Neutral-to-emotional voice conversion with cross-wavelet transform F0 using generative adversarial networks

Luo, Zhaojie
Chen, Jinhui
Takiguchi, Tetsuya
Ariki, Yasuo

SIP (2019), vol. 8, e10, page 1 of 11 © The Authors, 2019.

doi:10.1017/ATSIP.2019.3

ORIGINAL PAPER

Neutral-to-emotional voice conversion with cross-wavelet transform Fo using generative adversarial networks

ZHAOJIE LUO[1], JINHUI CHEN[2], TETSUYA TAKIGUCHI[1] AND YASUO ARIKI[1]

In this paper, we propose a novel neutral-to-emotional voice conversion (VC) model that can effectively learn a mapping from neutral to emotional speech with limited emotional voice data. Although conventional VC techniques have achieved tremendous success in spectral conversion, the lack of representations in fundamental frequency (Fo), which explicitly represents prosody information, is still a major limiting factor for emotional VC. To overcome this limitation, in our proposed model, we outline the practical elements of the cross-wavelet transform (XWT) method, highlighting how such a method is applied in synthesizing diverse representations of Fo features in emotional VC. The idea is (1) to decompose Fo into different temporal level representations using continuous wavelet transform (CWT); (2) to use XWT to combine different CWT-Fo features to synthesize interaction XWT-Fo features; (3) and then use both the CWT-Fo and corresponding XWT-Fo features to train the emotional VC model. Moreover, to better measure similarities between the converted and real Fo features, we applied a VA-GAN training model, which combines a variational autoencoder (VAE) with a generative adversarial network (GAN). In the VA-GAN model, VAE learns the latent representations of high-dimensional features (CWT-Fo, XWT-Fo), while the discriminator of the GAN can use the learned feature representations as a basis for a VAE reconstruction objective.

Keywords: Continuous wavelet transform, Emotional voice conversion, Generative adversarial networks, Variational autoencoder, Fo features

Received 30 June 2018; Revised 5 February 2019

I. INTRODUCTION

Emotional voice conversion (VC) is a kind of VC technique for converting prosody in speech, which can represent different emotions while keeping the linguistic information unchanged. In a voice, the spectral and fundamental frequency (Fo) features can affect the acoustic and prosodic features, respectively. So far, spectral mapping mechanisms have achieved tremendous success in VC tasks [1–4], while, how to effectively generate prosody in the target voice remains a challenge. We know that prosody plays an important role in conveying various types of non-linguistic information that represent the mood of the speaker. Previous studies have shown that Fo is an important feature for prosody conversion that is affected by both short- and long-term dependencies, such as the sequence of segments, syllables, and words within an utterance [5]. However, it may be difficult to apply conventional deep learning-based VC to Fo conversion using simple representations of Fo, such as dynamic features (delta Fo).

In recent years, it has been shown that the continuous wavelet transform (CWT) method can effectively model Fo in different temporal scales and significantly improve speech synthesis performance [6]. For this reason, Suni *et al.* [7] applied CWT to intonation modeling in hidden Markov model (HMM) speech synthesis. Ming *et al.* [8] used CWT in Fo modeling within the non-negative matrix factorization (NMF) model for emotional VC. Our earlier work [9] systematically captured the Fo features of different temporal scales using adaptive scales CWT (AS-CWT), which transforms Fo features into high-dimensional CWT-Fo features containing more specifics. Conventional emotional VC tasks are focusing on converting one emotional voice to another fixed emotional voice. However, in the real world, the emotional voice is complex, with different prosody factors, and evolves over time. Thus, in order to better represent the interactions of components in emotional voices and generate a greater variety of representations of Fo features. In this study we want to go one step further and generate more emotional features using the cross-wavelet transform (XWT) [10]. The XWT method is a technique that characterizes the interaction between the

[1]Graduate School of System Informatics, Kobe University, 1-1 Rokkodai, Nada, Kobe 657-8501, Japan
[2]RIEB, Kobe University, 2-1 Rokkodai, Nada, Kobe 657-8501, Japan

Corresponding author:
Zhaojie Luo
Email: luozhaojie@me.cs.scitec.kobe-u.ac.jp

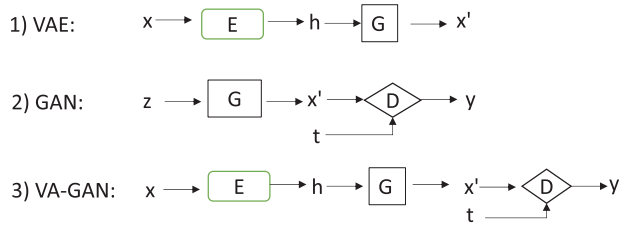

Fig. 1. Illustration of the structure of VAE [17], GAN [18] and the proposed VA–GAN. Here x and x' are input and generated features. z is the latent vector and *t* are target features. *E*, *G*, *D* are the encoder, generative, and discriminative networks, respectively. h is the latent representation processed by encoder network. y is a binary output which represents real/synthesized features.

wavelet transform of two individual time-series, which was first introduced by Hudgine *et al.* [11] and has been used in many areas such as economic analysis [12], geophysical time series [13] or electrocardiogram signals analysis [14]. Moreover, to improve the emotional VC effect with a limited amount of training data, we propose a new generative model, which will be described followed.

Many deep learning-based generative models exist including Restrictive Boltzmann Machine [15], Deep Belief Networks [16], Variational Autoencoder (VAE) [17] and Generative Adversarial Nets (GAN) [18]. Most of them have been applied in spectral conversion. Nakashika *et al.* [19] proposed a VC method using DBNs to achieve non-linear deep transformation. Hsu *et al.* [20] applied a convolutional VAE to model the generative process of natural speech. Kaneko *et al.* [21] used GAN in sequence-to-sequence voice conversion. The illustration of the structure of VAE [17] and GAN [18] are shown in Fig. 1 (2) and Fig. 1 (2), respectively.

Owing to the success of VAE and GAN in VC tasks, Hsu *et al.* [3] proposed the Variational Autoencoding Wasserstein Generative Adversarial Networks (VAW-GAN) for non-parallel VC tasks, which combined VAE with Wasserstein GAN [22]. In this study, different from the VAW-GAN [3], which is mainly used in non-parallel data voice conversion, the main task for emotional VC is how to maintain good model training using limited parallel emotional data and the simple representations of Fo features. Therefore, we propose an emotional VC framework that combines VAE with a conditional GAN [23], which is named VA-GAN. The effectiveness of GAN is due to the fact that an adversarial loss forces the generated data to be indistinguishable from real data. This is particularly powerful for generation tasks, but the training of the GANs is difficult and unstable. This leads to results in which the generated voice is unnatural and of bad quality. Another popular generative model, VAE, suffers from the problem of fuzzy sound, which is caused by injected noise and imperfect element-wise measures, such as the squared error. Thus, by combining the VAE and GAN models, the VAE can provide the efficient approximated posterior inference of the latent factors for improved GAN learning. Meanwhile, GAN can enhance VAE with an adversarial mechanism for leveraging generated samples.

As shown in Fig. 1 (3), our VA-GAN consists of three parts: (1) the encoder network *E*, which maps the data sample $\mathbf{x}$ to a latent representation $\mathbf{h}$, (2) the generative network G, which generates features $\mathbf{x}'$ from the latent representation $\mathbf{h}$, (3) and the discriminative network D, which distinguishes real/fake ($\mathbf{t}/\mathbf{x}'$) features. Here, we use the $\mathbf{t}$ and $\mathbf{x}'$ to represent the target and converted emotional voice. These three parts are seamlessly cascaded together, and the whole pipeline is trained end-to-end.

In summary, we show two major advantages of using our method. First, we employ the XWT to combine CWT-Fo features of different emotions. By doing so, we can obtain the XWT-Fo features and reconstruct them to various Fo features combined with different emotions, offering a huge advantage in emotional voice synthesis. Additionally, the XWT-Fo features can also increase the representations of Fo features which can improve the emotional VC. Second, we designed VA-GAN based on VAE and GAN architectures to process the emotional VC tasks, which can effectively avoid over-smoothing as well as capture a highly non-linear network structure, by simultaneously integrating XWT-Fo features into network representation learning.

In the remaining part of this paper, feature extraction and processing are introduced in Section II. Then, we describe our proposed training model (VA-GAN) in Section III. Section IV gives the detailed process stages of our experimental evaluations, and Section V presents our conclusions.

II. FEATURE EXTRACTION AND PROCESSING

It is well known that prosody is influenced both at a supra-segmental level, by long-term dependencies, and at a segmental-level, by short-term dependencies. And, as has been proven in our recent studies [9, 24], the CWT can effectively model Fo in different temporal scales and significantly improve the system performance. So, in the proposed method, we adopt CWT to decompose the one-dimensional Fo features into high-dimensional CWT-Fo features, and then, use the XWT to combine the CWT-Fo features of different emotions. In the remaining part of this section, we will present the basic CWT and XWT theories, and illustrate how to use them in carrying out emotional VC.

A) Continuous wavelet transform

In what follows, $L^2(\mathbb{R})$ denotes the set of square-integrable functions, which satisfy $\int_{-\infty}^{+\infty} |x(t)|^2\, dt < \infty$. Given a time series $f_o(t) \in L^2(\mathbb{R})$, its continuous wavelet is, in respect to the mother wavelet $\psi \in L^2(\mathbb{R})$, defined as

$$\psi_{s,\tau} = s^{-1/2}\psi_o\left(\frac{t-\tau}{s}\right) dt, s, \tau \in \mathbb{R}, s \neq 0, \qquad (1)$$

where s is the scaling factor (which controls the width of the wavelet), and τ is the translating factor (which decides the location of the wavelet). $(s^{-1/2})$ is used for normalization, which can ensure unit variance of the wavelet and $\left\|\psi_{s,\tau}\right\|^2 = 1$. Thus, the CWT of an input signal $f_o(t)$ can be

https://doi.org/10.1017/ATSIP.2019.3

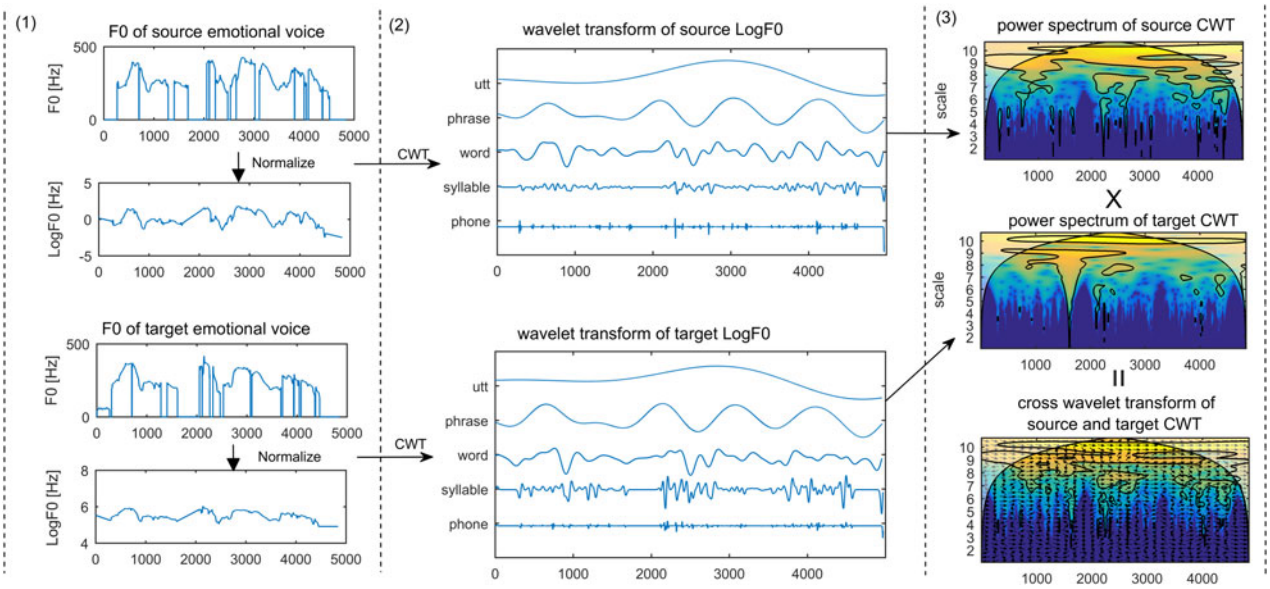

Fig. 2. (1) Linear interpolation and log-normalized processing for source and target emotional voice. (2) The CWT for normalized LogFo contours, the top pan and bottom pan show five examples of the different level's decomposed CWT-Fo features of the source and target Fo, respectively. (3) First and second pans show the continuous wavelet spectrums of the source and target CWT-Fo features and the bottom pan represents the cross-wavelet spectrums of the two CWT-Fo features. The relative phase relationship is shown as arrows in the cross-wavelet spectrum (with in-phase pointing right (→), anti-phase pointing left (←), and source emotional features leading target emotional features by 90° pointing straight down.

written as

$$W_{f_o;\psi}(s,\tau) = \langle f_o(t), \psi_{s,\tau}(t)\rangle = \int_{-\infty}^{\infty} f_o(t)\,\psi_{s,\tau}(t)dt$$
$$= s^{-1/2}\int_{-\infty}^{\infty} f_o(t)\,\psi_o\left(\frac{t-\tau}{s}\right)dt \quad (2)$$

$$\psi_o(t) = \frac{2}{\sqrt{3}}\pi^{-1/4}\left(1-t^2\right)e^{-t^2/2}, \quad (3)$$

where $f_o(t)$ is the input signal and mother wavelet $\psi(t)$ used in our model is the Mexican hat mother wavelet [10]. The wavelet transform can give us information simultaneously on time-frequency space by mapping the original time series into the function of s and τ. Because both s and τ are real values and vary continuously, $W_{f_o;\psi}(s,\tau)$ is named a CWT. To be a mother wavelet of the CWT, $\psi(t)$ must fulfill admissibility conditions [25], which can be written as follows:

$$0 < C_\psi = \int_{-\infty}^{+\infty}\frac{|\hat{\psi}(f)|^2}{|f|} < +\infty, \quad (4)$$

where $\hat{\psi}(f)$ is the Fourier transform of the mother wavelet $\psi(t)$ and f is the Fourier frequency. It is clear that C_ψ is independent of f and determined only by the wavelet $\psi(t)$. This means that C_ψ is a constant for each given mother wavelet function. Thus, we can reconstruct $f_o(t)$ using its CWT, $W_{f_o;\psi}(s,\tau)$, by inverse function as follows:

$$f_o(t) = \frac{1}{C_\psi}\int_o^{+\infty}\left[\int_{-\infty}^{+\infty} W_{f_o;\psi}(s,\tau)\,\psi_{s,\tau}(t)d\tau\right]\frac{ds}{s^2}, \quad s>o \quad (5)$$

In this way, we can decompose Fo features $f_o(t)$ to the CWT-Fo features $W_{f_o;\psi}(s,\tau)$ and reconstruct them back to Fo features. Therefore, we have reason to believe that Fo and CWT-Fo are two different representations of the same mathematical entity. Fig. 2 (2) shows the examples of CWT features decomposed from the Fo features of source and target emotional voices. In this figure, we show the parts of the scales that can represent the sentence, phrase, word, syllable, and phone levels, respectively. Moreover, the energy of the examined time series is preserved by its CWT in the sense that

$$\|f_o(t)\|^2 = \frac{1}{C_\psi}\int_o^{+\infty}\left[\int_{-\infty}^{+\infty}|W_{f_o;\psi}(s,\tau)|^2\psi_{s,\tau}(t)d\tau\right]\frac{ds}{s^2}, \quad s>o \quad (6)$$

where $\|f_o(t)\|^2$ is defined as the energy of $f_o(t)$, and $|W_{f_o;\psi}(s,\tau)|^2$ is defined as a wavelet power spectrum that can interpret the degree of local variance of $f_o(t)$ scale by scale. The examples of wavelet power spectrum of the source and target CWT-Fo features are represented by the first pan and the second pan of Fig. 2 (3).

B) Cross wavelet transform

Given two time series, $f_{o_x}(t)$ and $f_{o_y}(t)$, with their CWT features, W_x, and W_y, XWT is defined as $W_{xy;\psi}(s,\tau) = W_{x;\psi}(s,\tau)W_{y;\psi}*(s,\tau)$, where $*$ denotes complex conjugation. And their cross-wavelet power spectrum is accordingly written as

$$|W_{xy;\psi}(s,\tau)|^2 = |W_{x;\psi}(s,\tau)|^2|W_{y;\psi}*(s,\tau)|^2. \quad (7)$$

https://doi.org/10.1017/ATSIP.2019.3

An example of the cross-wavelet power spectrum is shown in the last pan of Fig. 2 (3). The XWT of a two-time series depicts the local covariance between them at each time and frequency and shows the area in the time-frequency space where the two signals exhibit high common power. Thus, we can do the cross-transform for different emotional voices to synthesize the new XWT-Fo features, and then reconstruct them to new synthesized Fo features. The new synthesized Fo features can be used in processing a new emotional voice or increasing the training data for emotional VC training.

C) Applying CWT and XWT in Fo features extraction

The STRAIGHT [26] is frequently used to extract features from a speech signal. Generally, the smoothing spectrum and instantaneous-frequency-based Fo are derived as excitation features for every 1 ms from the STRAIGHT. As shown in Fig. 2, there are three steps for processing the XWT-Fo features.

1) To explore the perceptually-relevant information, Fo contour is transformed from the linear to logarithmic semitone scale, which is referred to as logFo. As shown in the first pan and second pan of Fig. 2 (2), the Fo features extracted by STRAIGHT from the source and target emotional voice are discrete. As the wavelet method is sensitive to the gaps in the Fo contours, we must fill in the unvoiced parts in the logFo via linear interpolation to reduce discontinuities in voice boundaries. Finally, we normalize the interpolated logFo contour to zero mean and unit variance. Examples of interpolated LogFo contours of source and target emotional voice are depicted in the second and last pan of Fig. 2 (2), respectively.
2) Then, we adopt CWT to decompose the continuous logFo into 30 discrete scales, each separated by one-third of an octave. Fo is thus approximately represented by 30 separate components given by [24]

$$W_i(f_o)(t) = W_i(f_o)(2^{(i/3)+1}\tau_o, t)\,((i/3) + 2.5)^{-5/2}, \quad (8)$$

where $i = 1,\ldots,30$ and $\tau_o = 1$ ms. In Fig. 2 (2), we just show five examples of the decomposed CWT-Fo features which represent the utterance, phrase, word, syllable and phone levels, respectively.
3) Finally, we do the XWT for CWT-Fo features of the two emotional voices. First, we process the power spectrums of the source and target CWT-Fo features, and then use the cross-wavelet function to combine the two CWT-Fo features to cross-wavelet Fo features (XWT-Fo). Fig. 2 (3) shows the power spectrums of CWT-Fo features and the XWT-Fo feature. The XWT-Fo features can be used for generating the combinative emotional voice. Moreover, they can also be used for increasing the training data of the emotional VC tasks.

III. TRAINING MODEL: VA-GAN

A) Background: VAE and GAN

1) Variational autoencoder

VAE defines a probabilistic generative process between observation x and latent variable h as follows:

$$\mathbf{z} \sim Enc(\mathbf{x}) = q_{\boldsymbol{\phi}}(\mathbf{h}|\mathbf{x}), \tilde{\mathbf{x}} \sim Dec(\mathbf{h}) = p_{\boldsymbol{\theta}}(\mathbf{h}|\mathbf{x}), \quad (9)$$

where (Enc) represents encode networks that encode a data sample $\mathbf{x}$ to a latent representation $\mathbf{h}$ and decode networks (Dec) decode the latent representation back to data space. In the VAE, the recognition model $q_{\boldsymbol{\phi}}(\mathbf{h}|\mathbf{x})$ approximates the true posterior $p_{\boldsymbol{\theta}}(\mathbf{h}|\mathbf{x})$. The VAE regularizes the encoder by imposing a prior over the latent distribution $p_{\boldsymbol{\theta}}(\mathbf{h})$, which is assumed to be a centered isotropic multivariate Gaussian $p_{\boldsymbol{\theta}}(\mathbf{h}) \sim N(\mathbf{h}; \mathbf{o}, \mathbf{I})$. The VAE loss $L_{\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{x}}$ is minus the sum of the expected log-likelihood L_{like} (the reconstruction error) and a prior regularization term L_{prior} represented as:

$$L_{\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}} = -E_{q_{\boldsymbol{\phi}}(\mathbf{h}|\mathbf{x})}[\log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{h})p_{\boldsymbol{\theta}}(h)}{q_{\boldsymbol{\phi}}(\mathbf{h}|\mathbf{x})}] = L_{like} + L_{prior}, \quad (10)$$

$$L_{like} = -E_{q_{\boldsymbol{\phi}}(\mathbf{h}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{h})], \quad (11)$$

$$L_{prior} = KL(q_{\boldsymbol{\phi}}(\mathbf{h}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{h})), \quad (12)$$

where KL is the Kullback-Leibler divergence. We use the KL loss to reduce the gap between the prior $P(\mathbf{h})$ and the proposal distributions. The loss of KL is only related to the encoder network Enc. It represents whether or not the distribution of the latent vector is under expectation. Here, we want to optimize $L_{\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{x}}$ in respect to $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

In the VC area, VAE is used to learn latent representations of speech segments to model the conversion process. Refer to applying the VAE in the non-parallel VC tasks proposed by Hsu *et al.* [20]. We can also apply them in the emotional VC with parallel data. If we want to estimate the latent representations for each emotional feature, we can estimate latent attributes by taking the mean latent representations as shown in Fig. 3 (a). Supposing we want to convert the prosody attribute of a speech segment ${\mathbf{x}_\mathbf{A}}^i$, from being a source emotional voice e_A to being a target emotional voice e_B. Given the means of latent attribute representations $\boldsymbol{\mu}_{e_A}$ and $\boldsymbol{\mu}_{e_B}$ for emotional e_A and e_B, respectively, we can transform the latent attribute computed as $\mathbf{v}_{e_A \to e_B} = \boldsymbol{\mu}_{e_B} - \boldsymbol{\mu}_{e_A}$. We can then modify the speech x_A^i as follows:

$${\mathbf{h}_\mathbf{A}}^i \sim q_{\boldsymbol{\phi}}(\mathbf{h}|{\mathbf{x}_\mathbf{A}}^{(i)}), \quad (13)$$

$${\mathbf{h}_\mathbf{B}}^{(i)} = {\mathbf{h}_\mathbf{A}}^{(i)} + \mathbf{v}_{e_A \to e_B}, \quad (14)$$

$${\mathbf{x}_\mathbf{B}}^{(i)} \sim p_{\boldsymbol{\theta}}(\mathbf{x}|{\mathbf{h}_\mathbf{B}}^{(i)}), \quad (15)$$

Figure 3 (b) shows an illustration of the conversion phase.

2) Generative adversarial networks

GAN has obtained impressive results for image generation [27, 28], image editing [29], and representation learning [30]. The key to the success of the GAN is learning

https://doi.org/10.1017/ATSIP.2019.3

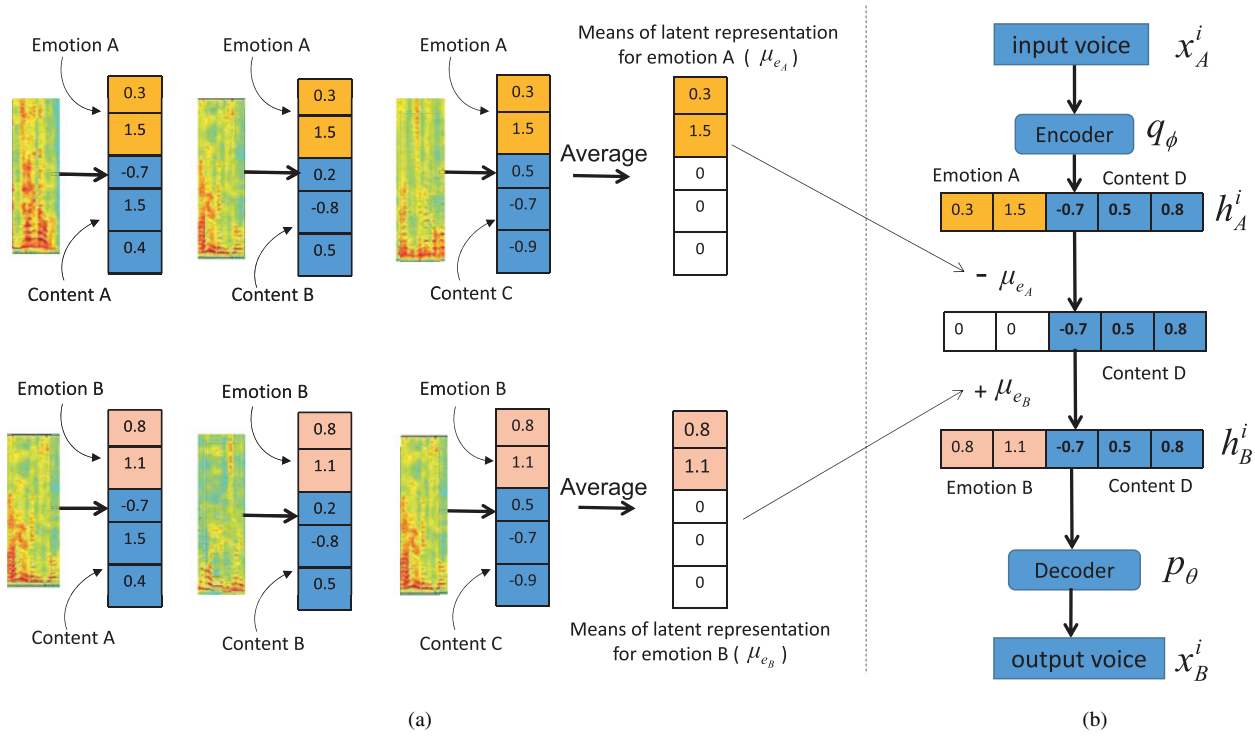

Fig. 3. (left) Examples of processing the latent representations for emotion A and emotion B using Variational Autoencoder (VAE). (right) Examples of modifying the emotional voice from emotion A to emotion B. $\boldsymbol{\mu}_{e_A}$ and $\boldsymbol{\mu}_{e_B}$ represent the means of latent attribute representations for emotional e_A and e_B, respectively. $q_{bm\phi}$ and $p_{bm\theta}$ mean the encode and decode functions, respectively. $\mathbf{x_A}^i$ and $\mathbf{x_B}^i$ represent the speech segment of emotion A and emotion B.

a generator distribution $P_G(\mathbf{x})$ that matches the true data distribution. It consists of two networks: a generator, G, which transforms noise variables $\mathbf{z} \sim P_{Noise}(\mathbf{z})$ to data space $\mathbf{x} = G(\mathbf{z})$ and a discriminator D. This discriminator assigns probability $p = D(\mathbf{x})$ when $\mathbf{x}$ is a sample from the $P_{Data}(\mathbf{x})$ and assigns probability $1 - p$ when $\mathbf{x}$ is a sample from the $P_G(\mathbf{x})$. In a GAN, D and G play the following two-player minimax game with the value function $V(G, D)$:

$$\min_G \max_D V(D,G) = E_{\mathbf{x}\sim p_{data}}[\log D(\mathbf{x})] + E_{z\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]. \quad (16)$$

This enables the discriminator, D, to find the binary classifier that provides the best possible discrimination between true and generated data and simultaneously enables the generator, G, to fit $P_{Data}(\mathbf{x})$. Both G and D can be trained using back-propagation.

The effectiveness of GAN is due to the fact that an adversarial loss forces the generated data to be indistinguishable from real data. This is particularly powerful for image generation tasks, and, owing to GAN's ability to generate a high-fidelity image which can mitigate the over-smoothing problem caused in the low-level data space when converting the speech features. GAN models have also begun to be applied in speech synthesis [31] and VC [21].

Figure 4 gives an overview of the training procedure when using the GAN model for VC tasks. In conventional studies [21, 31], in the VC tasks, they used not only a sample from the "G" but also a sample from the "C" as the discriminator input. Here, "C" represents the conversion

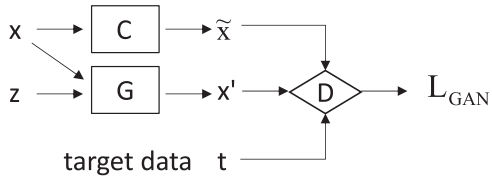

Fig. 4. Illustration of calculating the loss of GAN in voice conversion.

function. This is because the samples converted from the source voice are more similar to the target voice than the samples generated by noise. Thus, using the converted samples can improve the updating of D. The value function used in VC is rewritten as

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}}[\log D(\mathbf{x})] + E_{z\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{x},\mathbf{z})))] + E_{z\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(C(\mathbf{x})))]. \quad (17)$$

B) VA-GAN training model

As described above, there is a clear and recognized way to evaluate the quality of the VAE model. However, due to the injected noise and imperfect reconstruction, the generated samples are much more blurred than those coming from GAN. GAN models tend to be much more finicky to train than VAE and, therefore, can obtain a high-fidelity image. But, in practice, the GAN discriminator D can distribute the "real" and "fake" images easily, especially at the early stage of the training process. This will cause the problem of an

https://doi.org/10.1017/ATSIP.2019.3

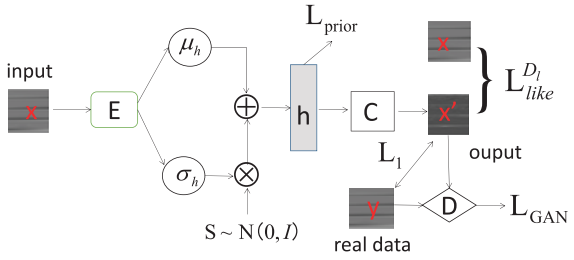

Fig. 5. Illustration of calculating the loss of VA-GAN.

unstable gradient of *G* when training GAN [22, 32]. To solve the problems associated with VAE and GAN models, we combine the VAE and GAN models to form one advanced model, which we named VA-GAN. As shown in Fig. 5, our model contains an encoder (*Enc*), conversion function ($C : \mathbf{x} \rightarrow \mathbf{y}$) and a discriminator (*D*). To resolve the instability of training GAN, we extract the representative features from a pre-trained *Enc*. Also, we can obtain better results when using the latent representation (*h*) from the encoder *Enc*. When dealing with the training of emotional VC, every two sets of labeled and paired feature matrices were sampled from domains source emotional voice *X* and target emotional voice *Y*, respectively. For the conversion function $C : x \rightarrow y$ and its discriminator *D* with pre-trained *Enc*, we express the objective of adversarial loss as:

$$L_{GAN}(C, D, X, Y) = E_{\mathbf{y}\sim P_{data}(\mathbf{y})}[\log D(\mathbf{y}, \mathbf{x})] + E_{\mathbf{x}\sim P_{data}(\mathbf{x})}[\log(1 - D(C(Enc(\mathbf{x}))))]. \quad (18)$$

The goal of emotional VC is to learn a converted emotional voice distribution $P_C(\mathbf{x})$ that matches the target emotional voice distribution $P_{data}(\mathbf{y})$. Equation (18) enables *D* to find the binary classifier that provides the best possible discrimination between a true and a converted voice and simultaneously enables the function "*C*" to fit the $P_{data}(y)$. We also mix the GAN objective with a traditional L1 distance loss as:

$$L_1(C) = E_{\mathbf{x},\mathbf{y},\mathbf{z}}[\|y - C(\mathbf{x}, \mathbf{z})\|_1]. \quad (19)$$

Our final objective is maximized and minimized with respect to "*D*" and "*C*", respectively.

$$G^* = arg \max_D \min_C L_{GAN}(C, D, X, Y) + \lambda L_1(C). \quad (20)$$

In the VAE for the emotional VC, emotion-independent encoder (Enc) infers a latent content $z \sim Enc(\mathbf{x})$, and then the emotion-dependent decoder (Dec) mixes $\mathbf{z}$ with an emotion-specific variable $\mathbf{y}$ to reconstruct the input as a variational approximation: $\tilde{x} \approx Dec(z, y)$. The aim of a VAE is to learn a reduced representation of the given data. Consequently, feature spaces learned by the VAE are powerful representations for reconstructing the $P_{data}(\mathbf{y})$ distribution. However, VAE models tend to produce unrealistic, blurry samples [33], and Zhao *et al.* [34] provide a formal explanation for this problem. To solve the blurriness issue, similar to some VAE and GAN combined models [35, 36], we replaced the reconstruction error term from equation (11) with a reconstruction error expressed in the discriminator *D*. To achieve this, we let $D_l(\mathbf{x})$ denote the hidden representation of the *l*th layer of the discriminator, and we introduce a Gaussian observation model for $D_l(\mathbf{x})$ with mean $D_l(\tilde{\mathbf{x}})$ and identity covariance:

$$p(D_l(\mathbf{x})|\mathbf{z}) = N(D_l(\mathbf{x})|D_l(\tilde{\mathbf{x}}), \mathbf{I}), \quad (21)$$

where $\tilde{\mathbf{x}} \sim Dec(\mathbf{h})$ in equation (9) is now the sample from the conversion function "*C*" of "*x*". We can now replace the VAE error of equation (11) with

$$L^{D_l}_{like} = -E_{q(\mathbf{z}|\mathbf{x})}[\log p(D_l(\mathbf{x})|\mathbf{z})]. \quad (22)$$

The goal of our approach is to minimize the following loss function:

$$L = L_{GAN} + L^{D_l}_{like} + L_{prior} + \lambda L_1(C), \quad (23)$$

where λ controls the relative importance of the distance loss function. The details of how to train VA-GAN are indicated in Algorithm 1.

Algorithm 1 Training procedure of VA-GAN

Require: 2D-features (30 × 30) datasets processed from source emotional voice *X* and target neutral voice *Y*, encoder *Enc*, conversion function *C* and discriminator *D* with enconder parameter θ_E, conversion parameter θ_C and θ_D, respectively, batch size *m*, and the epcohs *n*.

1: Initialize the parameters θ_E, θ_C and θ_D, randomly.
2: **repeat**
3: **for**$(i = 1; i < n + 1; i = i + 1)$ **do**
4: sample images $x_k \subseteq X, k \in \{1, \ldots, m\}$
5: latent representations $z_k \leftarrow Enc(x_k)$
6: update θ_E to minimize L_{prior}
7: update θ_E, θ_C to minimize $\frac{1}{m}\sum_{k=1}^{m} L^{D_l}_{like}(y_k, z_k)$
8: update θ_E, θ_C to minimize $\frac{1}{m}\sum_{k=1}^{m} L_1(x_k, z(x_k))$
9: update $\theta_E, \theta_C, \theta_D$ to minimize $\frac{1}{m}\sum_{k=1}^{m} L_{GAN}(y_k, z(x_k))$
10: **until** convergence

Throughout the training process, conversion functions *C* is optimized to minimize the VAE loss and to learn the converted emotional voice, which cannot be distinguished from the target emotional voice by corresponding discriminators *D*.

IV. EXPERIMENTS

In our experiments, we used a database of emotional Japanese speech constructed in a previous study [37]. In the database, there is a total of 50 different kinds of content sentences, and each sentence was read in four different emotions (angry, happy, sad, and neutral) by the same speaker. Thus, there 50 × 4 × *N* waveforms in the database, where N is the number of speakers, the waveforms were sampled at 16 kHz.

https://doi.org/10.1017/ATSIP.2019.3

In the emotional VC experiment, we classified the three datasets into the following voice types: neutral to angry voices (N2A), neutral to sad voices (N2S), and neutral to happy voices (N2H). For each data set, 40 sentences were chosen as basic training data and 10 sentences were chosen for the VC evaluation.

To do the evaluation of prosody conversion using our proposed **VA-GAN** method, we compared the results with several state-of-the-art methods. **Logarithm Gaussian (LG)** normalized transformation is often used for Fo features conversion in deep learning VC tasks [38, 39]. Our previous work [9] dealt with a **NNs** model that used the pre-trained NNs to convert the CWT-Fo features. We also compared **VA-GAN** with the **GAN** and **VAE**.

Before applying the XWT-Fo features to the emotional VC, we need to calculate the correlates of different emotional voices. In other words, we do the cross-matching of each two emotional voices for the same content by the same speaker, which will result in six datasets (N, A), (N, S), (N, H), (A, H), (A, S), (H, S). For each dataset, we do the XWT to synthesize the XWT-Fo features, and then reconstruct them to the Fo features that can be used for generating the complex emotional voice. Finally, we carry out a subjective experiment to decide the synthesized emotional voices that belong to which kind of emotion (neutral, angry, happy, sad, and unknown). Here, due to some generated emotional voices may sound strange, we added the unknown option.

To evaluate the effectiveness of XWT-Fo features used as the increased training data for emotional VC, we carry out the experiments by adding the XWT-Fo features, which are selected by the subjective experiment, to the corresponding CWT-Fo features to increase the training data.

Moreover, to evaluate the relationship between different data size and the conversion methods described above, we conducted our experiment using a larger open database [40], which contains 100 sentences for each emotional voice. Among them, 80 sentences were chosen for training data and the existing 20 sentences were chosen for the testing data. In these experiments, the spectral features were converted using the same models as their CWT-Fo features training model (For the NNs model, we used the DBNs model to train spectral features, which has been proposed in [9]).

A) Selecting cross wavelet features

To evaluate the synthesized emotional voice with the XWT-Fo features, we devised a subjective evaluation framework as shown in Fig. 6. We first perform CWT on the neutral, angry, happy, and sad voice to decompose their Fo contours into CWT-Fo features. Then do the XWT for each pair of emotional CWT-Fo features, to process the XWT-Fo features. Then we reconstruct the XWT-Fo features to Fo features to synthesize the complex emotional voice. Then we ask 10 subjects to listen and choose the emotion that is close to the synthesized speeches.

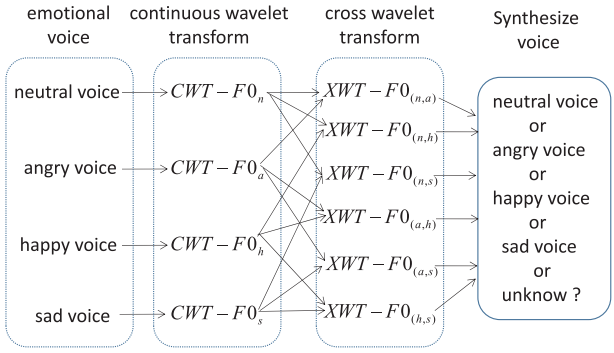

Fig. 6. The listening experiment setup for evaluating emotional voice generated by different XWT-Fo features. The small letters in the lower right of CWT-Fo and XWT-Fo represent the emotion. For instance, (n) and (a) represent the neutral and angry. (n,a) represents the XWT-Fo features generated by CWT-Fo features of angry and neutral.

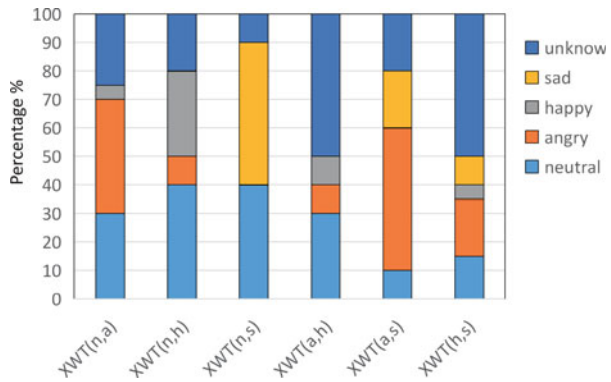

Fig. 7. Similarity to different emotional voices (neutral, angry, happy, and sad) of synthesized emotional voices with different XWT-Fo features.

Figure 7 shows the similarity to different emotions of the synthesized emotional voices using the different XWT-Fo features. For example, XWT (n, a) represents the synthesized emotional voice cross transformed with the neutral and angry CWT-fo features. As the figure shows, the emotional voice generated from the neutral and sad voices XWT (n, s) got the most votes for the similarity to a sad voice. The voting results of XWT (n, a) and XWT (a, s) indicate that the angry voice can be generated by angry and sad voice or angry and neutral voice. However, not all the generated emotional voices are good. For instance, the happy voice can be generated by XWT (n, h), but half of the subjects voted for the XWT (n, h) generated voice over the neutral voice. The XWT (h, s) and XWT (a, h) got the most votes for unknown, which means there was a strange prosody of the generated emotional voice.

In summary, depending on the emotional classification results of the emotional voice generated by the XWT-Fo (n, s), XWT-Fo (n, a), XWT-Fo (a, s) and XWT-Fo (n, h) features, we can use XWT-Fo (n, s) features as the CWT-Fo features of a sad voice, XWT-Fo (n, a) and XWT-Fo (a, s) as an angry voice, and the XWT-Fo (n, h) as a happy voice. Then, we can use both CWT-Fo features and their corresponding XWT-Fo features as the input training data, and the output is the CWT-Fo features.

https://doi.org/10.1017/ATSIP.2019.3

Table 1. Fo-RMSE results for different emotions. N2A, N2S and N2H represent the datasets from neutral to angry, sad and happy voice, respectively. (40) and (80) denote the number of training examples. "+" means using both the CWT-Fo features and selected XWT-Fo features for training model. U/V-ER represents the unvoiced/voiced error rate of the proposed method.

	Source	LG	NN	VAE	GAN	VA-GAN	NN+	VAE+	GAN+	VA-GAN+	U/V-ER (%)
N2A (40)	58.5	48.3	38.8	28.3	30.2	21.2	43.8	35.2	27.4	20.3	8.1
N2S (40)	59.9	54.7	40.3	25.5	25.3	22.5	38.2	23.1	24.3	20.8	8.6
N2H (40)	83.4	65.3	46.5	36.8	34.7	29.1	44.5	34.1	33.3	26.4	8.9
N2A (80)	58.5	47.5	37.2	27.5	28.4	20.4	42.4	33.3	25.2	18.9	8.5
N2S (80)	59.9	54.4	39.3	23.5	24.1	21.1	37.7	21.2	22.3	19.4	9.1
N2H (80)	83.4	63.2	45.2	32.1	33.5	28.1	43.4	32.1	32.7	24.8	9.4

Table 2. Details of network architectures of *Enc*, *C* and *D*

Enc (Input: 30 × 30 2D features, Output: latent representations *h*)
2×2 15 conv.↓, ReLU
2×2 30 conv.↓, BNorm, ReLU
120 fully connected, BNorm, ReLU
C (Input: latent representations processed by source features *X*, Output: 30 × 30 converted matrices)
120 fully-connected, BNorm, ReLU
2×2 15 conv.↓, ReLU
residual blocks $\begin{bmatrix} 2\times 2 & 15 & conv. & ReLU \\ 2\times 2 & 15 & conv. & ReLU \end{bmatrix} \times 2$
2×2 15 conv.↑, ReLU
2×2 30 conv.↑, BNorm, ReLU
D (Input: 30 × 30 2D features, Output: 1 Probability)
2×2 15 conv.↓, ReLU
2×2 30 conv.↓, BNorm, ReLU
120 fully connected, BNorm, ReLU
1 fully connected, sigmoid

B) Training procedure

Table 2 details the network architectures of the encoder (*Enc*), the converter (*C*) and discriminator (*D*). The symbols ↓ and ↑ indicate down sampling and up sampling, respectively. To upscale and downscale, we respectively used convolutions and backward convolutions with stride 2.

In the converter (*C*), similar to the generator used in Johnson *et al.* [41], we use batch normalization (BNorm) [42] and all convolutional layers are followed by ReLU nonlinearities [43] with the exception of the output layer. The *C* contains one stride-2 convolution to downsample the input, followed by two residual blocks [44], and two fractionally-strided upsampling convolutional layers with stride 1/2.

For discriminator (*D*), we refer to a convolutional PatchGAN classifier [45] but use a smaller patch size. The patch size at which the discriminator operates is fixed at 20 × 20, and two stride-2 convolutions and a fully connected layer make up the (*D*). After the last layer, we apply a sigmoid activation function to obtain a probability for sample classification.

When training the (*Enc*) (*C*) and (*D*), we use the Adam optimizer [46] with a mini-batch of size 8, which represents 8 input matrices (30 × 30), including 8 × 30 frames. The learning rate was set to 0.0002 for the *Enc* and *C*, and 0.0001 for the *D*, respectively. The momentum term was set to 0.5.

To clarify the characteristics of our framework with VA-GAN, as described above, we implemented three kinds of generative networks for comparison. The detailed units of the NNs are set to [32, 64, 64, 32]. The compared VAE model consists of the *E* and *C*, while the compared GAN model consists of the *C* and *D*, which are used in VA-GAN.

C) Objective experiment

To evaluate Fo conversion, we used the root-mean-square error (RMSE) which is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left(Fo_i^t\right) - \left(Fo_i^c\right)\right)^2}, \qquad (24)$$

where Fo_i^t and Fo_i^c denote the target and the converted interpolated Fo features, respectively. A lower Fo-RMSE value indicates a smaller distortion or predicting error. Unlike the RMSE evaluation function used in one previous study [8], which evaluated the Fo conversion by calculating logarithmic scaled Fo, we used the target interpolated Fo and converted interpolated Fo of a complete sentence including voiced and unvoiced (0 value), to calculate the RMSE values. Due to the alignment preprocessing in the parallel data, we used the unvoiced part of the source Fo in the converted Fo. Considering that Fo-RMSE evaluates complete sentences that contain both voiced and unvoiced Fo features instead of the voiced-logarithmic scaled Fo, the RMSE values are expected to be high. For emotional voices, the unvoiced features also include some emotional information, such as the pause in a sentence which can sometimes express nervousness or some other negative emotion. Therefore, we choose the Fo of complete sentences for evaluation as opposed to only voiced logarithmic-scaled Fo.

As shown in Table 1, LG represents the conventional linear Fo conversion method using Fo features, directly. NNs, VAE, and GAN represent the comparison method using CWT-Fo features only. "+" represents using both the CWT-Fo features and the corresponding selected XWT-Fo features for the training model. As the results indicated, the conventional linear conversion LG has almost no effect in all the emotional VC datasets. The other four methods can affect the conversion of all emotional voice datasets. In addition, the GAN and VA-GAN can obtain significant improvement in Fo conversion. That proves that GAN

https://doi.org/10.1017/ATSIP.2019.3

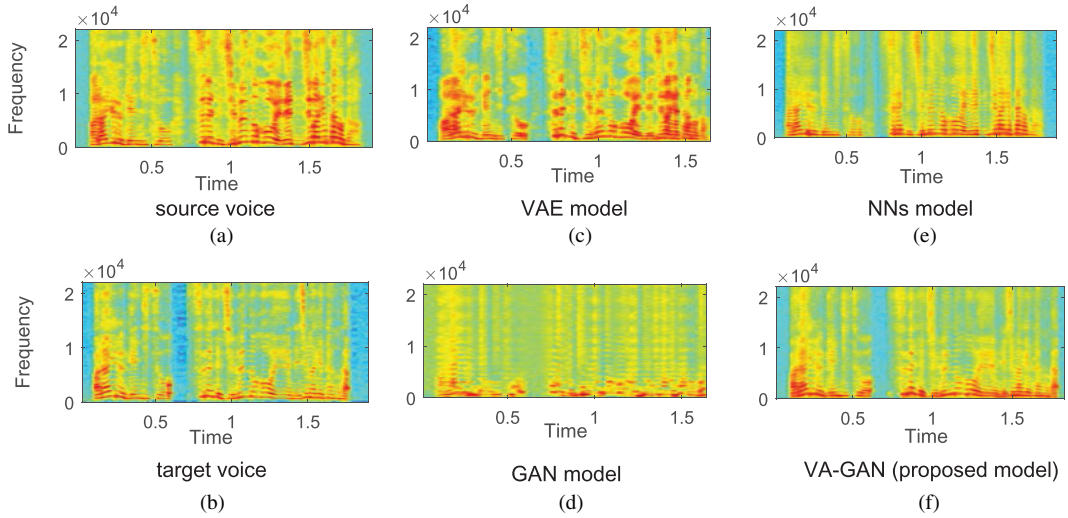

Fig. 8. Spectrograms of source voice, target voice and converted voices reconstructed by converted spectral features and Fo features using different methods.

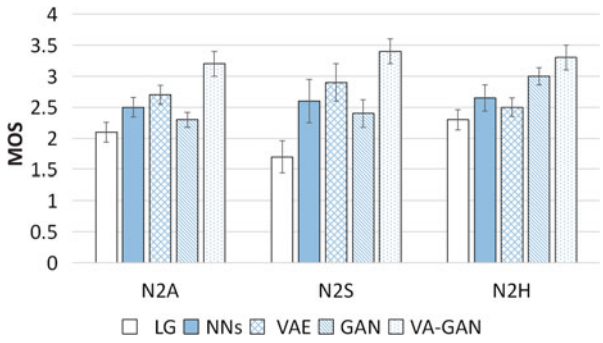

Fig. 9. MOS evaluation of emotional voice conversion.

and our proposed VA-GAN are effective in mapping emotional VC. Comparing the NN and NN+, VAE and VAE+, GAN and GAN+, and VA-GAN and VA-GAN+, we can see that adding the XWT-Fo features cannot improve the GAN model, but doing so can improve the training effectiveness for NN, VAE and the proposed VA-GAN training model.

Comparing the results of the small size training data and the larger size training data shown in Table 1, The RMSE of NN, GAN, and VAE increase significantly when using the smaller training data size, which means the conversion effect degrades. On the other hand, the RMSE of VA-GAN increased only a little when using a smaller data size, which shows the strength of VA-GAN when using limited training data.

Figure 8 shows the spectrograms of the source voice, target voice, and converted voices with different methods. The converted emotional voices are reconstructed by both converted Fo features and converted spectral features using STRAIGHT.

As discussed above, the RMSE values obtained using GAN without VAE are also slightly better than NNs model. However, as shown in the Fig. 8 (D), there is a significant loss in the high-frequency part of the spectrogram of the converted voice. We predict this is due to the fact that, unlike normal object image generation, the dissimilarities between the source and target emotional voices are very small. Therefore, without VAE, it is sometimes hard to obtain a good result due to the problem of insufficient data and the difficulty of training GAN. We can clearly see that only using VAE will result in fuzzy voice features. When combining the GAN with VAE, the obtained converted features are very similar to the target.

D) Subjective experiment

We conducted a subjective emotion evaluation using a mean opinion score test to measure naturalness. The opinion score was set to a five-point scale ranging from totally unnatural (2) to completely natural (5). Here, we tested the neutral to emotional pairs (N2H, N2S, N2A). In each test, 50 utterances, converted by the five methods using both CWT-Fo and XWT-Fo features were selected, and 10 listeners were involved. The subjects listened to the speech that was converted using the five methods before being asked to assign a point value to each conversion. Fig. 9 shows the results of the MOS test. The error bar shows the 95% confidence interval. As the figure shows, the conventional LG method shows poor performance in all emotional VC pairs. Although using GAN without VAE obtained a slightly better result than the NN method in the objective experiment, due to the instability and non-regularization of some converted features, it got worse scores in a MOS test. The effects of the VAE model and NN model is similar in the naturalness evaluation. The VA-GAN method obtained the best score for every emotional VC.

For the similarity evaluation, we carry out a subjective emotion classification test for neutral to emotional pairs (N2H, N2S, N2A) comparing different methods (LG, NN, GANs, VA-GAN). For each test model, 30 utterances (10 for angry, 10 for sad, 10 for happy) are selected, and 10 listeners are involved. The listeners are asked to label a converted voice as Angry (Ang), Sad, Happy (Hap) or Neutral (Neu). As shown in Table 3 (a), when evaluating the

https://doi.org/10.1017/ATSIP.2019.3

Table 3. Results of emotion classification for recorded (original) voices and converted voice by different methods [%].

Tar./ Percept	(a) original voice				(b) LG				(c) NN				(d) GAN				(e) VA-GAN			
	Ang	Sad	Hap	Neu	Ang	Sad	Hap	Neu	Ang	Sad	Hap	Neu	Ang	Sad	Hap	Neu	Ang	Sad	Hap	Neu
Ang	93	5	0	2	22	17	0	61	62	10	5	23	30	24	1	45	68	5	0	27
Sad	2	97	0	1	12	34	3	51	6	55	2	37	40	45	13	2	15	65	5	15
Hap	0	0	98	2	3	5	19	73	3	3	67	27	3	2	37	58	2	3	73	22

original recorded emotional speech utterances, the classifier performed quite strongly; thus, the corpus is sufficient for recognizing emotion.

As shown in Table 3 (b), the conventional LG method shows poor performance in all emotional VC tasks. Comparing the results of Table 3 (c) and Table 3 (d), it is clear that NNs method obtains better classification results than the GANs model. With reference to the results of VA-GAN shown in Table 3 (e), the proposed method yielded about 70% classification accuracy on average, which has indicated a significant improvement over the other models.

V. CONCLUSIONS

In this paper, we propose an effective neutral-to-emotional VC model, using the training model VA-GAN, which consists of two effective generator models (GAN and VAE). For the feature extraction and processing, we use CWT to systematically capture the Fo features of different temporal scales (CWT-Fo features), which are suitable for the VA-GAN model. Moreover, to increase the training data for the emotional VC, we carry out the cross-wavelet transforming for CWT-Fo features of different emotional voices to get the XWT-Fo features. A comparison between the proposed VA-GAN and conventional methods shows that our proposed model can effectively change the voice prosody better than other models, and adding the XWT-Fo features can improve the training effectiveness of the training model.

FINANCIAL SUPPORT

This work was supported in part by PRESTO, JST (Grant No. JPMJPR15D2).

REFERENCES

[1] Nakashika, T.; Takiguchi, T.; Minami, Y.: Non-parallel training in voice conversion using an adaptive restricted Boltzmann machine. *IEEE/ACM Trans. Audio, Speech, Language Process.*, **24** (11) (2016), 2032–2045.

[2] Wu, Z.; Chng, E.S.; Li, H.: Conditional restricted boltzmann machine for voice conversion, in *Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit & Int. Conf. on*. IEEE, 2013, 104–108.

[3] Hsu, C.-C.; Hwang, H.-T.; Wu, Y.-C.; Tsao, Y.; Wang, H.-M.: Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. *arXiv preprint arXiv:1704.00849*, 2017.

[4] Toda, T. *et al.*: The voice conversion challenge 2016, in *Interspeech*, 2016, 1632–1636.

[5] Ribeiro, M.S.; Clark, R.A.: A multi-level representation of fo using the continuous wavelet transform and the discrete cosine transform, in *ICASSP*, 2015, 4909–4913.

[6] Vainio, M. *et al.*: Continuous wavelet transform for analysis of speech prosody, in *TRASP 2013-Tools and Resources for the Analysis of Speech Prosody*, 2013.

[7] Suni, A.S. *et al.*: Wavelets for intonation modeling in HMM speech synthesis, in *8th ISCA Workshop on Speech Synthesis, Proceedings, Barcelona, August 31-September 2, 2013*, 2013.

[8] Ming, H.; Huang, D.; Dong, M.; Li, H.; Xie, L.; Zhang, S.: Fundamental frequency modeling using wavelets for emotional voice conversion, in *Affective Computing and Intelligent Interaction (ACII)*, 2015, 804–809.

[9] Luo, Z.; Chen, J.; Takiguchi, T.; Ariki, Y.: Emotional voice conversion with adaptive scales fo based on wavelet transform using limited amount of emotional data, in *Proc. Interspeech 2017*, 2017, 3399–3403.

[10] Torrence, C.; Compo, G.P.: A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.*, **79** (1) (1998), 61–78.

[11] Hudgins, L.; Friehe, C.A.; Mayer, M.E.: Wavelet transforms and atmospheric turbulence. *Phys. Rev. Lett.*, **71** (20) (1993), 3279.

[12] Cai, X.J.; Tian, S.; Yuan, N.; Hamori, S.: Interdependence between oil and east Asian stock markets: evidence from wavelet coherence analysis. *J. Int. Fin. Mark, Inst. Money*, **48** (2017), 206–223.

[13] Grinsted, A.; Moore, J.C.; Jevrejeva, S.: Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear Processes Geophys.*, **11** (5/6) (2004), 561–566.

[14] Banerjee, S.; Mitra, M.: Application of cross wavelet transform for ecg pattern analysis and classification. *IEEE Trans. Instrum. Meas.*, **63** (2) (2014), 326–333.

[15] Smolensky, P.: Information processing in dynamical systems: foundations of harmony theory. COLORADO UNIV AT BOULDER DEPT OF COMPUTER SCIENCE, Tech. Rep., 1986.

[16] Hinton, G.E.; Osindero, S.; Teh, Y.-W.: A fast learning algorithm for deep belief nets. *Neural Comput.*, **18** (7) (2006), 1527–1554.

[17] Kingma, D.P.; Welling, M.: Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[18] Goodfellow, I. *et al.*: Generative adversarial nets, in *Advances in Neural Information Processing Systems*, 2014, 2672–2680.

[19] Nakashika, T.; Takashima, R.; Takiguchi, T.; Ariki, Y.: Voice conversion in high-order eigen space using deep belief nets, in *INTERSPEECH*, 2013, 369–372.

[20] Hsu, W.-N.; Zhang, Y.; Glass, J.: Learning latent representations for speech generation and transformation. *arXiv preprint arXiv:1704.04222*, 2017.

[21] Kaneko, T.; Kameoka, H.; Hiramatsu, K.; Kashino, K.: Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks, in *Proc. INTERSPEECH*, 2017, 1283–1287.

[22] Arjovsky, M.; Chintala, S.; Bottou, L.: Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

https://doi.org/10.1017/ATSIP.2019.3

[23] Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A.: Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[24] Luo, Z.; Chen, J.; Nakashika, T.; Takiguchi, T.; Ariki, Y.: Emotional voice conversion using neural networks with different temporal scales of fo based on wavelet transform, in *9th ISCA Speech Synthesis Workshop*, 140–145.

[25] Daubechies, I: *Ten lectures on wavelets* Siam, vol. **61**, 1992.

[26] Kawahara, H.: Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical Sci. Technol.*, **27** (6) (2006), 349–353.

[27] Denton, E.L. *et al.*: Deep generative image models using a Laplacian pyramid of adversarial networks, in *Advances in Neural Information Processing Systems*, 2015, 1486–1494.

[28] Radford, A.; Metz, L.; Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[29] Zhao, J.; Mathieu, M.; LeCun, Y.: Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

[30] Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X.: Improved techniques for training GANs, in *Advances in Neural Information Processing Systems*, 2016, 2234–2242.

[31] Kaneko, T.; Kameoka, H.; Hojo, N.; Ijima, Y.; Hiramatsu, K.; Kashino, K.: Generative adversarial network-based postfilter for statistical parametric speech synthesis, in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE Int. Conf. on*. IEEE, 2017, 4910–4914.

[32] Arjovsky, M.; Bottou, L.: Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

[33] Dosovitskiy, A.; Brox, T.: Generating images with perceptual similarity metrics based on deep networks, in *Advances in Neural Information Processing Systems*, 2016, 658–666.

[34] Zhao, S.; Song, J.; Ermon, S.: Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658*, 2017.

[35] Larsen, A.B.L.; Sønderby, S.K.; Larochelle, H.; Winther, O.: Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

[36] Rosca, M.; Lakshminarayanan, B.; Warde-Farley, D.; Mohamed, S.: Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*, 2017.

[37] Kawanami, H.; Iwami, Y.; Toda, T.; Saruwatari, H.; Shikano, K.: GMM-based voice conversion applied to emotional speech synthesis, in *IEEE Trans Speech Audio Proc*, 2003, 2401–2404.

[38] Liu, K.; Zhang, J.; Yan, Y.: High quality voice conversion through phoneme-based linear mapping functions with straight for mandarin. *Fuzzy Syst. Knowledge Discovery*, **4** (2007), 410–414.

[39] Chen, L.-H.; Ling, Z.-H.; Liu, L.-J.; Dai, L.-R.: Voice conversion using deep neural networks with layer-wise generative training. *Audio, Speech, Language Process. IEEE/ACM Transactions on*, **22** (12) (2014), 1859–1872.

[40] Sonobe, R.; Takamichi, S.; Saruwatari, H.: Jsut corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis. *arXiv preprint arXiv:1711.00354*, 2017.

[41] Johnson, J.; Alahi, A.; Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution, in *European Conf. on Computer Vision*. Springer, 2016, 694–711.

[42] Ioffe, S.; Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift, in *Int. Conf. on Machine Learning*, 2015, 448–456.

[43] Maas, A.L.; Hannun, A.Y.; Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models, in *Proc. ICML*, vol. 30 (1), 2013.

[44] He, K.; Zhang, X.; Ren, S.; Sun, J.: Deep residual learning for image recognition, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, 770–778.

[45] Li, C.; Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks, in *European Conf. on Computer Vision*, Springer, 2016, 702–716.

[46] Kinga, D.; Adam, J.B.: A method for stochastic optimization, in *Int. Conf. on Learning Representations (ICLR)*, 2015.

Zhaojie Luo was born in Fujian, China, in 1989. He received his M.E.degree from Kobe University in Japan. He is currently a Ph.D. student in computer science at the same university. His research interest is voice conversion, facial expression recognition, multimodal emotion recognition and statistical signal processing. He is a member of IEEE, ISCA and ASJ. He has published more than 10 publications in major journals and international conferences, such as IEEE Trans. Multimedia, EURASIP JASMP, EURASIP JIVP, INTERSPEECH, SSW, ICME, ICPR, *etc.*

Jinhui Chen received his Ph.D degrees at Kobe University, Japan, in 2016. He is currently an assistant professor at Kobe University. His research interests include pattern recognition and machine learning. He is a member of IEEE, ACM, and IEICE. He has published more than 20 publications in major journals and international conferences, such as IEEE Trans. Multimedia, EURASIP JIVP, SIViP, ACM MM, ACM ICMR, *etc.*

Tetsuya Takiguchi received the M. Eng. and Dr. Eng. degrees in information science from Nara Institute of Science and Technology, Japan, in 1996 and 1999, respectively. From 1999 to 2004, he was a Researcher in IBM Research, Tokyo Research Laboratory. From 2004 to 2016, he was an associate professor at Kobe University. Since 2016 he has been a professor at Kobe University. From May 2008 to September 2008, he was a visiting scholar at the Department of Electrical Engineering, University of Washington. From March 2010 to September 2010, he was a visiting scholar at the Institute for Learning & Brain Sciences, University of Washington. From April 2013 to October 2013, he was a visiting scholar at the Laboratoire d'InfoRmatique en Image et Systèmes d'information, INSA Lyon. His research interests include speech, image, and brain processing, and multimodal assistive technologies for people with articulation disorders. He is a member of the Information Processing Society of Japan and the Acoustical Society of Japan.

Yasuo Ariki received his B.E., M.E., and Ph.D. in information science from Kyoto University in 1974, 1976, and 1979, respectively. He was an assistant professor at Kyoto University from 1980 to 1990, and stayed at Edinburgh University as visiting academic from 1987 to 1990. From 1990 to 1992 he was an associate professor and from 1992 to 2003 a professor at Ryukoku University. From 2003 to 2016, he was a professor at Kobe University. He is mainly engaged in speech and image recognition and interested in information retrieval and database. He is a member of IEEE, IEICE, IPSJ, JSAI, and ASJ.

https://doi.org/10.1017/ATSIP.2019.3