Neutral-to-Emotional Voice Conversion with Latent Representations of F0 using Generative Adversarial Networks *

☆ Zhaojie Luo, Tetsuya Takiguchi, and Yasuo Ariki (Kobe University)

1 Introduction

Emotional VC is a kind of voice conversion technique for converting prosody in speech, which can represent different emotions, while keeping the linguistic information unchanged. In a voice, the spectral and F0 features can affect the acoustic and prosodic features, respectively. So far, spectral mapping mechanisms have achieved tremendous success in VC tasks, while, how to effectively generate prosody in the target voice remains a challenge. Previous studies have shown that F0 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. However, it may be difficult to apply conventional deep learning-based VC to F0 conversion using simple representations of F0, such as dynamic features (delta F0).

In recent years, it has been shown that the CWT method can effectively model F0 in different temporal scales and significantly improve speech synthesis performance. Our earlier work [1] systematically captures the F0 features of different temporal scales using AS-CWT, which transforms F0 features into high-dimensional CWT-F0 features containing more specifics. Thus, building on top of the success of using CWT-F0 features for prosody conversion, in this study, we want to go one step further to generate emotional voice more similar to target emotion using a generative model.

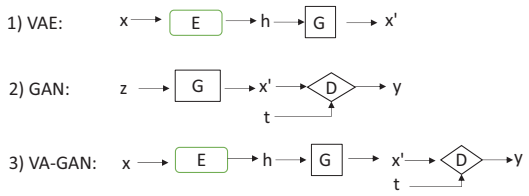

Fig. 1 Illustration of the structure of VAE [2], GAN [3] and the proposed VA-GAN.

In this study, inspired by the success of VAE and GAN in VC tasks, we propose an emotional VC framework that combines a VAE with a GAN, which 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, however, a generative adversarial model only discriminates between "real" and "fake" features. There are no constraints that the generated features have to sound like a human voice. This leads to results in which the generated voice is unnatural of bad quality. Another popular generative model, VAE, suffers from the problem of fuzzy sound, which is caused by the 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 improving GAN learning. Meanwhile, GAN can enhance VAE with an adversarial mechanism for leveraging generated samples.

As shown in Fig. 1, 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. our VA-GAN consists of three parts: 1) the encoder network E, which maps the x to a latent representation h, 2) the generative network G, which generates features x' from the latent representation h, 3) and the discriminative network D, which distinguishes real/fake (t/x') features. Here, we use the t and 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.

2 Features 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 dependen-

*Generative Adversarial Networks を用いた感情声質変換, 羅兆傑, 滝口哲也, 有木康雄 (神戸大)

This work was supported in part by PRESTO, JST (Grant No. JPMJPR15D2) and JSPS KAKENHI (Grant No. JP17H01995).

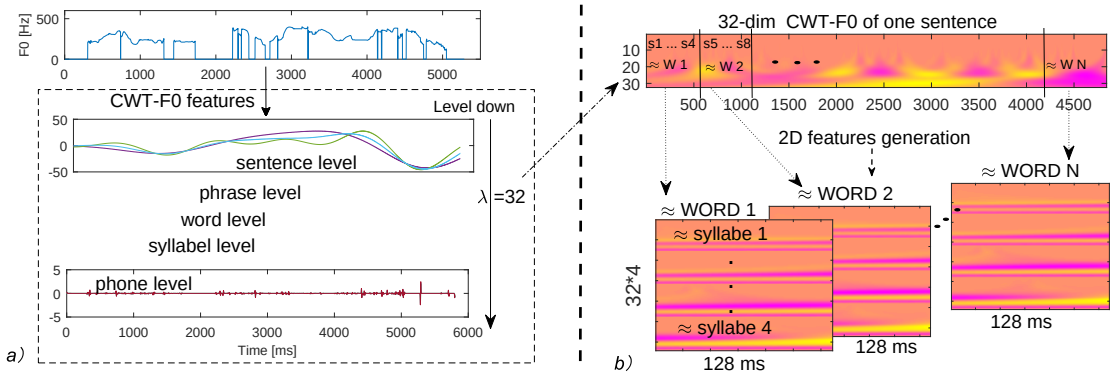

Fig. 2 CWT-F0 features extraction and processing.

cies. And, as has been proven in our recent studies [1], the CWT can effectively model F0 in different temporal scales and significantly improve the system performance. We adopt CWT to decompose the one-dimensional F0 features into high-dimensional CWT-F0 features. The continuous wavelet transform of F0 is defined by

$$W(f_0)(\tau,t) = \tau^{-1/2}\int_{-\infty}^{\infty} f_0(x)\,\psi\left(\frac{x-t}{\tau}\right)dx \quad (1)$$

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

where $f_0(x)$ is the input signal and ψ is the Mexican hat mother wavelet. We decompose the continuous F0 with 32 discrete scales, each one third of an octave apart. Our F0 is thus represented by 32 separate components given by

$$W_i(f_0)(t) = W_i(f_0)(2^{(i/3)+1}\tau_0, t) \quad (3)$$

where $i = 1, ..., 32$ and τ_0=1 ms. Fig. 2 (a) shows several CWT-F0 feature examples of decomposed components, which can represent the utterance, phrase, word, syllable, and phone levels, respectively.

GAN have been successful in image generation. So, before training in the VA-GAN model, we reshaped CWT-F0 matrix to 128 × 128 size images shown in Fig. 2 (b). As described in [1], the average duration of non-emphasized syllables was found to be 50 ms to 180 ms, and the words from 300 ms to 650 ms. Thus, one sentence of CWT-F0 features was reshaped into several 128×128 size images, and one image approximately represents one word (32×512) composed of four syllables $(32 \times 128 \times 4)$. By doing this, the learning rate can be improved and a higher overall VA-GAN accuracy can be achieved.

3 Emotional VC using VA-GAN

3.1 Background: VAE and GAN

3.1.1 Variational autoencoder

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

$$z \sim Enc(x) = q_\phi(h|x), \tilde{x} \sim Dec(h) = p_\theta(h|x) \quad (4)$$

where (Enc) represents encode networks that encode a data sample x to a latent representation h and decode networks (Dec) decode the latent representation back to data space. In the VAE, the recognition model $q_\phi(h|x)$ approximates the true posterior $p_\theta(h|x)$. The VAE regularizes the encoder by imposing a prior over the latent distribution $p\theta(h)$, which is assumed to be a centered isotropic multivariate Gaussian $p_\theta(h) \sim N(h;0,I)$. The VAE loss $L_{\theta,\phi;x}$ is minus the sum of the expected log-likelihood L_{like} (the reconstruction error) and a prior regularization term L_{prior} represented as:

$$L_{\theta,\phi;x} = -E_{q_\phi(h|x)}[\log\frac{p_\theta(x|h)p_\theta(x)}{q_\phi(h|x)}] = L_{like} + L_{prior} \quad (5)$$

$$L_{like} = -E_{q_\phi(h|x)}[\log p_\theta(x|h)] \quad (6)$$

$$L_{prior} = KL(q_\phi(h|x)||p_\theta(h)) \quad (7)$$

where KL is the Kullback-Leibler divergence. We use the KL loss to reduce the gap between the prior $P(h)$ and the proposal distributions. The loss of KL is only related to the encoder network Enc. It

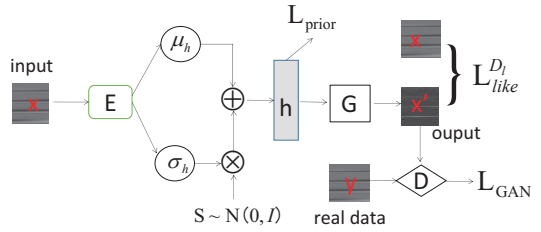

Fig. 3 Illustration of calculating the loss of VA-GAN.

represents whether or not the distribution of the latent vector is under expectation. Here, we want to optimize $L_{\theta,\phi;x}$ in respect to θ and ϕ.

3.1.2 Generative Adversarial Networks

GAN has obtained impressive results for image generation. The key to the success of the GAN is learning a generator distribution $P_G(x)$ that matches the true data distribution. In a GAN, D and G play the following two-player minimax game with the value function $V(G, D)$:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{x\sim p_z(z)}[\log(1 - D(G(z)))] \quad (8)$$

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 $P_{Data}(x)$. Both G and D can be trained using back-propagation.

3.2 The VA-GAN training model

When dealing with the training of emotional VC, each two sets of labeled and paired 128×128 feature images were sampled from domains source emotional voice X and target emotional voice Y, respectively. As shown in Fig. 3, our model contains an encode (Enc), conversion function ($G : x \to y$) and a discriminator (D). In practice, the 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 unstable gradient of G when training GAN. To resolve the instability of training GAN, we extract the representative features from a pre-trained Enc. We observe better results when using the latent representation (h) from the encoder Enc. For the conversion function $G : x \to y$ and its discriminator D with pre-trained Enc, we express the objective as:

$$L_G(G, D, X, Y) = E_{y\sim P_{data}(y)}[\log D(y)] + E_{x\sim P_{data}(x)}[\log(1 - D(G(Enc(x))))] \quad (9)$$

The goal of emotional VC is to learn a converted emotional voice distribution $P_G(x)$ that matches the target emotional voice distribution $P_{data}(y)$. Equation (9) 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 G to fit the $P_{data}(y)$. $L_G(G, D, X, Y)$ is maximized and minimized with respect to D and G, respectively.

$$G^* = arg \max_D \min_G L_G(G, D, X, Y) \quad (10)$$

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 $P_{data}(y)$ distribution. Meanwhile, replacing the reconstruction error term from Equation (6) with a reconstruction error expressed in the discriminator D can solve the blurry problem of VAE. To achieve this, let $D_l(x)$ denote the hidden representation of the lth layer of the discriminator. We introduce a Gaussian observation model for $D_l(x)$ with mean $D_l(\tilde{x})$ and identity covariance:

$$p(D_l(x)|z) = N(D_l(x)|D_l(\tilde{x}), I) \quad (11)$$

where $\tilde{x} \sim Dec(h)$ in Equation (4) now is the sample from the generator (G) of x. We can now replace the VAE error of Equation (6) with

$$L_{like}^{D_l} = -E_{q(z|x)}[\log p(D_l(x)|z)] \quad (12)$$

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

$$L = L_{GAN} + L_{like}^{D_l} + L_{prior} \quad (13)$$

4 Experiments

In our experiments, we used a database of emotional Japanese speech. The waveforms used were sampled at 16 kHz. Input and output data had the

Table 1 F0-RMSE results for different emotions. N2A, N2S and N2H represent the datasets from neutral to angry, sad and happy voice, respectively.

	Source	LG	NN	VAE	GAN	VA-GAN
N2A	76.8	76.3	70.4	73.4	59.5	**51.2**
N2S	73.7	72.0	62.3	77.5	**56.1**	58.5
N2H	100.4	99.1	75.2	85.8	65.5	**62.1**

same speaker, but the speaker was expressing different emotions. We classified the three data sets 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, 50 sentences were chosen as training data and 10 sentences were chosen for the VC evaluation.

To evaluate the effectiveness 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 F0 features conversion in deep learning VC tasks. **NNs** is our previous work [1] that used the pre-trained NNs to convert the CWT-F0 features. We also compared VA-GAN with the **GAN** and **VAE**. In these experiments, we focused on the conversion of F0, therefore, in these compared methods, the spectral features were converted using the same DBN-based model.

4.1 Objective Experiment

To evaluate F0 conversion, we used the root-mean-square error (RMSE). As shown in Table 1, the conventional linear conversion LG can only affect the conversion of neutral to happy, but only slightly affects the other conversions. The other three methods can affect the conversion of all emotional voice datasets. In addition, the GAN and VA-GAN can obtain significant improvement in F0 conversion.

4.2 Subjective Experiment

We conducted a subjective emotion evaluation using a mean opinion score test. The opinion score was set to a five-point scale (the more similar to the emotion of the sample voice the target speech sounded, the higher the point value). Here, we tested the neutral to emotional pairs (N2H, N2S, N2A). In each test, 50 utterances (10 for source speech, 10 for target speech, and 30 for converted speech by the three methods) were selected, and 10 listeners were involved. Each subject listened to the source and target speech samples. The subject then listened to the speech that was converted using the four methods before beening asked to assign a point value to each conversion. Fig. 4 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 the conversion of neutral to angry voice. Although using GAN without VAE obtianed 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 MOS test. The VA-GAN obtained the best score in every emotional VC.

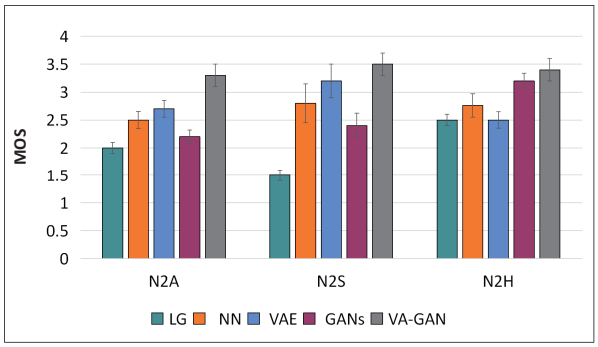

Fig. 4 MOS evaluation of emotional VC

5 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). Meanwhile, for the feature extraction and processing, we use CWT to systematically capture the F0 features of different temporal scales, and transform them to 2D-features, which are suitable for the VA-GAN model.

参考文献

[1] Z. Luo *et al.*, "Emotional voice conversion with adaptive scales F0 based on wavelet transform using limited amount of emotional data," Proc. Interspeech 2017, pp. 3399–3403, 2017.

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

[3] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.