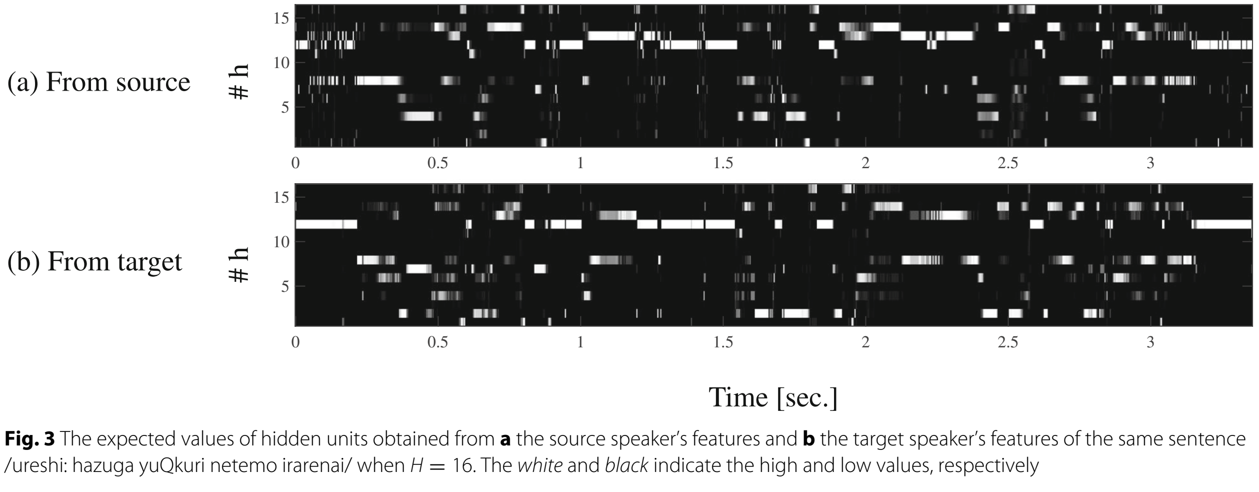

Fig. 3 The expected values of hidden units obtained from **a** the source speaker's features and **b** the target speaker's features of the same sentence /ureshi: hazuga yuQkuri netemo irarenai/ when $H = 16$. The *white* and *black* indicate the high and low values, respectively

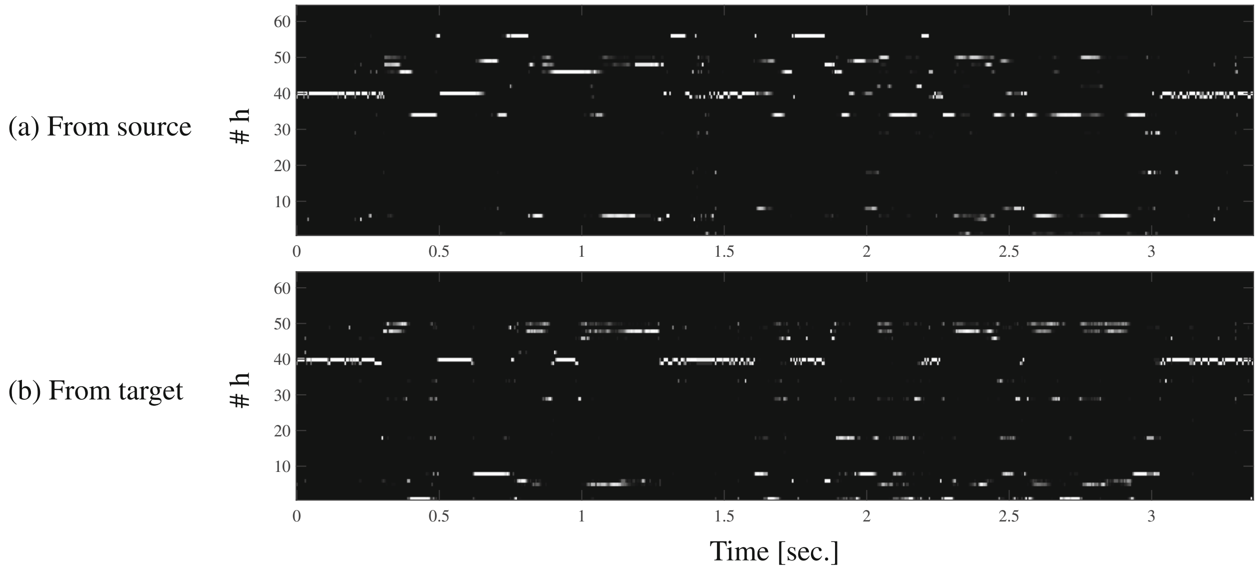

Fig. 4 The expected values of hidden units obtained from **a** the source speaker's features and **b** the target speaker's features of the same sentence /ureshi: hazuga yuQkuri netemo irarenai/ when $H = 64$. The *white* and *black* indicate the high and low values, respectively

as tridiagonal, pentadiagonal, and heptadiagonal) of the adaptation matrix are sufficient for warping the cepstral features. Therefore, it does not make sense to use adaptation matrices with many diagonal elements (more than seven diagonals) in terms of efficient learning for this speaker pair. For more discussion using the various speaker pairs, see Section 5.4.

The average MDIR of the GMM-based approach was 3.93 with 32 mixtures, which was the best in the GMM-based approach. Unfortunately, it performed better than our approach. However, again, such an approach benefits from using parallel data and should not be compared with the non-parallel approach, just in terms of VC quality.

5.3 Subjective comparison

We also conducted subjective experiments, comparing our method with the ARBM and the linear-based approaches. We decoded the converted mel cepstra back to the WORLD spectra using the filter theory [35] and generated signals using the original F0 and aperiodic features of the target speaker since we wanted to compare each method in the spectra. In this experiment, seven participants listened to 10 sentences of converted speech using the linear-based, the ARBM, and the SATBM approaches accompanied with the target speech and voted for the most preferable method for each sentence in terms of the speaker specificity of the target speaker. The number of votes for each method are shown in Fig. 6. The SATBM performed the best of all. It can be said that the SATBM has the ability to produce sounds auditorily closer to the target speech than the other methods.

For reference, we also compared the proposed method with the GMM-based VC in a subjective manner. In this experiment as well, the seven participants were asked to rank the converted speech using a 5-scale (1 poor, 2 fair, 3 good, 4 very good, and 5 excellent) in terms of speech quality (naturalness) and speaker specificity (similarity).

Table 1 Euclidean distance and cosine similarity between the hidden distributions obtained from the source and target speakers' speech of Figs. 3 ($H = 16$) and 4 ($H = 64$)

	Euclidean dist.	Cosine sim.
$H = 16$	0.660	0.526
$H = 64$	0.869	0.337

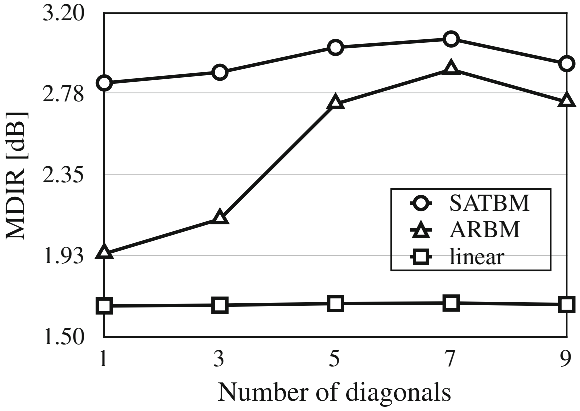

Fig. 5 Comparison of VC performance of non-parallel approaches

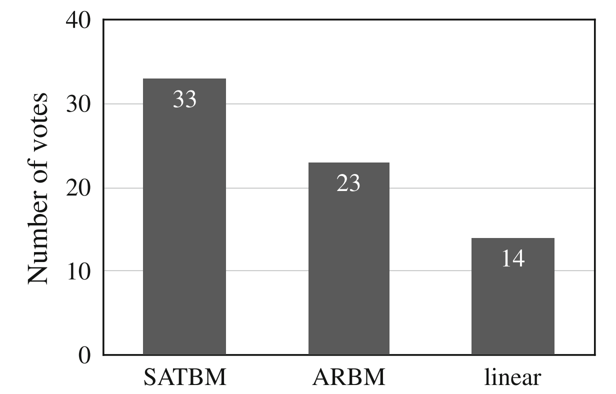

Fig. 6 Subjective performance of non-parallel VC methods

Table 3 MDIR and MCD from various speaker pairs

Source → target	Type	MDIR [dB]	MCD [dB]
ECL0001 → ECL1003	M2F	3.12	6.87
MIT0001 → ECL1003	M2F	4.03	7.00
ECL1003 → CAN0001	F2M	2.77	7.32
NEC1002 → MIT0002	F2M	3.48	6.98
ECL0001 → MIT0001	M2M	1.48	6.09
ECL0001 → MIT0002	M2M	2.02	6.46
CAN1001 → ECL1003	F2F	1.28	6.85
NEC1001 → ECL1003	F2F	1.55	6.14

The results are shown in Table 2. From Table 2, the GMM-based VC outperformed the proposed method in both criteria; however, again, the GMM-based VC uses parallel corpora, while the proposed method does not.

5.4 Evaluation using various speaker pairs

We also investigated the performance of the proposed SATBM using various speaker pairs that include four gender types: male-to-female (M2F), female-to-male (F2M), male-to-male (M2M), and female-to-female (F2F). For this evaluation, we randomly chose eight pairs from the corpus that were not included in the training, whose identities and the gender types are listed in the first and the second columns of Table 3, respectively. Table 3 compares the performance for each speaker pair with different criteria of the MDIR and MCD, using the SATBM with heptadiagonal adaptation matrices. When we compare the VC performance of different speaker pairs using the MCD, we may conclude that the model performed best when converting "ECL0001" to "MIT0001" because this conversion provided the smallest MCD. However, the speech of "ECL0001" and "MIT0001" was already close to each other. To see how effective the model is, we should focus on how much the model improves the original speech. The MDIR measures the extent of how much the model improves, and we can say that the VC model was the most effective when converting "MIT0001" to "ECL1003" according to the MDIR. The MDIR of the cross-gender conversion is higher than that of the within-gender conversion. This is natural because, in general, we can feel the extent of how much the source speech was converted from the converted speech of the cross-gender more than from that of the within-gender.

Table 2 Subjective comparison of parallel VC (GMM) and non-parallel VC (SATBM) using the 5-scale tests in terms of speech quality and speaker specificity

	Speech quality	Speaker specificity
GMM	3.37	3.66
SATBM	2.11	1.91

Finally, Fig. 7 shows the VC performance of the proposed method for each gender type when changing the types of the adaptation matrix. Each MDIR was averaged within the same gender type. As shown in Fig. 7, we achieved better results as the number of diagonals in the adaptation matrices increased and the best when using the heptadiagonal adaptation matrices, except for F2F conversion. The nonadiagonal adaptation matrices performed the worse for all gender types. This is because too many parameters are included in the adaptation matrices even though it is not needed for the linear transformation across the speakers and caused overfitting, as discussed in Section 5.2. We also noticed from Fig. 7 that when comparing the cases with F2F and M2M, the diagonal matrix was not so bad for the F2F conversion. This is because the female voice in general varies with the speaker less than the male voice, and only the diagonal elements in the adaptation matrix were enough to represent the differences in the female speakers.

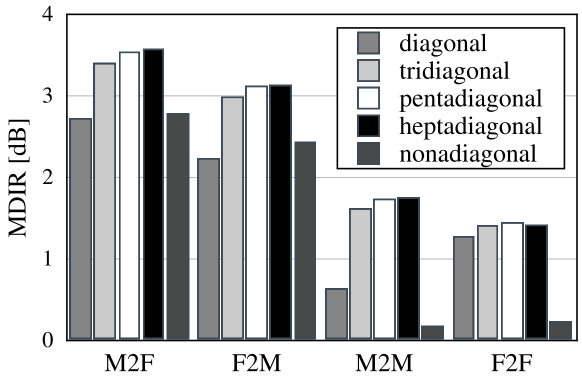

Fig. 7 Comparison of VC performance of the proposed model with changing the type of adaptation matrices for each gender type

6 Conclusions

In this paper, we presented a VC method that does not require any parallel data during training and adaptation according to the basic idea of dividing a speech signal into phoneme-relevant and speaker-relevant information and replacing only the speaker-relevant information with the desired one. To model this, we assumed that the neutral speaker's acoustic features are normally distributed, and its mean is affine-transformed from the latent phonological features that are Bernoulli-distributed. As a result, we showed that the joint probability of the acoustic features and the phonological features forms a type of Boltzmann machine. We also showed the method of estimating the target speaker's features given the source speaker's features in a probabilistic manner. In our VC experiments, we obtained better performance with our model than the other non-parallel VC approaches in both objective and subjective criteria. However, we still have concerns that the proposed approach fell short of the GMM-based approach that uses parallel data in training. In the future, we will continue to improve the system (hopefully, to the performance level of the GMM-based approach) in the non-parallel VC because non-parallel training has several merits, e.g., we can freely use most of the existing speech data.

Endnotes

[1] Note that they still require parallel data among the reference speakers.

[2] It means that the method requires neither the parallel data of a source speaker and target speaker nor the parallel data of the reference speakers.

[3] In our experiments, we used mel cepstra as the acoustic feature vector.

[4] http://research.nii.ac.jp/src/ASJ-JIPDEC.html

Authors' contributions
TN designed the speaker-adaptive-trainable Boltzmann machine, performed the experimental evaluation, and drafted the manuscript. YM reviewed the paper and provided some advice. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 20 October 2016 Accepted: 12 June 2017
Published online: 29 June 2017

References

1. A Kain, MW Macon, in *ICASSP*. Spectral voice conversion for text-to-speech synthesis (IEEE, 1998), pp. 285–288
2. C Veaux, X Robet, in *INTERSPEECH*. Intonation conversion from neutral to expressive speech (ISCA, 2011), pp. 2765–2768
3. K Nakamura, T Toda, H Saruwatari, K Shikano, Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. Speech Commun. **54**(1), 134–146 (2012)
4. L Deng, A Acero, L Jiang, J Droppo, X Huang, in *ICASSP*. High-performance robust speech recognition using stereo training data (IEEE, 2001), pp. 301–304
5. A Kunikoshi, Y Qiao, N Minematsu, K Hirose, in *INTERSPEECH*. Speech generation from hand gestures based on space mapping (ISCA, 2009), pp. 308–311
6. R Gray, Vector quantization. IEEE ASSP Mag. **1**(2), 4–29 (1984)
7. H Valbret, E Moulines, J-P Tubach, Voice transformation using PSOLA technique. Speech Commun. **11**(2), 175–187 (1992)
8. Y Stylianou, Cappé, E Moulines, Continuous probabilistic transform for voice conversion. IEEE Trans. Speech Audio Process. **6**(2), 131–142 (1998)
9. T Toda, AW 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)
10. 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)
11. D Saito, H Doi, N Minematsu, K Hirose, in *INTERSPEECH*. Application of matrix variate Gaussian mixture model to statistical voice conversion (ISCA, 2014), pp. 2504–2508
12. R Takashima, T Takiguchi, Y Ariki, in *SLT*. Exemplar-based voice conversion in noisy environment (IEEE, 2012), pp. 313–317
13. R Takashima, R Aihara, T Takiguchi, Y Ariki, in *SSW8*. Noise-robust voice conversion based on spectral mapping on sparse space (SynSIG, 2013), pp. 71–75
14. S Desai, EV Raghavendra, B Yegnanarayana, AW Black, K Prahallad, in *ICASSP*. Voice conversion using artificial neural networks (IEEE, 2009), pp. 3893–3896
15. LH Chen, ZH Ling, Y Song, LR Dai, in *INTERSPEECH*. Joint spectral distribution modeling using restricted Boltzmann machines for voice conversion (ISCA, 2013), pp. 3052–3056
16. Z Wu, ES Chng, H Li, in *ChinaSIP*. Conditional restricted Boltzmann machine for voice conversion (IEEE, 2013)
17. T Nakashika, R Takashima, T Takiguchi, Y Ariki, in *INTERSPEECH*. Voice conversion in high-order eigen space using deep belief nets (ISCA, 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. A Mouchtaris, J Van 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)
20. C-H Lee, C-H Wu, in *INTERSPEECH*. Map-based adaptation for speech conversion using adaptation data selection and non-parallel training (ISCA, 2006), pp. 2254–2257
21. T Toda, Y Ohtani, K Shikano, in *INTERSPEECH*. Eigenvoice conversion based on Gaussian mixture model (ISCA, 2006), pp. 2446–2449
22. Y Ohtani, T Toda, H Saruwatari, K Shikano, in *NTERSPEECH*. Many-to-many eigenvoice conversion with reference voice (ISCA, 2009), pp. 1623–1626
23. D Saito, K Yamamoto, N Minematsu, K Hirose, in *INTERSPEECH*. One-to-many voice conversion based on tensor representation of speaker space (ISCA, 2011), pp. 653–656
24. T Masuda, M Shozakai, in *ICASSP*. Cost reduction of training mapping function based on multistep voice conversion (IEEE, 2007), pp. 693–696
25. T Nakashika, T Takiguchi, Y Ariki, in *MLSLP 2015*. Parallel-data-free, many-to-many voice conversion using an adaptive restricted Boltzmann machine (ISCA, 2015), pp. 1–4
26. K Cho, A Ilin, T Raiko, in *ICANN*. Improved learning of Gaussian-Bernoulli restricted Boltzmann machines (ENNS, 2011), pp. 10–17
27. R Salakhutdinov, Learning and evaluating Boltzmann machines. Technical report, Technical Report UTML TR 2008-002, Department of Computer Science, University of Toronto (2008)
28. T Anastasakos, J McDonough, R Schwartz, J Makhoul, in *ICSLP 96*. A compact model for speaker-adaptive training, vol. 2 (WASET, 1996), pp. 1137–1140
29. GE Hinton, S Osindero, YW Teh, A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2006)

30. M Morise, in *Proc. the Stockholm Music Acoustics Conference (SMAC)*. An attempt to develop a singing synthesizer by collaborative creation (Logos Verlag, Berlin, 2013), pp. 287–292
31. T Kawahara, A Lee, T Kobayashi, K Takeda, N Minematsu, K Itou, A Ito, M Yamamoto, A Yamada, T Utsuro, K Shikano, Japanese dictation toolkit. J. Acoust. Soc. Japan (E). **20**(3), 233–239 (1999)
32. M Pitz, H Ney, Vocal tract normalization equals linear transformation in cepstral space. IEEE Trans. Speech Audio Process. **13**(5), 930–944 (2005)
33. E Variani, T Schaaf, in *INTERSPEECH*. VTLN in the MFCC domain: band-limited versus local interpolation (ISCA, 2011), pp. 1273–1276
34. T Emori, K Shinoda, Vocal tract length normalization using rapid maximum-likelihood estimation for speech recognition. Syst. Comput. Japan. **33**(5), 30–40 (2002)
35. B Milner, X Shao, in *INTERSPEECH*. Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model (ISCA, 2002), pp. 2421–2424