IEICE TRANS. INF. & SYST., VOL.E103–D, NO.11 NOVEMBER 2020

PAPER

Speech Chain VC: Linking Linguistic and Acoustic Levels via Latent Distinctive Features for RBM-Based Voice Conversion

Takuya KISHIDA[†a)], ***Nonmember and*** **Toru NAKASHIKA**[†], ***Member***

SUMMARY This paper proposes a voice conversion (VC) method based on a model that links linguistic and acoustic representations via latent phonological distinctive features. Our method, called speech chain VC, is inspired by the concept of the speech chain, where speech communication consists of a chain of events linking the speaker's brain with the listener's brain. We assume that speaker identity information, which appears in the acoustic level, is embedded in two steps—where phonological information is encoded into articulatory movements (linguistic to physiological) and where articulatory movements generate sound waves (physiological to acoustic). Speech chain VC represents these event links by using an adaptive restricted Boltzmann machine (ARBM) introducing phoneme labels and acoustic features as two classes of visible units and latent phonological distinctive features associated with articulatory movements as hidden units. Subjective evaluation experiments showed that intelligibility of the converted speech significantly improved compared with the conventional ARBM-based method. The speaker-identity conversion quality of the proposed method was comparable to that of a Gaussian mixture model (GMM)-based method. Analyses on the representations of the hidden layer of the speech chain VC model supported that some of the hidden units actually correspond to phonological distinctive features. Final part of this paper proposes approaches to achieve one-shot VC by using the speech chain VC model. Subjective evaluation experiments showed that when a target speaker is the same gender as a source speaker, the proposed methods can achieve one-shot VC based on each single source and target speaker's utterance.

key words: *voice conversion, restricted Boltzmann machine, speech chain, one-shot voice conversion*

1. Introduction

A deeper understanding of how humans identify individual speech would provide useful insight when designing a voice conversion (VC) system capable of speaker identity conversion—a system enables listeners to perceive converted speech as if it was uttered by a target speaker. Developing an effective VC system will also offer useful insights into psychological and physiological mechanisms of human speech communication.

One state-of-the-art VC framework is based on generative adversarial nets (GANs) [1]–[5], which were originally developed for image generation [6] and were also devised for image-to-image translation [7], [8]. CycleGAN-VC [2], [3] and StarGAN-VC [4], [5] are successful VC models incorporating GAN variants. These VC models perform high quality voice conversion; speaker identity conversion quality is comparable to that of a Gaussian mixture model (GMM)-based VC [9], which is a well-known approach for training with parallel data of target and source speaker recordings. Many other successful VC frameworks are also based on neural networks (NNs) having complex network structures, such as variational autoencoders [10], [11] and recurrent NNs [12], [13].

However, the low interpretability of these models hinders the acquisition of any relevant scientific insights. In this paper, model interpretability indicates the degree to which a human can understand the cause of an obtained decision or intermediate representation [14]. If the model is designed upon the interpretable way, we can associate model architectures with actual phenomena the model attempts to represent. Recent work provides some techniques for interpreting complex machine learning models [15], [16] but interpreting deep networks is still a challenging field. The above-mentioned models consist of complex structures, which improve the expressiveness of these models at the cost of interpretability.

An adaptive restricted Boltzmann machine (ARBM)-based VC model was proposed as a relatively simple and interpretable model [17]. An ARBM-based model consists of a visible layer and a hidden layer having undirected connections between visible-hidden units. The weights of the connections are designed to be adaptable to speakers by introducing an adaptation matrix for each speaker. Linear transformation of acoustic features (e.g., Mel-cepstrum) by an adaptation matrix is a speaker normalization technique used in automatic speech recognition systems [18]. The speaker-adapted connection weights can be interpreted as spectral templates characterizing voice of the speaker. An ARBM-based VC approach assumes that speaker-related information is mainly represented in the adaptation matrix and linguistic information is represented in the hidden layer. Speaker-independent and speaker-dependent parameters in the model are simultaneously trained with non-parallel data.

The ARBM-based VC approach still requires a number of improvements to obtain higher performances in both similarity and intelligibility of converted speech. The similarity to a target speaker of converted speech by the ARBM-based VC is slightly inferior to that of a GMM-based VC. The intelligibility of converted speech is also less intelligible than natural speech. These shortcomings seem to be caused by failing to preserve linguistic information when converting a source speaker's acoustic features to those of a target

Manuscript received February 20, 2020.
Manuscript revised June 19, 2020.
Manuscript publicized August 6, 2020.

[†]The authors are with the Graduate School of Informatics and Engineering, The University of Electro-Communications, Choufu-shi, 182–8585 Japan.
a) E-mail: kishida@uec.ac.jp

DOI: 10.1587/transinf.2020EDP7032

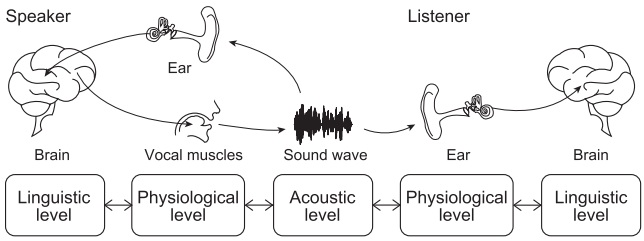

Fig. 1 The speech chain in speech communication.

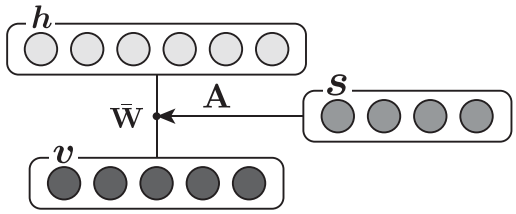

Fig. 2 Graphical representation of an ARBM.

speaker.

In this paper, we propose speech chain VC, which is an extended method of the ARBM-based VC. Speech chain VC is inspired by the concept of the speech chain [19], in which speech communication consists of a chain of events linking the speaker's brain with the listener's brain as shown in Fig. 1.

For the basis of this framework, we assume that speaker identity information is first embedded when motor control signals are sent to articulatory organs from a speaker's brain (linguistic level to physiological level) and then when the articulatory movements generate sound waves (physiological level to acoustic level). The conventional ARBM model links the acoustic and linguistic levels directly and represents the linguistic features with hidden layers. Linguistic features can be regarded as visible features by using descriptions such as phoneme, thus, speech chain VC represents linguistic features and acoustic features with visible layers and latent distinctive features associated with articulatory movements with a hidden layer.

We further propose methods to apply the speech chain VC to one-shot VC tasks—performing VC across arbitrary speakers based on only one each utterance of the speakers. A one-shot voice conversion task is a very challenging task, but it is highly convenient for users of voice conversion applications, and its technological development is required.

2. ARBM-Based Voice Conversion

In this section, we will introduce an ARBM-based VC method as a baseline method. A graphical representation of an ARBM is shown in Fig. 2. In an ARBM model, observed acoustic features and latent phonological features are represented as visible units $\boldsymbol{v} \in \mathbb{R}^I$ and hidden units $\boldsymbol{h} \in \{0,1\}^J$, respectively ($I$ and J denote the number of dimensions in the visible and hidden units, respectively). In addition to visible and hidden units, this model has speaker identity units $\boldsymbol{s} \in \{0,1\}^R, \sum_{r=1}^R s_r = 1$ that represent which speaker utters the sentence (R is the number of speakers used in the training). Usually $\boldsymbol{s}$ is used as a one-hot vector. For example, if we have one-hot vector $\boldsymbol{s}$, whose elements are $s_r = 1, \forall s_{r'} = 0\ (r' \neq r)$, the rth speaker is of interest. In this model, the connection weights between visible and hidden units and the bias terms of the visible and hidden units are controlled by $\boldsymbol{s}$. We define the speaker-dependent visible-hidden connections $\mathbf{W}(\boldsymbol{s})$, visible biases $\boldsymbol{b}(\boldsymbol{s})$, and hidden biases $\boldsymbol{c}(\boldsymbol{s})$ as follows.

$$\mathbf{W}(\boldsymbol{s}) = \sum_r \mathbf{A}_r s_r \bar{\mathbf{W}} \tag{1}$$

$$\boldsymbol{b}(\boldsymbol{s}) = \bar{\boldsymbol{b}} + \sum_r \boldsymbol{b}_r s_r = \bar{\boldsymbol{b}} + \mathbf{B}\boldsymbol{s} \tag{2}$$

$$\boldsymbol{c}(\boldsymbol{s}) = \bar{\boldsymbol{c}} + \sum_r \boldsymbol{c}_r s_r = \bar{\boldsymbol{c}} + \mathbf{C}\boldsymbol{s}, \tag{3}$$

where $\bar{\mathbf{W}} \in \mathbb{R}^{I \times J}$ and $\bar{\boldsymbol{b}}$ are speaker-independent parameters, and $\mathbf{A}_r \in \mathbb{R}^{I \times I}$, $\boldsymbol{b}_r \in \mathbb{R}^I (\mathbf{B} = [\boldsymbol{b}_1\ \boldsymbol{b}_2\ \cdots\ \boldsymbol{b}_R] \in \mathbb{R}^{I \times R})$ and $\boldsymbol{c}_r \in \mathbb{R}^J (\mathbf{C} = [\boldsymbol{c}_1\ \boldsymbol{c}_2\ \cdots\ \boldsymbol{c}_R] \in \mathbb{R}^{J \times R})$ are speaker-specific parameters of the rth speaker. If s is a one-hot vector where only the rth element is switched on, $\mathbf{A}_r$ is viewed as an adaptation matrix that adapts the speaker-independent weight matrix $\bar{\mathbf{W}}$ (phoneme-related features) to the rth speaker. $\boldsymbol{b}_r$ and $\boldsymbol{c}_r$ denote the speaker-specific bias of the rth speaker for the visible and hidden units, respectively. For convenience, we use a symbol $\mathcal{A} = \{\mathbf{A}_r\}_{r=1}^R$ for a collection of the speaker adaptation matrices.

Given the speaker information $\boldsymbol{s}$, the joint probability of visible and hidden units $p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s})$ as follows.

$$p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s}) = \frac{1}{Z} e^{-E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{s})} \tag{4}$$

$$E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s}) = \frac{1}{2} \left\| \frac{\boldsymbol{v} - \boldsymbol{b}(\boldsymbol{s})}{\boldsymbol{\sigma}} \right\|^2 - \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2} \right)^\top \mathbf{W}(\boldsymbol{s})\boldsymbol{h} - \boldsymbol{c}(\boldsymbol{s})^\top \boldsymbol{h} \tag{5}$$

$$Z = \int_{\boldsymbol{v}} \sum_{\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h}|\boldsymbol{s})} d\boldsymbol{v}, \tag{6}$$

where $\|\cdot\|^2$ denotes L2 norm. The fraction bar in Eq. (5) denotes the element-wise division. The parameters $\boldsymbol{\Theta} = \{\bar{\mathbf{W}}, \mathcal{A}, \mathbf{B}, \mathbf{C}, \bar{\boldsymbol{b}}, \bar{\boldsymbol{c}}, \boldsymbol{\sigma}\}$ are simultaneously estimated on the basis of maximum likelihood.

The lack of connections between visible units or between hidden units enable the conditional probabilities $p(\boldsymbol{h}|\boldsymbol{v}, \boldsymbol{s})$ and $p(\boldsymbol{v}|\boldsymbol{h}, \boldsymbol{s})$ to form simple equations as follows:

$$p(v_i = v|\boldsymbol{h}, \boldsymbol{s}) = \mathcal{N}(v \mid b_i(\boldsymbol{s}) + \boldsymbol{w}_{i:}(\boldsymbol{s})\boldsymbol{h}, \sigma_i^2) \tag{7}$$

$$p(h_j = 1|\boldsymbol{v}, \boldsymbol{s}) = \mathcal{S}\left(c_j(\boldsymbol{s}) + \boldsymbol{w}_{:j}(\boldsymbol{s})^\top \left(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2} \right) \right), \tag{8}$$

where $\boldsymbol{w}_{i:}(\boldsymbol{s})$ and $\boldsymbol{w}_{:j}(\boldsymbol{s})$ denote the ith row vector and jth column vector of $\mathbf{W}(\boldsymbol{s})$, respectively. $\mathcal{N}(\cdot|\mu, \sigma^2)$ and $\mathcal{S}(\cdot)$ denote a Gaussian probability density function with the mean μ and variance σ^2 and a sigmoid function, respectively.

In the converting step, the source speaker's acoustic features $\boldsymbol{x}^{(t)}$ at frame t can be converted to those of the target speaker $\boldsymbol{y}^{(t)}$ via latent phonological features $\hat{\boldsymbol{h}}^{(t)}$ so as to

maximize the probability $p(\boldsymbol{y}^{(t)}|\boldsymbol{x}^{(t)})$ as

$$\begin{aligned}\hat{\boldsymbol{y}}^{(t)} &\triangleq \underset{\boldsymbol{y}^{(t)}}{\operatorname{argmax}}\, p(\boldsymbol{y}^{(t)}|\boldsymbol{x}^{(t)}) \\ &\simeq \bar{\boldsymbol{b}} + \boldsymbol{b}_y + \mathbf{A}_y \bar{\mathbf{W}} \hat{\boldsymbol{h}}^{(t)},\end{aligned} \tag{9}$$

where

$$\begin{aligned}\hat{\boldsymbol{h}}^{(t)} &\triangleq \underset{\boldsymbol{h}^{(t)}}{\operatorname{argmax}}\, p(\boldsymbol{h}^{(t)}|\boldsymbol{x}^{(t)}) \\ &\simeq \mathcal{S}\left(\bar{\boldsymbol{c}} + \boldsymbol{c}_x + \bar{\mathbf{W}}^\top \mathbf{A}_x^\top \left(\frac{\boldsymbol{x}^{(t)}}{\boldsymbol{\sigma}^2}\right)\right).\end{aligned} \tag{10}$$

As Eq. (10) indicates, the (optimum) latent phonological features are approximated as the expectation values of $p(\boldsymbol{h}^{(t)}|\boldsymbol{x}^{(t)})$, which results in the sigmoidal outputs of affine-transformed acoustic features of the source speaker projected with the matrix $\bar{\mathbf{W}}^\top \mathbf{A}_x^\top$. As the column vectors of this matrix are similar to the patterns that appear in the source speaker's acoustic features, the obtained latent features $\hat{\boldsymbol{h}}$ represent speaker-independent information that is potentially phonological. Eq. (9) shows that the converted speech is generated from the phonological information that is projected to the acoustic feature space using the weight matrix adapted to the target speaker.

3. Speech Chain Voice Conversion

As shown in Fig. 3, speech chain VC is based on an ARBM model, which consists of two visible layers and one hidden layer. We represent observed acoustic features and linguistic features manually labeled parallel to the acoustic features as two classes of the visible units $\boldsymbol{a} \in \mathbb{R}^I$ and $\boldsymbol{l} \in \{0,1\}^J$, respectively, and latent phonetic distinctive features as hidden units $\boldsymbol{d} \in \{0,1\}^K$ (I, J, and K denotes the number of dimensions in acoustic features, linguistic features, and latent phonetic distinctive features, respectively).

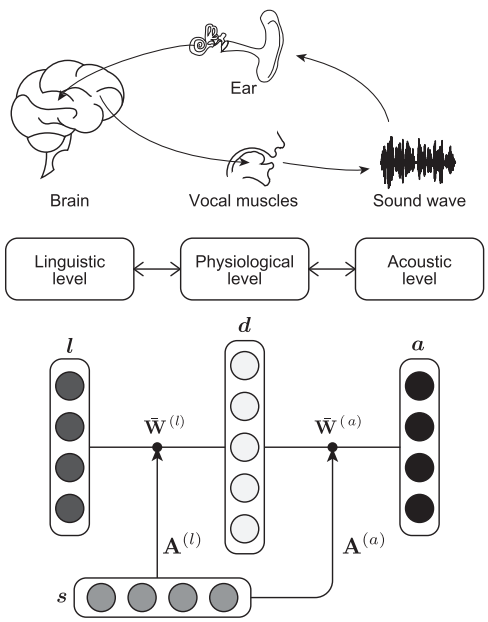

Fig. 3 Graphical representation of the proposed model. A chain of events in a speech production is represented by an adaptive restricted Boltzmann machine.

We assume speaker identity information is embedded in both the linguistic-physiological link and the physiological-acoustic link in the framework of the speech chain, and thus speaker identity unit $\boldsymbol{s}$ controls both weights of linguistic-distinctive and distinctive-acoustic feature connections. We define the first visible-hidden (linguistic-distinctive) connections $\mathbf{W}^{(l)}(\boldsymbol{s})$, the second visible-hidden (acoustic-distinctive) connections $\mathbf{W}^{(a)}(\boldsymbol{s})$, the first and second visible biases $\boldsymbol{b}^{(l)}(\boldsymbol{s})$, $\boldsymbol{b}^{(a)}(\boldsymbol{s})$, and the hidden biases $\boldsymbol{b}^{(d)}(\boldsymbol{s})$ as follows.

$$\mathbf{W}^{(\cdot)}(\boldsymbol{s}) = \sum_r \mathbf{A}_r^{(\cdot)} s_r \bar{\mathbf{W}}^{(\cdot)} \tag{11}$$

$$\boldsymbol{b}^{(\cdot)}(\boldsymbol{s}) = \bar{\boldsymbol{b}}^{(\cdot)} + \sum_r \boldsymbol{b}_r^{(\cdot)} s_r = \bar{\boldsymbol{b}}^{(\cdot)} + \mathbf{B}^{(\cdot)} \boldsymbol{s} \tag{12}$$

where s is a one-hot vector where only the rth element is switched on, and $\mathbf{A}_r^{(\cdot)}$ is an adaptation matrix that adapts the speaker-independent weight matrix $\bar{\mathbf{W}}^{(\cdot)}$ to the rth speaker. $\boldsymbol{b}_r^{(\cdot)}$ denotes the speaker-specific bias of the rth speaker. For convenience, we use a symbol $\mathcal{A}^{(\cdot)} = \{\mathbf{A}_r^{(\cdot)}\}_{r=1}^R$ for a collection of the speaker adaptation matrices.

Referring to Sone and Nakashika [20] and Cho *et al.* [21], we define the joint probability of visible and hidden units $p(\boldsymbol{a}, \boldsymbol{l}, \boldsymbol{d}|\boldsymbol{s})$ as follows:

$$p(\boldsymbol{a}, \boldsymbol{l}, \boldsymbol{d}|\boldsymbol{s}) = \frac{1}{Z} e^{-E(\boldsymbol{a},\boldsymbol{l},\boldsymbol{d}|\boldsymbol{s})} \tag{13}$$

$$\begin{aligned}E(\boldsymbol{a}, \boldsymbol{l}, \boldsymbol{d}|\boldsymbol{s}) = \frac{1}{2}\left\|\frac{\boldsymbol{a} - \boldsymbol{b}^{(a)}(\boldsymbol{s})}{\boldsymbol{\sigma}}\right\|^2 - \left(\frac{\boldsymbol{a}}{\boldsymbol{\sigma}^2}\right)^\top \mathbf{W}^{(a)}(\boldsymbol{s})\boldsymbol{d} \\ - \boldsymbol{b}^{(l)}(\boldsymbol{s})^\top \boldsymbol{l} - \boldsymbol{d}^\top \mathbf{W}^{(l)}(\boldsymbol{s})\boldsymbol{l} - \boldsymbol{b}^{(d)}(\boldsymbol{s})^\top \boldsymbol{d}\end{aligned} \tag{14}$$

$$Z = \int_{\boldsymbol{a}} \sum_{\boldsymbol{l},\boldsymbol{d}} e^{-E(\boldsymbol{a},\boldsymbol{l},\boldsymbol{d}|\boldsymbol{s})} d\boldsymbol{a}, \tag{15}$$

where $\|\cdot\|^2$ denotes L2 norm. The fraction bar in Eq. (14) denotes the element-wise division. $\boldsymbol{\sigma}$ is the deviation parameter of the acoustic feature units $\boldsymbol{a}$.

The lack of connections between visible units or between hidden units enable the conditional probabilities $p(\boldsymbol{a}|\boldsymbol{d}, \boldsymbol{s})$, $p(\boldsymbol{l}|\boldsymbol{d}, \boldsymbol{s})$, and $p(\boldsymbol{d}|\boldsymbol{a}, \boldsymbol{l}, \boldsymbol{s})$ to form simple equations as follows:

$$p(a_i = a|\boldsymbol{d}, \boldsymbol{s}) = \mathcal{N}(a \mid b_i^{(a)}(\boldsymbol{s}) + \boldsymbol{w}_{i:}^{(a)}(\boldsymbol{s})\boldsymbol{d}, \sigma_i^2) \tag{16}$$

$$p(l_j = 1|\boldsymbol{d}, \boldsymbol{s}) = \mathcal{S}(b_j^{(l)}(\boldsymbol{s}) + \boldsymbol{w}_{j:}^{(l)}(\boldsymbol{s})\boldsymbol{d}) \tag{17}$$

$$\begin{aligned}p(d_k = 1|\boldsymbol{a}, \boldsymbol{l}, \boldsymbol{s}) = \mathcal{S}\left(b_k^{(d)}(\boldsymbol{s}) + \boldsymbol{w}_{:k}^{(a)}(\boldsymbol{s})^\top \left(\frac{\boldsymbol{a}}{\boldsymbol{\sigma}^2}\right)\right. \\ \left. + \boldsymbol{w}_{:k}^{(l)}(\boldsymbol{s})^\top \boldsymbol{l}\right).\end{aligned} \tag{18}$$

where $\boldsymbol{w}_{i:}^{(\cdot)}(\boldsymbol{s})$, $\boldsymbol{w}_{j:}^{(\cdot)}(\boldsymbol{s})$, and $\boldsymbol{w}_{:k}^{(\cdot)}(\boldsymbol{s})$ denote the ith and jth row vectors, and kth column vector of $\mathbf{W}^{(\cdot)}(\boldsymbol{s})$, respectively.

Because we usually perceive linguistic features categorically in speech communication, we can add further constraints of $\sum_{j=1}^J l_j = 1$ to our model, resulting in a one-hot

vector $\boldsymbol{l}$, which indicates that only certain linguistic features are activated. If the constraints are activated, the probability distribution of the linguistic features turns into a categorical.

Given a collection of N speech and phoneme label data $\{\boldsymbol{a}^{(n)}, \boldsymbol{l}^{(n)}, \boldsymbol{s}^{(n)}\}_{n=1}^{N}$ that is composed of R speakers, the parameters $\boldsymbol{\Theta} = \{\bar{\mathbf{W}}^{(a)}, \bar{\mathbf{W}}^{(l)}, \mathcal{A}^{(a)}, \mathcal{A}^{(l)}, \mathbf{B}^{(a)}, \mathbf{B}^{(l)}, \mathbf{C}, \bar{\boldsymbol{b}}^{(a)}, \bar{\boldsymbol{b}}^{(l)}, \bar{\boldsymbol{c}}, \boldsymbol{\sigma}\}$, which include speaker-dependent and speaker-independent parameters, are simultaneously estimated to maximize the conditional log likelihood as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Theta}) &= \log \prod_{n} p(\boldsymbol{a}^{(n)}, \boldsymbol{l}^{(n)} | \boldsymbol{s}^{(n)}) \\ &= \sum_{n} \log \sum_{\boldsymbol{d}^{(n)}} p(\boldsymbol{a}^{(n)}, \boldsymbol{l}^{(n)}, \boldsymbol{d}^{(n)} | \boldsymbol{s}^{(n)}).\end{aligned} \tag{19}$$

As in the ARBM-based VC method, We can convert the source speaker's acoustic features $\boldsymbol{x}^{(t)}$ at time frame t to those of the target speaker $\boldsymbol{y}^{(t)}$ via latent phonetic distinctive features $\hat{\boldsymbol{d}}^{(t)}$ to maximize the probability $p(\boldsymbol{y}^{(t)}|\boldsymbol{x}^{(t)})$ as

$$\begin{aligned}\hat{\boldsymbol{y}}^{(t)} &\triangleq \operatorname*{argmax}_{\boldsymbol{y}} p(\boldsymbol{y}^{(t)}|\boldsymbol{x}^{(t)}) \\ &= \operatorname*{argmax}_{\boldsymbol{y}} \sum_{\boldsymbol{d}} p(\boldsymbol{d}^{(t)}|\boldsymbol{x}) p(\boldsymbol{y}^{(t)}|\boldsymbol{d}^{(t)}) \\ &\simeq \operatorname*{argmax}_{\boldsymbol{y}} p(\hat{\boldsymbol{d}}^{(t)}|\boldsymbol{x}) p(\boldsymbol{y}|\hat{\boldsymbol{d}}^{(t)}) \\ &= \operatorname*{argmax}_{\boldsymbol{y}} p(\boldsymbol{y}^{(t)}|\hat{\boldsymbol{d}}^{(t)}) \\ &= \bar{\boldsymbol{b}}^{(a)} + \boldsymbol{b}_{y}^{(a)} + \mathbf{A}_{y}^{(a)} \bar{\mathbf{W}}^{(a)} \hat{\boldsymbol{d}}^{(t)},\end{aligned} \tag{20}$$

where

$$\begin{aligned}\hat{\boldsymbol{d}}^{(t)} &\triangleq \operatorname*{argmax}_{\boldsymbol{d}} p(\boldsymbol{d}^{(t)}|\boldsymbol{x}^{(t)}) \\ &\simeq \mathbb{E}[\boldsymbol{d}^{(t)}|\boldsymbol{x}^{(t)}] \\ &= \mathcal{S}\left(\bar{\boldsymbol{b}}^{(d)} + \boldsymbol{b}_{\boldsymbol{x}}^{(d)} + \bar{\mathbf{W}}^{\top(a)} \mathbf{A}_{x}^{\top(a)} \left(\frac{\boldsymbol{x}^{(t)}}{\boldsymbol{\sigma}^2}\right) \right. \\ &\qquad \left. + \bar{\mathbf{W}}^{\top(l)} \mathbf{A}_{x}^{\top(l)} \boldsymbol{l}_{\boldsymbol{x}}^{(t)}\right).\end{aligned} \tag{21}$$

The source speaker's linguistic features $\boldsymbol{l}_{\boldsymbol{x}}^{(t)}$ required in Eq. (21) can be substituted with $\hat{\boldsymbol{l}}_{\boldsymbol{x}}^{(t)}$ obtained by repeating Gibbs sampling of Eqs. (17) and (18).

4. Experimental Evaluation I: Across Seen Speakers

To evaluate the performance of the proposed VC method, we conducted both objective and subjective evaluation experiments. The objective evaluation experiment was conducted to determine the best system configuration. The number of hidden units and type of probability distributions for the linguistic feature units were decided in the experiment. In the subjective evaluation experiment, subjective intelligibility and perceptual speaker similarity among source, target, and converted speech were evaluated.

4.1 System Configuration

We used ATR 503 sentences, which consists of recordings of five female and five male Japanese native speakers. Phonemes and their durations were manually labeled in each recording. Original recordings were sampled at 20,000 Hz. We downsampled the files to 16,000 Hz for the experiments. We also reduced the number of unique labels to 36 because a number of the original labels were redundant. For training, we used 50 sentences uttered by four speakers (two females and two males) from set A in the corpus. For evaluation, we selected one female and one male speaker as the source speakers, with the remaining used as target speakers. We used 64-dimensional acoustic features that consist of 32-dimensional Mel-cepstral features and their dynamic features. The Mel-cepstral features were calculated from 513-dimensional STRAIGHT [22] spectra every 5 ms. Phoneme label vectors were parallel to the acoustic feature vectors and used as linguistic feature vectors. To express the transitions of phonemes in sentences, we smoothed the linguistic feature vectors in each sentence by a moving average filter of a 15-ms-long rectangular window. Bernoulli and categorical distributions were chosen for a probability distribution function of the linguistic feature vectors.

For training, we used up to 512 hidden units. We trained the model for 200 iterations using Adam optimizer [23] with a batch size of 100, a learning rate of 0.001, and a momentum term β_1 of 0.9.

4.2 Methods to Compare

We compared our model with a conventional ARBM-based VC in the objective and subjective evaluations and with a GMM-based VC in the subjective evaluation. System configuration for the conventional ARBM-based VC was almost same with the proposed VC except for using no phoneme label data, using softmax constrains for hidden units and using a stochastic gradient descent (SGD) optimizer, which were suitable for training the ARBM model.

The GMM-based VC is a commonly-used method using parallel data for training. We synthesized converted speech with the parallel method for a reference condition in the subjective evaluation. We set the number of mixtures to 32, which was found to be the best in our preliminary experiment.

4.3 Objective Evaluation

We used the perceptual evaluation of speech quality (PESQ) [24], which is designed for end-to-end speech quality assessment. Source and target speech were uttered by different speakers, resulting in a very low PESQ score between them. If the VC system worked effectively, the PESQ score would improve.

For the evaluation, we used 53 sentences from set J in the corpus. To calculate the PESQ, we use parallel data of

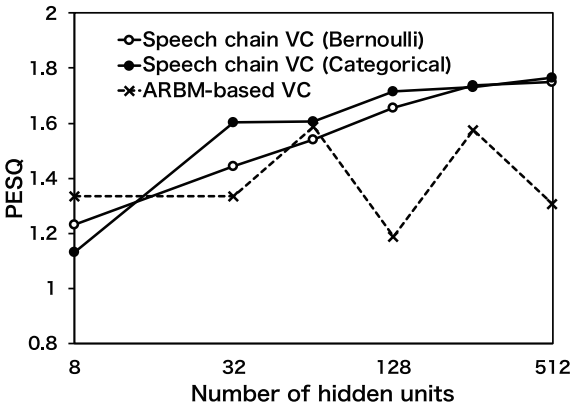

Fig. 4 Average PESQ score of our method and conventional ARBM-based method with varying numbers of hidden units.

the source and target speech that was aligned using dynamic programming. There were four variations of source and target pairs to synthesized converted speech. Half were intra-gender pairs and the remaining were inter-gender pairs. The total number of converted speech utterances was 212 for each VC model. The utterances were synthesized from the converted Mel-cepstrum, target F_0 contours and the target aperiodicities using the STRAIGHT vocoder. Phoneme labels were NOT supervised in the conversion, thus the proposed model estimated the phoneme labels from input acoustic features. Target speech utterances were also synthesized from their Mel-cepstrum, F_0 contours, and the aperiodicities to eliminate the effect of the vocoder on PESQ scores.

Figure 4 shows the effect of changing the number of hidden units in the conventional and proposed models when the number of hidden units were 8, 32, 64, 128, 256, and 512. The results of the ARBM-based method were unstable when the number of hidden units were changing, and the PESQ score increased with number of hidden units and plateaued around 256 units in the speech chain method. In the many cases, the PESQ scores of speech chain VC were higher when linguistic feature units were assumed a categorical distribution than when the units were assumed a Bernoulli distribution. The results were consistent with our expectation that we perceive linguistic features such as phonemes categorically in speech communication. On the basis of these results, we decided to use 256 hidden units for the speech chain VC, and 64 hidden units for the ARBM model in the succeeding subjective evaluations.

4.4 Subjective Evaluation

We conducted listening experiments to evaluate intelligibility and speaker similarity of the converted speech. The intelligibility was evaluated by a mean opinion score (MOS) test and the similarity was evaluated by same/different paradigm [25], [26]. Ten Japanese native speakers participated in the tests.

For the intelligibility test, 160 sentences were selected from sets D–G in the corpus. The assignment of sentences to the conversion methods was randomized over participants. 120 sentences were assigned to the converted speech conditions and the remaining 40 sentences were assigned to the original source and target speech conditions.

Participants evaluated intelligibility of the converted and original speech on a five-point Likert scale after listening to each of the speech stimuli. The scale ranged from (1) highly unintelligible to (5) highly intelligible.

In the same/different paradigm, 53 sentences were selected from set J in the corpus. In each trial, speech stimuli were presented to a participant in a pair-wise format. The pairs consisted of source/target speech and source/target/convert speech. Each participant was required to judge whether the speech pair was uttered by the same speaker and to indicate the confidence of his/her decision from four options: "Same: absolutely sure," "Same: not sure," "Different: not sure," and "Different: absolutely sure."

Table 1 shows the results of the MOS test for intelligibility. Our method significantly outperformed both the GMM-based and ARBM-based methods. The results indicate that one of the shortcomings of the ARBM-based method was overcome by adding linguistic features as visible units to the ARBM model. This extension implies that linguistic information is preserved over the conversion procedure.

Figure 5 shows the results of the similarity test. Both

Table 1 MOS for intelligibility with 95% confidence intervals. n indicates the number of measurements.

Method	MOS	n
Original	4.935±0.027	400
GMM	2.693±0.089	400
ARBM	1.970±0.099	400
SC	**3.635±0.096**	400

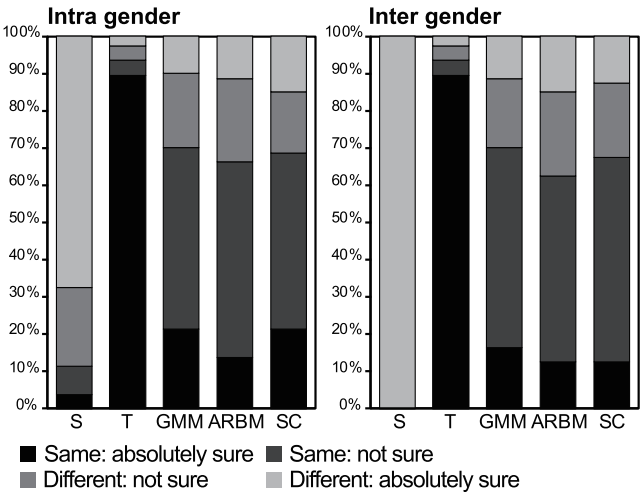

Fig. 5 Similarity (with listener confidence) to target speaker (S: Source, T: Target, GMM: GMM-based VC, ARBM: ARBM-based VC, and SC: Speech chain VC.)

Table 2 Unique phoneme clusters having a relatively high phi coefficient value ($|\phi| > 0.4$) with hidden units.

No.	ϕ	Phoneme cluster
1	0.614	/a/
2	0.599	/ç/, /ʤ/, /s/, /ʃ/, /ʧ/, /z/
3	0.548	/ç/, /s/, /ʃ/, /ʧ/
4	0.546	/a/, /e/, /o/
5	0.542	/a/, /e/, /o/, /u/, /g/, /h/, /m/, /ŋ/, /n/
6	0.535	/ʤ/, /s/, /ʃ/, /ʧ/, /z/
7	0.531	/a/, /o/, /u/, /h/
8	0.529	/ç/, /s/, /ʃ/, /ʧ/
9	0.523	/s/, /ʧ/
10	0.511	/e/, /i/, /ç/, /ʤ/, /kj/, /ʃ/, /j/
11	0.467	/a/, /e/, /i/, /o/, /u/, /h/, /j/
12	0.465	/ç/, /s/, /ʃ/, /ʧ/
13	0.461	/ç/, /hj/, /ʃ/
14	0.459	/a/, /o/, /ŋ/, /n/, /m/, /g/, /d/
15	0.446	/i/, /ç/, /ʤ/, /s/, /ʃ/, /ʧ/
16	0.443	/o/
17	0.420	/ç/, /ʤ/, /s/, /ʃ/, /ʧ/, /z/
18	0.417	/i/, /hj/, /j/

in the intra- and inter-gender conversions, around 70% of converted speech by the proposed method were perceived as being uttered by the target speakers. Chi-square tests showed that there was no significant difference in the distributions of participants' response among the VC methods: $\chi^2(2, N = 240) = 0.27, p = .874$ in the intra-gender conversion, and $\chi^2(2, N = 240) = 1.05, p = .592$ in the inter-gender conversion. The results suggest that our method can exhibit the equivalent speaker-identity conversion performance to the GMM-based method without parallel data training.

4.5 Interpretation of the Hidden Layer

If the hidden layer in the speech chain VC model corresponds to phonological distinctive features, we can find phoneme clusters sharing a same phonological distinctive feature and being associated with a certain hidden unit. For this purpose, we sought phoneme clusters of which appearance pattern being correlated with time series variation of a certain hidden unit given the training acoustic features and linguistic features. Because the two time series sequences are binary sequences, we used a phi coefficient [27], which is a similar measure to a correlation coefficient, to evaluate the correlation. The clustering procedure for each hidden unit is as follows:

1. Calculating an absolute phi coefficient value $|\phi|$ between time series variation of the hidden unit and an appearance pattern of each phoneme
2. Joining the highest-valued phoneme into the phoneme cluster as the first member
3. Calculating improvement of the $|\phi|$ when adding an appearance pattern of one of the remaining phonemes into that of the phoneme cluster.
4. Joining the new phoneme into the cluster if the $|\phi|$ improvement is over a small value ϵ (We use $\epsilon = 0.005$ in this paper)
5. Repeating 3. and 4. until reaching the last phoneme to be checked

Table 2 shows 18 unique phoneme clusters each having a relatively high phi coefficient value ($|\phi| > 0.4$) with one of hidden units. We found that some clusters consisted of phonemes sharing same phonological distinctive features. For example, the phonemes of which manners of articulation are affricate or fricative and points of articulation are alveolar or post-alveolar, i.e. /ç/ /ʤ/, /hj/, /kj/, /s/, /ʃ/, /ʧ/, and /z/, had a tendency to form clusters. These phonemes are categorized as obstruent consonants in the field of phonology [28]. Also some of sonorant consonants such as /m/, /n/, and /j/, which are more vowel-like consonants than obstruents, had a tendency to form clusters with vowels. The results of analysis support the idea that some of the hidden units correspond to phonological distinctive features.

5. Application to One-Shot VC

In this section, we describe an application of our method to one of the most challenging VC tasks: a one-shot VC task. A one-shot VC task is to perform VC across an arbitrary unseen speaker and another arbitrary unseen speaker based on only one each utterance of the speakers. One possible approach to perform a one-shot VC based on our method is blending speaker-dependent parameters and adapting them to new speakers. Another approach is training the speech chain VC model with speaker embeddings.

5.1 Speaker Identity Estimation

The first approach to perform a one-shot VC based on our proposed method is blending speaker-dependent parameters and adapting them to new speakers. If we allow each element of the speaker identity units s_r to be ranged from 0 to 1 ($0 \leq s_r \leq 1$) and sum of them to be 1 ($\sum_{r=1}^{R} s_r = 1$), we can obtain continuous weight matrices $W^{(\cdot)}(\boldsymbol{s})$ and biases $\boldsymbol{b}^{(\cdot)}(\boldsymbol{s})$ from Eq. (11) and (12). In this case, the speaker identity unit $\boldsymbol{s}$ can be regarded as the blending weight of speaker-dependent parameters.

When we have acoustic features $\boldsymbol{a}^{(t)}$, linguistic features $\boldsymbol{l}^{(t)}$, and latent distinctive features $\boldsymbol{d}^{(t)}$ at time frame t, the conditional probabilities $p(s_r^{(t)} = 1|\boldsymbol{a}^{(t)}, \boldsymbol{l}^{(t)}, \boldsymbol{d}^{(t)})$ can be formed as,

$$\begin{aligned} &p(s_r^{(t)} = 1|\boldsymbol{a}^{(t)}, \boldsymbol{l}^{(t)}, \boldsymbol{d}^{(t)}) \\ &= \frac{p(s_r^{(t)} = 1)p(\boldsymbol{a}^{(t)}, \boldsymbol{l}^{(t)}, \boldsymbol{d}^{(t)}|s_r^{(t)} = 1)}{\sum_{s'} p(s_r'^{(t)} = 1)p(\boldsymbol{a}^{(t)}, \boldsymbol{l}^{(t)}, \boldsymbol{d}^{(t)}|\boldsymbol{s}'^{(t)})} \\ &= \frac{p(\boldsymbol{a}^{(t)}, \boldsymbol{l}^{(t)}, \boldsymbol{d}^{(t)}|s_r^{(t)} = 1)}{\sum_{s'} p(\boldsymbol{a}^{(t)}, \boldsymbol{l}^{(t)}, \boldsymbol{d}^{(t)}|\boldsymbol{s}')}. \end{aligned} \tag{22}$$

We assume that $p(s_r^{(t)} = 1|\boldsymbol{a}^{(t)}, \boldsymbol{l}^{(t)}, \boldsymbol{d}^{(t)})$ indicates a likeness of the rth speaker in given features. Thus, we define the blending weight of speaker-dependent parameters $\hat{\boldsymbol{s}}^{(t)}$ at time frame t as

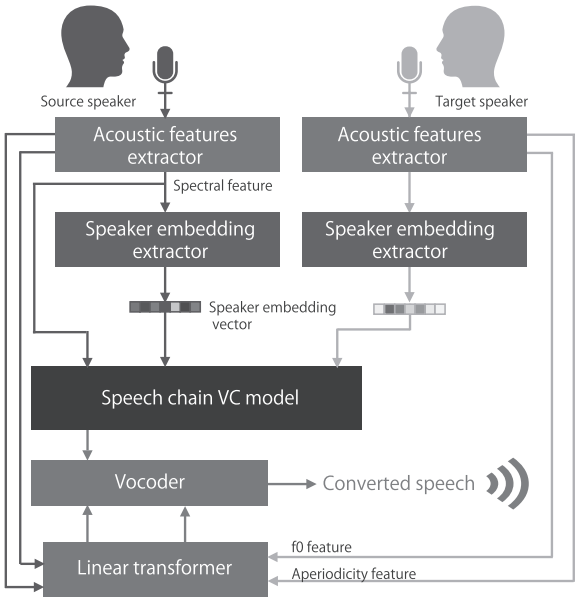

Fig. 6 Proposed one-shot voice conversion method.

$$\begin{aligned}\hat{\boldsymbol{s}}^{(t)} &\triangleq \underset{\boldsymbol{s}}{\operatorname{argmax}}\, p(\boldsymbol{s}^{(t)}|\boldsymbol{a}^{(t)}, \boldsymbol{l}^{(t)}, \boldsymbol{d}^{(t)}) \\ &\simeq \mathbb{E}[\boldsymbol{s}^{(t)}|\boldsymbol{a}^{(t)}, \boldsymbol{l}^{(t)}, \boldsymbol{d}^{(t)}] \\ &= g(\boldsymbol{a}'^{(t)\top}\boldsymbol{b}_r^{(a)} - \boldsymbol{l}^{(t)\top}\boldsymbol{b}_r^{(l)} - \boldsymbol{d}^{(t)\top}\boldsymbol{b}_r^{(d)} \\ &\qquad - \boldsymbol{a}'^{(t)\top}\mathbf{W}_r^{(a)}\boldsymbol{d}^{(t)} - \boldsymbol{l}^{(t)\top}\mathbf{W}_r^{(l)}\boldsymbol{d}^{(t)}),\end{aligned} \tag{23}$$

where $g(\cdot)$ denotes a softmax function. Equation (23) requires linguistic features $\boldsymbol{l}^{(t)}$ and latent distinctive features $\boldsymbol{d}^{(t)}$. Since we only have new speaker's acoustic features $\boldsymbol{a}^{(t)}$ in a one-shot VC task, we need to estimate $\hat{\boldsymbol{s}}^{(t)}$, $\hat{\boldsymbol{l}}^{(t)}$, and $\hat{\boldsymbol{d}}^{(t)}$ by repeating Eqs. (17), (18), and (23). The final $\hat{\boldsymbol{s}}$ is obtained by averaging $\hat{\boldsymbol{s}}^{(t)}$ over time frame t. After obtaining the blending weights of new source and target speakers: $\hat{\boldsymbol{s}}_{src}$ and $\hat{\boldsymbol{s}}_{tar}$ respectively, we can perform voice conversion in the same manner as described in Sect. 3.

This method can be adapted to the new source and target speakers using each single utterance of the speakers. However, its process needs iterative calculations to estimate parameters such as $\hat{\boldsymbol{s}}$, $\hat{\boldsymbol{l}}$, and $\hat{\boldsymbol{d}}$, resulting in a low computational efficiency and a low estimation accuracy.

5.2 Speech Chain VC Trained with Speaker Embeddings

The second approach is to train the speech chain VC model with speaker embeddings. This approach is simply done by using speaker embedding vectors as speaker identity units $\boldsymbol{s}$ in the training phase. The speaker embedding vectors are extracted from an additional model trained to embed speaker identity features into a fixed-length vector from an acoustic feature sequence.

Conversion phase is shown in Fig. 6. For an arbitrary source-target speakers pair, acoustic features—spectral features, F0 features, and aperiodicity features are extracted from both source and target speaker's utterances. The source and target spectral features are applied to compute speaker embedding vectors from the speaker embedding extractor used in the training phase. Then the speech chain VC model is driven by the obtained source spectral features and source and target speaker embedding vectors to get the converted spectral features. Finally, a vocoder synthesize the converted speech from the obtained converted spectral features, linear transformed f0 and aperiodicity features.

Compared with the first approach, the second approach has three advantages. Firstly, the size of speaker-dependent parameters is independent from the number of the training speakers. In the original speech chain VC model, the number of the parameters for the speaker adaptation matrices $\mathcal{A}^{(\cdot)}$ is I^2R, where I and R denote the size of visible units and the number of the training speakers respectively. We can reduce it by choosing the number of dimensions of speaker embedding vectors K to be smaller than R. Secondly, we can utilize training samples more efficiently. Even if some speakers in the training corpus having similar voice each other, the original speech chain VC model have to prepare different speaker-dependent parameters for each speaker. The resulting trained parameters might be redundant among the similar speakers. While in the model trained with speaker embeddings, since each speaker-dependent parameter is linked with a certain dimension of the speaker embedding vectors, the model training is less susceptible to a low variability of the training datasets. Thirdly, in a one-shot VC task, since the speech chain VC model trained with speaker embeddings is not required to estimate speaker identity units $\boldsymbol{s}$, a computational efficiency and estimation accuracies of $\hat{\boldsymbol{l}}$ and $\hat{\boldsymbol{d}}$ should be improved.

6. Experimental Evaluation II: Across Unseen Speakers

In this section, we evaluate our proposed one-shot VC method. We employed i-vectors [29] for speaker embedding vectors used in the second approach described in Sect. 5. There are several studies using i-vector extractor in VC frameworks (e.g. [30]–[32]). Liu *et al.* [32] reported that their multi-speaker VC system using i-vector extractor to obtain a speaker embedding as a conditional input can achieve voice conversion across arbitrary speakers based on a single target speaker's utterance. We use the abbreviation SCVC for the speech chain VC method utilizing one-hot vectors to identify a speaker and ISCVC for the speech chain VC trained with i-vectors.

6.1 System Configuration

To evaluate the performance of our methods, we used a dataset of 80 speakers from JVS corpus [33] for the model training. We used another four unseen speaker's datasets from the corpus as test sets.

Before the training of ISCVC, we trained the i-vector extractor using the training dataset. For the acoustic features in speaker modeling, 32-dimensional Mel-cepstrum were calculated every 5 ms by using the WORLD analyzer [34].

The 512 Gaussian universal background models, i-vector extractor, linear discriminant analysis (LDA) and probabilistic linear discriminant analysis (pLDA) [35] with whitening and length normalization were trained on all the training dataset. The dimension of i-vectors was set to 200. The rank of LDA and pLDA projection matrix is set to 79 and 40 respectively. The extracted i-vectors per each speaker were averaged and the resulting i-vector was used to represent each speaker.

The speech chain VC model of the ISCVC has the same structures as that of the SCVC. For the acoustic features, 32-dimensional Mel-cepstrum plus delta were calculated every 5 ms by using the WORLD analyzer. For the linguistic features, we used 39 phoneme labels for 3 state phones (previous, current and next phonemes.) Using 3 state phones was the minor modification from the experimental evaluation I. The purpose of the modification was to enrich the linguistic information. The 3 state phonemes were indicated by 3 sets of one-hot vectors every 5 ms. So the dimensions of acoustic features and linguistic features used in the training were 64 and 117 respectively. The number of hidden units were set to 256. In the training stage, 80 dimensional one-hot vectors were used as the speaker identity vectors $\boldsymbol{s}$ for the SCVC. For the ISCVC, to make the speaker identity vectors $\boldsymbol{s}$ were non-zero unit vectors, we modified the extracted i-vectors. The final dimension for the speaker identity vectors of the ISCVC became 80. This ensured that the model complexities for the ISCVC and the SCVC were the same. We trained the model for 80 iterations using Adam optimizer with a batch size of 8000, a learning rate of 0.001, and a momentum term β_1 of 0.9.

To make converted speech, we used 4 speakers from JVS corpus: jvs087 (male), jvs090 (female), jvs095 (female), and jvs097 (male), none of these speaker's data has appeared in the training set. We set 8 conversion pairs: male-to-male (jvs087 to jvs097 and jvs097 to jvs087), female-to-female (jvs090 to jvs095 and jvs095 to jvs090), male-to-female (jvs087 to jvs095 and jvs097 to jvs090), female-to-male (jvs090 to jvs097 and jvs095 to jvs087). There were 20 utterances for each speakers in the test sets. The sentences of the utterances were same among speakers. Each utterance was converted for each conversion setting and the target speaker's utterance was randomly selected from the 20 utterances excluding the same sentence one with the source speaker's utterance[†].

6.2 Subjective Evaluation

We conducted three subjective evaluations: a mean opinion score (MOS) listening test for naturalness, a same/different (SD) paradigm to measure speaker similarity, and a speaker similarity XAB test. In each trial of the naturalness MOS listening test, the speech sample, which was randomly selected from converted speech and the analysis-synthesized speech, was presented to a participant. Each participant evaluated the presented speech sample on a scale from 1 (highly unnatural) to 5 (highly natural). Each participant evaluated 20 samples in the MOS test. The same/different paradigm was conducted in the same manner as described in the Sect. 4.4. Each participant evaluated 20 pair samples in the SD paradigm. In the XAB test, X indicates the target reference speech. Paired speech (A and B) from the proposed and baseline methods with the same text content as the reference were presented and the participants were asked to determine which one was closer to the reference speaker. The baseline method was the GMM-based VC method which was trained with two speech samples of the same sentence uttered by source and target speakers. The number of mixtures was 8, which was found to be the best to train the GMM with such a small amount of sentences in our preliminary experiment. The number of trials was 15 for each test.

Table 3 MOS for naturalness with 95% confidence intervals. n indicates the number of measurements. Both the source and the target speakers were unseen in the training phases.

Method	MOS	n
Source	4.28±0.19	86
Target	4.26±0.18	123
SCVC	2.02±0.19	112
ISCVC	**2.33±0.21**	99

The three tests were conducted on an online evaluation system we developed. Twenty-one listeners participated in the naturalness MOS test and 20 listeners participated in the SD paradigm and 29 listeners participated in the speaker similarity XAB test.

The results of the MOS listening test for naturalness are summarized in Table 3. The ISCVC method was rated to have 2.33, which is slightly but statistically significantly higher than that of the SCVC method; $t(209) = 2.26, p = 0.0246$. We could infer that using an independent i-vector model helped the ISCVC model to estimate $\hat{\boldsymbol{l}}$ and $\hat{\boldsymbol{d}}$, and causing better naturalness of the converted speech.

Figure 7 shows the results of the SD paradigm. In the inter-gender conversion task, only 2% and 6% of converted speech based on the SCVC and the ISCVC respectively were recognized as target speaker's voices. In the intra-gender conversion task the percentages of converted speech recognized as target speaker's voices were 50% for the SCVC and 44% for the ISCVC. We further investigated the reason for the performance differences between inter- and intra-gender conversions. Table 4 shows Mel-cepstral distrotions (MCDs) between target and source/converted speech used in the experiment. We could confirm the improvements of the MCDs both in the inter- and intra-gender conversions but in the inter-gender conversion, the MCDs of the converted speech were larger than 7.02, which was the MCD between the source and target speech of the same genders. It is conceivable that the MCDs in the inter-gender conversion were improved but too large for the listeners to perceive the converted speech as being spoken by the target

[†]Some audio samples can be found in "http://sp.lab.uec.ac.jp/scvc_demo.html"

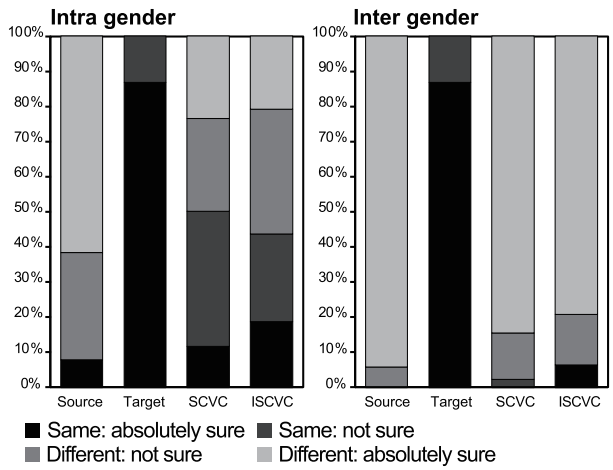

Fig. 7 Similarity (with listener confidence) to target speaker. Both the source and the target speakers were unseen in the training phases.

Table 4 MCDs (dB) between target and source/converted speech.

Method	Intra	Inter	All
Source	7.02	8.50	7.76
SCVC	6.40	7.24	6.82
ISCVC	6.96	7.91	7.43

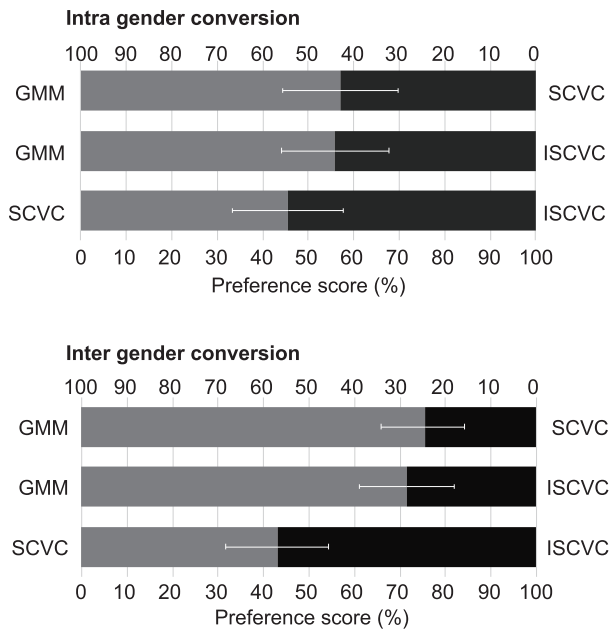

Fig. 8 Speaker similarity XAB test results for intra-gender conversion (upper) and inter-gender conversion (lower). Error bars indicate 95% confidence intervals.

speakers.

Figure 8 shows the results of the speaker similarity XAB test. In the inter-gender conversion, the preference scores of the proposed methods were significantly lower than the that of the baseline method. In the intra-gender conversion, there were no significant differences between the proposed and the baseline methods. We can see that the performance of the proposed methods are comparable to that of the GMM-based method in the intra-gender conversion. Please note that the GMM-based method is more advantageous in using parallel speech samples in the training.

It should be noted that there is a much room for improvement of performances for the method training the speech chain VC model with speaker embeddings because the method is more flexible in deciding hyper parameters. An additional examination confirmed that the MCDs were improved when we set the number of the i-vector dimension to 60: 6.28 dB in the intra-gender conversion and 6.93 dB in the inter-gender conversion. Future work will therefore investigate the best setting of the model parameters including what types of speaker embedding models and how many number of dimensions of the embedding vector should be employed.

7. Conclusion

In this paper, we proposed a voice conversion method called speech chain VC, which is based on an ARBM, introducing linguistic features and acoustic features as two classes of visible units and latent phonological distinctive features associated with articulatory movements as hidden units. The model is designed to be interpretable by associating its architecture with a chain of events in speech production. Our experimental results showed that the proposed method produced results comparable to that of a parallel-training approach utilizing GMMs in speaker similarity and the converted speech is highly intelligible. We analyzed the representations of the hidden layer of speech chain VC model and found that some of the hidden units correspond to phonological distinctive features. We further proposed two approaches to achieve one-shot VC by using the speech chain VC model. One of them was estimating the blending weight of parameters from acoustic inputs and the other one was training the speech chain VC model with speaker embeddings. Both the two approaches were moderately effective in the intra-gender conversion task, but the naturalness was slightly higher in the second approach. Our proposed method is not limited to VC and may be applicable to other tasks, such as speaker identification, speaker recognition, automatic speech recognition, and text-to-speech. Furthermore, our proposed method is extensible: one potential extension is modeling the whole speech chain framework. Such a model could be useful to obtain scientific insights into mechanisms of human speech communication.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Numbers 18K18069, 19K20618.

References

[1] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks,"

IEEE/ACM Trans. Audio Speech Lang. Process., vol.26, no.1, pp.84–96, 2017.
[2] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," Proc. EUSIPCO, pp.2100–2104, IEEE, 2018.
[3] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion," Proc. ICASSP, pp.6820–6824, IEEE, 2019.
[4] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," 2018 IEEE Spoken Language Technology Workshop (SLT), pp.266–273, IEEE, 2018.
[5] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "StarGAN-VC2: Rethinking conditional methods for StarGAN-based voice conversion," Proc. Interspeech, pp.679–683, 2019.
[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Proc. NIPS, pp.2672–2680, 2014.
[7] P. Isola, J.-Y. Zhu, T. Zhou, and A.A. Efros, "Image-to-image translation with conditional adversarial networks," Proc. CVPR, pp.5967–5976, 2017.
[8] J.-Y. Zhu, T. Park, P. Isola, and A.A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," Proc. ICCV, pp.2223–2232, 2017.
[9] 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.
[10] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," Proc. APSIPA, pp.1–6, IEEE, 2016.
[11] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," Proc. ICASSP, pp.5274–5278, IEEE, 2018.
[12] 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 Language Process., vol.23, no.3, pp.580–587, 2015.
[13] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," Proc. ICASSP, pp.4869–4873, IEEE, 2015.
[14] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.
[15] G. Montavon, W. Samek, and K.R. Müller, "Methods for interpreting and understanding deep neural networks," Digital Signal Processing, vol.73, pp.1–15, 2018.
[16] N. Liu, M. Du, and X. Hu, "Adversarial machine learning: An interpretation perspective," arXiv preprint arXiv:2004.11488, 2020.
[17] T. Nakashika, T. Takiguchi, and Y. Minami, "Non-parallel training in voice conversion using an adaptive restricted Boltzmann machine," IEEE/ACM Trans. Audio, Speech and Language Process., vol.24, no.11, pp.2032–2045, 2016.
[18] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," IEEE Trans. Speech and Audio Process., vol.13, no.5, pp.930–944, 2005.
[19] P.B. Denes and E.N. Pinson, The Speech Chain, 2 ed., W.H. Freeman and Co., New York, 1993.
[20] K. Sone and T. Nakashika, "Pre-training of DNN-based speech synthesis based on bidirectional conversion between text and speech," IEICE Trans. Inf.& Syst., vol.E102-D, no.8, pp.1546–1553, 2019.
[21] K. Cho, A. Ilin, and T. Raiko, "Improved learning of Gaussian-Bernoulli restricted Boltzmann machines," Proc. ICANN, vol.6791, pp.10–17, Springer, 2011.
[22] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," Acoust. Sci. & Tech., vol.27, no.6, pp.349–353, 2006.
[23] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
[24] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," Proc. ICASSP, pp.749–752, IEEE, 2001.
[25] M. Wester, Z. Wu, and J. Yamagishi, "Analysis of the Voice Conversion Challenge 2016 Evaluation Results," Proc. Interspeech, pp.1637–1641, 2016.
[26] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," arXiv preprint arXiv:1804.04262, 2018.
[27] K. Pearson and D. Heron, "On theories of association," Biometrika, vol.9, no.1-2, pp.159–315, 1913.
[28] A. Spencer, Phonology: Theory and Description, Blackwell, Oxford, 1996.
[29] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," IEEE Trans. Audio, Speech, and Language Process., vol.19, no.4, pp.788–798, 2010.
[30] J. Wu, Z. Wu, and L. Xie, "On the use of i-vectors and average voice model for voice conversion without parallel data," Proc. APSIPA, pp.1–6, IEEE, 2016.
[31] T. Kinnunen, L. Juvela, P. Alku, and J. Yamagishi, "Non-parallel voice conversion using i-vector plda: Towards unifying speaker verification and transformation," Proc. ICASSP, pp.5535–5539, IEEE, 2017.
[32] S. Liu, J. Zhong, L. Sun, X. Wu, X. Liu, and H. Meng, "Voice conversion across arbitrary speakers based on a single target-speaker utterance.," Proc. Interspeech, pp.496–500, 2018.
[33] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JVS corpus: free Japanese multi-speaker voice corpus," arXiv preprint arXiv:1908.06248, 2019.
[34] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," IEICE Trans. Inf.& Syst., vol.E99-D, no.7, pp.1877–1884, 2016.
[35] D. Garcia-Romero and C.Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," Proc. Interspeech, pp.249–252, 2011.

Takuya Kishida received his B.E. and M.E. degrees in design from Kyushu University in 2013 and 2015, respectively. He received Ph.D. (Design) from Kyushu University in 2018. To March 2019, he was a postdoctoral researcher at Kyushu University. He is currently a project researcher at the University of Electro-Communications. He is a member of the ASJ.

Toru Nakashika received his B.E. and M.E. degrees in computer science from Kobe University in 2009 and 2011, respectively. On the summer in 2010, he was a student researcher at IBM Research, Tokyo Research Laboratory. From September 2011 to August 2012, he was a visiting researcher in the image group at INSA de Lyon in France. In the same year, he continued his research as a doctoral student at Kobe University, and received his Dr.Eng. degree in computer science in 2014. To April 2015, he was an assistant professor at Kobe University. In 2015, he joined the University of Electro-Communications as an assistant professor. In 2020, he became an associate professor with the Graduate School of Information Systems. He received the Young Researcher's Award in IEICE Speech Field in 2013, the Best Paper Award in SIGMUS Ongaku Symposium 2016, the 44th Awaya Prize Young Researcher Award from the Acoustical Society of Japan, the 15th Itakura Prize Innovative Young Researcher Award from the Acoustical Society of Japan. He is a member of the IEEE, the IEICE, the ASJ, the JSAI, and the ISCA.