3WRBM-Based Speech Factor Modeling for Arbitrary-Source and Non-Parallel Voice Conversion

Toru Nakashika and Yasuhiro Minami
Graduate School of Informatics and Engineering
The University of Electro-Communications
Chofu, Japan 182-8585
Email: nakashika@uec.ac.jp, minami.yasuhiro@is.uec.ac.jp

***Abstract*—In recent years, voice conversion (VC) becomes a popular technique since it can be applied to various speech tasks. Most existing approaches on VC must use aligned speech pairs (parallel data) of the source speaker and the target speaker in training, which makes hard to handle it. Furthermore, VC methods proposed so far require to specify the source speaker in conversion stage, even though we just want to obtain the speech of the target speaker from the other speakers in many cases of VC. In this paper, we propose a VC method where it is not necessary to use any parallel data in the training, nor to specify the source speaker in the conversion. Our approach models a joint probability of acoustic, phonetic, and speaker features using a three-way restricted Boltzmann machine (3WRBM). Speaker-independent (SI) and speaker-dependent (SD) parameters in our model are simultaneously estimated under the maximum likelihood (ML) criteria using a speech set of multiple speakers. In conversion stage, phonetic features are at first estimated in a probabilistic manner given a speech of an arbitrary speaker, then a voice-converted speech is produced using the SD parameters of the target speaker. Our experimental results showed not only that our approach outperformed other non-parallel VC methods, but that the performance of the arbitrary-source VC was close to those of the traditional source-specified VC in our approach.**

***Index Terms*—Voice conversion, three-way restricted Boltzmann machine, unsupervised learning, speaker adaptation, non-parallel training.**

I. Introduction

Voice conversion (VC), where speaker-specific information in the speech of a source speaker is changed into that of a target speaker while retaining linguistic information, garners much attention nowadays because the VC techniques can be applied to various tasks such as speech synthesis, aid for people with articulation disorders, entertainment as voice changer, etc [1], [2], [3], [4], [5]. Most of the existing approaches rely on statistical models such as Gaussian mixture model (GMM) [6], [7], [8], [9], which is one of the mainstream, non-negative matrix factorization (NMF) [10], [11], neural networks (NNs) [12], restricted Boltzmann machines (RBMs) [13], [14], and deep learning [15], [16], etc.

However, these methods require parallel data (speech data of the source and the target speakers aligned so that each frame of the source speaker's data corresponds to that of the target speaker) for training the models, which hinders ease of use; 1) the data is limited to pre-defined articles (both speakers must utter the same articles), 2) 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. Furthermore, the aligned data is not the original speech anymore because the speech data is stretched and modified in the time axis when aligned, and it is not guaranteed that each frame is aligned perfectly, and such mismatching may cause some errors in training. Several approaches, such as eigenvoice and MAP [17], [18], [19], that do not use parallel data between the source and the target speakers have been proposed, although such methods still require parallel data between reference speakers to obtain the speaker-independent space. Erro *et al* proposed a non-parallel training method even on reference speakers based on an iterative approach called an INCA (iterative combination of a nearest neighbor search step and a conversion step alignment) algorithm [20]. Our earlier works also tackled with the non-parallel training using probabilistic models named adaptive Boltzmann machine (ARBM) [21], and speaker-adaptive-trainable Boltzmann machine (SATBM) [22]. Although the speech quality produced by the non-parallel approaches may fall short of that of the parallel approaches, the non-parallel approaches improve convenience and practicality since the models can be trained using existing speech data as it is.

In this paper, we propose a VC method that enhances convenience (easy-to-handle) in VC. Our approach requires neither parallel data in the training, nor specification of the source speaker in the conversion. As far as we know, one must specify the source speaker on the conversion stage in all of the conventional VC approaches. However, in many cases when using VC, we only have to convert given speech into that of the desired speaker; therefore, the VC that does not require the specification of the source speaker will be more convenient than the VC that does[1]. The former VC can be achieved by combining the existing VC and speaker recognition techniques; meanwhile our approach tries to achieve this in a probabilistic manner using a single model. In our approach, a three-way restricted Boltzmann machine (3WRBM) [23] is used to model the relationships between fundamental speech factors of acoustic, phonetic, and speaker features. The 3WRBM is a energy-based probabilistic model that extends the well-known two-layer RBM [24], [25] so that it represents up to three-order

[1] We refer to the former and the latter types as arbitrary-source VC and source-specified VC, respectively.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

potentials among three different factors. It is assumed that there are undirected connection weights between the different factors, but no connections between the same factors like an RBM. The connection weights may represent the strength of the relationships between the factors. We further add several constraints on the connection weights under the assumption that an observed acoustic features are from the *neutral* acoustic features that are not dependent on any speakers but on the latent, phonetic features, multiplied with the speaker-specific adaptation matrix. In other words, a speech signal of an arbitrary speaker is considered to be composed of neutral speech that only includes phonetic information, accompanied with the speaker specific information. Our VC scheme can be formulated as MAP estimation, which results in two steps: 1) to estimate phonetic features given acoustic features by marginalizing over speakers, and 2) to estimate the desired acoustic features from the phonetic features and the SD parameters of the target speaker.

The proposed approach may resemble our previous works [21], [22] in terms of unsupervised learning to decompose a speech signal into phonetic- and speaker-related information. The most significant difference is that our approach regards speaker-identity features as variables that can be sampled, and hence makes it possible to convert the voice from arbitrary speakers, while the previous approaches do not.

II. Modelling speech using 3WRBM

A well-known energy-based probabilistic model of visible and hidden variables, restricted Boltzmann machine (RBM), can be generally extended so as to represent more than two variables [26]. Especially we call the model of three variables three-way RBM. In this paper, we define the relationships among three types of variables (descriptors) of acoustic features (mainly cepstrum-based features) $\boldsymbol{v} = [v_1, \cdots, v_D] \in \mathbb{R}^D$, latent features $\boldsymbol{h} = [h_1, \cdots, h_H] \in \{0,1\}^H, \sum_j h_j = 1$, and speaker features $\boldsymbol{s} = [s_1, \cdots, s_R] \in \{0,1\}^R, \sum_k s_k = 1$ using a 3WRBM, where D, H, and R indicate the numbers of the acoustic features, the latent features, and the speakers. In our approach, we only target on modelling clean speech by various speakers; therefore, the latent features $\boldsymbol{h}$ may represent phonetic-related information[2] that are not observable but exist behind the speech, since the variation caused by speakers is captured by the speaker features $\boldsymbol{s}$. $\boldsymbol{h}$ and $\boldsymbol{s}$ are defined as one-hot vectors, and have values of 1 if only the element of interest is activated. For example, the statements $h_j = 1, \forall h_{j'} = 0 \; (j' \neq j)$ and $s_k = 1, \forall s_{k'} = 0 \; (k' \neq k)$ indicate that the jth phonetic feature acts on the speech at that time, and that the kth speaker uttered, respectively. The joint probability of the three descriptors is defined as follows:

$$p(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) = \frac{1}{N} e^{-E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s})}, \tag{1}$$

where N denotes the normalization term. The energy function $E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s})$ is defined as:

$$E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) = U(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) + P(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) + T(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) \tag{2}$$

$$U(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) = \frac{1}{2} \boldsymbol{v}^\top \bar{\boldsymbol{v}} - \boldsymbol{b}^\top \bar{\boldsymbol{v}} - \boldsymbol{c}^\top \boldsymbol{h} - \boldsymbol{d}^\top \boldsymbol{s} \tag{3}$$

$$P(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) = -\bar{\boldsymbol{v}}^\top \mathbf{W} \boldsymbol{h} - \boldsymbol{h}^\top \mathbf{V} \boldsymbol{s} - \boldsymbol{s}^\top \mathbf{U} \bar{\boldsymbol{v}} \tag{4}$$

$$T(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) = -\sum_{i,j,k} \bar{v}_i h_j s_k Z_{ijk}, \tag{5}$$

where we denote $\bar{\boldsymbol{v}}$ as the normalized acoustic features ($\bar{\boldsymbol{v}} = [\bar{v}_i] = [\frac{v_i}{\sigma_i^2}]$). $U(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s})$, $P(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s})$, and $T(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s})$ describe the unary potentials, the pairwise potentials, and the three-way potentials of the three descriptors, respectively, where $\boldsymbol{b} \in \mathbb{R}^D$, $\boldsymbol{c} \in \mathbb{R}^H$, $\boldsymbol{d} \in \mathbb{R}^R$, and $\boldsymbol{\sigma} = [\sigma_i] \in \mathbb{R}^D$ are bias terms of the acoustic features, of the phonetic features, of the speaker features, and variance terms of the acoustic features, $\mathbf{W} \in \mathbb{R}^{D \times H}$, $\mathbf{V} \in \mathbb{R}^{H \times R}$, and $\mathbf{U} \in \mathbb{R}^{R \times D}$ are pairwise weights of $\boldsymbol{v}$ and $\boldsymbol{h}$, $\boldsymbol{h}$ and $\boldsymbol{s}$, and $\boldsymbol{s}$ and $\boldsymbol{v}$, and $\mathcal{Z} \in \mathbb{R}^{D \times H \times K}$ is the three-way weights, whose element Z_{ijk} is of v_i, h_j, and s_k. Like an RBM, because there are no connections between visible features, between phonetic features, or between speaker features, the conditional probabilities $p(\boldsymbol{v}|\boldsymbol{h}, \boldsymbol{s})$, $p(\boldsymbol{h}|\boldsymbol{s}, \boldsymbol{v})$, and $p(\boldsymbol{s}|\boldsymbol{v}, \boldsymbol{h})$ form simple equations as follows:

$$p(\boldsymbol{v}|\boldsymbol{h}, \boldsymbol{s}) = \mathcal{N}(\boldsymbol{v} \mid \boldsymbol{b} + \mathbf{W}\boldsymbol{h} + \mathbf{U}^\top \boldsymbol{s} + \sum_{j,k} h_j s_k \mathcal{Z}_{:jk}, \boldsymbol{\sigma}^{\mathbf{2}})$$

$$p(\boldsymbol{h}|\boldsymbol{s}, \boldsymbol{v}) = \mathcal{B}(\boldsymbol{h} \mid \boldsymbol{f}(\boldsymbol{c} + \mathbf{V}\boldsymbol{s} + \mathbf{W}^\top \bar{\boldsymbol{v}} + \sum_{i,k} \bar{v}_i s_k \mathcal{Z}_{i:k}))$$

$$p(\boldsymbol{s}|\boldsymbol{v}, \boldsymbol{h}) = \mathcal{B}(\boldsymbol{s} \mid \boldsymbol{f}(\boldsymbol{d} + \mathbf{U}\bar{\boldsymbol{v}} + \mathbf{V}^\top \boldsymbol{h} + \sum_{i,j} \bar{v}_i h_j \mathcal{Z}_{ij:}))$$

where $\mathcal{N}(\cdot|\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, $\mathcal{B}(\cdot|\boldsymbol{\pi})$, and $\boldsymbol{f}(\cdot)$ indicate an element-wise Gaussian probability density function with the means $\boldsymbol{\mu}$ and variances $\boldsymbol{\sigma}^2 = [\sigma_i^2]$, a multivariate Bernoulli distribution with the probabilities $\boldsymbol{\pi}$ of each taking the value of 1, and an element-wise softmax function, respectively. $\mathcal{Z}_{:jk}$, $\mathcal{Z}_{i:k}$, and $\mathcal{Z}_{ij:}$ denote the partial vectors of $\mathcal{Z}$ along the first, the second, and the third modes, respectively. The model defined in Eq. (1) closely resembles a factored 3WRBM found in [23]. The significant difference is that a factored 3WRBM deals with *one* visible descriptor with a hidden descriptor and models the third-order relationships among two visible units and a hidden unit, while our model deals with two visible descriptors and a hidden descriptor to capture the relationships among three units of the first visible, the second visible and the hidden descriptors. Note that there are no connections between units belonging to the same descriptors in our model unlike a factored 3WRBM.

A. Constraints on phonetic- and speaker-related factors

The model defined in the previous section has a large number of parameters and no constraints no parameters, which causes overfitting and difficulties in training. Therefore, it would be better to add some constraints to the model. In this paper, we redefine the 3WRBM with structured parameters,

[2] So, we may call $\boldsymbol{h}$ as phonetic features.

motivated by the well-known speech modeling with affine-transformation.

When we look at the parameters of three-way potentials $\mathcal{Z}_{:jk}$, we may notice that the energy related to these parameters when a phoneme j and a speaker k are activated is calculated as negative inner product of $\boldsymbol{v}'$ and $\mathcal{Z}_{:jk}$, which is $T(\boldsymbol{v}, h_j = 1, s_k = 1) = -\bar{\boldsymbol{v}}^\top \mathcal{Z}_{:jk}$. The negative inner product takes a small value when the normalized acoustic features are close to the parameter vector $\mathcal{Z}_{:jk}$. In other words, under the stable (low-energy) condition, $\mathcal{Z}_{:jk}$ represents the acoustic pattern that often appears in the training data and that depends on the jth phoneme and the kth speaker. Considering decomposing the pattern $\mathcal{Z}_{:jk}$ into phoneme-related and speaker-related factors, we define

$$\mathcal{Z}_{:jk} = \mathbf{A}_k \boldsymbol{m}_j, \tag{6}$$

where $\boldsymbol{m}_j \in \mathbb{R}^D$ and $\mathbf{A}_k \in \mathbb{R}^{D \times D}$ denote the factors related to the phoneme j and to the speaker k, respectively. Eq. (6) indicates that $\mathcal{Z}_{:jk}$ is obtained by projecting the feature vector $\boldsymbol{m}_j$ of the phoneme j into the speaker k's space with his/her own matrix $\mathbf{A}_k$. Since it is generally known that the speaker-induced modification is formulated as affine-transformation in the cepstrum-based domain [27], [28], the formulation in Eq. (6) is considered to be reasonable. Therefore, $\boldsymbol{m}_j$ and $\mathbf{A}_k$ indicate the acoustic pattern of the phoneme j that does not depend on any speakers (*neutral* acoustic pattern) and the adaptation matrix of the speaker k that projects neutral acoustic patterns into the speaker-specific space, respectively. The $\boldsymbol{m}_j$ can represent the relationships between the phoneme j and the acoustic features; hence, we set $\mathbf{W}_{:j} = \mathbf{0}$.

In addition, the bias d_k of the speaker k may represent something such as *frequency* of the speaker k in the training data. In this study, we do not use such biases on speakers, i.e., $\boldsymbol{d} = \mathbf{0}$, in order to treat speakers impartially.

Summarizing the above discussion, we redefine the energy function for modeling speech as follows:

$$\begin{aligned} & E(\boldsymbol{v}, \boldsymbol{h}, \boldsymbol{s}) \\ & = \frac{1}{2}\boldsymbol{v}^\top \bar{\boldsymbol{v}} - \boldsymbol{b}^\top \bar{\boldsymbol{v}} - \boldsymbol{c}^\top \boldsymbol{h} - \boldsymbol{h}^\top \mathbf{V} \boldsymbol{s} - \boldsymbol{s}^\top \mathbf{U} \bar{\boldsymbol{v}} - \bar{\boldsymbol{v}}^\top \mathbf{A}_{\boldsymbol{s}} \mathbf{M} \boldsymbol{h}, \end{aligned} \tag{7}$$

where we use $\mathbf{A}_{\boldsymbol{s}} = \sum_k \mathbf{A}_k s_k$ and $\mathbf{M} = [\boldsymbol{m}_1 \ \cdots \ \boldsymbol{m}_H]$. With this reformulation, we obtain

$$p(\boldsymbol{v}|\boldsymbol{h}, \boldsymbol{s}) = \mathcal{N}(\boldsymbol{v} \mid \boldsymbol{b} + \mathbf{U}^\top \boldsymbol{s} + \mathbf{A}_{\boldsymbol{s}} \mathbf{M} \boldsymbol{h}, \boldsymbol{\sigma}^2) \tag{8}$$

$$p(\boldsymbol{h}|\boldsymbol{s}, \boldsymbol{v}) = \mathcal{B}(\boldsymbol{h} \mid \boldsymbol{f}(\boldsymbol{c} + \mathbf{V}\boldsymbol{s} + \mathbf{M}^\top \mathbf{A}_{\boldsymbol{s}}^\top \bar{\boldsymbol{v}})) \tag{9}$$

$$p(\boldsymbol{s}|\boldsymbol{v}, \boldsymbol{h}) = \mathcal{B}(\boldsymbol{s} \mid \boldsymbol{f}(\mathbf{U}\bar{\boldsymbol{v}} + \mathbf{V}^\top \boldsymbol{h} + [\bar{\boldsymbol{v}}^\top \mathbf{A}_k]\mathbf{M}\boldsymbol{h})). \tag{10}$$

Letting $\mathcal{A} \in \mathbb{R}^{D \times D \times R}$ be a third order tensor whose elements are $\mathbf{A}_k$ in the third mode, the proposed model defined in Eq. (7) is graphically represented as shown in Fig. 1.

B. Parameter estimation

Given a collection of training speech data $\boldsymbol{X} = \{\boldsymbol{v}_t, \boldsymbol{s}_t\}_{t=1}^T$ that has T frames composed of R speakers, the parameters of the proposed model $\Theta = \{\mathbf{M}, \mathcal{A}, \mathbf{U}, \mathbf{V}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\sigma}\}$ are simultaneously estimated so as to maximize the log-likelihood as

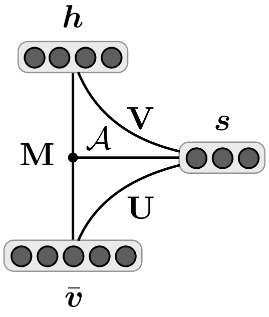

Fig. 1. Graphical representation of the proposed speech factor modeling.

$$\mathcal{L} = \log p(\boldsymbol{X}) = \sum_t \log \sum_{\boldsymbol{h}} p(\boldsymbol{v}_t, \boldsymbol{h}_t, \boldsymbol{s}_t). \tag{11}$$

In this paper, the parameters are iteratively updated using stochastic gradient descent in the similar way to the training of an RBM. Partially-differentiating the log-likelihood $\mathcal{L}$ in terms of each parameter, we obtain

$$\frac{\partial \mathcal{L}}{\partial \mathbf{M}} = \langle \sum_k \mathbf{A}_k^\top \bar{\boldsymbol{v}} \boldsymbol{h}^\top s_k \rangle_{\text{data}} - \langle \sum_k \mathbf{A}_k^\top \bar{\boldsymbol{v}} \boldsymbol{h}^\top s_k \rangle_{\text{model}} \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{A}_k} = \langle \bar{\boldsymbol{v}} \boldsymbol{h}^\top s_k \mathbf{M}^\top \rangle_{\text{data}} - \langle \bar{\boldsymbol{v}} \boldsymbol{h}^\top s_k \mathbf{M}^\top \rangle_{\text{model}} \tag{13}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{U}} = \langle \boldsymbol{s} \bar{\boldsymbol{v}}^\top \rangle_{\text{data}} - \langle \boldsymbol{s} \bar{\boldsymbol{v}}^\top \rangle_{\text{model}} \tag{14}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{V}} = \langle \boldsymbol{h} \boldsymbol{s}^\top \rangle_{\text{data}} - \langle \boldsymbol{h} \boldsymbol{s}^\top \rangle_{\text{model}} \tag{15}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{b}} = \langle \bar{\boldsymbol{v}} \rangle_{\text{data}} - \langle \bar{\boldsymbol{v}} \rangle_{\text{model}} \tag{16}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{c}} = \langle \boldsymbol{h} \rangle_{\text{data}} - \langle \boldsymbol{h} \rangle_{\text{model}} \tag{17}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\sigma}} = & \frac{1}{\boldsymbol{\sigma}^3} \circ \Big(\langle \boldsymbol{v} \circ \boldsymbol{v} - 2\boldsymbol{v} \circ \big(\boldsymbol{b} + \mathbf{U}^\top \boldsymbol{s} + \mathbf{A}_{\boldsymbol{s}} \mathbf{M} \boldsymbol{h}\big) \rangle_{\text{data}} \\ & - \langle \boldsymbol{v} \circ \boldsymbol{v} - 2\boldsymbol{v} \circ \big(\boldsymbol{b} + \mathbf{U}^\top \boldsymbol{s} + \mathbf{A}_{\boldsymbol{s}} \mathbf{M} \boldsymbol{h}\big) \rangle_{\text{model}} \Big), \end{aligned}$$

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) [24] and efficiently approximate them with the expectations of the reconstructed data $\langle \cdot \rangle_{\text{recon.}}$. In the CD scheme, the reconstructed data is calculated using randomly-sampled variables, starting from the original data. In this paper, we sample variables of each descriptor $\boldsymbol{h}$, $\boldsymbol{s}$, $\boldsymbol{v}$ in order using Gibbs chain; e.g., we first sample $\tilde{\boldsymbol{h}}$ as $\tilde{\boldsymbol{h}} \sim p(\boldsymbol{h}|\boldsymbol{s}, \boldsymbol{v})$, secondly sample $\tilde{\boldsymbol{s}}$ as $\tilde{\boldsymbol{s}} \sim p(\boldsymbol{s}|\boldsymbol{v}, \tilde{\boldsymbol{h}})$, thirdly sample $\tilde{\boldsymbol{v}}$ as $\tilde{\boldsymbol{v}} \sim p(\boldsymbol{v}|\tilde{\boldsymbol{h}}, \tilde{\boldsymbol{s}})$, etc.

III. Application to arbitrary-source VC

The goal of arbitrary-source VC is to change the input speech of any persons as if the particular target speaker spoke. After the training discussed in the previous section, we have the model parameters $\Theta = \{\mathbf{M}, \mathcal{A}, \mathbf{U}, \mathbf{V}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{\sigma}\}$ that includes the parameters of the target speaker o. Now given the frame-wise acoustic features $\boldsymbol{v}^{(i)}$ of the arbitrary speaker's speech that we want to convert to that of the target speaker $\boldsymbol{v}^{(o)}$ with the identity vector $\boldsymbol{s}^{(o)}$ where only oth element takes the value of 1 (otherwise 0), we estimate $\boldsymbol{v}^{(o)}$ using MAP (maximum a posteriori) as follows:

$$\begin{aligned}\hat{\boldsymbol{v}}^{(o)} &\triangleq \underset{\boldsymbol{v}^{(o)}}{\operatorname{argmax}}\, p(\boldsymbol{v}^{(o)}|\boldsymbol{v}^{(i)}, \boldsymbol{s}^{(o)}) \\ &= \underset{\boldsymbol{v}^{(o)}}{\operatorname{argmax}} \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}^{(i)}, \boldsymbol{s}^{(o)}) p(\boldsymbol{v}^{(o)}|\boldsymbol{h}, \boldsymbol{v}^{(i)}, \boldsymbol{s}^{(o)}) \\ &\simeq \underset{\boldsymbol{v}^{(o)}}{\operatorname{argmax}}\, p(\hat{\boldsymbol{h}}|\boldsymbol{v}^{(i)}, \boldsymbol{s}^{(o)}) p(\boldsymbol{v}^{(o)}|\hat{\boldsymbol{h}}, \boldsymbol{v}^{(i)}, \boldsymbol{s}^{(o)}) \\ &= \underset{\boldsymbol{v}^{(o)}}{\operatorname{argmax}}\, p(\boldsymbol{v}^{(o)}|\hat{\boldsymbol{h}}, \boldsymbol{s}^{(o)}) \\ &= \boldsymbol{b} + \mathbf{U}_{o:}^{\top} + \mathbf{A}_o \mathbf{M} \hat{\boldsymbol{h}}, \end{aligned} \tag{18}$$

where we use $\hat{\boldsymbol{h}} \triangleq \mathbb{E}[\boldsymbol{h}|\boldsymbol{v}^{(i)}]$, which is regarded as the most likely phonetic features calculated from the input acoustic features $\boldsymbol{v}^{(i)}$. Thanks to the definition of $\boldsymbol{s}$ as variables, we can rewrite $\hat{\boldsymbol{h}}$ as follows:

$$\begin{aligned}\hat{\boldsymbol{h}} &\triangleq \mathbb{E}[\boldsymbol{h}|\boldsymbol{v}^{(i)}] \\ &= \left[p(h_j = 1|\boldsymbol{v}^{(i)})\right] \\ &= \left[\frac{\sum_{\boldsymbol{s}} p(\boldsymbol{v}^{(i)}, h_j = 1, \boldsymbol{s})}{\sum_{\boldsymbol{h}'} \sum_{\boldsymbol{s}'} p(\boldsymbol{v}^{(i)}, \boldsymbol{h}', \boldsymbol{s}')}\right] \\ &= \boldsymbol{f}(\boldsymbol{c} + \boldsymbol{g}(\mathbf{V} + \bar{\boldsymbol{v}}^{(i)\top}\mathbf{U}^{\top} + \mathbf{M}^{\top}[\mathbf{A}_k^{\top}\bar{\boldsymbol{v}}^{(i)}])), \end{aligned} \tag{19}$$

where $\bar{\boldsymbol{v}}^{(i)} = [\frac{v_i^{(i)}}{\sigma_i^2}]$ and $\boldsymbol{g}(\mathbf{X}) = \log \sum_k e^{\mathbf{X}_{:k}}$ indicates an element-wise generalized softplus function. In short, the proposed VC scheme has two steps: 1) calculate Eq. (19) to obtain speaker-independent phonetic features included in the input acoustic vector, and 2) calculate Eq. (18) to obtain desired acoustic features using the phonetic features and the target speaker's parameters. As Eq. (19) indicates, our VC method does not specify the source speaker's parameters. In conventional methods, such as SATBM, the speaker-identity vector $\boldsymbol{s}$ is not defined as variables; therefore, if we want to achieve the arbitrary-source VC in such methods, we have to first estimate the speaker, and then convert the speech specifying the source speaker. This approach is possibly largely affected by the accuracy of the speaker recognition.

IV. Experimental evaluation

A. System configuration

In our VC experiments, we evaluated the performance of the proposed model, using ASJ Continuous Speech Corpus for Research (ASJ-JIPDEC[3]). In the training stage, we randomly selected and used speech data of 5 sentences (approx. 160k frames) uttered by $R = 58$ speakers (27 males and 31 females) from the set A in the corpus. In the conversion stage, we randomly picked up a male speaker (identified with "ECL0001" in the dataset) and a female speaker ("ECL1003") from the training set as a source and a target speakers, respectively, unless otherwise stated. As an acoustic feature vector, we used 32-dimensional mel-cepstral features that were calculated from 513-dimensional WORLD [29] spectra without dynamic features. In the training of the system, we used 16 softmax hidden units (phonetic features), a learning rate of 0.01, a momentum of 0.9, and a batch-size of $R \times 100 (= 5800)$, and set the number of iterations as 200. For the evaluation of the proposed method, we used parallel data (of different 10 sentences from in the training 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 is NOT parallel.

TABLE I
Comparison of non-parallel yet speaker-specified VC methods.

Method	ARBM	SATBM	Proposed
MDIR [dB]	2.11	2.66	**3.07**

TABLE II
Performance of the proposed arbitrary-source VC method.

	MDIR [dB]
Correct speaker specified	3.07
Different speaker specified	2.79
Arbitrary source approach	3.03

Mel-cepstral distortion (MCD) is generally used for objective evaluation 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 MCD varies in the evaluation data. The MDIR is defined as follows:

$$MDIR[dB] = \frac{10\sqrt{2}}{\ln 10}(\left\|\boldsymbol{v}^{(o)} - \boldsymbol{v}^{(i)}\right\|^2 - \left\|\boldsymbol{v}^{(o)} - \hat{\boldsymbol{v}}^{(o)}\right\|^2)$$

where $\boldsymbol{v}^{(i)}$, $\boldsymbol{v}^{(o)}$, and $\hat{\boldsymbol{v}}^{(o)}$ are mel-cepstral features at a frame of the source speaker's speech, target speaker's speech, and converted speech, respectively. The MDIR measures how the input speech was improved toward the target speech in the mel-cepstum domain; 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.

B. Results and discussion

In the first experiment, we compared our method with the conventional VC methods, the ARBM [21] and the SATBM [22], in the non-parallel yet speaker-specific paradigm. For the speaker-specific VC in the proposed method, we replace the calculation of the phonetic features in Eq. (19) with $\hat{\boldsymbol{h}} \triangleq \mathbb{E}[\boldsymbol{h}|\boldsymbol{s}^{(i)}, \boldsymbol{v}^{(i)}]$ where $\boldsymbol{s}^{(i)}$ is the one-hot vector that takes value of 1 only at the index of the source speaker, which can be easily calculated from Eq. (9); i.e., $\hat{\boldsymbol{h}} =$

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

$\boldsymbol{f}(\boldsymbol{c}+\mathbf{V}\boldsymbol{s}^{(i)}+\mathbf{M}^{\top}\mathbf{A}^{\top}_{\boldsymbol{s}^{(i)}}\bar{\boldsymbol{v}}^{(i)})$. The results are shown in Table I. As shown in Table I, our method outperformed the other conventional methods by a large margin. We can say that our model performed better because of the explicit modeling of acoustic, phonetic, and speaker features with considering up to three-way connections between the speech factors. Just for a reference, we also compared with a popular GMM-based VC with 64 mixtures using parallel data of 5 sentences, which got 3.86 MDIR. However, such approach takes a benefit from using parallel data and should not be directly compared with non-parallel approaches just in terms of VC quality.

In the second experiment, we investigated how well our arbitrary-source approach based on Eqs. (18)(19) worked. Table II compares the performance of the arbitrary-source VC and the source-specified VC that includes two cases where the correct source speaker is specified and a different speaker is specified as a source speaker. For the different speaker, we used "CAN0001", which was also a male speaker that was included in the training data. As shown in Table II, when we specified a wrong speaker as a source speaker, the performance degraded. Meanwhile, even though we did not specify the correct speaker in the arbitrary-source approach, we obtained similar results to the speaker-specified VC with the correct speaker specified.

V. Conclusion

In this paper, we presented an easy-to-handle VC method that does not require any parallel data during training and the specification of the source speaker during conversion. In our approach, we explicitly model the strength of the connections among fundamental speech factors: acoustic, phonetic, and speaker features, using three-way restricted Boltzmann machine (3WRBM). We also proposed the arbitrary-source VC formulation in the probabilistic framework, which results in two step estimation of the phonetic features given input acoustic features and the acoustic features of the target speaker. In our VC experiments, we obtained better performance with our model than the conventional non-parallel VC approaches in objective criteria. We also showed that our arbitrary-source method well performed, where the results were quite similar to those of the source-specified approach. In the future, we will investigate the arbitrary-source VC performance with variation in speakers, gender, age, etc.

References

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

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

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

[4] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang, "High-performance robust speech recognition using stereo training data," in *ICASSP*, 2001, pp. 301–304.

[5] A. Kunikoshi, Y. Qiao, N. Minematsu, and K. Hirose, "Speech generation from hand gestures based on space mapping," in *INTERSPEECH*, 2009, pp. 308–311.

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

[7] T. Toda, A. W. Black, and K. 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.

[8] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 5, pp. 912–921, 2010.

[9] N. M. Daisuke Saito, Hidenobu Doi and K. Hirose, "Application of matrix variate Gaussian mixture model to statistical voice conversion," in *INTERSPEECH*, 2014, pp. 2504–2508.

[10] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *SLT*, 2012, pp. 313–317.

[11] R. Takashima, R. Aihara, T. Takiguchi, and Y. Ariki, "Noise-robust voice conversion based on spectral mapping on sparse space," in *SSW8*, 2013, pp. 71–75.

[12] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *ICASSP*, 2009, pp. 3893–3896.

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

[14] Z. Wu, E. S. Chng, and H. Li, "Conditional restricted Boltzmann machine for voice conversion," in *ChinaSIP*, 2013.

[15] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets," in *INTERSPEECH*, 2013, pp. 369–372.

[16] T. Nakashika, T. Takiguchi, and Y. 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.

[17] A. Mouchtaris, J. Van der Spiegel, and P. 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.

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

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

[20] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2010.

[21] T. Nakashika, T. Takiguchi, and Y. Ariki, "Parallel-data-free, many-to-many voice conversion using an adaptive restricted Boltzmann machine," in *MLSLP 2015*, 2015, pp. 1–4.

[22] T. Nakashika and Y. Minami, "Speaker adaptive model based on Boltzmann machine for non-parallel training in voice conversion," in *ICASSP 2016 (to appear)*, 2016, pp. 1–5.

[23] A. Krizhevsky, G. E. Hinton *et al.*, "Factored 3-way restricted Boltzmann machines for modeling natural images," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 621–628.

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

[25] K. Cho, A. Ilin, and T. Raiko, "Improved learning of Gaussian-Bernoulli restricted Boltzmann machines," in *ICANN*. Springer, 2011, pp. 10–17.

[26] T. J. Sejnowski, "Higher-order Boltzmann machines," in *AIP Conference Proceedings*, vol. 151, no. 1, 1986, pp. 398–403.

[27] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 930–944.

[28] E. Variani and T. Schaaf, "VTLN in the MFCC domain: Band-limited versus local interpolation," *INTERSPEECH*, pp. 1273–1276, 2011.

[29] M. Morise, "An attempt to develop a singing synthesizer by collaborative creation," in *Proc. the Stockholm Music Acoustics Conference (SMAC)*, 2013, pp. 287–292.