Non-Parallel Voice Conversion Based on Free-Energy Minimization of Speaker-Conditional Restricted Boltzmann Machine

Takuya Kishida* and Toru Nakashika*
* The University of Electro-Communications, Tokyo, Japan
E-mail: kishida@uec.ac.jp, nakashika@uec.ac.jp

***Abstract*—In this paper, we propose a non-parallel voice conversion method based on the minimization of the free energy of a restricted Boltzmann machine (RBM). The proposed method uses an RBM that learns the generative probability of acoustic features conditioned on a target speaker, and it iteratively updates the input acoustic features until their free energy reaches a local minimum to obtain converted features. Since it is based on the RBM, only a few hyperparameters need to be set, and the number of training parameters is very small. Therefore, training is stable. In determining the step size of the update formula in accordance with the Newton-Raphson method to obtain the feature that gives the local minimum of the free energy, we found that the Hesse matrix of the free energy can be approximated by a diagonal matrix, and the update can be performed efficiently with a small amount of calculation. In objective evaluation experiments, the proposed method outperforms StarGAN-VC in Mel-cepstral distortions. In subjective evaluation experiments, the performance of the proposed method is comparable to that of StarGAN-VC in similarity MOS.**

I. Introduction

Voice conversion (VC) is a speech signal processing technique that converts certain aspects of the information conveyed by speech while preserving the linguistic content. The information to be converted by the VC method includes the speaker's identity, the speaker's emotion, and the impression the voice gives. Converting such information allows, for example, a synthesized speech sound to be made that sounds like a specific person, the identity of one's own voice to be changed to protect privacy, and accents to be changed to sound more fluent when speaking a foreign language.

Methods that do not require speech content to be aligned across domains to be converted are called non-parallel VC methods. The task of recording speech samples whose linguistic content is aligned across domains is labor-intensive. Non-parallel VC not only facilitates data collection but also allows the range of applications to be expanded, such as training speaker identity conversion models of different native languages or adapting a model to convert voice in a new domain by introducing a new dataset for a previously trained model.

Many methods that utilize generative models have been proposed for non-parallel VC. Examples include methods that use adversarial learning [1] to obtain a mapping function between source and target acoustic features [2]–[4] and a method that uses an encoder-decoder model to create a latent space in which the speech content and voice information of the input acoustic features is disentangled and the disentangled information is then recomposed before being input to the decoder to obtain acoustic features of the target voice [5]–[7].

As a different approach from the methods listed above, we propose a method that uses the minimization of the free energy of a restricted Boltzmann machine (RBM) that learns probability distributions of the acoustic features of target voices. The proposed method requires a relatively small number of model parameters and can be easily trained with a single model. The encoder-decoder model sometimes has a problem with feature quality degradation caused by using different combinations of latent variables and target voice codes from those used in training. The proposed method, in comparison, transforms the input acoustic features to eliminate mismatches between latent variables and target voice codes, so it is expected that similar quality degradation is unlikely to occur.

The contributions of our paper are as follows. We propose a voice conversion method with reasonable performance based on a lightweight generative model with few hyperparameters. Since the proposed method is based on RBM, the only hyperparameter is the number of hidden units. This not only makes the model easy to train but also reduces the number of parameters compared with other methods based on deep generative models. We also show that using a step size that approximates an inverse Hesse matrix in the Newton-Raphson method for the update formula for feature conversion allows for efficient updating, and that the update formula has a connection to mean-field approximation [8].

II. Related Works

Generative adversarial nets (GAN) [1] are a generative model widely used for various tasks due to the high quality of the generated data. GAN-based methods have also been proposed for VC methods, such as CycleGAN-VC [2], [3], and StarGAN-VC [4].

CycleGAN-VC is an application of CycleGAN [9], which was proposed as an image-to-image translation model, to the voice conversion task. CycleGAN is a model in which a generator that takes data from one domain as input and outputs fake data translated to another domain is trained by

adversarial loss, while the fake data is also translated in the reverse direction so that cycle consistency is maintained. This training framework yields a generator that translates from one domain to the other while preserving the common structure across domains in the data. CycleGAN-VC is a method for converting the voice of speech to a target voice while preserving its linguistic content by taking advantage of CycleGAN's tendency to preserve the common structure among domains. StarGAN [10] is a generalized model of CycleGAN. A domain label is input along with input data to a single generator, which specifies the target domain for conversion. The generator is trained simultaneously with a discriminator as well as a classifier that classifies the domains to which the real and fake data belong. The parameters of the generator are updated by adversarial loss to fool the discriminator, cycle consistency loss to maintain a common structure across domains before and after the conversion, and classification loss to ensure that the fake data is classified in the target domain by the classifier. This training framework allows us to build a single network that is capable of converting to multiple target domains regardless of input, rather than a network for one-to-one conversion. Thus, StarGAN-VC is a method that can convert to a variety of voice domains by training on a non-parallel data set.

Autoencoders (AEs) and their stochastic versions, variational autoencoders (VAEs) [11], are another group of generative models used as base models for voice conversion. AEs and VAEs consist of encoder-decoder networks, where the encoder maps the input data to the latent space and the decoder reconstructs the original data from the features on the latent space. By well-directed training, the features on the latent space can be trained to be factorial representations of the input data. When using these models for voice conversion, the decoder is conditioned on voice domain codes to facilitate the encoder in extracting information outside the conditioned domain. Voice conversion can be performed by replacing the original voice domain code with the target one and inputting it to the decoder along with latent features.

Voice conversion methods that have been proposed on the basis of these models include VQVC [5], [6] and AutoVC [12]. VQVC is a method that strongly encourages the latent variable to be the encoding of the linguistic content by vector quantization [7]. AutoVC is based on vanilla autoencoders and is a method that encourages the latent variable to contain linguistic content without over- or under-encoding by limiting the number of dimensions of the latent variable.

Both GAN- and VAE-based voice conversion methods have their own weaknesses. GANs are generally unstable in training because the generators and discriminators are trained in an adversarial manner, and it takes a lot of effort to adjust the hyperparameters for stabilization. It has also been noted that even with a high generation quality, the generated data tend to lack diversity [13]. In voice conversion, this problem becomes a concern because the GAN generators may not be able to represent a wide variety of phonemes of speech. Weaknesses in VAEs can occur during conversion. During conversion, VAEs condition the decoder on a domain code that is different from the original voice domain. Since this situation is not experienced during training, the mismatch between the latent variable and the domain code tends to lower the quality of the generated data.

VoiceGrad [14], a non-parallel VC method based on the denoising score matching method [15], [16], has been shown to have performance comparable to GAN-based VCs. The key idea of VoiceGrad [14] is considering VC as a problem of finding a path to a stationary point in the log-density distribution of the acoustic features of target voices, starting from input acoustic features. The gradient of the target log-density distribution is called the score function. Various levels of Gaussian noise are added to the data during training, and the score function of the noisy data is estimated by a noise conditional score network. VC is performed by using Langevin dynamics to gradually transform source speaker acoustic features along the direction of the estimated gradient. Training with various noise levels allows for a wide range of movement within the feature space during the transformation. This method is closely related to the proposed method in that the conversion is based on a common key idea. That is, it defines an update formula for the transformation of the features, and the update gradually transforms the features to achieve the target voice while preserving the linguistic content. Transforming the features in accordance with a score function is equivalent to transforming the features so that the free energy is reduced in the RBM. The proposed method, however, does not train on data with various noise levels as VoiceGrad does and can transform features only in accordance with the free energy for clean data. The reason the proposed method does not need to set multiple noise levels is probably because the step size of the update formula is chosen to be theoretically efficient.

We previously proposed RBM-based methods [17], [18] based on a different idea than that behind the proposed method. RBM [19], [20] is a generative model that consists of a network of visible layers and one hidden layer, with undirected connections only between the visible and hidden layers. The model is lightweight yet capable of learning the generative probabilities of data consisting of high dimensions as joint probabilities with hidden variables that cannot be observed but are certainly considered to exist. Adaptive RBM (ARBM) and its derivatives are designed so that universal phonological information is represented in the hidden layer by applying different adaptation matrices for each voice to the RBM weight parameters. Therefore, voice conversion can be performed by encoding phonological information from input acoustic features into the hidden layer using model parameters adapted to the input voice and then obtaining output acoustic features using model parameters adapted to the target. This method enables voice conversion in a simple way, but for the conversion to work, the hidden layer must be trained to represent universal phonological information. Therefore, the number of units in the hidden layer must be narrowed down, but this is

a trade-off against the quality of the features generated.

III. Proposed Model

In this section, we will introduce the training of the RBM used in the proposed method and how the voice conversion is performed using the trained RBM. Although the proposed method can be used for a variety of voice conversion tasks, we will assume a speaker identity conversion task in the following for the purpose of explaining the method.

A. Speaker-conditional RBM

To estimate the probability distribution of the acoustic features $\boldsymbol{x} \in \mathbb{R}^I$ of the speech of the target speaker, we consider an RBM with hidden variables $\boldsymbol{h} \in \{0,1\}^J$, conditioned on the speaker one-hot vector $\boldsymbol{s} \in \{0,1\}^K, \sum_{k=1}^K s_k = 1$, as:

$$\mathbb{P}_\theta(\boldsymbol{x}, \boldsymbol{h}|\boldsymbol{s}) = \frac{1}{Z(\boldsymbol{s})} e^{-E_\theta(\boldsymbol{x},\boldsymbol{h},\boldsymbol{s})} \tag{1}$$

$$\begin{aligned} E_\theta(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{s}) = &\frac{1}{2}\left(\frac{\boldsymbol{x}}{\boldsymbol{\sigma}}\right)^\mathsf{T}\left(\frac{\boldsymbol{x}}{\boldsymbol{\sigma}}\right) - \boldsymbol{x}^\mathsf{T}\mathbf{W}\boldsymbol{h} - \boldsymbol{b}^\mathsf{T}\boldsymbol{x} \\ &- \boldsymbol{c}^\mathsf{T}\boldsymbol{h} - \boldsymbol{s}^\mathsf{T}\mathbf{V}\boldsymbol{h} \end{aligned} \tag{2}$$

$$Z(\boldsymbol{s}) = \int \sum_{\boldsymbol{h}} e^{-E_\theta(\boldsymbol{x},\boldsymbol{h},\boldsymbol{s})} d\boldsymbol{x} \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{I\times J}, \mathbf{V} \in \mathbb{R}^{K\times J}, \boldsymbol{b} \in \mathbb{R}^I, \boldsymbol{c} \in \mathbb{R}^J$, and $\boldsymbol{\sigma} \in \mathbb{R}^I_+$ are a matrix of visible-hidden connection weights, a matrix of speaker-hidden connection weights, a visible bias vector, a hidden bias vector, and the deviation vector of $\boldsymbol{v}$, respectively; Z is a normalization term; $\frac{\cdot}{\cdot}$ is element-wise division.

Defining the free energy of this RBM as $F_\theta(\boldsymbol{x}|\boldsymbol{s}) \equiv -\log \sum_{\boldsymbol{h}} e^{-E_\theta(\boldsymbol{x},\boldsymbol{h},\boldsymbol{s})}$, the marginal probability distribution of the visible variable $\mathbb{P}_\theta(\boldsymbol{x}|\boldsymbol{s})$ can be written as:

$$\mathbb{P}_\theta(\boldsymbol{x}|\boldsymbol{s}) = \sum_{\boldsymbol{h}} \mathbb{P}_\theta(\boldsymbol{x}, \boldsymbol{h}|\boldsymbol{s}) = \frac{1}{Z(\boldsymbol{s})} \sum_{\boldsymbol{h}} e^{-E_\theta(\boldsymbol{x},\boldsymbol{h},\boldsymbol{s})} \tag{4}$$

$$= \frac{1}{Z(\boldsymbol{s})} e^{-F_\theta(\boldsymbol{x}|\boldsymbol{s})}. \tag{5}$$

Let $\mathbb{P}_{\text{data}}(\boldsymbol{x}|\boldsymbol{s})$ denote the empirical distribution of training data for the acoustic features of the target speaker $\boldsymbol{s}$; the model $\mathbb{P}_\theta(\boldsymbol{x}|\boldsymbol{s})$ is trained with the goal of minimizing the Kullback-Leibler divergence $\text{KL}(\mathbb{P}_{\text{data}}||\mathbb{P}_\theta)$. By extracting the term related to the parameter θ from $\text{KL}(\mathbb{P}_{\text{data}}||\mathbb{P}_\theta)$, model training is reduced to solving the maximization problem of the following objective function $f(\theta)$:

$$f(\theta) \equiv \sum_{\boldsymbol{x}\in\mathcal{D}} \mathbb{P}_{\text{data}}(\boldsymbol{x}|\boldsymbol{s}) \log \mathbb{P}_\theta(\boldsymbol{x}|\boldsymbol{s}), \tag{6}$$

where $\mathcal{D}$ is training dataset. Since θ that maximizes $f(\theta)$ cannot be obtained analytically, the parameter update is based on the gradient descent method in accordance with the following gradient:

$$\begin{aligned} \nabla_\theta f(\theta) &= \sum_{\boldsymbol{x}} \mathbb{P}_{\text{data}}(\boldsymbol{x}|\boldsymbol{s}) \nabla_\theta \log \mathbb{P}_\theta(\boldsymbol{x}|\boldsymbol{s}) \\ &= -\mathbb{E}_{\mathbb{P}_{\text{data}}(\boldsymbol{x}|\boldsymbol{s})}\left[\nabla_\theta F_\theta(\boldsymbol{x}|\boldsymbol{s})\right] + \mathbb{E}_{\mathbb{P}_\theta(\boldsymbol{x}|\boldsymbol{s})}\left[\nabla_\theta F_\theta(\boldsymbol{x}|\boldsymbol{s})\right]. \end{aligned} \tag{7}$$

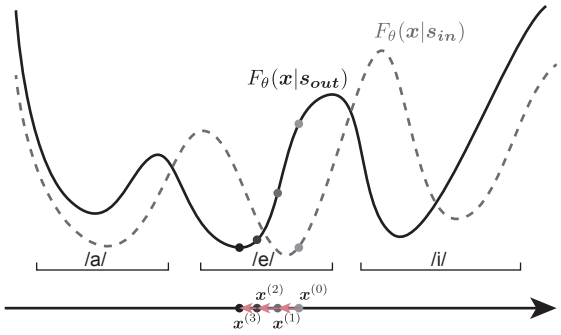

Fig. 1. Conceptual illustration of free-energy minimization. Input acoustic features $\boldsymbol{x}^{(0)}$ perceived as a certain phoneme is iteratively updated along the free-energy gradient of the RBM conditioned on the target speaker $\boldsymbol{s}_{out}$ without changing the phonological information.

The second term on the right-hand side of (7) is difficult to compute because it is an expected value based on the probability distribution given by the model, but it can be substituted by an approximation using the contrastive divergence method [21].

B. Voice Conversion Based on Free-energy Minimization

In this paper, we propose a voice conversion method that iteratively updates the input acoustic feature $\boldsymbol{x}_{\text{in}}$ along the free-energy gradient $\nabla_{\boldsymbol{x}} F_\theta(\boldsymbol{x}|\boldsymbol{s})$ of the RBM in the feature space in the direction where the free energy becomes lower. Obtaining features such that the free energy $F_\theta(\boldsymbol{x}|\boldsymbol{s})$ is lower means obtaining acoustic features that are more likely to be uttered by the target speaker $\boldsymbol{s}$. Therefore, as illustrated in Fig. 1, if $\hat{\boldsymbol{x}}$ gives a local minimum of $F_\theta(\boldsymbol{x}|\boldsymbol{s})$ in the neighborhood of $\boldsymbol{x}_{\text{in}}$, and if the speech content is preserved in such a change to $\hat{\boldsymbol{x}}$, voice conversion to the target speaker is achieved.

C. Step Size Based on Newton-Raphson Method

To update $\boldsymbol{x}^\tau$ to $\boldsymbol{x}^{\tau+1}$ in accordance with $\nabla_{\boldsymbol{x}} F_\theta(\boldsymbol{x}|\boldsymbol{s})$, let us consider the following update formula:

$$\boldsymbol{x}^{\tau+1} = \boldsymbol{x}^\tau - \alpha \nabla_{\boldsymbol{x}} F_\theta(\boldsymbol{x}^\tau|\boldsymbol{s}), \tag{8}$$

where α indicates step size. For efficient updating, α must be set appropriately. Therefore, we consider following the Newton-Raphson method, which chooses the inverse Hessematrix $\nabla^2_{\boldsymbol{x}} F_\theta(\boldsymbol{x}|\boldsymbol{s})^{-1}$ as the step size. First, $\nabla_{\boldsymbol{x}} F_\theta(\boldsymbol{x}|\boldsymbol{s})$ of the RBM based on (1) becomes:

$$\begin{aligned} \nabla_{\boldsymbol{x}} F_\theta(\boldsymbol{x}|\boldsymbol{s}) &= \sum_{\boldsymbol{h}} \mathbb{P}_\theta(\boldsymbol{h}|\boldsymbol{x}, \boldsymbol{s}) \nabla_{\boldsymbol{x}} E_\theta(\boldsymbol{x}, \boldsymbol{h}|\boldsymbol{s}) \\ &= \mathbb{E}_{\mathbb{P}_\theta(\boldsymbol{h}|\boldsymbol{x},\boldsymbol{s})}\left[\nabla_{\boldsymbol{x}} E_\theta(\boldsymbol{x}, \boldsymbol{h}|\boldsymbol{s})\right] \\ &= \frac{\boldsymbol{x}}{\boldsymbol{\sigma}^2} - \boldsymbol{b} - \mathbf{W}\mathbb{E}_{\mathbb{P}_\theta(\boldsymbol{h}|\boldsymbol{x},\boldsymbol{s})}\left[\boldsymbol{h}|\boldsymbol{x}, \boldsymbol{s}\right] \\ &= \frac{\boldsymbol{x}}{\boldsymbol{\sigma}^2} - \boldsymbol{b} - \mathbf{W}\mathcal{S}(\mathbf{W}^\mathsf{T}\boldsymbol{x} + \mathbf{V}^\mathsf{T}\boldsymbol{s} + \boldsymbol{c}), \end{aligned} \tag{9}$$

where $\cdot^2$ is an element-wise power, and $\mathcal{S}(\cdot)$ is an element-wise sigmoid function. If we then write $\hat{\boldsymbol{h}} = \mathcal{S}(\mathbf{W}^\mathsf{T}\boldsymbol{x} + \mathbf{V}^\mathsf{T}\boldsymbol{s} + \boldsymbol{c})$, we get

$$\nabla^2_{\boldsymbol{x}} F_\theta(\boldsymbol{x}|\boldsymbol{s}) = \Delta(\frac{1}{\boldsymbol{\sigma}^2}) - \mathbf{W}\Delta(\hat{\boldsymbol{h}})\Delta(\mathbf{1} - \hat{\boldsymbol{h}})\mathbf{W}^\mathsf{T}, \tag{10}$$

where $\Delta(\cdot)$ is a function that returns a diagonal matrix with vector $\cdot$ as its diagonal component. Here, each element of $\hat{\boldsymbol{h}}$ is the output of a sigmoid function, many of which take values close to 0 or 1. Therefore, by approximating $\hat{h}_j(1-\hat{h}_j)$ to 0, the Hesse matrix $\nabla^2_{\boldsymbol{x}} F_\theta(\boldsymbol{x}|\boldsymbol{s})$ is considered a diagonal matrix, and the inverse matrix is easy to calculate. Therefore, in this paper, we used $\nabla^2_{\boldsymbol{x}} F_\theta(\boldsymbol{x}|\boldsymbol{s})^{-1} \simeq \Delta(\boldsymbol{\sigma}^2)$ as the step size, and the update formula becomes:

$$\boldsymbol{x}^{\tau+1} = \Delta(\boldsymbol{\sigma}^2)\left(\mathbf{W}\mathcal{S}(\mathbf{W}^\mathsf{T}\boldsymbol{x}^\tau + \mathbf{V}^\mathsf{T}\boldsymbol{s} + \boldsymbol{c}) + \boldsymbol{b}\right). \quad (11)$$

The iterative update by (11) has the same operation as the alternating update of $\boldsymbol{x}$ and $\boldsymbol{h}$ by mean-field approximation [8]. The update by this method is based on the Newton-Raphson method, so the update is efficient, and the step size is positive, so the update is stable.

IV. Experiments

To evaluate the performance of the proposed VC method, we conducted both objective and subjective evaluation experiments. Among existing non-parallel VC methods, we chose the adaptive restricted Boltzmann machine-based VC (ARBM-VC) [17], which is RBM-based but with different conversion methods, and StarGAN-VC [4], which is a GAN-based method, as comparison methods. Note that StarGAN-VC is a method that uses two-dimensional data consisting of feature dimensions and time frames as input features, and its performance cannot be directly compared to the proposed method. VoiceGrad, which is based on an idea common to the proposed method, currently has no publicly available code, including unofficial open-source implementations. We also tried to implement VoiceGrad, but the quality of the converted speech was apparently worse than that reported in the paper, so we did not include it with the comparison methods in this experiment.

A. Materials & Configurations

For the experiments, we used the CMU ARCTIC dataset [22], which consists of recordings of 18 speakers each reading the same 1,132 English sentences. All the recorded speech was sampled at 16,000 Hz. For the target speakers, we used two female speakers, "clb" and "slt," and two male speakers, "bdl" and "rms." We used a set of 400 sentences for training data and used another set of 100 sentences for test data. We divided the training 400 sentences into 4 subsets of 100 sentences and assigned subsets to the 4 target speakers to simulate a non-parallel training scenario so that the training data would not contain the same sentences across speakers.

A 32-dimensional Mel-spectrum normalized over each dimension was used for input acoustic features. The Mel-spectrum was calculated from the smoothed spectrum obtained with the speech analysis and synthesis system WORLD [23] every 5 ms.

The number of hidden units both in the RBM and the ARBM was set at 400. Both models were trained in Adam [24] with a batch size of 100, learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and 100 epochs. For StarGAN-VC, we used the open-source implementation[1]. The total numbers of parameters were 14864 for RBM, 17108 for ARBM, and approximately 20M for StarGAN.

TABLE I
Mel-cepstral distortion comparisons.

Pair	StarGAN	ARBM	RBM
clb-to-bdl	**7.17**	7.58	7.24
clb-to-slt	6.84	6.38	**6.27**
clb-to-rms	6.93	7.83	**6.64**
bdl-to-clb	7.13	7.54	**6.97**
bdl-to-slt	7.44	7.75	**6.94**
bdl-to-rms	7.34	7.03	**6.39**
slt-to-clb	6.73	6.29	**6.17**
slt-to-bdl	**7.45**	7.90	7.48
slt-to-rms	7.35	7.92	**6.92**
rms-to-clb	**7.01**	8.00	7.40
rms-to-bdl	7.50	7.28	**7.00**
rms-to-slt	7.32	7.92	**7.26**

TABLE II
Mean opinion score with 95% confidence intervals for speech naturalness (NAT) and similarity (SIM).

Test	StarGAN	ARBM	RBM	Natural
NAT	3.24±.25	2.97±.20	2.66±.18	4.61±.14
SIM	3.52±.25	2.59±.30	3.18±.27	–

B. Objective Evaluations

Voice conversion was performed with each method on 100 test sentences. In the conversion by the proposed method, the feature updating by (11) was performed 10 times. In the conversion by ARBM-VC, the value of the hidden layer of the RBM for the input acoustic features was obtained by using the model parameters adapted for the input speaker, and the output acoustic features were then approximated by mean field approximation using the model parameters adapted for the target speaker.

Table I shows the Mel-cepstral distortion (MCD) between the converted speech and the target speech. For most of the conversion pairs, the best performance of the proposed method was observed. Compared with ARBM-VC, the MCDs of the proposed method were superior for all conversion pairs, confirming the effectiveness of the method using free energy minimization. The proposed method was more effective, especially for inter-gender conversions. Since the proposed method searches for features that minimize free energy in the neighborhood of the input features, better conversion results could have more likely be obtained between speakers who originally have somewhat similar voices.

C. Subjective Listening Tests

We conducted mean opinion score (MOS) tests to compare the speech naturalness and speaker similarity of the converted speech samples synthesized by the proposed and comparison methods. For these tests, eight listeners participated in both naturalness and similarity tests.

[1]https://github.com/liusongxiang/StarGAN-Voice-Conversion

We generated an acoustic waveform of converted speech using the WORLD vocoder from a converted Mel-spectrum, linear transformed F_0 contours, and aperiodicities.

For the naturalness test, we included a natural speech condition in which samples were synthesized from the acoustic features directly extracted from real speech samples. The listeners evaluated the naturalness by selecting from 5: Excellent, 4: Good, 3: Fair, 2: Poor, or 1: Bad for each utterance. In the similarity test, paired utterances of converted and natural speech of the corresponding target speaker were presented to the listeners. The listeners evaluated how likely they were to be produced by the same speaker by selecting from 5: Definitely, 4: Likely, 3: Fair, 2: Not very likely, or 1: Unlikely. In each test, the listeners evaluated 40 samples randomly selected from the test speech samples.

The results of the MOS test with 95% confidence intervals are shown in Table II. For the naturalness test, the proposed method was inferior to StarGAN-VC but was comparable to ARBM-VC. For the similarity test, the proposed method outperformed ARBM-VC and was comparable to StarGAN-VC. Since StarGAN-VC can use the time series information of features, it is thought that the naturalness of the converted speech is easily preserved. The proposed method is also expected to improve naturalness by allowing the model to train on the time-series information of features.

V. Conclusions

In this paper, we proposed a non-parallel voice conversion method that iteratively updates input acoustic features until the free energy of the RBM, conditioned on a target speaker, reaches a local minimum, and the updated acoustic features are used as conversion features. The proposed method has less than one-thousandth the number of parameters of StarGAN-VC, yet the performance of the target speaker similarity of the converted speech is comparable to that of StarGAN-VC. The proposed method is adaptable not only to RBM but also to energy-based models in general. In the future, we would like to investigate its performance with more expressive energy-based models.

VI. Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Numbers 19K20618, 21K11957.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[2] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017.

[3] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6820–6824.

[4] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks," *arXiv preprint arXiv:1806.02169*, 2018.

[5] D.-Y. Wu and H.-y. Lee, "One-shot voice conversion by vector quantization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7734–7738.

[6] D.-Y. Wu, Y.-H. Chen, and H.-Y. Lee, "VQVC+: One-shot voice conversion by vector quantization and U-net architecture," *arXiv preprint arXiv:2006.04154*, 2020.

[7] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[8] J. R. Anderson and C. Peterson, "A mean field theory learning algorithm for neural networks," *Complex Systems*, vol. 1, pp. 995–1019, 1987.

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[10] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.

[11] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[12] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[13] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with vq-vae-2," *Advances in neural information processing systems*, vol. 32, 2019.

[14] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, and S. Seki, "VoiceGrad: Non-parallel any-to-many voice conversion with annealed langevin dynamics," *arXiv preprint arXiv:2010.02977*, 2020.

[15] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Adv. Nerual Information Processing Systems (NeurIPS)*, 2019, pp. 11 918–11 930.

[16] ——, "Improved techniques for training score-based generative models," *arXiv preprint arXiv:2006.09011*, 2020.

[17] 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 (TASLP)*, vol. 24, pp. 2032–2045, 2016.

[18] T. Kishida and T. Nakashika, "Speech Chain VC: Linking linguistic and acoustic levels via latent distinctive features for RBM-based voice conversion," *IEICE TRANSACTIONS on Information and Systems*, vol. 103, pp. 2340–2350, 2020.

[19] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, pp. 1527–1554, 2006.

[20] K. Cho, A. Ilin, and T. Raiko, "Improved learning of Gaussian-Bernoulli restricted Boltzmann machines," in *International conference on artificial neural networks*. Springer, 2011, pp. 10–17.

[21] M. A. Carreira-Perpinan and G. Hinton, "On contrastive divergence learning," in *International workshop on artificial intelligence and statistics*. PMLR, 2005, pp. 33–40.

[22] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[23] 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. 99, pp. 1877–1884, 2016.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.