Emotional Voice Conversion by Learning Disentangled Representations with Spectrum and Prosody Features *

☆ Xunquan Chen[1], Jinhui Chen[2], Ryoichi Takashima[1], Tetsuya Takiguchi[1]
[1] Kobe University, [2]Prefectural University of Hiroshima

1 Introduction

Emotional voice conversion (EVC) is a voice conversion technique for converting prosody in speech, which can represent different emotions while retaining the linguistic information. Recently, the study of VC has attracted wide attention in the field of speech processing. This technology can be applied in various domains [1] [2]. Therefore, it has continued to motivate related studies each year.

Normal VC tasks are designed to transform the speech of the source speaker to that of the target speaker, making the conversion speech sound like the voice of the target speaker. Existing VC methods can be roughly divided into two categories: a) conventional shallow approaches and b) deep neural networks models. Specifically, among the shallow approaches, a Gaussian Mixture Model (GMM) has been commonly used and successfully built upon over many years [3]. Other VC methods, such as those based on non-negative matrix factorization (NMF) [4], have also been proposed. However, these approaches have some limitations, specifically, they are time-consuming and require a degree of background knowledge to implement.

In recent years, deep learning has also markedly improved the performance of VC systems [12] through learning hierarchies of features optimized for the task at hand. However, deep learning models are restricted to problems with moderate dimensions and sufficient available data. Therefore, most deep learning-based VC networks do not focus on the emotional VC, which is mainly affected by low-dimensional fundamental frequency (F0) features. Thus, conventional VC models usually convert F0 using Logarithm Gaussian (LG) normalized transformation [5].

However, in VC tasks, the spectral and F0 features can affect the acoustic and prosodic features, respectively. Particularly in emotional VC tasks where the prosody plays an important role in conveying various types of non-linguistic information that represent the mood of the speaker, such as identity, intention, and attitude. Previous studies [6] have shown that prosody conversion is affected by both short- and long-term dependencies in different temporal levels such as the phones, syllables, and words, within an utterance. The LG-based method is insufficient to convert prosody effectively because of the constraints of their linear models and low-dimensional F0 features.

It has been shown that Continuous Wavelet Transform (CWT) can effectively model F0 in different temporal scales and significantly improve speech synthesis performance [7]. Thus, in this paper, we also applied the CWT in the F0 features processing. Our proposed framework is based on GAN which is composed of a generator and a discriminator typically, and the generator is an encoder-decoder module in our work.

2 Related work

2.1 Continuous Wavelet Transform

The CWT was originally proposed by Goupillaud *et al.* [8], and for a 1-D input signal, the result is a 2-D description of the signal with respect to time-scale parameter (s, τ) of the CWT function

$$CWT\{x(t); s, \tau\} = \int x(t)\psi^*_{s,\tau}(t)dt$$
$$\psi_{s,\tau}(t) = s^{-1/2}\psi\left(\frac{t-\tau}{s}\right) \quad (1)$$

where $*$ stands for complex conjugate, $\psi_{s,\tau}(t)$ is mother wavelet with the scaling factor s and translating factor τ. Here, scaling factor s controls the width of the wavelet and translating factor τ decides the location of the wavelet. The scale s is inversely proportional to the central frequency (ω) of the rescaled mother wavelet.

* Emotional Voice Conversion by Learning Disentangled Representations with Spectrum and Prosody Features, 陳訓泉 [1], 陳金輝 [2], 高島遼一 [1], 滝口哲也 [1] ([1] 神戸大, [2] 広島県立大)

2.2 Generative Adversarial Networks

Moreover, to improve the emotional VC effect with non-parallel training data, our proposed framework is based on generative adversarial networks (GANs) [9] which is composed of a generator and a discriminator typically. The key to the success of the GANs is learning 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 $\mathbf{z} \sim P_{\text{Noise}}(\mathbf{z})$ to data space $\mathbf{x} = G(\mathbf{z})$ and a discriminator D. This discriminator assigns probability $p = D(\mathbf{x})$ when x is a sample from the $P_{\text{Data}}(\mathbf{x})$ 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 $V(G, D)$:

$$\begin{aligned}\min_G \max_D V(D,G) =& E_{\mathbf{x}\sim p_{\text{data}}}[\log D(\mathbf{x})] \\ &+ E_{z\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]\end{aligned} \tag{2}$$

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 $P_{\text{Data}}(\mathbf{x})$. Both G and D can be trained using backpropagation. The effectiveness of GANs is due to the fact that an adversarial loss forces the generated data to be indistinguishable from real data.

2.3 Disentangled Representation Learning

Disentangled representation learning aims to encode input data into separate independent embedding subspaces, where different subspaces represent different data attribute [13]. In the context of emotional voice conversion, if we are able to disentangle emotion from linguistic content and speaker identity, we can change the emotion independently of the linguistic content.

3 Proposed method

3.1 Model Architecture

Our proposed model, illustrated in Fig. 1, consists of two major modules: a generator and a discriminator. The generator is an encoder-decoder module in our work. The generator consists of three modules, a content encoder E_c, a style encoder E_s and a decoder De. The generator is made up entirely of convolution layers in order to operate in a non-autoregressive generative manner. In the training stage, the speaker encoder E_s accepts the acoustic features as input, and learns the reasonable representations relating to emotions in the embedding space. The E_s is built with stacks of convolutional layers followed an average pooling for downsampling. The content encoder E_c is adopted to predict reasonable linguistic representation. E_c is composed of convolution layers. In addition, we adopt Instance normalization after each convolution layer of the content encoder to eliminate speaking style information. Finally, the decoder D_e takes the concatenation of linguistic embeddings and emotion embeddings to synthesizes the converted speech by only changing the source style to the target one. We append a PixelShuffle 1d layer for upsampling.

Unlike the generator, the discriminator is constructed with 2d convolution layers like [12] to better capture the acoustic texture. There are 3 convolution layers to downsample the feature map gradually. Instance normalization and leaky ReLU are applied after each convolution layer except the final output layer.

3.2 Objective Function

Let x and y be acoustic feature sequences belonging to source speech S and target speech T, respectively. The training losses for the proposed VC are described as follows.

Adversarial loss: The adversarial loss is used to render the converted feature indistinguishable from the real target feature.

$$\begin{aligned}\mathcal{L}_{\text{adv}}\left(G_{S\to T}, D_T\right) = \mathbb{E}_{y\sim P_T(y)}[\log D_T\left(y\right)] + \\ \mathbb{E}_{x\sim P_S(x)} \log\left(1 - D_T\left(G_{S\to T}\left(x\right)\right)\right)\end{aligned} \tag{3}$$

Content loss: The content loss is used to preserve the linguistic content of the input speech.

$$\mathcal{L}_{\text{c}} = \left(\left\|E_c\left(G\left(x, y\right)\right) - E_c\left(x\right)\right\|_2\right) \tag{4}$$

Style loss: The style loss is used to achieve style transfer.

$$\begin{aligned}\mathcal{L}_s = \| &\left(E_c\left(G\left(x, y\right)\right), E_s\left(G\left(x, y\right)\right)\right) \\ &- \left(E_c\left(x\right), E_s\left(y\right)\right) \|_2\end{aligned} \tag{5}$$

A reconstruction loss is also adopted to generate reasonable speech using disentangled representations.

$$\mathcal{L}_{\text{rec}} = \left\|G\left(x, x\right) - x\right\|_1 \tag{6}$$

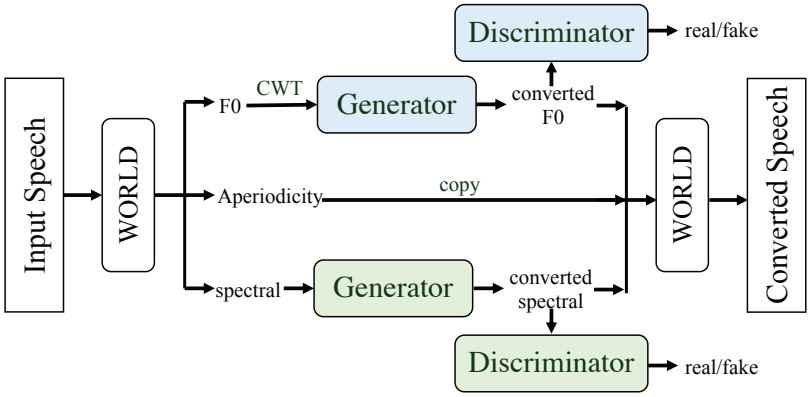

Fig. 1 Overview of proposed EVC system

The full objective function can be summarized as follows:

$$\mathcal{L}_{full} = \mathcal{L}_{adv} + \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s + \lambda_{rec} \mathcal{L}_{rec} \quad (7)$$

where λ_c, λ_s, and λ_{rec} are trade-off parameters to control the weights of the corresponding losses relative to the adversarial loss.

4 Experiments

4.1 Experimental Conditions

We conduct experimental evaluations on the IEMOCAP database [10], which contains about 12 hours of audio data recorded by ten actors with nine different emotions. The baseline model is a simple linear F0 conversion system with a neural network model Stargan-EVC [14]. In this paper, we only consider four emotional categories of them: angry, happy, neutral, sad. Input and output data had the same speaker, but expressing different emotions. We set the three datasets into the following: neutral to happy voice, neutral to angry voice, and neutral to sad voice. Training and testing sets are non-overlapping utterances randomly selected from the same speaker (80% for training, 20% for test). We use WORLD [25] vocoder to extract fundamental frequencies, spectral sequences (sps) and aperiodicities (aps) from raw audio waveforms sampled at 16KHz. During preprocessing, we normalized the source and target, MCC and CWT-F0, features to zero-mean and unit-variance for each dimension using their respective training sets. We trained the proposed model by ADAM optimizer with 0.0001 as learning rate. The weighting parameters are simply set as $\lambda_c = 5$, $\lambda_s = 5$ and $\lambda_3 = 10$ in Eq. (7).

4.2 Objective Evaluations

Mel Cepstral Distortion (MCD) was used for the objective evaluation of spectral conversion, MCD is defined below.

$$MCD = (10/\ln 10)\sqrt{2\sum_{i=1}^{24}(mc_i^t - mc_i^c)^2} \quad (8)$$

In Eq. 8, mc_i^t and mc_i^c represent the target and the converted mel-cepstral, respectively.
To evaluate the F0 conversion, we used the RMSE

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((F0_i^t) - (F0_i^c))^2}, \quad (9)$$

where $F0_i^t$ and $F0_i^c$ denote the target and the converted F0 features, respectively. A lower MCD and F0-RMSE value indicate smaller distortion or predicting error.

Table 1 shows the MCD and F0-RMSE results from the neutral to emotional pairs. As shown in Table 1, the proposed method can also obtain good results in spectral and F0 conversion. Through the objective experiments, we empirically confirm that the proposed method effectively brings the converted acoustic feature sequence closer to the target one than baseline.

5 Conclusions

In this paper, we present a non-parallel EVC method with disentangled linguistic and style representations. Our proposed framework is based

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

	MCD [dB]			F0-RMSE [Hz]		
	N2A	N2S	N2H	N2A	N2S	N2H
Source	6.54	5.36	6.84	77.5	74.6	101.7
Stargan-EVC	4.54	4.81	4.57	57.2	60.8	69.5
Proposed method	3.97	4.69	4.23	46.3	49.6	59.4

on GAN which is composed of a generator and a discriminator typically, and the generator is an encoder-decoder module in our work. The experimental results show the effectiveness of our proposed method.

References

[1] Krivokapić, Jelena, "Rhythm and convergence between speakers of American and Indian English," Laboratory Phonology, vol. 4, no. 1, pp. 39-65, 2013.

[2] Raitio *et al.*, "Phase Perception of the Glottal Excitation of Vocoded Speech," in *Proc. Interspeech*, pp. 254-258, 2015.

[3] Toda *et al.*, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 15, pp. 2222-2235, 2007.

[4] Takashima *et al.*, "Exemplar-based voice conversion in noisy environment," in *Proc. IEEE SLT*, pp. 313-317, 2012.

[5] Liu *et al.*, "High quality voice conversion through phoneme-based linear mapping functions with STRAIGHT for mandarin," in *Proc. 4th Int. Conf. Fuzzy Syst. Knowl. Discovery (FSKD)*. Vol. 4, pp. 410-414, 2007.

[6] Ribeiro *et al.*, "A multi-level representation of F0 using the continuous wavelet transform and the discrete cosine transform," in *Proc. ICASSP*, pp. 4909-4913, 2015.

[7] Vainio *et al.*, "Continuous wavelet transform for analysis of speech prosody," in *TRASP*, pp. 78-81, 2013

[8] Goupillaud *et al.*, "Cycle-octave and related transforms in seismic signal analysis," Geoexploration, Vol. 23, no. 1,pp. 85-102, 1984.

[9] Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NeurIPS*, 2014.

[10] Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, 2008.

[11] Morise *et al.*, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," IEICE TRANSACTIONS on Information and Systems, vol. 99, no.7, pp. 1877-1884, 2016.

[12] Kameoka *et al.*, "StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks," in *Proc. IEEE SLT*, pp. 266-273, 2018.

[13] Locatello *et al.*, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proc. PMLR*, pp. 4114-4124, 2019.

[14] Rizos *et al.*, "StarGAN for Emotional Speech Conversion: Validated by Data Augmentation of End-to-End Emotion Recognition," in *Proc. ICASSP*, pp. 3502-3506, 2020.