INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Simultaneous Conversion of Speaker Identity and Emotion Based on Multiple-Domain Adaptive RBM

Takuya Kishida, Shin Tsukamoto, Toru Nakashika

The University of Electro-Communications, Japan

kishida@uec.ac.jp, tsukamoto@sd.is.uec.ac.jp, nakashika@uec.ac.jp

Abstract

In this paper, we propose a multiple-domain adaptive restricted Boltzmann machine (MDARBM) for simultaneous conversion of speaker identity and emotion. This study is motivated by the assumption that representing multiple domains (e.g., speaker identity, emotion, accent) of speech explicitly in a single model is beneficial to reduce the effects from other domains when the model learns one domain's characteristics. The MDARBM decomposes the visible-hidden connections of an RBM into domain-specific factors and a domain-independent factor to make it adaptable to multiple domains of speech. By switching the domain-specific factors from the source speaker and emotion to the target ones, the model can perform a simultaneous conversion. Experimental results showed that the target domain conversion task was enhanced by the other in the simultaneous conversion framework. In a two-domain conversion task, the MDARBM outperformed a combination of ARBMs independently trained with speaker-identity and emotion units.

Index Terms: voice conversion, emotion conversion, speaker recognition, emotional speech recognition, generative model

1. Introduction

Thanks to the recent developments in automatic speech recognition (ASR) and text-to-speech (TTS) systems, interactions and cooperation between humans and machines are becoming more real. An ASR system can recognize the linguistic contents of speech precisely and a TTS system can produce speech that is just as natural and intelligible as that produced by humans (e.g., [1, 2]). However, speech contains not only linguistic information but also paralinguistic information [3] such as speaker identity and emotion. When it comes to achieving comfortable speech communication between humans and machines, handling the paralinguistic information in speech is still a challenging task.

One of the techniques to handle such paralinguistic information on speech interfaces is voice conversion (VC). The purpose of VC is to modify speaker identity and speaking type (emotion, style, accent, and character) while preserving the inherent linguistic information [4]. This technique can be applied to various scenarios. For example, VC has been used to improve the speech recognition accuracy in noisy environments [5] and to develop speaking-aid systems [6, 7].

Speaker-identity conversion and emotion conversion tasks are major variants of VC tasks, and there are many previously proposed approaches to both. Modification of the spectrum configurations and statistical parameters of the fundamental frequency is effective for both tasks. In other words, if an approach is effective for one task, it is also available for the other. Gaussian mixture model (GMM)-based VC, which is a successful traditional approach that has been proposed for speaker identity conversion [8, 9], has also been successfully applied to emotion conversion [10, 11]. Recent research trends have focused on artificial neural network models such as a restricted Boltzmann machine (RBM) [12], a variational autoencoder (VAE) [13, 14], and a generative adversarial net (GAN) [15, 16, 17, 18].

However, despite the several common points among the two tasks, there are currently no methods for simultaneous conversion of speaker identity and emotion. We feel that representing a speaker identity domain and an emotion domain explicitly in a single model would be beneficial because it leads to suppressing the influence of the other domain when the model represents one domain. With this assumption as our motivation, we propose a multiple-domain adaptive RBM (MDARBM) that enables simultaneous conversion of speaker identity and emotion by separating the two domains.

An MDARBM-based VC is an expansion of the adaptive restricted Boltzmann machine (ARBM)-based VC [12]. An ARBM is an energy-based model consisting of a visible layer and a hidden layer having undirected connections between visible-hidden units. By decomposing the weight matrix of the connections into a class-specific factor (adaptation matrix) and an independent factor (class-independent weight matrix), this model can encode the distributions of visible features from different classes into class-independent distributions of hidden features. Conversion is performed by switching the adaptation matrices from source class one to target class one when the model reconstructs visible features from the encoded distributions of hidden features. By decomposing the weight matrix in the same way as a mathematical formulation for an ARBM, we can further introduce adaptation matrices for multiple other domains. An MDARBM is a model in which the weight matrix is decomposed into two or more class-specific factors and a class-independent factor to make it adaptable to two or more domains.

An RBM (and its expansion models) consists of small network architectures compared to other neural networks such as VAEs or GANs; however, we assume that this is a critical advantage of using an expansion of an ARBM-based VC for developing a simultaneous conversion system. When developing such a system with VAEs or GANs, it might be necessary to implement additional generator or classifier modules for each added domain. As each of the modules usually consists of at least hundreds or thousands of parameters to be learned, training such a system might be unstable or fail, especially when the number of available training samples is limited. We argue that the expansion of ARBM is reasonable for the first attempt to develop a simultaneous conversion framework.

2. Conventional method

RBM-based [19, 20] probabilistic models are used for representing latent features that cannot be observed but certainly exist in the background. The RBM was originally introduced as an undirected graphical model that defines the distribution

http://dx.doi.org/10.21437/Interspeech.2020-2262

of binary-visible variables with binary-hidden variables, and was later extended to deal with real-valued data, known as the Gaussian-Bernoulli RBM (GBRBM) [20].

In an ARBM model, observed visible features and hidden 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 class identifier units $\boldsymbol{s} \in \{0,1\}^R, \sum_{r=1}^R s_r = 1$ (R is the number of classes in a domain). In an ARBM, the visible-hidden weights are adapted to a certain class in a domain using its adaptation matrix $\mathbf{A}_r$ controlled by $\boldsymbol{s}$. The class-specific visible-hidden connections $\mathbf{W}(\boldsymbol{s})$, visible biases $\boldsymbol{b}(\boldsymbol{s})$, and hidden biases $\boldsymbol{c}(\boldsymbol{s})$ are defined as

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

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

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

where $\bar{\mathbf{W}} \in \mathbb{R}^{I \times J}$, $\bar{\boldsymbol{b}} \in \mathbb{R}^I$, and $\bar{\boldsymbol{c}} \in \mathbb{R}^J$ are class-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 class-specific parameters of the rth class. $\boldsymbol{b}_r$ and $\boldsymbol{c}_r$ denote the class-specific bias of the rth class for the visible and hidden units, respectively. For convenience, we use the symbol $\mathcal{A} = \{\mathbf{A}_r\}_{r=1}^R$ to denote a collection of the adaptation matrices.

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

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

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

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

where $||\cdot||^2$ denotes the L2 norm. The fraction bar in Eq. (5) denotes the element-wise division. $\boldsymbol{\sigma}$ is the deviation parameter of the visible units. 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:

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

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

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 class's visible features $\boldsymbol{x}^{(t)}$ at frame t can be converted into those of the target class $\boldsymbol{y}^{(t)}$ via hidden features $\hat{\boldsymbol{h}}^{(t)}$ so as to maximize the probability $p(\boldsymbol{y}^{(t)}|\boldsymbol{x}^{(t)})$, as

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

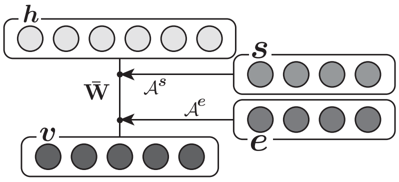

Figure 1: *Graphical representation of multiple-domain adaptive RBM.*

where

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

As Eq. (10) indicates, the (optimum) hidden features are approximated as the expectation values of $p(\boldsymbol{h}^{(t)}|\boldsymbol{x}^{(t)})$, which results in the sigmoidal outputs of affine-transformed visible features of the source class 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 class's visible features, the obtained hidden features $\hat{\boldsymbol{h}}$ represent class-independent information that is potentially phonological features when the acoustic features are input as visible features. 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 class of a certain domain (e.g., speaker or emotion).

3. Multiple-domain adaptive RBM

Our proposed multiple-domain adaptive RBM (MDARBM), depicted in Fig. 1, is a model that has multiple class identifier units. In this paper, the number of class identifier units $D (\geq 2)$ is set to two, and we use $\boldsymbol{s}$ for speaker-identity units and $\boldsymbol{e}$ for emotion units. The speaker-emotion-specific visible-hidden connections $\mathbf{W}(\boldsymbol{s}, \boldsymbol{e})$, visible biases $\boldsymbol{b}(\boldsymbol{s}, \boldsymbol{e})$, and hidden biases $\boldsymbol{c}(\boldsymbol{s}, \boldsymbol{e})$ are defined as

$$\mathbf{W}(\boldsymbol{s}, \boldsymbol{e}) = \sum_q \mathbf{A}_q^e e_q \sum_r \mathbf{A}_r^s s_r \bar{\mathbf{W}} \tag{11}$$

$$\begin{aligned} \boldsymbol{b}(\boldsymbol{s}, \boldsymbol{e}) &= \bar{\boldsymbol{b}} + \sum_r \boldsymbol{b}_r \boldsymbol{s}_r + \sum_q \boldsymbol{b}_q \boldsymbol{e}_q \\ &= \bar{\boldsymbol{b}} + \mathbf{B}^s \boldsymbol{s} + \mathbf{B}^e \boldsymbol{e} \end{aligned} \tag{12}$$

$$\begin{aligned} \boldsymbol{c}(\boldsymbol{s}, \boldsymbol{e}) &= \bar{\boldsymbol{c}} + \sum_r \boldsymbol{c}_r \boldsymbol{s}_r + \sum_q \boldsymbol{c}_q \boldsymbol{e}_q \\ &= \bar{\boldsymbol{c}} + \mathbf{C}^s \boldsymbol{s} + \mathbf{C}^e \boldsymbol{e}, \end{aligned} \tag{13}$$

where $\mathbf{A}_r^s$, $\mathbf{A}_q^e \in \mathbb{R}^{I \times I}$, $\boldsymbol{b}_r \in \mathbb{R}^I (\mathbf{B}^s = [\boldsymbol{b}_1^s, \boldsymbol{b}_2^s, \cdots, \boldsymbol{b}_R^s] \in \mathbb{R}^{I \times R}), \boldsymbol{b}_q \in \mathbb{R}^I (\mathbf{B}^e = [\boldsymbol{b}_1^e, \boldsymbol{b}_2^e, \cdots, \boldsymbol{b}_Q^e] \in \mathbb{R}^{I \times Q}), \boldsymbol{c}_r \in \mathbb{R}^J (\mathbf{C}^s = [\boldsymbol{c}_1^s, \boldsymbol{c}_2^s, \cdots, \boldsymbol{c}_R^s] \in \mathbb{R}^{J \times R})$, and $\boldsymbol{c}_q \in \mathbb{R}^J (\mathbf{C}^e = [\boldsymbol{c}_1^e, \boldsymbol{c}_2^e, \cdots, \boldsymbol{c}_Q^e] \in \mathbb{R}^{J \times E})$ are speaker and emotion-specific parameters (R and Q indicate the number of speakers and emotions, respectively). The decomposition order of the speech factors in Eq. (11) is derived from our assumption that an emotion would be a more global aspect of speech than a speaker identity.

Table 1: *Performance of ARBM-based VC in Exp. 1: Percentage distributions of listener's responses for speaker-identity converted speech.*

		Response			
		F1	M1	F2	M2
	F1	N/A			
	M1				
Target	F2	37.8	20.5	37.1	4.6
	M2	8.3	55.6	3.0	33.1

Table 2: *Performance of MDARBM-based VC in Exp. 1: Percentage distributions of listener's responses for speaker-identity converted speech.*

		Response			
		F1	M1	F2	M2
	F1	N/A			
	M1				
Target	F2	35.8	17.0	45.3	1.9
	M2	5.2	48.9	1.5	44.4

Given the speaker-identity and emotion units, the joint probability of visible and hidden units $p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s})$ is derived as

$$p(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s}, \boldsymbol{e}) = \frac{1}{Z} e^{-E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s}, \boldsymbol{e})} \tag{14}$$

$$\begin{aligned} E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s}, \boldsymbol{e}) = \| \frac{\boldsymbol{v} - \boldsymbol{b}(\boldsymbol{s}, \boldsymbol{e})}{2\boldsymbol{\sigma}} \|^2 \\ - \boldsymbol{c}(\boldsymbol{s}, \boldsymbol{e})^T \boldsymbol{h} - (\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2})^T \mathbf{W}(\boldsymbol{s}, \boldsymbol{e}) \boldsymbol{h} \end{aligned} \tag{15}$$

$$Z = \int_{\boldsymbol{v}} \sum_{\boldsymbol{h}, \boldsymbol{s}, \boldsymbol{e}} e^{-E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{s}, \boldsymbol{e})} d\boldsymbol{v}. \tag{16}$$

Therefore, conditional probabilities $p(\boldsymbol{h}|\boldsymbol{v}, \boldsymbol{s}, \boldsymbol{e})$ and $p(\boldsymbol{v}|\boldsymbol{h}, \boldsymbol{s}, \boldsymbol{e})$ are calculated as

$$p(h_j = 1|\boldsymbol{v}, \boldsymbol{s}, \boldsymbol{e}) = \mathcal{S}(c_j(\boldsymbol{s}, \boldsymbol{e}) + \mathbf{W}(\boldsymbol{s}, \boldsymbol{e})^T_{:j}(\frac{\boldsymbol{v}}{\boldsymbol{\sigma}^2})) \tag{17}$$

$$p(v_i = v|\boldsymbol{h}, \boldsymbol{s}, \boldsymbol{e}) = \mathcal{N}(v|b_i(\boldsymbol{s}, \boldsymbol{e}) + \mathbf{W}(\boldsymbol{s}, \boldsymbol{e})_{i:}\boldsymbol{h}, \sigma_i^2). \tag{18}$$

In the converting step, the same as with an ARBM, the source speaker and emotion visible features $\boldsymbol{x}^{(t)}$ at frame t can be converted into those of the target speaker and target emotion $\boldsymbol{y}^{(t)}$ via hidden features $\hat{\boldsymbol{h}}^{(t)}$ so as to maximize the probability $p(\boldsymbol{y}^{(t)}|\boldsymbol{x}^{(t)})$, as

$$\begin{aligned} \hat{\boldsymbol{y}}^{(t)} &\triangleq \operatorname*{argmax}_{\boldsymbol{y}^{(t)}} p(\boldsymbol{y}^{(t)}|\boldsymbol{x}^{(t)}, \boldsymbol{s}_x, \boldsymbol{s}_y, \boldsymbol{e}_s, \boldsymbol{e}_t) \\ &\simeq \operatorname*{argmax}_{\boldsymbol{y}^{(t)}} p(\boldsymbol{y}^{(t)}, \hat{\boldsymbol{h}}^{(t)}|\boldsymbol{x}^{(t)}, \boldsymbol{s}_x, \boldsymbol{s}_y, \boldsymbol{e}_s, \boldsymbol{e}_t) \\ &= \bar{\boldsymbol{b}} + \boldsymbol{b}^s_y + \boldsymbol{b}^e_t + \mathbf{A}^e_t \mathbf{A}^s_y \bar{\mathbf{W}} \hat{\boldsymbol{h}}^{(t)}, \end{aligned} \tag{19}$$

where $\boldsymbol{s}_x$, $\boldsymbol{s}_y$ and $\boldsymbol{e}_s$, $\boldsymbol{e}_t$ are speaker identity units and emotion units of source and target speakers/emotions, respectively, $\boldsymbol{b}^s_y$ and $\boldsymbol{b}^e_t$, $\mathbf{A}^e_t$ and $\mathbf{A}^s_y$ are target speaker- and emotion-specific visible biases and adaptation matrices for target emotion and speaker, respectively, and

$$\begin{aligned} \hat{\boldsymbol{h}}^{(t)} &\triangleq \operatorname*{argmax}_{\boldsymbol{h}^{(t)}} p(\boldsymbol{h}^{(t)}|\boldsymbol{x}^{(t)}, s_x, e_s) \\ &\simeq \mathcal{S}\left(\bar{\boldsymbol{c}} + \boldsymbol{c}^s_x + \boldsymbol{c}^e_s + \bar{\mathbf{W}}^\top \mathbf{A}^{s\top}_x \mathbf{A}^{e\top}_s \left(\frac{\boldsymbol{x}^{(t)}}{\boldsymbol{\sigma}^2}\right)\right). \end{aligned} \tag{20}$$

In Eq. (20), $\boldsymbol{c}^s_x$ and $\boldsymbol{c}^e_s$, $\mathbf{A}^e_s$ and $\mathbf{A}^s_x$ are source speaker- and emotion-specific hidden biases and adaptation matrices for source speaker and emotion, respectively.

4. Evaluation experiments

We conducted subjective experiments to evaluate the proposed model. Three aspects of performance were evaluated: speaker-identity conversion (Exp. 1), emotion conversion (Exp. 2), and simultaneous conversion (Exp. 3). An ARBM-based VC was utilized as a baseline method.

4.1. Experimental conditions

We trained three independent models: an ARBM for speaker-identity conversion, an ARBM for emotion conversion, and an MDARBM for simultaneous conversion. We used the Japanese Twitter-based Emotional Speech (JTES) [21] dataset, which consists of about 23 hours and 31 minutes of 50 spoken sentences representing different emotions acted out emotionally by 50 females and 50 males. There were four common emotion categories: *anger, joy, sadness,* and *neutral.* For each speaker, 40 sentences were used for training (0.1% of training data for validation) and ten sentences were used for testing.

We used 98-dimensional acoustic features consisting of 32-dimensional Mel-cepstral features and 33 time steps of F0 and power contours (1 step = 5 ms). Acoustic features were calculated every 5 ms using the WORLD analyzer [22]. The number of hidden units was 128 for each model. We trained the baseline and proposed models for 1,000 epochs using Momentum SGD [23] with a batch size of 40,000, $\eta = 0.001$, and $\alpha = 0.9$. The adaptation matrices for the neutral emotion were fixed identity matrices and were not updated while training.

In the synthesizing phase, the WORLD vocoder synthesized acoustic signals from the time series of original aperiodicity features, converted Mel-cepstral features, F0, and power. The time series of F0 and power were obtained from converted F0 and power contours using an overlap-add technique [24].

4.2. Performance of speaker-identity conversion (Exp. 1)

In the first experiment, we evaluated the performances of speaker-identity conversion. Two females (F1, F2) and two males (M1, M2) were randomly selected from the corpus. We set four conversion pairs: F1 to F2, F1 to M2, M1 to F2, and M1 to M2. The emotion of all the source speech was neutral. Eight listeners participated in the experiment. Before the experimental trial, listeners learned to identify the four speakers from the sample speech. Each speaker was associated with a distinct name: F1, F2, M1, and M2. Following the initial familiarization, listeners undertook a 128-item speaker identification test in which they identified the voices without feedback. The presented speech consisted of speech converted with the baseline method and the proposed method and original source and target speech. Emotion conversion was performed only in the proposed method (from neutral to neutral/joy/anger/sadness), but we didn't provide any instruction about speech emotion to the listeners.

Tables 1 and 2 show the results. The identification rates for target speakers were improved in the proposed method: from

Table 3: *Performance of ARBM-based VC in Exp. 2: Percentage distributions of listener's responses for emotion converted speech.*

		Response			
		Neutral	Joy	Anger	Sadness
Target	Neutral	39.7	10.3	15.5	34.5
	Joy	22.0	30.5	6.8	40.7
	Anger	22.8	21.1	7.0	49.1
	Sadness	31.5	14.8	14.8	38.9

Table 4: *Performance of MDARBM-based VC in Exp. 2: Percentage distributions of listener's responses for emotion converted speech.*

		Response			
		Neutral	Joy	Anger	Sadness
Target	Neutral	40.7	16.9	11.9	30.5
	Joy	33.9	33.9	12.5	19.6
	Anger	41.4	13.8	13.8	31.0
	Sadness	49.0	3.9	17.6	29.4

37.1% to 45.3% in the X to F2 conversion and from 33.1% to 44.4% in the X to M2 conversion. These results indicate that emotion conversion performed in the background does not disturb but rather enhances the main task (speaker-identity conversion in this case).

4.3. Performance of emotion conversion (Exp. 2)

In Exp. 2, we evaluated the performances of emotion conversion. The source and target speakers were the same as in Exp. 1. Converted speech of the proposed method was synthesized in the same manner as Exp. 1 and that of the baseline method was synthesized with the ARBM trained with emotion units. Seven listeners participated in the experiment. Listeners undertook a 128-item speech emotion identification test in which they identified emotions from neutral, joy, anger, and sadness.

Tables 3 and 4 show the results. Consistent with the results of Exp. 1, the performance of emotion conversion was improved in the proposed method except for the neutral-to-sadness conversion. These results also support the idea that one conversion task is enhanced by the other task when performed simultaneously.

4.4. Performance of simultaneous conversion (Exp. 3)

In Exp. 3, we evaluated the performances of speaker-identity and emotion conversion. The source and target speakers were the same as in Exp.1. Converted speech of the proposed method was synthesized in the same manner as Exp. 1. For the baseline method, we performed speaker-identity conversion using the ARBM trained with speaker-identity units followed by performing emotion conversion using another ARBM trained with emotion units.

Two types of XAB tests, speaker identity XAB and emotion XAB, were conducted. 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 listeners were asked to determine which one was closer to the reference speaker/emotion. The number of trials was 64 for each test. Nine and ten listeners participated in the speaker identity XAB test and the emotion XAB test, respectively.

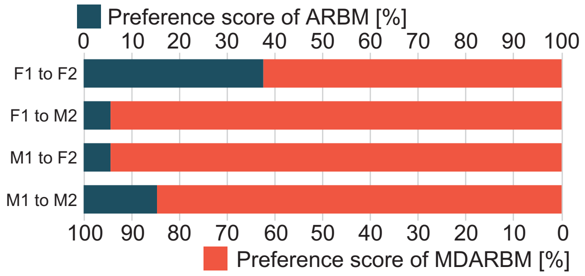

Figure 2: *Speaker identity XAB test results for each conversion pair in Exp. 3. Baseline vs. proposed.*

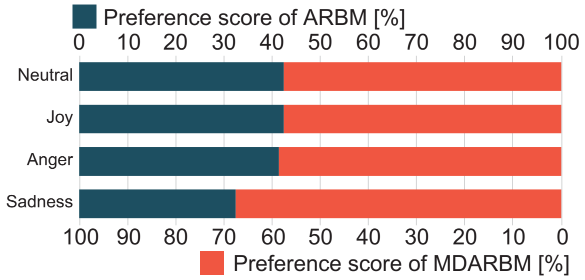

Figure 3: *Emotion XAB test results for each target emotion in Exp. 3. Baseline vs. proposed.*

Figures 2 and 3 show the results of the XAB tests. The proposed method outperformed the baseline method for both speaker-identity conversion and emotion conversion. These results strongly support the finding that the performance improvements in the proposed method observed in Exps. 1 and 2 were brought about by the simultaneous conversion.

5. Conclusion

In this paper, we have proposed a multiple-domain adaptive RBM that enables simultaneous conversion of multiple domains of speech by decomposing the visible-hidden connections of the RBM into two or more domain-specific factors and a domain-independent factor. To the best of our knowledge, this study is the first to develop a method for simultaneous conversion of speaker identity and emotion. The results of the two experiments showed that the performance of the one domain conversion task was enhanced by the other task. The results of an additional experiment confirmed that this enhancement was brought about by simultaneous conversion. We feel a simultaneous conversion framework is beneficial for learning the characteristics of a certain aspect of speech while reducing the effects from the other aspects. This will encourage the utilization of training speech samples consisting of diverse aspects such as speaker identity, speaking styles, and more.

6. Acknowledgements

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

7. References

[1] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[3] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan, "Paralinguistics in speech and language—State-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, pp. 4–39, 2013.

[4] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[5] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 1. IEEE, 1998, pp. 285–288.

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

[7] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "An evaluation of alaryngeal speech enhancement methods based on voice conversion techniques," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5136–5139.

[8] K. Nakamura, T. Toda, H. Saruwatari, K. Shikano *et al.*, "Enhancement of esophageal speech using statistical voice conversion," in *Proceedings: APSIPA ASC*. Asia-Pacific Signal and Information Processing Association, 2009 Annual …, 2009, pp. 805–808.

[9] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[10] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "Gmm-based voice conversion applied to emotional speech synthesis," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[11] B. Li, Z. Xiao, Y. Shen, Q. Zhou, and Z. Tao, "Emotional speech conversion based on spectrum-prosody dual transformation," in *2012 IEEE 11th International Conference on Signal Processing*, vol. 1. IEEE, 2012, pp. 531–535.

[12] 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.

[13] 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," in *Proc. APSIPA*. IEEE, 2016, pp. 1–6.

[14] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5274–5278.

[15] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5279–5283.

[16] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. EUSIPCO*. IEEE, 2018, pp. 2100–2104.

[17] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[18] 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.

[19] 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.

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

[21] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balanced emotional speech database," in *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2016, pp. 16–21.

[22] 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.

[23] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural networks*, vol. 12, no. 1, pp. 145–151, 1999.

[24] W. Verhelst, "Overlap-add methods for time-scaling of speech," *Speech Communication*, vol. 30, no. 4, pp. 207–221, 2000.