Emotional Voice Conversion with a Novel Content-Style Fusion Block *

☆ 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 technique for transforming the emotional state of a given utterance while keeping linguistic information and speaker identity unchanged. This technique can be used in a variety of real-world applications, including voice assistants, conversational agents, sound design as well as other entertainment applications. [1, 2].

Existing EVC methods can be roughly categorized into two types based on the use of training data. Early works [3, 4] mainly focused on using aligned parallel data, *i.e.*, any speech pairs from source and target speakers share the same linguistic content and are aligned in the temporal dimension. However, these data were difficult to collect and time-consuming to align. The restricted corpus availability limits the performance and generalizability of speech conversion.

These limitations have motivated research to explore non-parallel EVC approaches [5, 6, 7, 8, 12, 13]. An appealing solution to this problem is based on generative adversarial networks (GANs) [9]. Zhou *et al.* [5] proposed an EVC method based on CycleGAN [10], to model the spectrum and prosody mapping between source speech and target speech. This has been widely acknowledged as an effective way to achieve one-to-one conversion with non-parallel data. However, using only one model to achieve many-to-many conversions is more attractive for a wide range of applications. Inspired by StarGAN [11], Rizos *et al.* proposed StarGAN-EVC [6] to train the spectral mapping between multiple emotional domains as an improvement. Recently, several studies [7, 8, 12, 13] based on speech representation disentanglement have attempted to decompose the speech into different representations. These methods can easily achieve emotional voice conversion by simply replacing the emotion-related representations. Gao *et al.* [7] proposed a non-parallel EVC approach based on style transfer autoencoders, which consists of two encoders and decoder for each emotion domain. To use a limited amount of emotional speech data for unseen speakers, Zhou *et al.* [8] proposed a two-stage training strategy and used the corresponding phoneme transcription to guide the disentanglement of the emotional style and linguistic content. Choi *et al.* [12] used an emotion encoder and an additional speaker encoder to utilize various emotional characteristics of multiple speakers.

The above-mentioned methods can transfer the emotion states in non-parallel setting. However, these methods only learn an average representation or extract a fixed-length vector for each emotional style. It is a straightforward way to obtain the emotion information, but only global-level emotion information can be learned. Therefore, there still remains a gap between the converted speech and the real target in terms of quality and emotion fidelity.

In this paper, we present a novel EVC model, which can sufficiently learn the emotion information in both global-level and local-level. Since speech signals dynamically change in time, some parts of emotion information also would change in time. And silence parts of the signals, which hardly convey emotion information, should be treated differently. Unlike the previous studies, it assumes that the emotional style is dynamic and time-varying relevant to linguistic content in our study. Therefore, instead of only using a fixed-length vector to represent the global-level emotion information of the whole utterance, the local-level emotion information should rely on single phoneme content and change with time. For local-level emotion information, a novel content-style fusion block is proposed to implement the implicit alignment for emotion and phoneme content, further embedding the phoneme-level emotion representation. For global-level emotion information, we embed the complete set of time steps of speech emotion into a fixed-length vector to obtain the sentence-level emotion representation.

* Emotional Voice Conversion with a Novel Content-Style Fusion Block, 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)

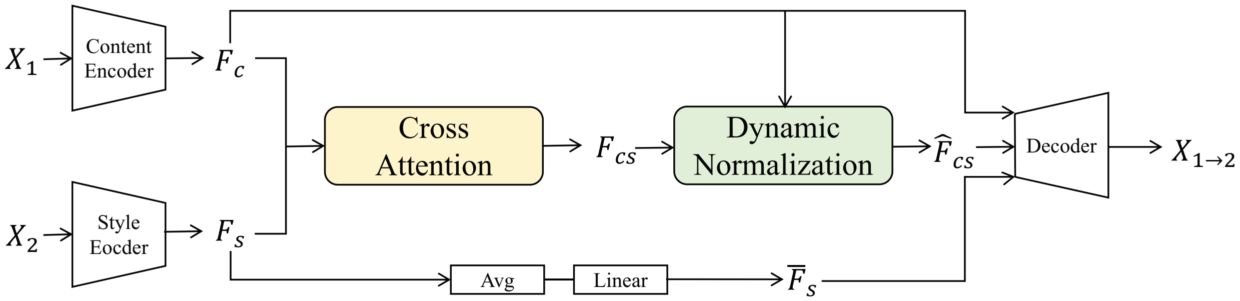

Fig. 1 The generator architecture of the proposed model. X_1 and X_2 indicate the mel-spectrogram of source and target speech respectively. IN is instance normalization.

2 Proposed method

2.1 Framework Overview

The proposed method employs an autoencoder framework with an adversarial training strategy [9] to disentangle the emotion information from the content information of each input speech into separated representation spaces. During the adversarial training procedure, we utilize the autoencoder framework as the GAN generator, which aims to fool the discriminator by generating high-quality and realistic audio signals. Figure 1 shows the generator architecture of the proposed model. The generator is an autoencoder framework in our work, which consists of four modules: a content encoder $E_c(\cdot)$, a style encoder $E_s(\cdot)$, a content-style fusion block $CSFB(\cdot,\cdot)$, and a decoder $D_e(\cdot,\cdot,\cdot)$. Here, the content-style fusion block $CSFB(\cdot,\cdot)$ is composed of a cross attention module and a dynamic normalization module [14], which will be described in detail in Section 2.2. The generator is composed entirely of convolution neural networks to achieve non-autoregressive generation. Unlike the generator, the discriminator is constructed with 2d convolution layers like [6] to better capture the acoustic texture.

The content encoder $E_c(\cdot)$ is used to extract the content representation F_c from the mel-spectrogram X_1 of source speech. The style encoder $E_s(\cdot)$ extracts the emotion representation F_s from the mel-spectrogram X_2 of target speech. Then the content-style fusion block $CSFB(\cdot,\cdot)$ can generate content-dependent emotion representation $\hat{F}_{cs}$. Finally, the decoder $D_e(\cdot,\cdot,\cdot)$ will takes the content representation F_c, the phoneme-level emotion representation $\hat{F}_{cs}$ and the averaged sentence-level emotion representation $\bar{F}_s$ as inputs, and then it synthesizes the converted mel-spectrogram $X_{1\to2}$ which only transfers the source emotion state to the target one.

The whole conversion process can be formulated as follows:

$$\begin{aligned}
&F_c = E_c\left(X_1\right), F_s = E_s\left(X_2\right),\\
&\bar{F}_s = AvgPool\left(F_s\right), \hat{F}_{cs} = CSFB\left(F_c, F_{cs}\right),\\
&X_{1\to2} = D_e\left(F_c, \hat{F}_{cs}, \bar{F}_s\right),
\end{aligned} \quad (1)$$

where X_1 and X_2 are the source and target speech respectively. To fuse the global-level emotion feature $\bar{F}_s$, we first use AvgPooling layer for different length utterances to obtain fixed-length representations, and then feed it into several linear transformations. The local-level emotion feature $\hat{F}_{cs}$ is the all time step for the output feature and its length is the same as F_c.

2.2 Content-style Fusion Block

The detailed structure of our proposed content-style fusion block is illustrated in Figure 2. As shown in this figure, the content-style fusion block is built with a cross attention module followed by a dynamic normalization module [14].

The speech signal can be considered a composition of content information and emotion information in EVC task. Moreover, there is a rich and subtle variation of emotions in human speech. Therefore, in order to generate a more natural emotional voice, global-level and local-level emotion information should be considered simultaneously. The global-level emotion information can be extracted by encoding the whole utterance into a fixed-length vector. For local-level emotion information, instead of only using a fixed-length vector to represent the global-level emotion information of the whole utterance, the local-level emotion information should rely on single phoneme content and change with time.

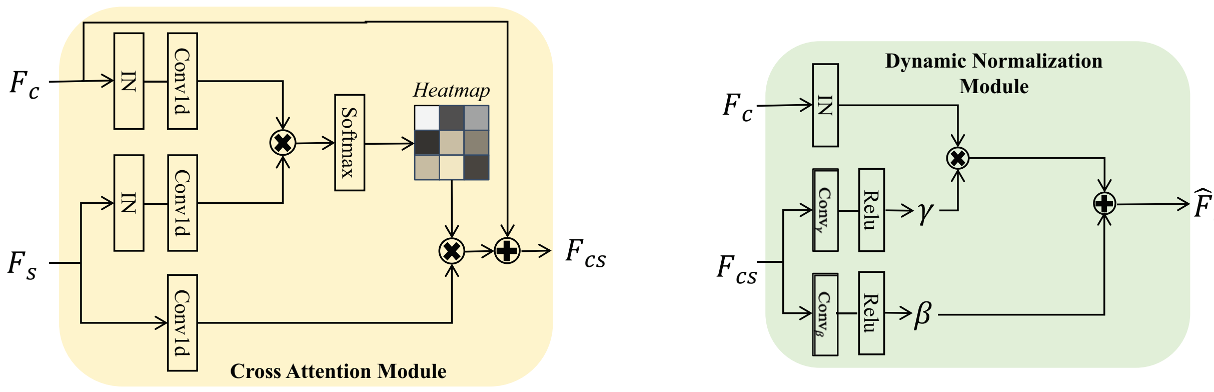

Fig. 2 Detailed structure of the proposed content-style fusion block, which is composed of a cross attention module and a dynamic normalization module.

Let F_s denote the emotion representation of target speech and it should depend on the source content representation F_c. First, the input features are normalized and transformed linearly, giving $Query(F_c)$, $Key(F_s)$ and $Value(F_s)$ respectively. Then we use $Query(F_c)$ and $Key(F_s)$ to calculate attention heatmap by aligning different phonemic speech content. Then we exploit a dynamic normalization module [14] to further improve the performance. Finally we can obtain the corresponding emotion feature $\hat{F}_{cs}$ which depends on F_c in the dynamic normalization module. Our content-style fusion block can appropriately embed an emotion feature which depends on the content information for another phoneme.

2.3 Objective Function

The training losses for the proposed method are described in this section.

Reconstruction loss: A reconstruction loss L_{REC} is calculated between the reconstructed mel-spectrogram and ground truth, which is adopted to generate reasonable speech using disentangled representations.

$$\mathcal{L}_{\text{rec}} = \|X_{1\to1} - X_1\|_1 \tag{2}$$

Adversarial loss: The adversarial loss is used to encourages the generator to generate realistic speech.

$$\mathcal{L}_{\text{adv}} = \mathbb{E}[\log D(X_2) + \log(1 - D(X_{1\to2}))] \tag{3}$$

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

$$\mathcal{L}_c = \|E_c(X_{1\to2}) - E_c(X_1)\|_1 \tag{4}$$

Style loss: The style loss is used for better emotion state transferring.

$$\begin{aligned}\mathcal{L}_s = \|&CSFB(E_c(X_{1\to2}), E_s(X_{1\to2})) \\ &- CSFB(E_c(X_1), E_s(X_2))\|_1\end{aligned} \tag{5}$$

The full objective function can be summarized as follows:

$$\mathcal{L}_{full} = \mathcal{L}_{adv} + \lambda_c\mathcal{L}_c + \lambda_s\mathcal{L}_s + \lambda_{rec}\mathcal{L}_{rec} \tag{6}$$

where λ_c, λ_s, and λ_{rec} are trade-off parameters.

3 Experiments

3.1 Experimental Conditions

We evaluated the proposed model with the Emotion Spseech Dataset (ESD) [13]. In this paper, we only consider four emotional categories of them: angry, happy, neutral, sad. 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 (300 utterances for training, 50 utterances for testing). We use MelGAN vocoder to generate audio waveforms from converted mel-spectrogram.

3.2 Objective Evaluations

In this paper, two comparative methods, StarGAN-EVC[6] and Atuo-EVC [7], were adopted for performance comparisons. Mel Cepstral Distortion (MCD) is used for the objective evaluation of spectral conversion. Moreover, Root Mean Square Error (RMSE) is used to evaluate the F0 conversion. For both MCD and F0-RMSE, a lower value indicates a smaller distortion or predicting error.

Figure 3 and Figure 4 show the MCD and F0-RMSE results from the neutral to emotional pairs respectively. Here, N2A, N2S, N2H represent the datasets neutral to angry voice, neutral to sad voice and neutral to happy voice, respectively. We can see that the proposed method can 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 comparative methods.

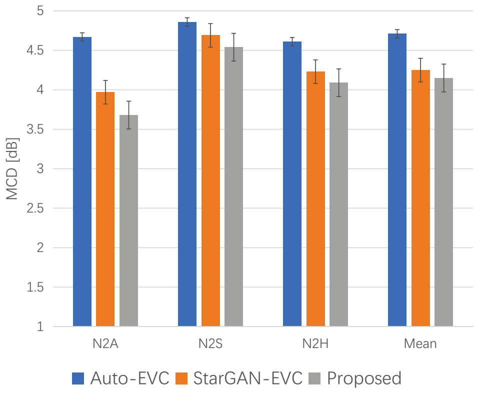

Fig. 3 MCD results for different emotions.

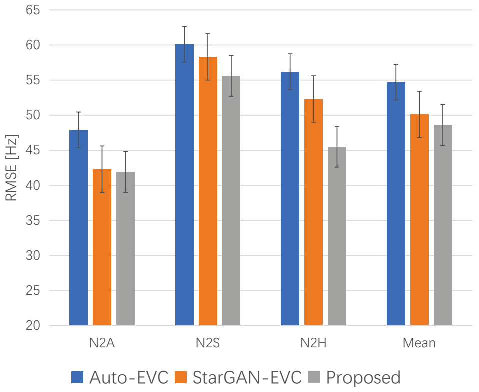

Fig. 4 F0-RMSE results for different emotions.

4 Conclusions

In this paper, we propose an emotional voice conversion framework with a novel content-style fusion block for rearranging the emotional style distribution. The proposed model can sufficiently learn the emotion information in both global-level and local-level. 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] Aihara *et al.*, "GMM-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing*, pp. 134-138, 2012.

[4] Luo *et al.*, "Emotional voice conversion using deep neural networks with MCC and F0 features," in *Proc. ICIS*, pp. 1-5, 2016.

[5] Zhou *et al.*, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," in *Proc. Odyssey*, pp. 230-237, 2020.

[6] 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.

[7] Gao *et al.*, "Nonparallel emotional speech conversion," in *Proc. INTERSPEECH*, pp. 2858-2862, 2019.

[8] Zhou *et al.*, "Limited data emotional voicec conversion leveraging text-to-speech: two-stage sequence-to-sequence training," in *Proc. INTERSPEECH*, pp. 811-815, 2021.

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

[10] Zhu *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, pp. 2223-2232, 2017.

[11] Zhu *et al.*, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. CVPR*, pp. 8789-8797, 2018.

[12] Zhou *et al.*, "Sequence-to-sequence emotional voice conversion with strength control," IEEE Access, vol. 9, pp. 42674-42687, 2021.

[13] Zhou *et al.*, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *Proc. ICASSP*, pp. 920-924, 2021.

[14] Jing *et al.*, "Dynamic instance normalization for arbitrary style transfer," in *Proc. AAAI*, pp. 4369-4376, 2020.