NONPARALLEL EMOTIONAL SPEECH CONVERSION

Jian Gao ⋆ *Deep Chakraborty* † *Olaitan Olaleye* ‡

⋆ Department of Computer Science and Engineering, New York University, USA
† College of Information and Computer Sciences, University of Massachusetts, Amherst, USA
‡ Affiliation Number Three

ABSTRACT

We propose a nonparallel data-driven emotional speech conversion framework. The goal is to change emotion-related characteristics of a speech signal while preserving the speaker identity and linguistic content. Most existing approaches require parallel data and time alignment, which is not only tedious work to prepare the data but also a source of quality degradation due to misalignment. We achieve nonparallel training based on an unsupervised style transfer technique, which learns a translation model between two distributions instead of a deterministic one-to-one mapping between paired data. The translation model consists of a encoder and a decoder for each emotion domain. We assume that the speech signal can be decomposed into an emotion-invariant content code and an emotion-related style code in some latent space. Emotion conversion is performed by recombining the content code of the source speech and the style code of the desired emotion. We tested our method on a nonparallel corpora with four emotions. Both subjective and objective evaluations show the effectiveness of our approach.

Index Terms— Emotional speech conversion, nonparallel data, style transfer, autoencoder, GAN

1. INTRODUCTION

Voice transformation (VT) is a technique to modify some properties of human speech while preserving its linguistic information. VT can be applied to change the speaker identity, i.e., voice conversion (VC) [1], or to transform the speaking style of a speaker, such as emotion and accent conversion [2]. In this work, we will focus on voice transformation for emotional speech. The goal is to change emotion-related characteristics of a speech signal while preserving its linguistic content and speaker identity. Emotion conversion techniques can be applied to various tasks (need one sentence, and references)

Traditional VC approaches cannot be directly applied because they change speaker identity by assuming pronunciation and intonation to be a part of the speaker-independent information. Since the speaker's emotion is mainly conveyed by prosodic aspects, some studies focus on modeling prosodic features such as pitch, pause, stress, tempo and volume [3][4][5]. In [6], a rule-based emotional voice conversion system was proposed by modifying prosody-related acoustic features of neutral speech to generate different kinds of emotional speech. It uses speech analysis-synthesis tool STRAIGHT [7] to extract fundamental frequency (F_0) and power envelope, and then parameterize them by Fujisaki model [8] and target prediction model [9]. The modified features are then fed into STRAIGHT to re-synthesis speech waveforms with desired emotions. However, this method not only needs parallel data for training, but also requires time modification to align the speech duration of each utterance pair. The duration-related features are extracted by manual segmentation of the speech signal at phoneme level. Though this approach did successfully convert emotions, it is hard to obtain adequate time-aligned utterance pairs and also inconvenient in real applications due to the manual operation.

To address this issue, we propose a method for nonparallel training. Instead of learning one-to-one mapping between paired emotional utterances, we switch to training a conversion model between two emotional domains. Let $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ be utterances drawn from two different emotional categories, our model learns a mapping between two distributions $p(x_1)$ and $p(x_2)$. Since the joint distribution $p(x_1, x_2)$ is unknown, we cannot directly estimate the conversion models $p(x_1|x_2)$ and $p(x_2|x_1)$.

Inspired by disentangled representation learning in image style transfer [10], we assume that each speech signal $x_i \in \mathcal{X}_i$ can be decomposed into a content latent code $c \in \mathcal{C}$ that represents emotional-independent information and a style latent code $s_i \in \mathcal{S}_i$ that represents emotion-related information. $\mathcal{C}$ is shared across domains that contains the information we want to preserve, and $\mathcal{S}_i$ is domain specific that contains the information we want to change. In conversion stage, we extract content code of the source speech and recombine it with style code of the desired emotion. A generative adversarial network (GAN) [16] is added to improve the quality of converted speech. Our approach is nonparallel, text-independent, and does not rely on any external data, preprocessing or manual operation. The model can be trained on a small amount of utterances ($\sim$ 8 min per emotion).

We evaluated our approach on IEMOCAP [17] with four

emotions: angry, happy, neutral, sad. An objective evaluation showed that our model can modify the speech to significantly increase the percentage of desired emotions. A subjective evaluation on Amazon MTurk showed that the converted speech had good quality and preserved the speaker identity.

The rest of the paper is organized as follows: Section 2 presents the relation to prior work. Section 3 gives a detailed description of our model. Experiment and evaluation results are reported in Section 4. Finally, we conclude in Section 5.

2. RELATED WORK

2.1. Emotion-related features

Previous emotion conversion methods directly modify parameterized prosody-related features that convey emotions. [11] first proposed to use Gaussian mixture models (GMM) for spectrum transformation. [12] introduced a data-driven emotion conversion system that combines independent parameter transformation techniques including HMM-based F_0 generation, F_0 segment selection, duration conversion and GMM-based spectral conversion. However, it requires large amounts of parallel data.

A recent work [6] explored four types of acoustic features: F_0 contour, spectral sequence, duration and power envelope, and investigated their impact on emotional speech synthesis by controlled feature replacement. They found that F_0 and spectral sequence are the dominant factors in emotion conversion, while power envelope and duration alone has little influence. They further claimed that all emotions can be synthesised by modifying the spectral sequence, but they did't know how to do it. In this paper, we focus on learning the conversion models for F_0 and spectral sequence. Since changing duration and power envelope requires manually segmenting the phoneme boundaries of vowels and consonants, we leave it for future work.

2.2. Nonparallel training approaches

Parallel data means utterances with the same linguistic content but vary in aspects to be studied. Since parallel data is hard to collect, nonparallel approaches have been developed. Some borrow ideas from image-to-image translation and create models suitable for speech, such as VAE-based SC [22], VC-VAW-GAN [18], VC-CycleGAN [20][21] and VC-StarGAN [25]. Another trend is to use WaveNet architecture [23], which is an autoregressive model for raw audio synthesis. Although it can generate high-quality speech, the huge amount of computation resource and training data is not affordable for most users.

2.3. Disentangled representation learning

Our work draws inspiration from recent studies in image style transfer. A basic idea is to find disentangled representations that can independently model image content and style. It is claimed in [10] that Convolutional Neural Network (CNN) is an ideal representation to factorize semantic content and artistic style. They introduced a method to separate and recombine content and style of natural images by matching feature correlations in different convolutional layers. For us, the task is to find disentangled representations for speech signal that can split emotion from speaker identity and linguistic content.

3. METHOD

3.1. Assumptions

The research of human emotion expression and perception has two major conclusions. First, human emotion perception is a multi-layered process. For example, [26] figured out that humans do not perceive emotion directly from acoustic features, but through an intermediate layer of semantic primitives. They introduced a three-layered model and built the connections by a fuzzy inference system. Some researchers found that adding middle layer can improve the emotion recognition accuracy. Based on this finding, we suggest to use multilayer perceptron (MLP) to encode emotion-related information in speech.

Second, the emotion production of human speech follows the opposite direction of emotion perception. This means the encoding process of the speaker is the inverse operation of the decoding process of the listener. We assume that emotion speech production and perception share the same representation methodology, therefore the encoder and the decoder have mirror structures.

Define $\mathcal{X}_i$ as the speech domain with emotion i. Given parallel data, we can draw sample pairs (x_1, x_2) from the joint distribution $P(x_1, x_2)$, $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2$. It is easy to learn the conditions $P(x_2|x_1), P(x_1|x_2)$. However, nonparallel data only give us the marginal distributions $P(x_1)$ and $P(x_2)$. Since estimating the joint distribution from marginal distribution is an ill-posed problem, additional conditions are required. We assume that the speech signal can be decomposed into a domain-invariant content code and a domain-dependent style code.

Fig. 1 shows the generative model of speech with a partially shared latent space. A pair of corresponding speech (x_1, x_2) is assumed to have a shared latent code $c \in \mathcal{C}$ and emotion-related style codes $s_1 \in \mathcal{S}_1, s_2 \in \mathcal{S}_2$. For any emotional speech x_i, we have a deterministic decoder $x_i = G_i(c_i, s_i)$ and its inverse encoders $c_i = E_i^c(x_i)$, $s_i = E_i^s(x_i)$. To convert emotion, we just recombine the content code of the source speech with the style code of the target emotion.

$$\begin{aligned} x'_{1\leftarrow 2} &= G_1(c_2, s_1) = G_1(E_2^c(x_2), s_1) \\ x'_{2\leftarrow 1} &= G_2(c_1, s_2) = G_2(E_1^c(x_1), s_2) \end{aligned} \tag{1}$$

It should be noted that the style code s_i is not inferred from

one utterance, but learnt from the entire emotion domain. Because the emotion style from a single utterance is ambiguous and may not capture the general character of the target emotion. This makes our assumption slightly different from the cycle consistent constraint [19]. When an utterance is converted to another emotion and converted back, it does not need to be exactly the same as the original signal, i.e., $x''_{1\leftarrow 2\leftarrow 1} \neq x_1$. Instead, we apply the cycle-consistency in the shared latent space by assuming that $E^c_1(x'_{1\leftarrow 2}) = c_1$.

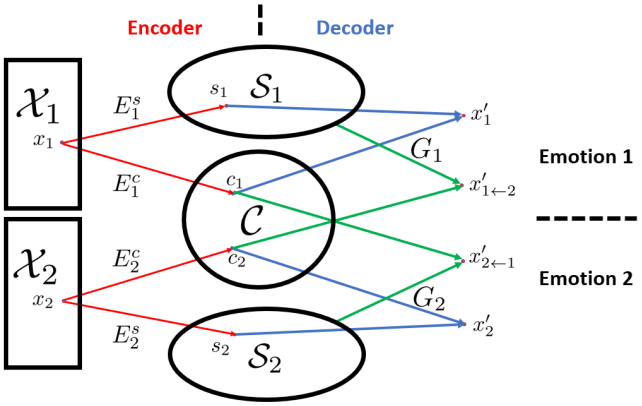

Fig. 1. Speech autoencoder model with partially shared latent space. Utterance with emotion i is decomposed into a emotion-dependent space $\mathcal{S}_i$ and a shared content space $\mathcal{C}$. Emotion conversion is conducted by recombining the content code of the input utterance and the style code randomly sampled from the target emotion space.

3.2. Model

Traditional emotional speech analysis mainly works on four types of acoustic features: fundamental frequency (F_0), spectral sequence, time duration and energy envelope. It was found in [6] that only F_0 and spectral sequence have significant influence, while the other two require manual segmentation and have little impact on changing the emotions. Therefore we focus on learning the conversion model for F_0 and spectral sequence. Fig. 2 shows an overview of our nonparallel emotional speech conversion system. The acoustic features are extracted and recombined by WORLD analysis system [27] and converted separately.

We convert F_0 by linear transform to match statistics of the fundamental frequencies in the target emotion domain. The conversion is performed by logarithm Gaussian normalization

$$f_2 = \exp((\log f_1 - \mu_1) \cdot \frac{\sigma_2}{\sigma_1} + \mu_2) \tag{2}$$

where μ_i, σ_i are the mean and variance obtained from the target emotion set. Aperiodicity (AP) is mapped directly since it does not contain emotion-related information.

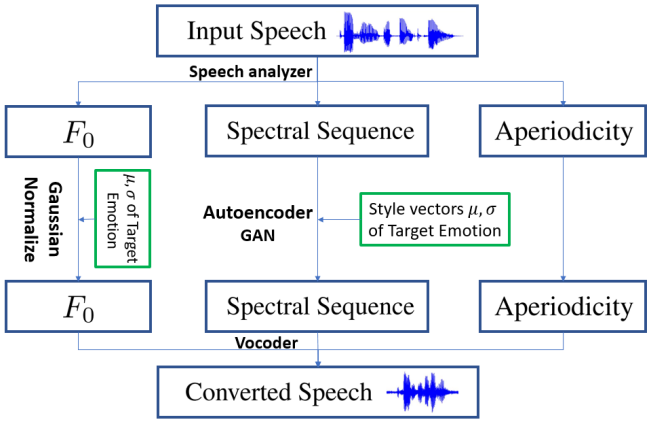

Fig. 2. Overview of nonparallel emotion conversion system

For spectral sequence, we use low-dimensional representation in mel-cepstrum (MCEP) domain to reduce computation complexity. [28] shows that 50 MCEP coefficients are enough to synthesize full-band speech without quality degeneration. Spectra conversion is learnt by the autoencoder model in Fig. 1. The encoders and decoders are implemented with gated convolutional neural networks [13]. In addition, two generative adversarial networks (GANs) were added to produce realistic spectral frames. Our model has 4 subnetworks E^c, E^s, G, D, in which D is the discriminator in GANs to distinguish between real samples and machine-generated samples.

3.3. Loss functions

We jointly train the encoders, decoders and GAN's discriminators with multiple losses displayed in Fig. 3. To keep encoder and decoder as inverse operations, we apply reconstruction loss in the direction $x_i \rightarrow (c_i, s_i) \rightarrow x'_i$. The spectral sequence should not change after encoding and decoding.

$$L^{x_i}_{recon} = \mathbb{E}_{x_i}(\|x_i - x'_i\|_1), \quad x'_i = G_i(E^c_i(x_i), E^s_i(x_i)) \tag{3}$$

In our model, the latent space is partially shared. Thus the cycle consistency constraint [19] is not preserved, i.e., $x''_{1\leftarrow 2\leftarrow 1} \neq x_1$. We apply a semi-cycle loss in the coding direction $c_1 \rightarrow x'_{2\leftarrow 1} \rightarrow c'_{2\leftarrow 1}$ and $s_2 \rightarrow x'_{2\leftarrow 1} \rightarrow s'_{2\leftarrow 1}$.

$$\begin{aligned} L^{c_1}_{cycle} &= \mathbb{E}_{c_1,s_2}(\|c_1 - c'_{2\leftarrow 1}\|_1), \quad c'_{2\leftarrow 1} = E^c_2(x'_{2\leftarrow 1}) \\ L^{s_2}_{cycle} &= \mathbb{E}_{c_1,s_2}(\|s_2 - s'_{2\leftarrow 1}\|_1), \quad s'_{2\leftarrow 1} = E^s_2(x'_{2\leftarrow 1}) \end{aligned} \tag{4}$$

Moreover, we add a GAN module to improve the speech quality. The converted samples should be indistinguishable from the real samples in the target emotion domain. GAN loss is computed between $x'_{i\leftarrow j}$ and x_i, $(i \neq j)$.

$$L^i_{GAN} = \mathbb{E}_{c_j,s_i}[\log(1 - D_i(x'_{i\leftarrow j}))] + \mathbb{E}_{x_i}[\log D_i(x_i)] \tag{5}$$

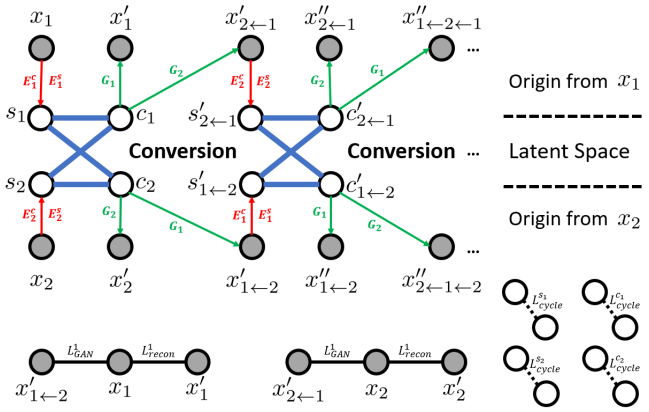

Fig. 3. Train on multiple loss functions

The full loss is the weighted sum of L_{recon}, L_{cycle}, L_{GAN}.

$$\begin{aligned}\min_{E_1^c,E_1^s,E_2^c,E_2^s,G_1,G_2} \max_{D_1,D_2} L(E_1^c, E_1^s, E_2^c, E_2^s, G_1, G_2, D_1, D_2) \\ = \lambda_s(L_{cycle}^{s_1} + L_{cycle}^{s_2}) + \lambda_c(L_{cycle}^{c_1} + L_{cycle}^{c_2}) \\ + \lambda_x(L_{recon}^1 + L_{recon}^2) + \lambda_g(L_{GAN}^1 + L_{GAN}^2)\end{aligned} \tag{6}$$

where $\lambda_s, \lambda_c, \lambda_x, \lambda_g$ control the weights of the components.

4. EXPERIMENTS

4.1. Corpus

We test the proposed method on IEMOCAP [17], which is a widely used corpus for emotion recognition. To our knowledge, this is the first work to use it for emotion conversion. IEMOCAP contains scripted and improvised dialogs in five sessions; each has labeled emotional sentences pronounced by two English speakers. The emotions in scripted dialogs have strong correlation with the lingual content. Since our task is to change emotion but keep the speaker identity and linguistic content, we only use the improvised dialogs of the same speaker. We train the conversion model on four emotions: angry, happy, neutral, sad. The acoustic features F_0, spectral sequence and AP are extracted by WORLD [27] every 5 ms, then coded to 24-dimension Mel-cepstral vectors of temporal size 128 as the autoencoder's input.

4.2. Network Architecture

Our network structure is illustrated in Fig. 4. The encoders and decoders are implemented with one-dimension CNNs to capture the temporal dependencies; the GAN discriminators are implemented with two-dimension CNNs to capture the spectra-temporal patterns. All networks are equipped with gated linear units (GLU) as activation functions. The emotion style is learnt by a MLP with 3 linear layers and output channel-wise mean and variance $\mu(s), \sigma(s)$. Then they are fed into the decoder by adding an adaptive instance normalization (AdaIN) [28] layer before activation. This mechanism is similar to the conversion model of F_0 (Eq.2).

$$\text{AdaIN}(c, s) = \sigma(s)\left(\frac{c - \mu(c)}{\sigma(c)}\right) + \mu(s) \tag{7}$$

4.3. Objective evaluation

4.4. Subjective evaluation

We conducted listening tests on Amazon MTurk to evaluate the converted speech [1].

5. CONCLUSION

We presented a nonparallel emotional speech conversion system. Objective and subjective evaluations showed that our model can successfully manipulate emotions to fool the emotion classifier as well as human listeners. As our approach does not require any paired data, transcripts or time alignment, it is easy to be applied in real-world situations. One limitation is that the current model is restricted to one specific speaker. Future work is to develop a general model that can do emotion conversion for unseen speakers. Additionally, we will build a larger database for training deep neural networks.

6. REFERENCES

[1] Seyed Hamidreza Mohammadi and Alexander Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[1] We provide converted samples at https://www.jian-gao.org/emovc

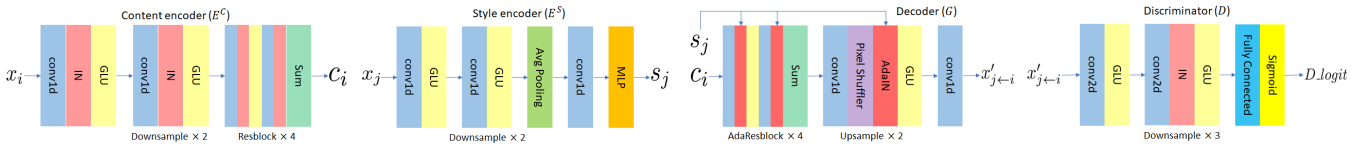

Fig. 4. The network architecture of content encoder, style encoder, decoder, and GAN discriminator.