Towards Natural Emotional Voice Conversion with Novel Attention Module *

☆ 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) technology aims to convert a neutral voice into an emotional one while keeping the linguistic information and speaker identity unchanged. This technology is widely used in many real-world applications, such as voice assistants, conversational agents and sound design [1, 2].

Previously developed methods for EVC can be roughly categorized into two types based on the use of training data. The early studies of EVC mainly focused on parallel training data. That is, the mapping function is trained on paired utterances of the same linguistic content spoken in different emotion state. Among these approaches, a Gaussian Mixture Model (GMM) has been commonly used, and many improvements have been proposed for GMM-based EVC [6]. Other EVC methods, such as those based on non-negative matrix factorization (NMF) [7] or deep belief networks (DBNs) [8], have also been proposed. While these methods have demonstrated their effectiveness, they require accurately-aligned parallel data. Collecting parallel data and aligning the source and target utterances can be costly and time-consuming. These limitations have motivated research to explore non-parallel EVC approaches.

Recently, many non-parallel EVC methods based on deep learning have be proposed, such as GAN-based models [9, 5, 10] and autoencoder-based [11, 13] models. CycleGAN-EVC [9] and StarGAN-EVC [5] have both employed cycle-consistency to ensure the invertible mapping that results is identical with the source input. Recent works have shown that disentangled representations learning achieve remarkable performance in style transfer tasks. Gao *et al.* [11] use an auto-encoder framework to disentangle emotional style and linguistic content from the speech, and thus the emotional style could be modified independently without changing the linguistic content.

Although deep learning made the breakthrough in the EVC task, there still remains a gap between the converted speech and the real target in terms of quality and emotion fidelity. In many EVC systems, it is assumed that the linguistic content is dynamic and time-varying while the emotion information is static and time-independent. Therefore, 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. 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. 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.

To overcome the limitation, we propose a novel EVC model, which can sufficiently learn the emotion information in both global-level and local-level. For local-level emotion information, we adopt the attention mechanism for implementing time varying emotion representation. Thus, a novel attention module 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. If we are able to disentangle emotion information from content information, we can change the emotion state independently of the content.

2 Proposed method

2.1 Model Architecture

The proposed model is based on Generative Adversarial Network (GAN) [3] to have better gener-

* Towards Natural Emotional Voice Conversion with Novel Attention Module, 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)

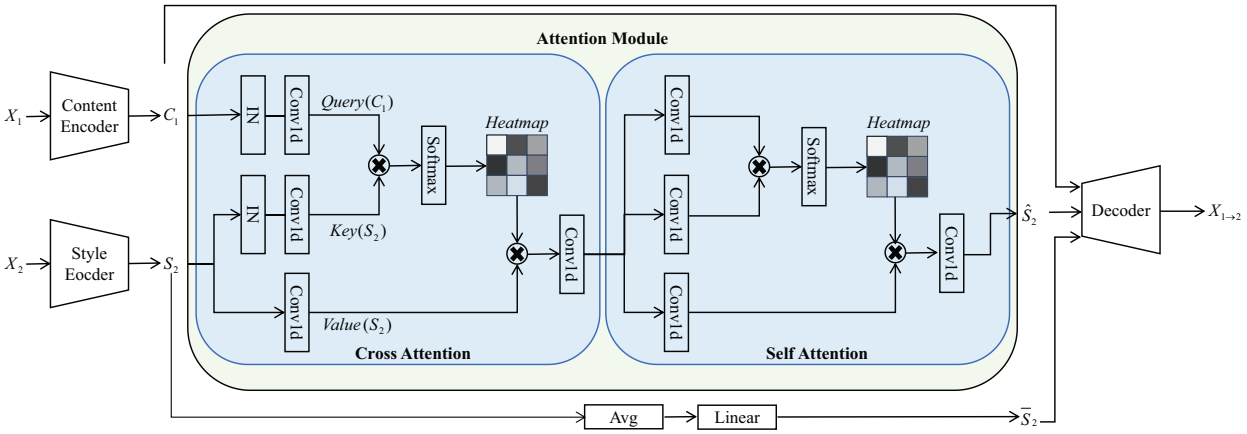

Fig. 1 The generator architecture of the proposed model. X_1 and X_2 indicate the mel-spectrogram of source and target speech respectively. IN is instance normalization.

alization on the converted speech. The generator is used to generate converted speech while the discriminator is adopted to distinguish real samples from machine-generated samples. 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 $E_s(\cdot)$, an attention module $Att(\cdot,\cdot)$, and a decoder $D_e(\cdot,\cdot,\cdot)$.

The content encoder $E_c(\cdot)$ is used to extract the content representation C_1 from the mel-spectrogram X_1 of source speech. The style encoder $E_s(\cdot)$ extracts the emotion representation S_2 from the mel-spectrogram X_2 of target speech. Then the attention module $Att(\cdot,\cdot)$ can generate content-dependent emotion representation $\hat{S}_2$, which will be explained in details in Section 2.2. Finally, the decoder $D_e(\cdot,\cdot,\cdot)$ will takes the content representation C_1, the phoneme-level emotion representation $\hat{S}_2$ and the averaged sentence-level emotion representation $\bar{S}_2$ 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 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 [4] to better capture the acoustic texture.

2.2 Attention Module

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. So 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, We assume that the emotion information is related to content. 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. The key idea is that using a novel attention module to perform local-level style embedding according to the content embedding.

The attention module is illustrated in Figure 1. As shown in this figure, the attention module is built with cross attention followed by self attention. Let S_2 denote the emotion representation of target speech and it should depend on the source content representation C_1. First, the input features are normalize and transformed linearly, giving $Query(C_1)$, $Key(S_2)$ and $Value(S_2)$ denoted by q, K and V respectively. Then we use q and K to calculate and attention heatmap by aligning different phonemic speech content. Then we exploit a self attention structure to further improve the performance. Finally we can obtain the corresponding emotion feature $\hat{S}_2$ which depends on C_1 by taking the dot product of value and the attention heatmap in self-attention.

The whole conversion process can be formulated

as follows:

$$\begin{aligned} &C_1 = E_c\left(X_1\right), S_2 = E_s\left(X_2\right), \\ &\bar{S_2} = AvgPooling\left(S_2\right), \hat{S_2} = Att\left(C_1, S_2\right), \\ &X_{1\to 2} = D_e\left(C_1, \hat{S_2}, \bar{S_2}\right), \end{aligned} \tag{1}$$

where X1 and X2 are the source and target speech respectively. To fuse the global-level emotion feature $\bar{S_2}$, 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 $\hat{S_2}$ is the all time step for the output feature and its length is the same as C_1.

Our attention module can appropriately embed an emotion feature which depends on the content information for another phoneme. For each time step of C_1, this attention mechanism can automatically align the most similar phonemic pronunciation of target speech S_2 and generate the target style features which depend on source speech content in a learnable manner.

2.3 Objective Function

Let X_1 and X_2 be mel-spectrogram belonging to source speech and target speech respectively. The training losses for the proposed method are described as follows:

Reconstruction loss: The reconstruction loss is adopted to generate reasonable speech using disentangled representations.

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

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

$$\mathcal{L}_{\text{adv}} = \mathbb{E}[\log D\left(X_2\right) + \log\left(1 - D\left(X_{1\to 2}\right)\right)] \tag{3}$$

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

$$\mathcal{L}_{\text{c}} = \left\|E_c\left(X_{1\to 2}\right) - E_c\left(X_1\right)\right\|_1 \tag{4}$$

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

$$\begin{aligned} \mathcal{L}_s = \|&Att\left(E_c\left(X_{1\to 2}\right), E_s\left(X_{1\to 2}\right)\right) \\ &- Att\left(E_c\left(X_1\right), E_s\left(X_2\right)\right)\|_1 \end{aligned} \tag{5}$$

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) [12]. The ESD dataset contains 350 sentences read by 10 native English speakers with five different emotions. 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. We trained the proposed model by ADAM optimizer with 0.0001 as learning rate. The baseline model is a StarGAN-based EVC model [5].

3.2 Objective Evaluations

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

$$\text{MCD} = (10/\ln 10)\sqrt{2\sum_{i=1}^{24}\left(mc_i^t - mc_i^c\right)^2} \tag{7}$$

Here, mc_i^t and mc_i^c represent the target and the converted mel-cepstral, respectively.
Morever, Root Mean Square Error (RMSE) was used to evaluate the F0 conversion.

$$\text{F0-RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left(F0_i^t\right) - \left(F0_i^c\right)\right)^2}. \tag{8}$$

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

Figure 2 and Figure 3 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 baseline.

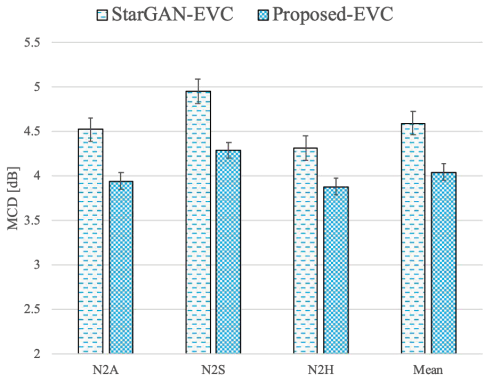

Fig. 2 MCD results for different emotions.

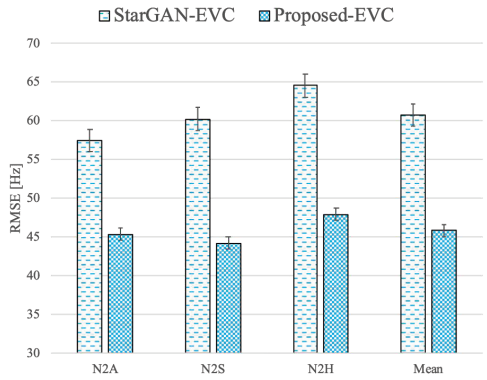

Fig. 3 F0-RMSE results for different emotions.

4 Conclusions

In this paper, we propose an emotional voice conversion framework with a novel attention module for realistic and natural speech conversion. The proposed model can sufficiently learn the emotion information in both global-level and local-level. For local-level emotion information, a novel attention module is proposed to implement the implicit alignment for emotion and phoneme content, further embedding the phoneme-level emotion representation. For global-level emotion information, the complete set of time steps of speech emotion is embedded into a fixed-length vector to obtain the sentence-level emotion representation. The experimental results show the effectiveness of our proposed method.

ACKNOWLEDGMENT This work was supported in part by JSPS KAKENHI (Grant No. 21H00906).

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

[4] 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.

[5] 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.

[6] Aihara *et al.*, "GMM-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing*, pp. 134-138, 2012.

[7] Aihara *et al.*, "Exemplar-based emotional voice conversion using non-negative matrix factorization," *APSIPA*, pp. 1-7, 2014.

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

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

[10] Cao *et al.*, 'Nonparallel Emotional Speech Conversion Using VAE-GAN," in *Proc. INTERSPEECH*, pp. 3406-3410, 2020.

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

[12] 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.

[13] 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.