INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

CAB: An Energy-Based Speaker Clustering Model for Rapid Adaptation in Non-Parallel Voice Conversion

Toru Nakashika[1]

[1]University of Electro-Communications, Tokyo, Japan
nakashika@uec.ac.jp

Abstract

In this paper, a new energy-based probabilistic model, called CAB (Cluster Adaptive restricted Boltzmann machine), is proposed for voice conversion (VC) that does not require parallel data during the training and requires only a small amount of speech data during the adaptation. Most of the existing VC methods require parallel data for training. Recently, VC methods that do not require parallel data (called non-parallel VCs) have been also proposed and are attracting much attention because they do not require prepared or recorded parallel speech data, unlike conventional approaches. The proposed CAB model is aimed at statistical non-parallel VC based on cluster adaptive training (CAT). This extends the VC method used in our previous model, ARBM (adaptive restricted Boltzmann machine). The ARBM approach assumes that any speech signals can be decomposed into speaker-invariant phonetic information and speaker-identity information using the ARBM adaptation matrices of each speaker. VC is achieved by switching the source speaker's identity into those of the target speaker while retaining the phonetic information obtained by decomposition of the source speaker's speech. In contrast, CAB speaker identities are represented as cluster vectors that determine the adaptation matrices. As the number of clusters is generally smaller than the number of speakers, the number of model parameters can be reduced compared to ARBM, which enables rapid adaptation of a new speaker. Our experimental results show that the proposed method especially performed better than the ARBM approach, particularly in adaptation.

Index Terms: voice conversion, non-parallel training, cluster adaptive training (CAT), rapid adaptation, energy-based model

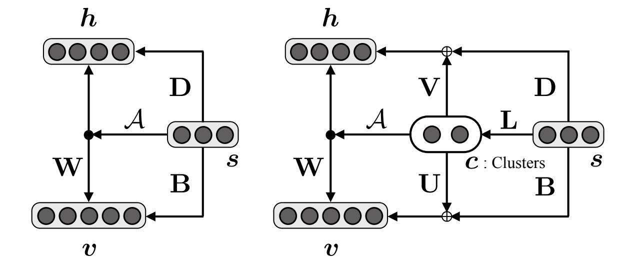
1. Introduction

Voice conversion (VC) is a technique where only speaker specific information from the source speaker's speech is converted while the phonological information is retained unchanged. In recent years, VC has been attracting attention because it can be applied to various speech tasks, not only for entertainment but also for welfare use. Most conventional VC models such as the Gaussian mixture model (GMM) [1, 2, 3], non-negative matrix factorization (NMF) [4], and deep learning [5, 6], must use parallel data (pairs of speech data generated by source and the target speakers uttering the same sentences) for training. However, several approaches do not rely on parallel data (termed non-parallel VC) [7, 8, 9, 10]. Although VC quality with parallel data often outperforms that without, these non-parallel VCs have been attracting much attention because not using parallel data is more convenient and practical.

Recently, Erro *et al.* proposed the INCA algorithm [7] for non-parallel VC. The algorithm begins by creating pseudo-parallel data by matching the nearest-neighbor frames of the source and target speakers from non-parallel corpora. This is followed by a conversion, using conventional VC models such as GMM, using the pseudo-parallel data. They showed that by gradually repeating these steps, the converted speech came close to the target voice. However, the INCA has several problems: e.g., high computational cost associated with the iteration of training VC models and the NN search from all the training data; and unexpected mismatch of phonemes caused by the NN search without considering phonetic information. Therefore, in previous works we have proposed statistical non-parallel VC methods using different probabilistic models: ARBM (adaptive restricted Boltzmann machine) [8], SATBM (speaker-adaptive-trainable Boltzmann machine) [9], and 3WRBM (three-way restricted Boltzmann machine) [10]. These approaches include speaker-dependent (SD) and speaker-independent (SI) parameters to facilitate decomposition of the speech into latent phonetic features and speaker identity features, and conversely synthesize speech from them. The parameters can be simultaneously estimated using speech data uttered by multiple speakers during training; estimation of only the SD parameters of a new speaker is also possible after the training (referred to as adaptation). Once the parameters are estimated, voice-converted speech is obtained from the target speaker's identity features and the latent phonetic features that are calculated from the source speaker's speech.

These models contain adaptation matrices for each speaker as the SD parameters. When the number of parameters is proportional to the square of the dimensions of the acoustic features such as mel-cepstra: 1) it increases computational costs as the number of dimensions of the acoustic features and the speakers increases, and 2) it requires sufficient data from the new speaker used in the adaptation, which makes it difficult to conduct voice conversion in a timely manner. In real situations where VC is used, it is useful and practicable if the voice is converted immediately just after starting to record the speech. In this paper, we propose a statistical non-parallel VC method, based on cluster adaptive training (CAT) [11], in which the model can be adapted for a new speaker with a small amount of data. CAT is a technique where speaker-specific characteristics are represented as the weighted sum of clusters that are constructed to categorize the characteristics of the speakers. The proposed model in this paper, the cluster adaptive restricted Boltzmann machine (CAB), extends our previous model ARBM, to allow the automatic creation of speaker clusters during training. In CAB, three types of parameters are estimated at the same time: the weights indicating which speaker-cluster each speaker is assigned to; the bidirectional weights between the observable acoustic features and the latent speaker-invariant (possibly phonetic) features; and the adaptation matrices of each cluster. Noting that the number of clusters K is, in general, set to be smaller than the number of speakers R, we can greatly reduce the number of model parameters associated with the adaptation matrices, which were prepared for each speaker in ARBM. In addi-

http://dx.doi.org/10.21437/Interspeech.2017-133



(a) Conventional model (ARBM)

(b) Proposed model (CAB)

Figure 1: *Comparison of structures, conventional model (a) and proposed model (b).*

tion, in the adaptation, only a few cluster weight parameters are estimated for a new person, which enables rapid adaptation of the model using a small amount of data.

It is also worth noting that by using the proposed model we can produce the speech of any speakers we are interested in just by adjusting the cluster weights, even if the speaker was not included in the training or adaptation. The automatically created clusters represent *representatives* that gather different types of speaker from each cluster; e.g., one cluster represents masculine speech, another represents bright speech, and another represents muffled speech.

2. Energy-based non-parallel VC models

Before going into the details of the proposed CAB model, in the following section we revisit our previous model, ARBM [8]. Both are defined as energy-based models for non-parallel VC.

2.1. Conventional model: ARBM

ARBM is an energy-based probabilistic model that extends RBM (restricted Boltzmann machine) [12, 13] so that the model can be adapted to a specific speaker. As Figure 1 (a) illustrates, ARBM defines the conditional probability of observable frame-wise acoustic features such as mel-cepstra $\boldsymbol{v} = [v_1, \cdots, v_I] \in \mathbb{R}^I$ (I is the number of dimensions of the acoustic features) and hidden features $\boldsymbol{h} = [h_1, \cdots, h_J] \in \{0,1\}^J, \sum_j h_j = 1$ (J is the number of hidden features) given another observable speaker identity features $\boldsymbol{s} = [s_1, \cdots, s_R] \in \{0,1\}^R, \sum_r s_r = 1$ where only the rth element in $\boldsymbol{s}$ is on when the speech data is of speaker r, assuming that all speech signals contain R speakers in total. In ARBM, there are bidirectional speaker-independent (SI) connections $\mathbf{W} \in \mathbb{R}^{I\times J}$ between $\boldsymbol{v}$ and $\boldsymbol{h}$, directional connections $\mathbf{B} \in \mathbb{R}^{I\times R}$ from $\boldsymbol{s}$ to $\boldsymbol{v}$, and directional connections $\mathbf{D} \in \mathbb{R}^{J\times R}$ from $\boldsymbol{s}$ to $\boldsymbol{h}$ without having connections among units in $\boldsymbol{v}$ or in $\boldsymbol{h}$, as in RBM. Interestingly, ARBM also has directional connections $\mathcal{A} = \{\mathbf{A}_r\}_{r=1}^R, \mathbf{A}_r \in \mathbb{R}^{I\times I}$ from $\boldsymbol{s}$ toward $\mathbf{W}$, each of which projects the canonical spectral templates $\mathbf{W}$ to adapt the model to the speaker, referred to as an adaptation matrix[1]. The probability distribution of ARBM is defined as follows:

$$p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s}) = \frac{1}{Z} e^{-E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{s})} \quad (1)$$

$$E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s}) = \frac{1}{2}\left\|\frac{\boldsymbol{v}-\tilde{\boldsymbol{b}}}{\boldsymbol{\sigma}}\right\|_2^2 - \tilde{\boldsymbol{d}}^\top \boldsymbol{h} - (\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2})^\top \tilde{\mathbf{W}}\boldsymbol{h} \quad (2)$$

$$Z = \int \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{s})} d\boldsymbol{v} \quad (3)$$

where $\dot{\overline{\cdot}}, \cdot^2$ denotes element-wise division and element-wise square, respectively, and $\boldsymbol{\sigma} \in \mathbb{R}^I$ indicates the standard deviation parameters of the acoustic features. $\tilde{\mathbf{W}} \in \mathbb{R}^{I\times J}$, $\tilde{\boldsymbol{b}} \in \mathbb{R}^I$, and $\tilde{\boldsymbol{d}} \in \mathbb{R}^J$ in Eq. (2) are not model parameters to be estimated, but adapted parameters of the connection weights between $\boldsymbol{v}$ and $\boldsymbol{h}$, the biases of $\boldsymbol{v}$, and the biases of $\boldsymbol{h}$ to the given speaker, respectively (terms with a tilde indicate *adapted* parameters). The adapted parameters are defined as:

$$\tilde{\mathbf{W}} \triangleq \sum_r \mathbf{A}_r s_r \mathbf{W} = (\mathcal{A} \circ_3^1 \boldsymbol{s})\mathbf{W} \quad (4)$$

$$\tilde{\boldsymbol{b}} \triangleq \boldsymbol{b} + \sum_r \boldsymbol{b}_r s_r = \boldsymbol{b} + \mathbf{B}\boldsymbol{s} \quad (5)$$

$$\tilde{\boldsymbol{d}} \triangleq \boldsymbol{d} + \sum_r \boldsymbol{d}_r s_r = \boldsymbol{d} + \mathbf{D}\boldsymbol{s} \quad (6)$$

where $\circ_i^j$ denotes the inner product of the left tensor along the ith mode and the right tensor along the jth mode. $\boldsymbol{b} \in \mathbb{R}^I$ and $\boldsymbol{d} \in \mathbb{R}^J$ are the SD parameters of biases of $\boldsymbol{v}$ and $\boldsymbol{h}$, respectively. $\boldsymbol{b}_r$ and $\boldsymbol{d}_r$ are the rth column vectors of $\mathbf{B}$ and $\mathbf{D}$, respectively. The SI parameters $\{\mathbf{W}, \boldsymbol{b}, \boldsymbol{d}, \boldsymbol{\sigma}\}$ and the SD parameters $\{\mathcal{A}, \mathbf{B}, \mathbf{D}\}$ are simultaneously estimated to maximize the log-likelihood of N training data of multi-speaker corpus $\{\boldsymbol{v}_n|\boldsymbol{s}_n\}_{n=1}^N$. Once the training is complete, the trained ARBM can be adapted for a new speaker by estimating only their SD parameters while fixing their SI parameters.

In this approach, we use mel-cepstra of *clean* speech without noise, reverberation, or emotion for the acoustic features $\boldsymbol{v}$, and assume that there are no variations in speech except for the changes in phonemes (allophones) and speakers. Therefore, because the variations invoked by the differences in speakers are caught by the the SD parameters associated with the speaker identity features $\boldsymbol{s}$, the latent features $\boldsymbol{h}$, which is defined as a vector where only an element is on at a certain frame and un-observable, possibly represents speaker-invariant, phonetic features[2].

2.2. Proposed model: CAB

The previously mentioned ARBM includes a large number of SD parameters, proportional to I^2R, which leads to expensive computational costs as the number of speakers increases. Also, in the adaptation stage, an estimation of the $I^2 + I + J$ parameters of a new speaker needs to be made, which may cause overfitting unless sufficient adaptation data is available. Therefore, we adopt the cluster adaptive training (CAT) [11] technique in the ARBM-based non-parallel VC scheme to reduce the number of SD parameters. CAT represents the characteristics of each speaker as the weighted sum of multiple clusters. As Figure 1 (b) shows, we introduce a cluster vector $\boldsymbol{c} \in \mathbb{R}^K$ (K is

[1] The matrices $\mathbf{B}$ and $\mathbf{D}$ are also used to adapt the model, but we do not refer to these matrices as adaptation matrices.

[2] Therefore, we call $\boldsymbol{h}$ latent phonetic features or just phonetic features in this paper.

the number of clusters) that takes the weighted sum of speaker features as:

$$\boldsymbol{c} \triangleq \mathbf{L}\boldsymbol{s} \tag{7}$$

where $\lambda_{k,r}$ in $\mathbf{L} \in \mathbb{R}^{K \times R} = [\boldsymbol{\lambda}_1 \cdots \boldsymbol{\lambda}_R]$ indicates the weight parameter of the rth speaker to the kth cluster, and satisfies $\|\boldsymbol{\lambda}_r\|_1 = 1, \lambda_{k,r} \geq 0, \forall k, r$. Note that the cluster vector is not defined as a variable, and is always determined by the speaker identity features $\boldsymbol{s}$ as in Eq. (7). While ARBM contains adaptation matrices for each speaker, CAB has adaptation matrices for each cluster (that is, $\mathcal{A} = \{\mathbf{A}_k\}_{k=1}^{K}, \mathbf{A}_k \in \mathbb{R}^{I \times I}$). In CAB, the adapted parameters $\tilde{\mathbf{W}}, \tilde{\boldsymbol{b}}, \tilde{\boldsymbol{d}}$ are also modified as:

$$\tilde{\mathbf{W}} \triangleq (\mathcal{A} \circ_3^1 \boldsymbol{c})\mathbf{W} \tag{8}$$

$$\tilde{\boldsymbol{b}} \triangleq \boldsymbol{b} + \mathbf{U}\boldsymbol{c} + \mathbf{B}\boldsymbol{s} \tag{9}$$

$$\tilde{\boldsymbol{d}} \triangleq \boldsymbol{d} + \mathbf{V}\boldsymbol{c} + \mathbf{D}\boldsymbol{s} \tag{10}$$

where $\mathbf{U} \in \mathbb{R}^{I \times K}$ and $\mathbf{V} \in \mathbb{R}^{J \times K}$ are the cluster-dependent (CD) parameters of biases regarding $\boldsymbol{v}$ and $\boldsymbol{h}$, respectively. The parameters included in CAB adaptation matrices reduce the size of I^2K from those in ARBM of I^2R (c.f., in our experiments, we used $R = 58, I = 32, K = 8$ as maximum, where CAB had $8,192$ parameters and ARBM had $59,392$ parameters in $\mathcal{A}$). In adaptation, when we have $H = 16$ phonetic features, the number of parameters for a new speaker is $I^2 + I + J = 1,072$ in ARBM, whereas it is only $K + I + J = 56$ in CAB. Therefore, CAB is capable of rapid adaptation.

The probability distribution of the proposed model $p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s})$ is defined as in Eqs. (1), (2), and (3). In this case, the conditional probabilities $p(\boldsymbol{v}|\boldsymbol{h}, \boldsymbol{s}), p(\boldsymbol{h}|\boldsymbol{v}, \boldsymbol{s})$ form simple distributions as:

$$p(\boldsymbol{v}|\boldsymbol{h}, \boldsymbol{s}) = \mathcal{N}(\boldsymbol{v}|\tilde{\boldsymbol{b}} + \tilde{\mathbf{W}}\boldsymbol{h}, \boldsymbol{\sigma}^2) \tag{11}$$

$$p(\boldsymbol{h}|\boldsymbol{v}, \boldsymbol{s}) = \mathcal{B}(\boldsymbol{h}|\boldsymbol{f}(\tilde{\boldsymbol{d}} + \tilde{\mathbf{W}}^\top \frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2})) \tag{12}$$

where $\mathcal{N}(\cdot)$, $\mathcal{B}(\cdot)$, and $\boldsymbol{f}(\cdot)$ indicate a multivariate normal distribution, a Bernoulli distribution, and an element-wise softmax function, respectively. Now let us consider the mean acoustic features $\boldsymbol{\mu}_r$ of a speaker r when $\boldsymbol{h}$ is known. From Eqs. (8), (9), and (11), the mean vector is calculated as:

$$\begin{aligned} \boldsymbol{\mu}_r &= \boldsymbol{b} + \boldsymbol{b}_r + \mathbf{U}\boldsymbol{\lambda}_r + (\mathcal{A} \circ_3^1 \boldsymbol{\lambda}_r)\mathbf{W}\boldsymbol{h} && (13) \\ &= \mathbf{M}\boldsymbol{\lambda}'_r + \boldsymbol{b}_r && (14) \end{aligned}$$

where we introduce the extended vector $\boldsymbol{\lambda}'_r = [\boldsymbol{\lambda}_r^\top\ 1]^\top$, and the CD acoustic templates $\mathbf{M} = [\boldsymbol{m}_1 \cdots \boldsymbol{m}_{K+1}]$, the column vectors of which are defined as:

$$\boldsymbol{m}_k \triangleq \begin{cases} \boldsymbol{u}_k + \mathbf{A}_k\mathbf{W}\boldsymbol{h} & (k = 1, \cdots, K) \\ \boldsymbol{b} & (k = K+1) \end{cases}. \tag{15}$$

As shown in Eq. (14), the vector $\boldsymbol{\mu}_r$ is represented in the CAT [11] scheme except for an additional term $\boldsymbol{b}_r$ and the phonetic-structured CD templates in Eq. (15).

2.3. Parameter estimation

The CAB parameters $\boldsymbol{\Theta}_{CAB} = \{\boldsymbol{\Theta}_{SI}, \boldsymbol{\Theta}_{SD}, \boldsymbol{\Theta}_{CD}\}$, where $\boldsymbol{\Theta}_{SI} = \{\mathbf{W}, \boldsymbol{b}, \boldsymbol{d}, \boldsymbol{\sigma}\}$ represents the SI parameters, $\boldsymbol{\Theta}_{SD} = \{\mathbf{B}, \mathbf{D}\}$ the SD parameters, and $\boldsymbol{\Theta}_{CD} = \{\mathcal{A}, \mathbf{U}, \mathbf{V}\}$ the CD parameters, can be estimated at the same time so as to maximize the log-likelihood of N frame training data $\{\boldsymbol{v}_n|\boldsymbol{s}_n\}_{n=1}^{N}$:

$$\mathcal{L}(\boldsymbol{\Theta}_{CAB}) = \log \prod_n p(\boldsymbol{v}_n|\boldsymbol{s}_n) = \sum_n \log \sum_{\boldsymbol{h}} p(\boldsymbol{v}_n, \boldsymbol{h}_n|\boldsymbol{s}_n)$$

using stochastic gradient ascent. We omit the gradients of each parameter due to spacelimitations. Each gradient will contain the expectations of the model that are difficult to compute; however, we employ contrastive divergence [12] to approximate the expectations as in normal RBM. To satisfy the non-negative constraints on the cluster weights, we replace as $\boldsymbol{\lambda}_r = e^{\boldsymbol{z}_r}$ and estimate the parameters with $\boldsymbol{z}_r$. We then give regularization on the cluster weights to satisfy $\|\boldsymbol{\lambda}_r\|_1 = 1$ for each update. It should be also noted that we can *simultaneously* optimize all the parameters in CAB, while in conventional models such as HMM or GMM using the CAT algorithm [11], this is difficult due to limitations in optimization methods.

After the training, the model completes the creatation of latent phonemes and speaker clusters, and the parameters $\Theta_{r'} = \{\boldsymbol{\lambda}_{r'}, \boldsymbol{b}_{r'}, \boldsymbol{d}_{r'}\}$ for a new speaker r' can be estimated while the other parameters are retained unchanged. This step is referred to as adaptation.

3. Application to non-parallel VC

In this section, VC using the proposed model, CAB, is presented. Assuming that training or adaptation has finished, the voice-converted acoustic features $\hat{\boldsymbol{v}}^{(o)}$, which are supposed to be of the target speaker o, are formulated as the most likely acoustic features derived from the acoustic features $\boldsymbol{v}^{(i)}$ of the source speaker i frame-by-frame; that is:

$$\hat{\boldsymbol{v}}^{(o)} \triangleq \underset{\boldsymbol{v}}{\operatorname{argmax}}\, p(\boldsymbol{v}|, \boldsymbol{v}^{(i)}, \boldsymbol{s}^{(i)}, \boldsymbol{s}^{(o)}) \tag{16}$$

where $\boldsymbol{s}^{(i)}$ is the speaker identity features of the source speaker where only the ith element is on, and similarly $\boldsymbol{s}^{(o)}$ is the speaker identity features of the target speaker where only the oth element is on. Eq. (16) can be further rewritten as:

$$\begin{aligned} \hat{\boldsymbol{v}}^{(o)} &= \underset{\boldsymbol{v}}{\operatorname{argmax}} \sum_{\boldsymbol{h}} p(\boldsymbol{h}|\boldsymbol{v}^{(i)}, \boldsymbol{s}^{(i)}, \boldsymbol{s}^{(o)}) p(\boldsymbol{v}|\boldsymbol{h}, \boldsymbol{v}^{(i)}, \boldsymbol{s}^{(i)}, \boldsymbol{s}^{(o)}) \\ &\simeq \underset{\boldsymbol{v}}{\operatorname{argmax}}\, p(\hat{\boldsymbol{h}}|\boldsymbol{v}^{(i)}, \boldsymbol{s}^{(i)}) p(\boldsymbol{v}|\hat{\boldsymbol{h}}, \boldsymbol{s}^{(o)}) && (17) \\ &= \boldsymbol{b} + \mathbf{B}\boldsymbol{s}^{(o)} + \mathbf{U}\mathbf{L}\boldsymbol{s}^{(o)} + (\mathcal{A} \circ_3^1 \mathbf{L}\boldsymbol{s}^{(o)})\mathbf{W}\hat{\boldsymbol{h}} && (18) \end{aligned}$$

where we introduce $\hat{\boldsymbol{h}}$ as the expected phonetic features obtained from the source acoustic features $\boldsymbol{v}^{(i)}$ and the speaker identity features $\boldsymbol{s}^{(i)}$ as:

$$\begin{aligned} \hat{\boldsymbol{h}} &\triangleq \mathbb{E}[\boldsymbol{h}|\boldsymbol{v}^{(i)}, \boldsymbol{s}^{(i)}] \\ &= \boldsymbol{f}(\boldsymbol{d} + \mathbf{V}\mathbf{L}\boldsymbol{s}^{(i)} + \mathbf{D}\boldsymbol{s}^{(i)} + \mathbf{W}^\top(\mathcal{A} \circ_3^1 \mathbf{L}\boldsymbol{s}^{(i)})^\top \frac{\boldsymbol{v}^{(i)}}{\boldsymbol{\sigma}^2}). \end{aligned}$$

4. Experiments

4.1. System configuration

We evaluated the performance of the proposed CAB through VC experiments using the ASJ Continuous Speech Corpus for Research (ASJ-JIPDEC[3]). For the training, we randomly selected and used speech data from 40 sentences uttered by $R =$

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

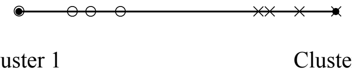


Figure 2: *Estimated cluster weight distribution when $K = 2, R = 8$. The two clusters (●) are at the opposite sides of a straight line. ○ and × indicate the estimated cluster weights of male and female speakers, respectively.*

8, 16, and 58 speakers from the corpus. In the first experiment, the source speaker (identified as "ECL0001" in the corpus) and the target speaker ("ECL1003") used in the evaluation were also included in the training set (no adaptation case). In the evaluation, 10 sentences, different from the training sentences uttered by the two speakers, were used. In the second experiment, a source speaker ("ECL1004") and a target speaker ("ECL0002") were used in the evaluation but not included in the training set, and were adapted using their speech by increasing the number of sentences up to 40 (adaptation case). In addition, 10 sentences, different from those used in training and adaptation, were also used in this evaluation. In both experiments, 32-dimensional mel-cepstra were used as acoustic features calculated from 513-dimensional WORLD [14] spectra without dynamic features. We evaluated the model by changing the number of phonetic features ($J = 8$, 16, and 24) and the number of clusters ($K = 2$, 3, 4, 6, and 8) and chose the most performed configurations (the numbers of phonetic features and of clusters) for each test. For training the system, we used a learning rate of 0.01, a momentum of 0.9, and a batch-size of $100 \times R$ for stochastic gradient ascent with 100 iterations. We also compared CAB with the conventional model, ARBM, using the same configurations.

As an objective criterion, we evaluated the methods using the average of the mel-cepstral distortion improvement ratio (MDIR) [8], defined as follows:

$$MDIR[dB] = \frac{10\sqrt{2}}{\ln 10}(\left\|\boldsymbol{v}^{(o)} - \boldsymbol{v}^{(i)}\right\|_2^2 - \left\|\boldsymbol{v}^{(o)} - \hat{\boldsymbol{v}}^{(o)}\right\|_2^2).$$

The MDIR measures how input speech is improved toward the target speech in the mel-cepstral domain; the higher the MDIR value, the better the VC performance.

Table 1: *VC performance (MDIR [dB]) when changing the number of training speakers, where test speakers were included in the training.*

# speakers	**8**	**16**	**58**
ARBM	3.70	2.64	3.02
CAB	3.21	3.06	3.23

Table 2: *VC performance (MDIR [dB]) when changing the number of adaptation sentences, where test speakers were not included in the training.*

# sentences	**0.2**	**0.5**	**1**	**10**	**40**
ARBM	2.48	3.25	3.21	3.41	3.45
CAB	3.14	3.54	3.63	3.60	3.58

4.2. Visualization of estimated cluster weights

The estimated cluster weights $\boldsymbol{\lambda}_r$ of each speaker when $K = 2, R = 8$ and $K = 3, R = 16$ are plotted in Figures 2 and 3, respectively. In Figure 2, the first cluster (Cluster 1) may serve as the representative of "male" that groups the male speakers and the second cluster (Cluster 2) may be the representative of "female." It is interesting that these clusters were automatically created even though gender information was not provided in the training. In Figure 3, the other gender-mixed cluster (Cluster 3) was created aside from the "male" cluster (Cluster 1) and the "female" cluster (Cluster 2). In both Figures 2 and 3, we can see that each cluster was created to get as far away from the other clusters as possible, and occurs at the same location as the real speakers. This is preferable when we want to generate speech with arbitrary voices by adjusting the cluster weights as it has a large margin within which to adjust the weights.

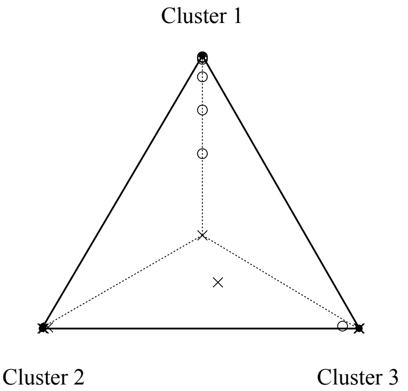


Figure 3: *Estimated cluster weight distribution when $K = 3, R = 16$. The three clusters (●) are illustrated at the vertices of a regular triangle. ○ and × indicate the estimated cluster weights of male and female speakers, respectively.*

4.3. Comparison with conventional model

First, we compared the models when the speakers were included in the training, as shown in Table 1. This shows that although ARBM performs well when the number of speakers is small ($R = 8$), the performance declines and unstable as the number of speakers increases. Meanwhile, the performance of CAB is stable and can even be improved with a large number of speakers. Second, we compared the adaptation performance, see Table 2. We see that CAB performed better overall in adaptation than ARBM, and that its performance was sufficient even when a vary short speech was given (half a sentence, around 2 sec). This means that CAB can be adapted more rapidly than the ARBM. This is considered to be a result of the reduction in the number of SD parameters in CAB by the introduction of speaker clusters.

5. Conclusions

In this paper, as an extension of ARBM, we proposed a non-parallel VC method that can be rapidly adapted to a new speaker. In our experiments, the proposed model showed its effectiveness, particularly when a small amount of adaptation data was given, compared with the conventional model. In the future, we will further investigate the performance of this proposed model in depth, e.g., performance with other test speaker pairs, performance with more phonetic features, and performance with a large number (> 100) of training speakers.

6. References

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

[2] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[3] D. Saito, H. Doi, N. Minematsu, and K. Hirose, "Application of matrix variate Gaussian mixture model to statistical voice conversion," in *Interspeech*, 2014, pp. 2504–2508.

[4] R. Aihara, T. Takiguchi, and Y. Ariki, "Individuality-preserving voice conversion for articulation disorders using phoneme-categorized exemplars," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 6, no. 4, p. 13, 2015.

[5] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion using RNN pre-trained by recurrent temporal restricted Boltzmann machines," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 580–587, 2015.

[6] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2009, pp. 3893–3896.

[7] 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.

[8] T. Nakashika, T. Takiguchi, and Y. Minami, "Non-parallel training in voice conversion using an adaptive restricted Boltzmann machine," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2032–2045, 2016.

[9] T. Nakashika and Y. Minami, "Speaker adaptive model based on Boltzmann machine for non-parallel training in voice conversion," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5530–5534.

[10] T. Nakashika, , and Y. Minami, "Generative acoustic-phonemic-speaker model based on three-way restricted Boltzmann machine," in *Interspeech 2016*, 2016, pp. 1487–1491.

[11] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, Jul. 2000.

[12] 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.

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

[14] M. Morise, "An attempt to develop a singing synthesizer by collaborative creation," in *SMAC2013*, 2013, pp. 287–292.