NONPARALLEL EMOTIONAL SPEECH CONVERSION USING STYLE TRANSFER

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

⋆ Department of Computer Science and Engineering, New York University, USA
† College of Information and Computer Sciences, University of Massachusetts Amherst, USA
‡ Signify (formerly Philips Lighting) Research, North America, USA

ABSTRACT

We propose a nonparallel data-driven emotional speech conversion method. It enables the transfer of emotion-related characteristics of a speech signal while preserving the speaker's identity and linguistic content. Most existing approaches require parallel data and time alignment, which is not available in most real applications. 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 examples. The conversion model consists of an 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 latent space. Emotion conversion is performed by extracting and recombining the content code of the source speech and the style code of the target 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 training, Style Transfer, Autoencoder, GANs

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 emotion voice transformation. 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, such as hiding negative emotions for customer service agents, helping film dubbing, and creating more expressive voice messages on social media.

Traditional VC approaches cannot be applied directly 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 have focused on modelling prosodic features such as pitch, tempo, and volume [3, 4]. In [5], a rule-based emotional voice conversion system was proposed. It modifies prosody-related acoustic features of neutral speech to generate different types of emotions. A speech analysis-synthesis tool STRAIGHT [6] was used to extract fundamental frequency (F_0) and power envelope from raw audio. These features were parameterized and modified based on Fujisaki model [7] and target prediction model [8]. The converted features were then fed back into STRAIGHT to resynthesize speech waveforms with desired emotions. However, this method requires temporal aligned parallel data that is difficult to obtain in real applications; and the accurate time alignment needs manual segmentation of the speech signal at phoneme level, which is very time consuming.

To address these issues, we propose a nonparallel training method. Instead of learning one-to-one mapping between paired emotional utterances (x_1, x_2), we switch to training a conversion model between two emotional domains $(\mathcal{X}_1, \mathcal{X}_2)$.

Inspired by disentangled representation learning in image style transfer [9], we assume that each speech signal $x_i \in \mathcal{X}_i$ can be decomposed into a content code $c \in \mathcal{C}$ that represents emotion-invariant information and a style code $s_i \in \mathcal{S}_i$ that represents emotion-dependent information. $\mathcal{C}$ is shared across domains and contains the information we want to preserve. $\mathcal{S}_i$ is domain-specific and 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 target emotion. A generative adversarial network (GAN) [10] is added to improve the quality of converted speech. Our approach is nonparallel, text-independent, and does not rely on any manual operation.

We evaluated our approach on IEMOCAP [11] for four emotions: angry, happy, neutral, sad, which are widely studied in emotion speech analysis literatures. 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. [12] first proposed to use Gaussian mixture models (GMM) for spectrum transformation. A recent work [5] explored four types of acoustic features: F_0 contour, spectral sequence, duration and power envelope, and investigated their impact on emotional speech synthesis The authors 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 synthesized by modifying the spectral sequence, but did not provide a method to do it. In this paper, we focus on learning the conversion models for F_0 and spectral sequence.

2.2. Nonparallel training approaches

Parallel data means utterances with the same linguistic content but varying 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 VC-VAW-GAN [13] and VC-CycleGAN [14]. Another trend is based on WaveNet [15]. Although it can train directly on raw audio without feature extraction, the huge amount of computation resources and training data required 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 [9] 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 on human emotion expression and perception has two major conclusions. First, human emotion perception is a multi-layered process. [16] 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 learnt the connections by a fuzzy inference system. Some researchers found that adding middle layers can improve emotion recognition accuracy. Based on this finding, we suggest to use multilayer perceptrons (MLP) to extract emotion-related information in speech signals.

Second, the emotion generation process 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 generation and perception share the same representation methodology. This means the encoder and decoder are inverse operations with mirror structures.

Let $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ be utterances drawn from two different emotional categories. Our goal is to learn a mapping between two distributions $p(x_1)$ and $p(x_2)$. Since the joint distribution $p(x_1, x_2)$ is unknown for nonparallel data, the conversion models $p(x_1|x_2)$ and $p(x_2|x_1)$ cannot be directly estimated. To solve this problem, we make two assumptions: (i). The speech signal can be decomposed into an emotion-invariant content code and an emotion-dependent style code; (ii). The encoder E and decoder G are inverse functions.

3.2. Model

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 extract and 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. This is because the emotion style from a single utterance is ambiguous and may not capture the general characteristics of the target emotion. It makes our assumption slightly different from the cycle consistent constraint [17], which assumes that an example converted to another domain and converted back should remain the same as the original, i.e., $x''_{1\leftarrow 2\leftarrow 1} = x_1$. Instead, we apply a semi-cycle consistency in the latent space by assuming that $E_1^c(x'_{1\leftarrow 2}) = c_1$ and $E_1^s(x'_{1\leftarrow 2}) = s_1$.

Traditional emotional speech analysis mainly focuses on four types of acoustic features: fundamental frequency (F_0), spectral sequence, time duration and energy envelope. It was found in [5] that only F_0 and spectral sequence have significant influence, while the other two require manual segmentation and have little impact on changing 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 features are extracted and recombined by WORLD [18] and converted separately. We modify F_0 by linear transform to match statistics of the

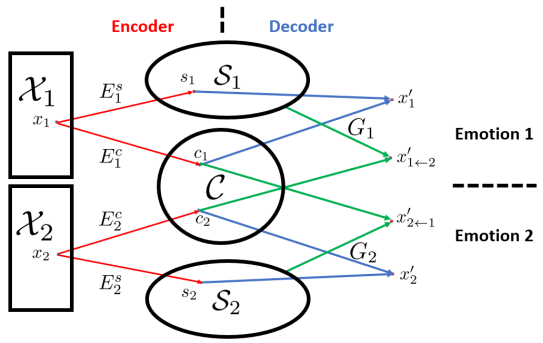

Fig. 1. Speech autoencoder model with partially shared latent space. Speech with emotion i is decomposed into an emotion-specific space $\mathcal{S}_i$ and a shared content space $\mathcal{C}$. Corresponding speech (x_1, x_2) are encoded to the same content code c.

fundamental frequencies in the target emotion domain. The conversion is performed by log 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 source and target emotion set. Aperiodicity (AP) is mapped directly since it does not contain emotion-related information.

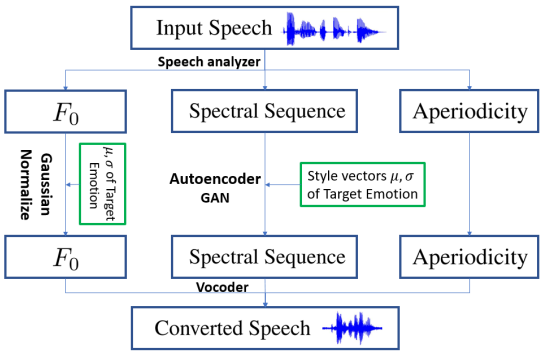

Fig. 2. Overview of nonparallel emotion conversion system

For spectral sequence, we use low-dimensional representation in mel-cepstrum domain to reduce 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 CNN [19]. In addition, a GAN module is added to produce realistic spectral frames. Our model has 4 subnetworks E^c, E^s, G, D, in which D is the discriminator in GAN 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 $x_i \to (c_i, s_i) \to x'_i$. The spectral sequence should not change after encoding and decoding.

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

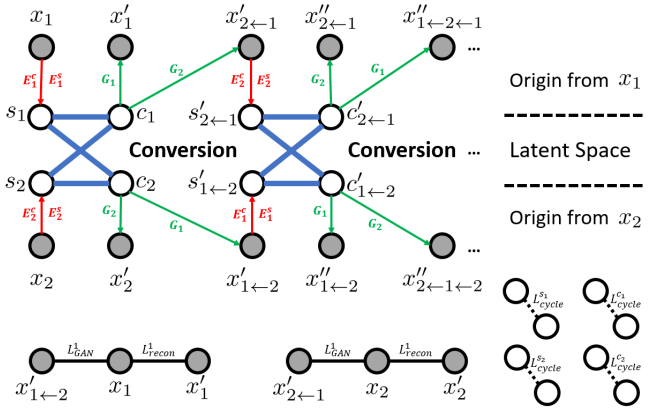

Fig. 3. Train on multiple loss functions

In our model, the latent space is partially shared. Thus the cycle consistency constraint [17] is not preserved, i.e., $x''_{1\leftarrow 2\leftarrow 1} \neq x_1$. We apply a semi-cycle loss in the coding direction $c_1 \to x'_{2\leftarrow 1} \to c'_{2\leftarrow 1}$ and $s_2 \to x'_{2\leftarrow 1} \to s'_{2\leftarrow 1}$.

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

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, $(i \neq j)$.

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

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

$$\begin{aligned} &\min_{E^c_1,E^s_1,E^c_2,E^s_2,G_1,G_2} \max_{D_1,D_2} L(E^c_1, E^s_1, E^c_2, E^s_2, G_1, G_2, D_1, D_2) \\ &= \lambda_s(L^{s_1}_{cycle} + L^{s_2}_{cycle}) + \lambda_c(L^{c_1}_{cycle} + L^{c_2}_{cycle}) \\ &\quad + \lambda_x(L^1_{recon} + L^2_{recon}) + \lambda_g(L^1_{GAN} + L^2_{GAN}) \end{aligned} \tag{6}$$

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 [11], 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

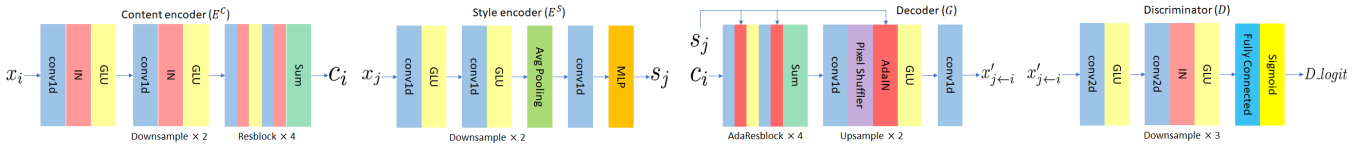

Fig. 4. The network structure of content encoder, style encoder, decoder, and GAN discriminator.

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 [18] every 5 ms, then encoded to 24-dimension mel-cepstral vectors of temporal size $w = 128$ as the autoencoder's input.

4.2. Network Structure

Our network structure is illustrated in Fig. 4. The encoders and decoders are implemented with 1-dimension CNNs to capture the temporal dependencies; the GAN discriminators are implemented with 2-dimension CNNs to capture the spectra-temporal patterns. All networks are equipped with gated linear units (GLU) [19] as activation functions. The emotion style is learnt by a 3-linear MLP that outputs 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 in eq. (2).

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

We use Adam optimizer and set $\beta_1 = 0.5$. The learning rate is initialized as 0.0001 and linearly decayed to 0 from the $200K$-th iteration. We set $\lambda_s = \lambda_c = \lambda_g = 1$ and $\lambda_x = 10$. For training, we randomly sample fixed length frames (128) from the input audio with 16KHz frequency. Conversion was conducted on speech sequences with arbitrary length.

4.3. Results

The results were evaluated on three matrices: voice quality, emotion correctness, and the ability to keep speaker identity.

Subjective evaluation We conducted listening tests on Amazon MTurk to evaluate the converted speech [1]. Each example was listened by 5 random evaluators. They were asked to manually classify the emotion, and give 1-to-5 opinion scores on voice quality and the similarity with the original speaker. The mean opinion score (MOS) of the latter two metrics were listed in Fig. 5. For subjective emotion classification, we found results consistent with the objective evaluations, therefore omitted for space constraints.

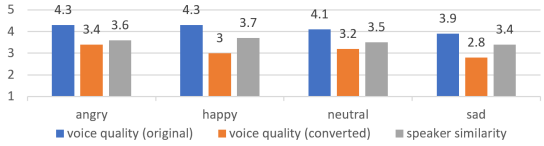

Fig. 5. MOS for voice quality and speaker similarity

Objective evaluation We applied a state-of-the-art speech emotion classifier [20] for objective evaluation. The results listed in Table 1 show that our model can effectively increase the percentage of desired emotions. Note that neutral and sad speech often get mixed up even by humans.

Table 1. Emotion classification by method in [20]

	emotion percentage % origin (converted)			
model	**Angry**	**Happy**	**Neutral**	**Sad**
neu2ang	**5(26)**	7(14)	**19(9)**	70(51)
ang2neu	**84(16)**	5(16)	**11(63)**	0(5)
hap2sad	12(7)	**51(31)**	34(56)	**2(5)**
sad2hap	7(4)	**21(24)**	33(45)	**39(27)**
ang2sad	**62(16)**	15(17)	21(58)	**2(9)**
sad2ang	**7(17)**	21(16)	33(32)	**39(35)**
ang2hap	**72(32)**	**8(16)**	10(53)	10(0)
hap2ang	**12(63)**	**65(16)**	12(22)	12(0)

5. CONCLUSION AND FUTURE WORK

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. To our knowledge, this is the first work for nonparallel emotion conversion using style transfer. Future work is to develop a multi-domain emotion conversion model for unseen speakers.

Acknowledgements: This research was supported by U.S. Air Force under grant number FA9550-17-1-0259.

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

6. REFERENCES

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

[2] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent conversion using phonetic posteriorgrams," in *ICASSP*. IEEE, 2018, pp. 5314–5318.

[3] M. Wang, M. Wen, K. Hirose, and N. Minematsu, "Emotional voice conversion for mandarin using tone nucleus model–small corpus and high efficiency," in *Speech Prosody 2012*, 2012.

[4] Z. Wang and Y. Yu, "Multi-level prosody and spectrum conversion for emotional speech synthesis," in *Signal Processing (ICSP)*. IEEE, 2014, pp. 588–593.

[5] Y. Xue, Y. Hamada, and M. Akagi, "Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space," *Speech Communication*, vol. 102, pp. 54–67, 2018.

[6] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds1," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[7] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of japanese," *ASJs Japan (E)*, vol. 5, no. 4, pp. 233–242, 1984.

[8] Y. Xue and M. Akagi, "A study on applying target prediction model to parameterize power envelope of emotional speech," in *RISP workshop NCSP'16*. , 2016.

[9] L.A. Gatys, A.S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *CVPR*. IEEE, 2016, pp. 2414–2423.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[11] C. Busso, M. Bulut, CC Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[12] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "Gmm-based voice conversion applied to emotional speech synthesis," in *Eurospeech*, 2003.

[13] CC Hsu, HT Hwang, YC Wu, Y. Tsao, and HM Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv preprint arXiv:1704.00849*, 2017.

[14] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," *arXiv preprint arXiv:1804.00425*, 2018.

[15] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A.W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio.," in *SSW*, 2016, p. 125.

[16] CF Huang and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, vol. 50, no. 10, pp. 810–828, 2008.

[17] JY Zhu, T. Park, P. Isola, and A.A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE ICCV*, Oct 2017.

[18] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[19] Y.N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *ICML*, 2017, pp. 933–941.

[20] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *ICASSP*. IEEE, 2017, pp. 2227–2231.