INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Generative Acoustic-Phonemic-Speaker Model Based on Three-Way Restricted Boltzmann Machine

Toru Nakashika, Yasuhiro Minami

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

Abstract

In this paper, we argue the way of modeling speech signals based on three-way restricted Boltzmann machine (3WRBM) for separating phonetic-related information and speaker-related information from an observed signal automatically. The proposed model is an energy-based probabilistic model that includes three-way potentials of three variables: acoustic features, latent phonetic features, and speaker-identity features. We train the model so that it automatically captures the undirected relationships among the three variables. Once the model is trained, it can be applied to many tasks in speech signal processing. For example, given a speech signal, estimating speaker-identity features is equivalent to speaker recognition; on the other hand, estimated latent phonetic features may be helpful for speech recognition because they contain more phonetic-related information than the acoustic features. Since the model is generative, we can also apply it to voice conversion; i.e., we just estimate acoustic features from the phonetic features that were estimated given the source speakers acoustic features along with the desired speaker-identity features. In our experiments, we discuss the effectiveness of the speech modeling through a speaker recognition, a speech (continuous phone) recognition, and a voice conversion tasks.

Index Terms: speech modeling, three-way restricted Boltzmann machine, speaker-adaptive training, voice conversion, speech recognition, speaker recognition.

1. Introduction

One of the most typical and widely-used speech modeling methods is hidden Markov model (HMM). HMM consists of state transition probabilities and state-wise output probability distribution, and Gaussian mixture model (GMM) is usually used as the output distribution. In other words, in the most of the speech signal processing, the observed acoustic features at a certain frame (state) is modeled as GMM. However, when we model the observations using GMM, we do not capture the inner structures (latent features) that exist behind the observations. On the other hand, modeling based on deep learning that stacks several hidden layers has an ability to represent latent features, and outperformed GMM-based modeling. Nevertheless, such approaches have too much free parameters and tend to be overfit. Furthermore, it is inevitable that the gradient-descent-based updates cause local minima and make it difficult to train correctly without proper constraints.

In this paper, we propose a structural speech modeling method[1] that includes three variables of the fundamental speech factors: acoustic features such as mel-cepstral features, latent phonetic features, and speaker features using three-way restricted Boltzmann machine (3WRBM) [1]. The 3WRBM is a energy-based probabilistic model that extends the well-known two-layer RBM [2, 3] so that it represents up to three-order potentials among the three speech 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 among the speech factors, and are optimized so as to maximize the likelihood of the training data. In our approach, 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.

Since our proposed model is generative distribution that takes phonetic-related and speaker-related information separately into account, we can apply the model to various tasks in speech signal processing. For example, we can estimate the speaker who spoke the sentences just by calculating conditional probability distribution of the speaker features given the acoustic features. We can also estimate the phonetic features from the conditional probability distribution of the phonetic features, which may be more effective inputs of HMM for speech recognition than the acoustic features because the phonetic features include less speaker-related information than the acoustic features.

Especially our model shows its effectiveness on voice conversion (VC) task. Most of the existing VC approaches [4, 5, 6, 7, 8, 9] 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) in the training stage[2], 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 3) mismatch in alinement may cause some errors in training. Several approaches [13, 14, 15] do not require any parallel data in the training, and neither does our VC approach. Therefore, our VC approach improves convenience and practicality since the models can be trained using existing speech data without limitations. Our VC scheme can be formulated as MAP estimation, which results in two steps: 1) to estimate phonetic features given the acoustic features and the speaker features that indicate that the speech data is of the

[1] We do not focus on time series modeling but frame-wise acoustic modeling in this paper.

[2] Several approaches, such as eigenvoice and MAP [10, 11, 12], that do not use parallel data between the source and the target speakers has been proposed, although such methods still require parallel data between reference speakers to obtain the speaker-independent space.

http://dx.doi.org/10.21437/Interspeech.2016-1105

source speaker, and 2) to estimate the acoustic features from the obtained phonetic features and the speaker features that indicate that the speech data is of the target speaker.

2. Modeling 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 [16]. 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 modeling clean speech by various speakers; therefore, the latent features $\boldsymbol{h}$ may represent phonetic-related information[3] 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. The model defined in Eq. (1) closely resembles a factored 3WRBM found in [1]. 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.

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

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

When we look at the parameters of three-way potentials $\mathcal{Z}_{:jk}$ which denotes the partial vector of $\mathcal{Z}$ along the first mode, 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 $\bar{\boldsymbol{v}}$ 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 [17, 18], 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 $\mathbf{W} = \mathbf{0}$.

In addition, the bias d_k of the speaker k may represent something such as *frequency* of the speaker k appearing 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 $\mathbf{M} = [\boldsymbol{m}_1\ \cdots\ \boldsymbol{m}_H]$. Like an RBM, because there are no connections between acoustic features, between phonetic features, or between speaker features, each conditional probabilities form simple equations as

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

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

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

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 success probabilities $\boldsymbol{\pi}$, and an element-wise softmax function, respectively. 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.

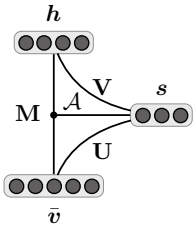

Figure 1: Graphical representation of the proposed speech factor modeling.

2.2. 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

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

In this paper, the parameters are iteratively updated using stochastic gradient descent in the similar way to the training of an RBM. We can derive partial gradients of each parameter in the similar forms of an RBM, although we omit the equations here due to space limitation. The expectations will appear in the derivatives; however; we can still use contrastive divergence (CD) [2] and efficiently approximate them with the expectations of the reconstructed data just like the training of an RBM.

3. Application to speech tasks

3.1. Speech/speaker recognition

After the training of a 3WRBM, we can calculate the following conditional probabilities that the kth speaker and the jth phoneme are activated given the acoustic features for test:

$$p(s_k = 1|\boldsymbol{v}) = f(-g(\frac{\boldsymbol{c}}{R}) + \mathbf{U}_{k:}\bar{\boldsymbol{v}} + g(\frac{\boldsymbol{c}}{R} + \mathbf{V}_{:k} + \mathbf{M}^\top \mathbf{A}_k^\top \bar{\boldsymbol{v}})) \quad (12)$$

$$p(h_j = 1|\boldsymbol{v}) = f(c_j + g(\mathbf{V}_{j:}^\top + \mathbf{U}\bar{\boldsymbol{v}} + [\bar{\boldsymbol{v}}^\top \mathbf{A}_k]\boldsymbol{m}_j)), \quad (13)$$

where $g(\boldsymbol{x}) = \log \sum_k e^{x_k}$ indicates a generalized softplus function, and $\mathbf{U}_{k:}$, $\mathbf{V}_{:k}$, and $\mathbf{V}_{j:}$ are the kth row vector of $\mathbf{U}$, the kth column vector of $\mathbf{V}$, and the jth row vector of $\mathbf{V}$, respectively. For speaker recognition, we can use the expectations of the speaker features $\mathbb{E}[\boldsymbol{s}|\boldsymbol{v}] = [p(s_k = 1|\boldsymbol{v})]$ as an input of a speaker recognizer (tandem approach), or just estimate the speaker as $\hat{k} = \operatorname{argmax}_k p(s_k = 1|\boldsymbol{v})$ (direct approach). For speech (phoneme) recognition, we can use $\mathbb{E}(\boldsymbol{h}|\boldsymbol{v}) = [p(h_j = 1|\boldsymbol{v})]$ as an input vector of a speech recognizer such as HMM. Since the phonetic features $\boldsymbol{h}$ do not indicate the real (supervised) phonemes but latent features obtained in the unsupervised training, we just take the tandem approach for speech recognition.

3.2. Voice conversion

Given the acoustic features $\boldsymbol{v}^{(i)}$ of the source speaker's speech that we want to convert to that of the target speaker $\boldsymbol{v}^{(o)}$ with the identity vector $\boldsymbol{s}^{(i)}$ an $\boldsymbol{s}^{(o)}$ where only ith and oth elements take the value of 1 (otherwise 0), respectively, we estimate $\boldsymbol{v}^{(o)}$ using MAP (maximum a posteriori) as follows:

$$\begin{aligned}\hat{\boldsymbol{v}}^{(o)} &\triangleq \operatorname*{argmax}_{\boldsymbol{v}^{(o)}} p(\boldsymbol{v}^{(o)}|\boldsymbol{v}^{(i)}, \boldsymbol{s}^{(i)}, \boldsymbol{s}^{(o)}) \\ &= \operatorname*{argmax}_{\boldsymbol{v}^{(o)}} \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}^{(i)}, \boldsymbol{s}^{(i)}, \boldsymbol{s}^{(o)}) p(\boldsymbol{v}^{(o)}|\boldsymbol{h}, \boldsymbol{v}^{(i)}, \boldsymbol{s}^{(ij)}, \boldsymbol{s}^{(o)}) \\ &\simeq \operatorname*{argmax}_{\boldsymbol{v}^{(o)}} p(\hat{\boldsymbol{h}}|\boldsymbol{v}^{(i)}, \boldsymbol{s}^{(i)}) p(\boldsymbol{v}^{(o)}|\hat{\boldsymbol{h}}, \boldsymbol{s}^{(o)}) \\ &= \boldsymbol{b} + \mathbf{U}_{o:}^\top + \mathbf{A}_o \mathbf{M} \hat{\boldsymbol{h}}, \end{aligned} \quad (14)$$

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

$$\hat{\boldsymbol{h}} \triangleq \mathbb{E}[\boldsymbol{h}|\boldsymbol{v}^{(i)}, \boldsymbol{s}^{(i)}] = \boldsymbol{f}(\boldsymbol{c} + \mathbf{V}\boldsymbol{s}^{(i)} + \mathbf{M}^\top \mathbf{A}_{\boldsymbol{s}^{(i)}}^\top \bar{\boldsymbol{v}}^{(i)}). \quad (15)$$

In short, the proposed VC scheme has two steps: 1) calculate Eq. (15) to obtain the phonetic features included in the input acoustic vector, and 2) calculate Eq. (14) to obtain desired acoustic features using the phonetic features and the target speaker's parameters.

Table 1: Speaker recognition accuracies of each method.

Method	GMM	UBM	3WRBM	3WRBM (ideal)
Acc. [%]	85.9	83.2	78.1	90.6

Table 2: SVM-based speaker recognition of the proposed method with various features.

Features	$\boldsymbol{v}$ (mcep)	$\boldsymbol{h}$	$\boldsymbol{s}$
# dims.	32	20	8
Acc. [%]	82.0	42.2	**78.7**

4. Experimental evaluation

4.1. System configuration

In order to evaluate our speech modeling method, we conducted speech recognition, speaker recognition, and voice conversion experiments. Through all the experiments, we used ASJ Continuous Speech Corpus for Research (ASJ-JIPDEC[4]). In the training stage, we randomly selected and used speech data of 5 sentences (approx. 160k frames) uttered by $R = 8$ speakers (4 males and 4 females) from the set A in the corpus. For the evaluation, we used the speech data spoken by the same 10 speakers of the different 10 sentences from the training data. As an acoustic feature vector, we used 32-dimensional mel-cepstral features that were calculated from 513-dimensional WORLD [19] spectra without dynamic features. In the training of the system, we used 20 hidden units (phonetic features), a learning rate of 0.01, a momentum of 0.9, and a batch-size of 800, and set the number of iterations as 200.

4.2. Speech/speaker recognition task

Firstly, we examined the effectiveness of our model in speaker recognition. In this experiment, we compared our method in direct approach with the conventional GMM-based method and universal background model (UBM), by calculating the frame-wise recognition accuracy $100 \cdot N_{\mathrm{corr.}}/N_{\mathrm{all}}$ where $N_{\mathrm{corr.}}$ and

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

Table 3: Continuous phone recognition (correct rate [%]) with changing the number of phonemes to be recognized.

Phones	5 vowels	+5 cons.	+10 cons.
$\boldsymbol{v}$ (mcep)	53.53	43.18	36.52
$\boldsymbol{h}$	**59.34**	41.61	33.03

N_{all} indicate the numbers of the corrected frames and the total frames of the test speech data, respectively. In the GMM-based method, we trained GMMs of 64 mixtures for each speaker, and estimated the speakers by calculating the likelihood of each GMM and choosing the most likely GMM. In the UBM approach, we first trained a single GMM (UBM) of 64 mixtures using the whole speech data by all the speakers, copied the parameters of the UBM to those of speaker-dependent GMMs, and then trained each GMM using speaker-wise training data.

The speaker recognition results are shown in Table 1. When we compare our method of direct approach with the conventional methods, the GMM performed best of the three, although the GMM and the UBM approaches are discriminative whereas our approach is generative and they should not be compared directly. It should be also noted that if we first calculated the expected phonetic features given the acoustic features and the correct speaker features in Eq. (15) then estimated the speakers using the conditional probability in Eq. (10) rather than in Eq. (12), we got much better accuracy up to 90.6 with our model (3WRBM (ideal)). Therefore, we can say that our model has a potential of being a comparable speaker recognizer to GMM even though our method is a generative approach.

Secondly, we evaluated the performance of speaker recognition in our model using a support vector machine (SVM) with a linear kernel. In this experiment, we compared the feature type of input to the SVM recognizer as the original acoustic features (mel-cepstral features), the conditional expectation values of phonetic features calculated in Eq. (13), and the conditional expectation values of speaker features calculated in Eq. (12), respectively, as shown in Table 2. Interestingly, the speaker features $\boldsymbol{s}$ produced much better accuracy than the phonemic features $\boldsymbol{h}$, which is close to that of the acoustic features, despite of the compact size of 8 dimensions.

Thirdly, we conducted a speech (continuous phoneme) recognition test. We trained triphone HMMs that have five states with three distributions as a speech recognizer. Each distribution was represented with 32-mixture Gaussians. For an input of the HMM recognizer, we used the phonetic features (E q. (13)), and the traditional acoustic features of mel-cepstra for comparison. We evaluated three cases: the 5 vowels (/a/, /e/, /i/, /o/, and /u/), the 5 vowels and the 5 consonants, the 5 vowels and the 10 consonants. Table 3 shows the speech recognition results. As shown in Table 3, we obtained better performance from the phonetic features than the acoustic features in case of 5 vowels. This is due to the fact that the phonetic features exclude the speaker-related information and consequently include the remaining phonetic-related information, which were more helpful for speech recognition than the acoustic features. In the cases that consonants were considered, the acoustic features outperformed the phonetic features because our model does not include dynamic features. In general, consonant sounds, such as /s/ and /k/, should be represented as dynamics or multiple frame features; nevertheless, in our model, each unit j of the phonetic features is related with the corresponding acoustic pattern $\boldsymbol{m}_j$, which represents static features.

Table 4: Comparison of non-parallel VC methods.

Method	ARBM	SATBM	3WRBM
MDIR [dB]	2.11	2.66	**3.35**

4.3. Voice conversion task

In the VC experiment, 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. Just for the evaluation, we converted the test speech in parallel data (of 10 sentences) of the source and the target speakers, which was created using dynamic programming. As an objective criteria, we used mel-cepstral distortion improvement ratio (MDIR) that is defined as follows:

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

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-cepstal domain; the higher the value of MDIR is, the better the performance of the VC is.

The results are shown in Table 4, which compares the proposed model with the conventional non-parallel VC methods[5], the ARBM [14] and the SATBM [15]. As shown in Table 4, 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.

5. Conclusion

In this paper, we presented a generative speech modeling method based on a three-way restricted Boltzmann machine. In our approach, we explicitly model the strength of the connections among fundamental speech factors: acoustic, phonetic, and speaker features, which enables us to apply the model to many speech signal processing tasks. In order to evaluate our speech modeling method, we conducted speech recognition, speaker recognition, and voice conversion experiments. In the speaker recognition, we showed the potential of our generative approach which was comparable to discriminative approach based on GMM. In the speech recognition, the latent phonetic features obtained from our model outperformed the traditional mel-cepstral features in case of vowels were evaluated. In the voice conversion experiment, we obtained better performance with our model than the conventional non-parallel VC approaches.

In this paper, we focused on the formulation of our model and the basic evaluation. In the future, we will further investigate our method more deeply. For example, we want to evaluate our model when changing the number of training sentences and the number of the reference speakers.

[5]For the fair comparison, we used the same configuration (the number of hidden units, the type of adaptation matrices, the number of reference speakers, etc.) to the conventional approaches.

6. References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] T. Nakashika and Y. Minami, "Speaker adaptive model based on Boltzmann machine for non-parallel training in voice conversion," in *ICASSP 2016*, 2016, pp. 5530–5534.

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

[17] 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.

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

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