SPEAKER ADAPTIVE MODEL BASED ON BOLTZMANN MACHINE FOR NON-PARALLEL TRAINING IN VOICE CONVERSION

Toru Nakashika, Yasuhiro Minami

The University of Electro-Communications
Graduate School of Information Systems
nakashika@uec.ac.jp, minami.yasuhiro@is.uec.ac.jp

ABSTRACT

In this paper, we present a voice conversion (VC) method that does not use any parallel data while training the model. VC is a technique where only speaker specific information in 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) mismatch in alignment may happen. Although it is, thus, fairy preferable in VC not to 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 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 a voice-converted speech is obtained by mixing the two. Our experimental results showed that our approach unfortunately fell short of the popular conventional GMM-based method that used parallel data, but outperformed the conventional non-parallel approach.

Index Terms— Voice conversion, Boltzmann machine, unsupervised training, speaker adaptation, 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, 2, 3, 4, 5]. Most of the existing approaches rely on statistical models [6, 7], and the approach based on Gaussian mixture model (GMM) [8, 9, 10, 11] is one of the mainstream 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, 20, 21, 22]. In [19], for example, they model the spectral relationships between two arbitrary speakers (reference speakers) using GMMs, and convert the source speaker's speech using the matrix that projects the feature space of the source speaker into that of the target speaker through that of reference speakers. As a result, parallel data from the source and target speakers is not required. In [21], codebooks (eigenvoice) are obtained using the parallel data of reference speakers, and many-to-many VC is achieved by mapping the source speaker's speech into eigenvoice and the eigenvoice into target speaker's speech.

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 considered to be composed of neutral speech that only includes phonological information and is belong 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 speaker information replaced. The proposed model, called a speaker adaptive trainable Boltzmann machine (SATBM), is designed to help such decomposition.

We have tackled the non-parallel training using another probabilistic model named adaptive Boltzmann machine (ARBM) [23] in our previous work, too. The architecture is different from the proposed model in this paper, which makes some differences; e.g. while an ARBM is based on model-space transformation, a SATBM is based on constrained model-space transformation. In the following sections we will discuss more about this.

2. FORMULATION

In general, it is known that the differences of speech signals in terms of speakers can be represented as multiplication. After the general

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

form, we define an acoustic feature vector $\hat{\boldsymbol{x}}_{rt} \in \mathbb{R}^D$ (D is the number of dimensions) of a speaker r at the time t as follows:

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

where $\boldsymbol{x}_t \in \mathbb{R}^D$, $\mathbf{A}_r \in \mathbb{R}^{D \times D}$ and $\boldsymbol{b}_r \in \mathbb{R}^D$ denote the speaker-normalized acoustic feature vector (acoustic features of the neutral speaker), a speaker adaptation matrix and a bias vector of the speaker r, respectively. Here, we assume that $\boldsymbol{x}_t$ is normally distributed with time-varying mean $\boldsymbol{\mu}_t \in \mathbb{R}^D$ and time-invariant diagonal variance $\boldsymbol{\Sigma} = \mathrm{diag}\,(\boldsymbol{\sigma}^2), \boldsymbol{\sigma}^2 = [\sigma_1^2, \cdots, \sigma_D^2] \in \mathbb{R}^D$. At this time $\hat{\boldsymbol{x}}_{rt}$ is also normally distributed; that is

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

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. Therefore, we assume that the mean vector of the neutral speaker $\boldsymbol{\mu}_t$ is determined using a latent phonological vector $\boldsymbol{h}_t \in \mathbb{B}^H$ ($\mathbb{B}$ is a binary space and H is the number of dimensions of the latent vector) 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 that project the phnological space into the acoustic space. Incidentally, the conditional probability $p(\hat{\boldsymbol{x}}_{rt}|\boldsymbol{h}_t)$ given $\boldsymbol{h}_t$ can be calculated as follows:

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

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, phonological information should be determined by acoustic features as well. It means $h_{jt} \in \boldsymbol{h}_t$ is Bernoulli distributed and its parameter $\pi_{jt} \in \boldsymbol{\pi}_t$ ($j = 1, \cdots, H$) that represents the probablity $p(h_{jt} = 1)$ should be a function of $\boldsymbol{x}_t$. When it comes to formulize this, it is beneficial in reducing the number of parameters to use the already-defined parameters. We define $\boldsymbol{\pi}_t$ as follows:

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

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

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

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

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

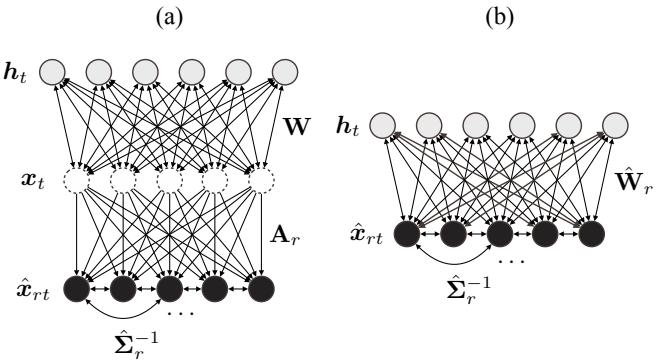

Fig. 1. (a) Proposed model: SATBM (speaker-adaptive-trainable Boltzmann machine) and (b) its simplified representation, which can be seen as a sort of semi-RBM.

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

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 (1) for (7) forms

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

which is nothing else but the definition of a Gaussian-Bernoulli restricted Boltzmann machine (GB-RBM) [24]. 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 and the hidden units of latent phonological features, as shown in Fig. 1. In another viewpoint, it can be regarded as a sort of semi-RBM [25] since there are shared connections $\hat{\mathbf{W}}_r$ between $\hat{\boldsymbol{x}}_{rt}$ and $\boldsymbol{h}_t$, and connections $\hat{\boldsymbol{\Sigma}}_r^{-1}$ among $\hat{\boldsymbol{x}}_{rt}$ but no connections among $\hat{\boldsymbol{h}}_t$ (Fig. 1 (b)). 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 [23], 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 model-space transformation, whereas the SATBM is based on model-space transformation and also feature-space transformation (i.e., constrained model-space transformation), as Eqs. (1) and (2) indicate. In another perspective, the SATBM directly models the correlations between the dimensions in observed features while the ARBM does not. For this reason, we expect the SATBM would be superior in acoustic modeling to the ARBM.

3. PARAMETER ESTIMATION BASED ON SAT

In this section, we describe the way of the parameter estimation in the previously-defined model, a SATBM, based on speaker adaptive training (SAT) [26]. For convenience, we use symbols $\mathbf{\Theta}^{SD} = \{\mathbf{A}_r, \boldsymbol{b}_r\}_{r=1}^{R}$ for speaker-dependent (SD) parameters and $\mathbf{\Theta}^{SI} = \{\mathbf{W}, \boldsymbol{\sigma}^2, \boldsymbol{b}, \boldsymbol{c}\}$ for speaker-independent (SI) parameters. Given a collection of the 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

$$(\hat{\mathbf{\Theta}}^{SD}, \hat{\mathbf{\Theta}}^{SI}) \triangleq \underset{(\mathbf{\Theta}^{SD}, \mathbf{\Theta}^{SI})}{\operatorname{argmax}} \prod_{r=1}^{R} \prod_{t=1}^{T_r} p(\hat{\boldsymbol{x}}_{rt}). \tag{9}$$

According to the SAT paradigm, the SD parameters $\mathbf{\Theta}^{SD}$ undertake the speaker-induced variation, and the SI parameters $\mathbf{\Theta}^{SI}$ 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 more 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 $\theta \in \{\mathbf{\Theta}^{SD}, \mathbf{\Theta}^{SI}\}$ is derived as follows:

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

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) [27] 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:

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

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

When it comes to use the proposed model for VC, we follow three stages of training, adaptation, and conversion. In the training stage, speaker-independent parameters $\hat{\mathbf{\Theta}}^{SI}$ are obtained as in Eq. (9) using R reference speakers' speech (We discard the speaker-dependent parameters $\hat{\mathbf{\Theta}}^{SD}$). In the adaptation stage, new speaker-dependent parameters $\mathbf{\Theta}_i^{SD} = \{\mathbf{A}_i, \boldsymbol{b}_i\}$ and $\mathbf{\Theta}_o^{SD} = \{\mathbf{A}_o, \boldsymbol{b}_o\}$ are estimated using adaptation data of the source and the target speakers $\{\hat{\boldsymbol{x}}_{it}\}_{t=1}^{T_i}$, $\{\hat{\boldsymbol{x}}_{ot}\}_{t=1}^{T_o}$ with keeping $\hat{\mathbf{\Theta}}^{SI}$ fixed. That is,

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

In order 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

$$\begin{aligned}
\boldsymbol{x}_{ot} &\triangleq \underset{\boldsymbol{x}_{ot}}{\operatorname{argmax}}\, p(\boldsymbol{x}_{ot}|\boldsymbol{x}_{it}) \\
&= \underset{\boldsymbol{x}_{ot}}{\operatorname{argmax}} \sum_{\boldsymbol{h}_t} p(\boldsymbol{h}_t|\boldsymbol{x}_{it}) p(\boldsymbol{x}_{ot}|\boldsymbol{h}_t) \\
&\simeq \underset{\boldsymbol{x}_{ot}}{\operatorname{argmax}}\, p(\hat{\boldsymbol{h}}_t|\boldsymbol{x}_{it}) p(\boldsymbol{x}_{ot}|\hat{\boldsymbol{h}}_t) \\
&= \underset{\boldsymbol{x}_{ot}}{\operatorname{argmax}}\, p(\boldsymbol{x}_{ot}|\hat{\boldsymbol{h}}_t) \\
&= \mathbf{A}_o\mathbf{W}\phi(\mathbf{W}^{\top}\mathbf{\Sigma}^{-1}\mathbf{A}_i^{-1}(\boldsymbol{x}_{it} - \boldsymbol{b}_i) + \boldsymbol{c}) + \mathbf{A}_o\boldsymbol{b} + \boldsymbol{b}_o,
\end{aligned} \tag{11}$$

where we give $\hat{\boldsymbol{h}}_t \triangleq \underset{\boldsymbol{h}_t}{\operatorname{argmax}}\, p(\boldsymbol{h}_t|\boldsymbol{x}_{it})$. It is worth noting that the conversion function is based on non-linear tranformation.

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[3]). In the training stage where the SI parameters are estimated, we randomly selected and used speech data of 5 sentences (approx. 160k frames) uttered by 56 speakers (26 males and 30 females) from the set A in the corpus. For adaptation and evaluation, a male and a female speakers that were not included in the training were used as a source and a target speakers, respectively. The amount of the adaptation data was 5 sentences for each person. As an acoustic feature vector, we used 64-dimensional mel-cepstral features that were calculated from 513-dimensional STRAIGHT [28] spectra without dynamic features. In the training of the system, we used 96 hidden units, a learning rate of 0.01, a momentum of 0.9, and a batch-size of 1000, and set the number of iterations as 15 in order to avoid overfitting (already converged). For the evaluation of the proposed method, we used parallel data (of different 10 sentences from in the training and adaptation data) of the source and the target speakers, which was created using dynamic programming. But again, note that every speech data used for the training and the adaptation is NOT parallel.

Mel-cepstral distortion (MCD) is generally used for objective evalulation in VC. However, we used mel-cepstral distortion improvement ratio (MDIR) instead in this paper because it does not make sense to see the distance between the spectral features in mel-scale of the source and the target speakers when we want to recognize the differences in speaker identities, and because the scale of

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

Table 1. Average MDIR [dB] of each method in non-parallel VC

adapt. matrix	full	tridiag	diagonal
linear	1.27	2.07	0.07
ARBM [23]	—	—	1.17
SATBM	2.24	**2.60**	2.19

MCD varies in the evaluation data. The MDIR is defined as follows:

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

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.

5.2. Comparison methods

It is difficult to evaluate the proposed method because most of the existing VC approaches use parallel data in training and comparing our method that does not use parallel data with those methods is not fair. Nevertheless, we can still compare the proposed method with our earlier model, ARBM [23]. 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

$$\boldsymbol{x}_{ot} \triangleq \mathbf{A}_o\mathbf{A}_i^{-1}(\boldsymbol{x}_{it} - \boldsymbol{b}_i) + \boldsymbol{b}_o, \tag{12}$$

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 $\mathbf{A}_r$, $\boldsymbol{b}_r$ are estimated in SAT using gradient decent just the same as our proposed method. So the difference between the linear-transform approach and the proposed model is whether latent phonological features are modeled or not.

Just for a reference, we also compared with a popular GMM-based VC with 64 mixtures using parallel data of 5 sentences.

5.3. Results and discussion

The VC performance of the linear-transform-based approach, the ARBM, and the proposed model is summarized in Table. 1. Each method is compared with changing the type of the adaptation matrix $\mathbf{A}_r$ as a full-rank matrix, a tridiagonal matrix, and a diagonal matrix. As shown in Table. 1, the proposed model with a tridiagonal adaptation matrix performed best of all with any types of the adaptation matrix. When we see the results of a diagonal matrix, the linear approach hardly improved the source speech closed to the target one because it was considered that the diagonal matrix could not capture the correlations between dimensions of the mel-cepstrum, which makes impossible to match the vocal tracts. On the other hand, the ARBM and the SATBM could get the source speech closed to the target voice more or less even when a diagonal adaptation matrix was used due to modeling latent phonological information. The reason why the full rank matrix degrades the MDIR was due to large number of parameters that caused overfiting. Furthermore, it is known that the tridiagonal elements are sufficient for warping vocal tracts [29]; hence we obtained better results from the case with a tridiagonal matrix in the linear and SATBM approaches.

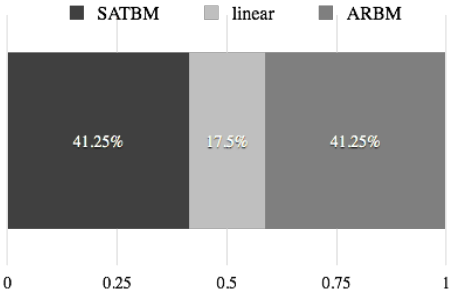

Fig. 2. The ratios of voting for preference in subjective evaluation.

The average MDIR of the GMM-based approach was 3.86. Unfortunately, it was better performance than our approach. However, such approach takes a benefit from the parallel data that restricts to match the frames of the source and the target features. It can be considered the influence of using no parallel data.

5.4. Subjective evaluation

We also conducted subjective experiments, comparing our method with the ARBM and the linear-based approach using a diagonal adaptation matrix for all methods. We decoded the the converted mel-cepstra back to STRAIGHT spectra using filter-theory [30], and generated signals using the original F0 and aperiodic features of the target speaker since we wanted to compare each method in spectra. In this experiments, 7 participants listened 10 sentences of converted speech by the linear-based, the ARBM, and the SATBM approaches accompanied with the target speech and voted for the most preferable one for each sentence in terms of the speaker specificity of the target speaker. The ratios of voting of each method are shown in Fig. 2. The SATRM and the ARBM obtained were both outperformed the linear-based approach but produced the same ratios unexpectedly. It can bs said that the SATBM and the ARBM have the similar pottential in modeling speaker-specificity when a diagonal adaptation matrix is used.

6. CONCLUSION

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 affin-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 sort of a 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 objective 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 up to around the performance of the GMM-based approach) in non-parallel VC because non-parallel training has a lot of merits; e.g. we can freely use the most of existing speech data.

7. REFERENCES

[1] Alexander Kain and Michael W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *ICASSP*, 1998, pp. 285–288.

[2] Christophe Veaux and X. Robet, "Intonation conversion from neutral to expressive speech," in *INTERSPEECH*, 2011, pp. 2765–2768.

[3] Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, "Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[4] Li Deng, Alex Acero, Li Jiang, Jasha Droppo, and Xuedong Huang, "High-performance robust speech recognition using stereo training data," in *ICASSP*, 2001, pp. 301–304.

[5] Aki Kunikoshi, Yu Qiao, Nobuaki Minematsu, and Keikichi Hirose, "Speech generation from hand gestures based on space mapping," in *INTERSPEECH*, 2009, pp. 308–311.

[6] Robert Gray, "Vector quantization," *IEEE ASSP Magazine*, vol. 1, no. 2, pp. 4–29, 1984.

[7] H. Valbret, E. Moulines, and Jean-Pierre Tubach, "Voice transformation using PSOLA technique," *Speech Communication*, vol. 11, no. 2, pp. 175–187, 1992.

[8] Yannis Stylianou, Olivier Cappé, and Eric Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Process.*, vol. 6, no. 2, pp. 131–142, 1998.

[9] Tomoki Toda, Alan W. Black, and Keiichi Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[10] Elina Helander, Tuomas Virtanen, Jani Nurminen, and Moncef Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 5, pp. 912–921, 2010.

[11] Nobuaki Minematsu Daisuke Saito, Hidenobu Doi and Keikichi Hirose, "Application of matrix variate Gaussian mixture model to statistical voice conversion," in *INTERSPEECH*, 2014, pp. 2504–2508.

[12] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Exemplar-based voice conversion in noisy environment," in *SLT*, 2012, pp. 313–317.

[13] Ryoichi Takashima, Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki, "Noise-robust voice conversion based on spectral mapping on sparse space," in *SSW8*, 2013, pp. 71–75.

[14] Srinivas Desai, E. Veera Raghavendra, B. Yegnanarayana, Alan W. Black, and Kishore Prahallad, "Voice conversion using artificial neural networks," in *ICASSP*, 2009, pp. 3893–3896.

[15] L. H. Chen, Z. H. Ling, Yan Song, and L. R. Dai, "Joint spectral distribution modeling using restricted Boltzmann machines for voice conversion," in *INTERSPEECH*, 2013, pp. 3052–3056.

[16] Zhizheng Wu, Eng Siong Chng, and Haizhou Li, "Conditional restricted Boltzmann machine for voice conversion," in *ChinaSIP*, 2013.

[17] Toru Nakashika, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Voice conversion in high-order eigen space using deep belief nets," in *INTERSPEECH*, 2013, pp. 369–372.

[18] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, "Voice conversion using RNN pre-trained by recurrent temporal restricted Boltzmann machines," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 23, no. 3, pp. 580–587, 2015.

[19] Athanasios Mouchtaris, Jan Van der Spiegel, and Paul Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE Trans. Audio, Speech, and Lang. Processs*, vol. 14, no. 3, pp. 952–963, 2006.

[20] Chung-Han Lee and Chung-Hsien Wu, "Map-based adaptation for speech conversion using adaptation data selection and non-parallel training," in *INTERSPEECH*, 2006, pp. 2254–2257.

[21] Tomoki Toda, Yamato Ohtani, and Kiyohiro Shikano, "Eigenvoice conversion based on Gaussian mixture model," in *INTERSPEECH*, 2006, pp. 2446–2449.

[22] Daisuke Saito, Keisuke Yamamoto, Nobuaki Minematsu, and Keikichi Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in *INTERSPEECH*, 2011, pp. 653–656.

[23] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, "Parallel-data-free, many-to-many voice conversion using an adaptive restricted boltzmann machine," in *MLSLP 2015*, 2015, pp. 1–4.

[24] KyungHyun Cho, Alexander Ilin, and Tapani Raiko, "Improved learning of Gaussian-Bernoulli restricted Boltzmann machines," in *ICANN*, pp. 10–17. Springer, 2011.

[25] Ruslan Salakhutdinov, "Learning and evaluating Boltzmann machines," Tech. Rep., Technical Report UTML TR 2008-002, Department of Computer Science, University of Toronto, 2008.

[26] Tasos Anastasakos, John McDonough, Richard Schwartz, and John Makhoul, "A compact model for speaker-adaptive training," in *ICSLP 96*. IEEE, 1996, vol. 2, pp. 1137–1140.

[27] Geoffrey E. Hinton, Simon Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[28] Hideki Kawahara, Masanori Morise, Toru Takahashi, Ryuichi Nisimura, Toshio Irino, and Hideki Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *ICASSP*, 2008, pp. 3933–3936.

[29] Tadashi Emori and Koichi Shinoda, "Vocal tract length normalization using rapid maximum-likelihood estimation for speech recognition," *Systems and Computers in Japan*, vol. 33, no. 5, pp. 30–40, 2002.

[30] Ben Milner and Xu Shao, "Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model," in *INTERSPEECH*, 2002, pp. 2421–2424.