Paper ID: 4278

Seen and Unseen Emotional Style Transfer for Voice Conversion with a New Emotional Speech Database

NUS National University of Singapore

Kun Zhou [1], Berrak Sisman [2], Rui Liu [2], Haizhou Li [1]

[1] Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore
[2] Singapore University of Technology and Design, Singapore

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Introduction

- Emotional voice conversion (EVC): transform the emotional prosody while preserving the linguistic content and speaker identity;

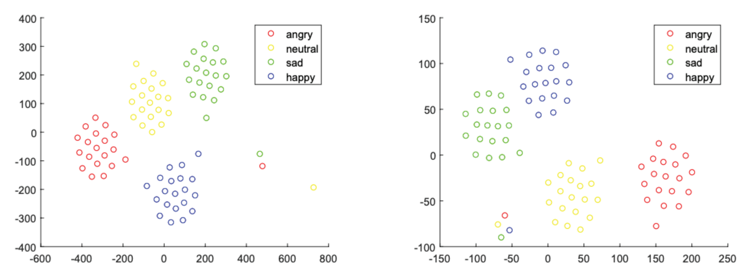
- Prior studies propose to disentangle the emotional prosody using an encoder-decoder network conditioned on discrete representation, such as one-hot emotion labels, but only learn to remeber a fixed set of emotional styles;

Our contributions:

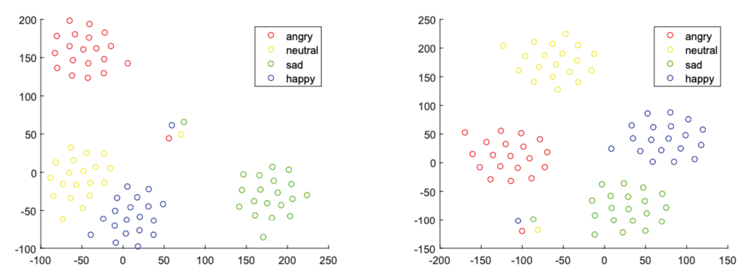
In this paper, we propose:
- a one-to-many EVC framework that does not need parallel data;
- to use deep emotional features to describe different emotional styles;
- release a multi-speaker and multi-lingual emotional speech dataset;

To our best knowledge, it is the first reported study on emotional style transfer for unseen emotions!

Analysis of deep emotional features



(a) Data in left and right panels are from two different male speakers.



(b) Data in left and right panels are from two different female speakers.

Fig. 1. t-SNE plot of deep emotional features for 20 utterances with the same content but spoken by different speakers.

- Recent advances of deep learning have led to a shift from human-crafted representations to deep features learnt by neural network;

- As shown in Fig. 1, deep emotional features form clear emotion groups in terms of feature distribution. It suggests the potential to use deep emotional features as the emotion descriptor to encode an emotion class.

One-to-many emotional style transfer

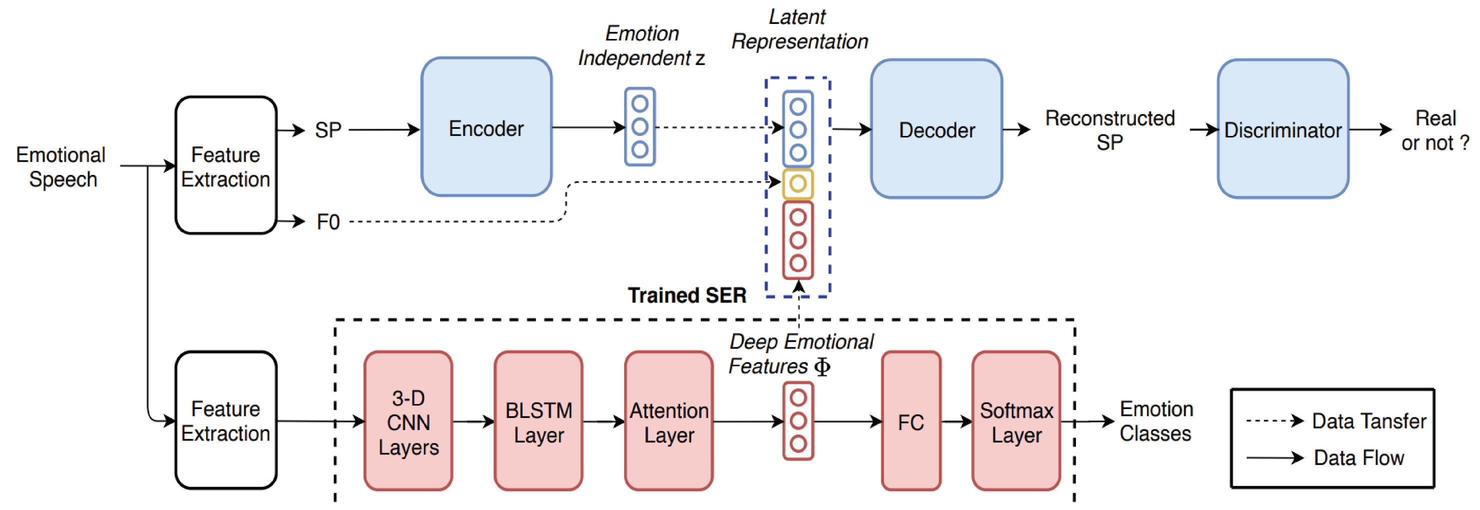


Fig. 2. The training phase of the proposed DeepEST framework. Blue boxes represent the networks that involved in the training and red boxes represent the networks that are already trained.

1) Stage I: Emotion Descriptor Training

- We propose to use a SER model to extract deep emotional features from the input utterances;
- The SER serves as the emotion descriptor to describe the emotional prosody in a continous space;

2) Stage II: Encoder-Decoder Training with VAW-GAN [1]

- The encoder learns to generate emotion-independent latent representation from the input features;
- The decoder learns to reconstruct the input features with latent representation, F0 contour and deep emotional features;
- The discriminator learns to determine whether the gerenated features real or not;

3) Stage III: Run-time Conversion

- We first use pre-trained SER to generate the deep emotional features from the reference set;
- We then concatenate the deep emotional features with the converted F0 and emotion-independent latent representation;
- We synthesis the converted speech with the reference emotion type using converted spectral features.

Codes & Speech Samples:
https://kunzhou9646.github.io/controllable-evc/
For any inquiries:
Please email: zhoukun@u.nus.edu

Experiments

1) ESD: A new multi-speaker and multi-lingual emotional speech dataset

- 350 parallel utterances spoken by Mandarin and English speakers
- For each language, there are 5 male and 5 female speakers in 5 emotions: a) happy, b) sad, c) neutral, d) angry, e) surprise

- During conversion, we use one universal model to conduct emotion conversion from neutral to both seen emotions (happy, sad) and unseen emotion (angry);

- Baseline: VAW-GAN-EVC [2]: one-hot emotion label, one-to-one conversion

2) Objective Evaluation

Table 1. MCD values of the baseline framework VAW-GAN-EVC and the proposed framework DeepEST in a comparative study.

MCD [dB]	Male			Female		
	Zero Effort	VAW-GAN-EVC	**DeepEST**	Zero Effort	VAW-GAN-EVC	**DeepEST**
neutral-to-happy	6.769	4.738	4.569	7.088	4.284	4.260
neutral-to-sad	6.306	4.284	4.127	8.287	5.464	4.916
neutral-to-angry	6.649	4.482	4.564 *(unseen)*	6.690	4.204	4.451 *(unseen)*

3) Subjective Evaluation

Table 2. MOS results with 95 % confidence interval to assess the speech quality.

MOS	N2H	N2S	N2A
Reference	4.95 ± 0.11	4.88 ± 0.22	4.87 ± 0.22
VAW-GAN-EVC	3.23 ± 0.71	2.80 ± 0.55	3.11 ± 0.57
DeepEST	3.24 ± 0.72	2.94 ± 0.57	3.15 ± 0.63

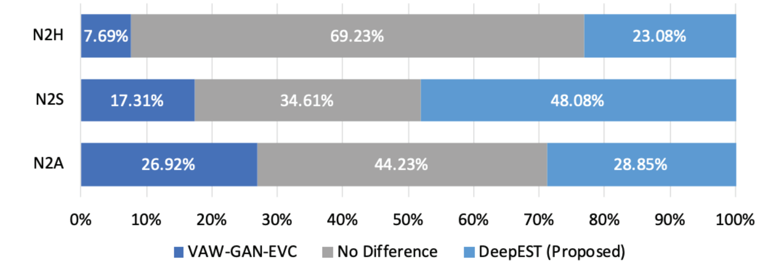


Fig. 3. AB preference test results for the speech quality.

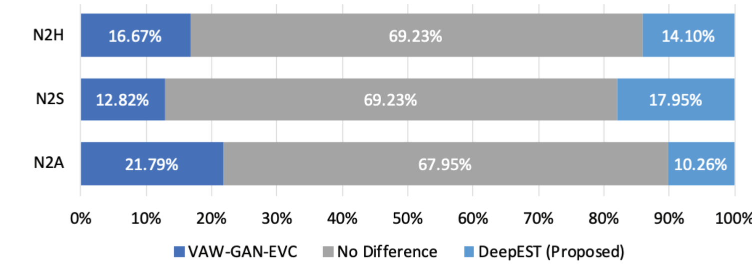


Fig. 4. XAB preference test results for the emotion similarity.

Conclusions

- We propose to build a one-to-many emotional style transfer framework that does not require parallel data
- We propose to leverage deep emotional features from SER to describe emotional prosody in a continuous space
- By conditioning the decoder with controllable attributes such as deep emotional features and F0 values, we achieve competitive results for both seen and unseen emotions over the baseline framework;
- We release a multi-speaker and multi-lingual emotional speech

References

[1] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and HsinMin Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in Proc. Interspeech, 2017.

[2] Kun Zhou, Berrak Sisman, Mingyang Zhang, and Haizhou Li, "Converting Anyone's Emotion: Towards Speaker-Independent Emotional Voice Conversion," in Proc. Interspeech, 2020.