Emotional Voice Conversion Using Disentangled Representation Learning and Attention Mechanism *

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

1 Introduction

The human speech is a complex signal that contains rich information. A listener perceives not only linguistic information from a speech but also speaker identity, emotional information, etc. Emotional voice conversion (EVC) is the task of converting speech from one emotion state into another one while keeping the linguistic information and speaker identity unchanged. It is an enabling technique for many real-world applications, such as voice assistants, conversational agents and sound design [1] [2]. Therefore, there has been tremendous active research in EVC recently.

Many statistical approaches have been proposed for EVC in the past few decades. Among these approaches, a Gaussian Mixture Model (GMM) has been commonly used, and many improvements have been proposed [6] for GMM-based EVC. Other EVC methods, such as those based on non-negative matrix factorization (NMF) [7], have also been proposed. Meanwhile, some deep learning approaches construct nonlinear mapping relationships using neural networks (NNs) to train the mapping dictionaries between the source and target features [8], whereas others use deep belief networks (DBNs) to achieve non-linear deep transformation [9]. 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.

There have been studies on deep learning approaches for EVC that do not require parallel training data, such as CycleGAN-based [12], StarGAN-based [5] and autoencoder-based [13] frameworks. These works are inspiring, but 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 the emotion representation is often modeled as a fixed-size vector. This would be a reasonable modeling strategy. However, only using fixed-size representation of the emotion by an utterance has two issues to be considered. First, since speech signals dynamically change in time, some parts of emotion information also would change in time. Considered the differences of mechanisms of speech production, vowels and consonants would convey different aspects of emotion information. From a view point of applications, silence parts of the signals, which hardly convey emotion information, should be treated differently. Second, only using a fixed-size vector as emotion representation causes a loss of information, and rich emotion information in speech would be compressed into a predefined capacity.

With consideration for these issues, we propose to use time-varying emotion representation for EVC. For extraction of the time-varying emotion information, the functions of content and emotion extractors should interlock each other. The novel feature of the proposed is to simultaneously embed sentence-level and phoneme-level emotion information. To achieve the concept, we adopt an 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 a phoneme-level emotion representation. In addition, we consider embedding 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 linguistic content information, we can change the emotion state independently of the linguistic content. It should be noted that the proposed method does not require any pre-trained models, and is only trained with non-parallel speech data.

---

*Emotional Voice Conversion Using Disentangled Representation Learning and Attention Mechanism, 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)

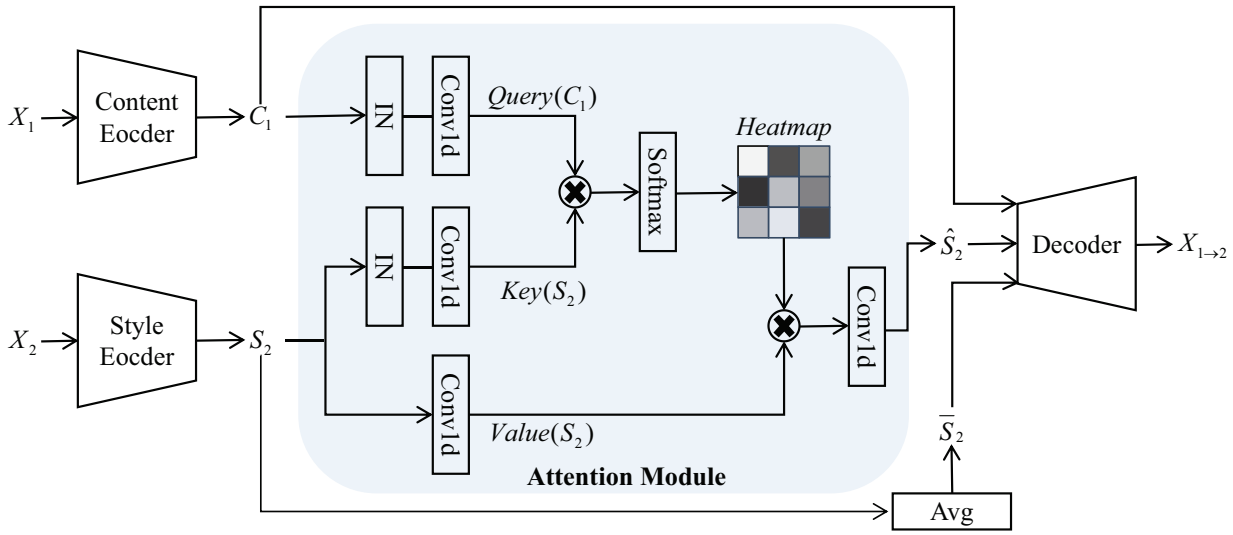

Fig. 1 The generator architecture of the proposed model. X_1 and X_2 indicate the mel-spectrogram of source and target speech respectively. IN is instance normalization.

2 Proposed method

2.1 Model Architecture

We propose a GAN-based [3] to have better generalization on the converted speech. The effectiveness of GANs is due to the fact that an adversarial loss forces the generated data to be indistinguishable from real data. The generator is used to generate converted speech while the discriminator is adopted to distinguish real samples from machine-generated samples. As show in Figure 1, the generator is an encoder-decoder module in our work. The generator 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 generator is made up entirely of convolution layers in order to operate in a non-autoregressive generative manner.

In the conversion process, the content encoder $E_c(\cdot)$ captures the linguistic content information C_1 from the mel-spectrogram X_1 of source speech. The style encoder $E_s(\cdot)$ is adopted to produce an 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 information $\hat{S}_2$, which will be explained in details in Section 2.2. Finally, the decoder $D_e(\cdot,\cdot,\cdot)$ will takes the content embedding C_1, the phoneme-level emotion representation $\hat{S}_2$ and the averaged sentence-level emotion representation $\overline{S}_2$ 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 $E_s(\cdot)$ 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. $E_c(\cdot)$ is composed of convolution layers. In addition, we adopt Instance normalization (IN) after each convolution layer of the content encoder to eliminate emotion information. A PixelShuffle layer is used in $D_e(\cdot,\cdot,\cdot)$ for upsampling. Unlike the generator, the discriminator is constructed with 2d convolution layers like [4] to better capture the acoustic texture.

2.2 Attention Mechanism

To obtain phoneme-level emotion information, the key idea is an attention mechanism relating emotion to content. Our approach assumes that the emotion information is related to content, so instead of only using a fixed-length vector to represent the emotion of the whole utterance, the emotion information should rely on single phoneme content and change with time.

As shown in Figure 1.(a), a novel attention module has been developed to meet the hypothesis above. Accordingly, let S_2 denote the emotion information of target speech and it should depend on the source content representation C_1. First, we normalize the input features and transform them 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. Subsequently we calculate the corresponding emotion feature $\hat{S}_2$ which depends on C_1 by taking the dot product of V and the attention heatmap. Mathematically, we express this as follows:

$$\hat{S}_2(t) = \frac{\sum_{t'=1}^{T'} \exp\left(q^T(t)K\left(t'\right)\right) V\left(t'\right)}{\sum_{t'=1}^{T'} \exp\left(q^T(t)K\left(t'\right)\right)} \cdot W_o \quad (1)$$

where $q = W_f \cdot IN\left(C_1\right)$, $K = W_h \cdot IN\left(S_2\right)$ and $V = W_g \cdot \left(S_2\right)$, and IN indicates a mean–variance channel-wise normalization to eliminate emotion information. Here $\hat{Z}_2(t)$ is the t^{th} time step for the output feature and its length is the same as C_1. If T' is the length of S_2, then t' is the index that enumerates all time steps of the target speech. Further, W_f, W_g, W_h and W_o above denote the learned weight matrices, which are implemented as Conv1d layer in which both kernel and stride are of unit length.

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.

$$\mathcal{L}_{\text{rec}} = \left\| G\left(X_1, X_1\right) - X_1 \right\|_1 \quad (2)$$

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

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

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

$$\mathcal{L}_{\text{c}} = \left(\left\| E_c\left(X_{1\to2}\right) - E_c\left(X_1\right) \right\|_2\right) \quad (4)$$

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

$$\begin{aligned}\mathcal{L}_s = \| &Att\left(E_c\left(X_{1\to2}\right), E_s\left(X_{1\to2}\right)\right) \\ &- Att\left(E_c\left(X_1\right), E_s\left(X_2\right)\right) \|_2\end{aligned} \quad (5)$$

The full objective function can be summarized as follows:

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

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

3 Experiments

3.1 Experimental Conditions

We conduct experimental evaluations on the ESD database [14], which contains 350 parallel utterances audio data recorded by 10 native English speakers with five different emotions. The baseline model is a StarGAN-based EVC model [5]. 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 (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 weighting parameters are simply set as $\lambda_c = 2$, $\lambda_s = 2$ and $\lambda_{rec} = 5$ in Eq. (6).

3.2 Objective Evaluations

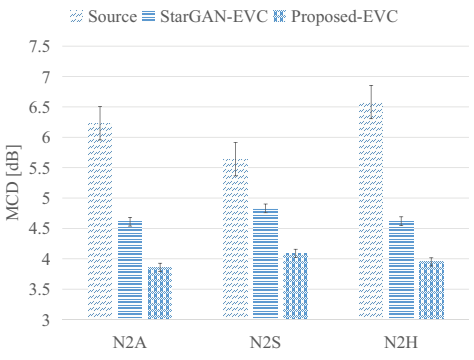

Fig. 2 MCD results for different emotions.

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

defined below.

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

In Eq. 7, 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

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((F0_i^t) - (F0_i^c))^2}, \quad (8)$$

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.

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.

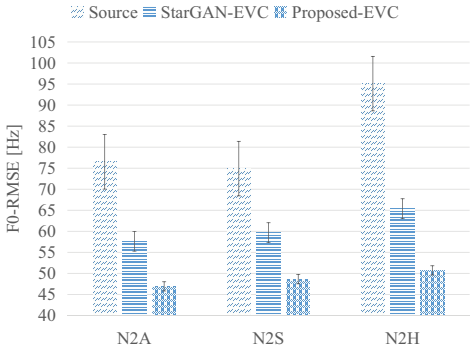

Fig. 3 F0-RMSE results for different emotions.

4 Conclusions

In this paper, we propose an emotional voice conversion framework with a novel attention module for realistic and natural speech conversion. 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. 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] 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] Lorenzo-Trueba *et al.*, "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis," *Speech Communication*, pp. 135-143, 2018.

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

[10] Desai *et al.*, "Voice conversion using artificial neural networks," in *Proc. ICASSP*, pp. 3893-3896, 2009.

[11] Nakashika *et al.*, "Voice conversion in high-order eigen space using deep belief nets," in *Proc. INTERSPEECH*, pp. 369-372, 2013.

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

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

[14] 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.