Seen and Unseen emotional style transfer for voice conversion with a new emotional speech dataset

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

[1]National University of Singapore
[2]Singapore University of Technology and Design

ICASSP 2021
TORONTO
Canada
June 6-11, 2021
Metro Toronto Convention Centre

NUS
National University of Singapore

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Outline

1. Introduction
2. Related Work
3. Contributions
4. Proposed Framework
5. Experiments
6. Conclusions

Outline

1. Introduction
2. Related Work
3. Contributions
4. Proposed Framework
5. Experiments
6. Conclusions

1. Introduction

1.1 Emotional voice conversion

Emotional voice conversion aims to transfer the emotional style of an utterance from one to another, while preserving the linguistic content and speaker identity.

Applications:* *emotional text-to-speech, conversational agents

Compared with speaker voice conversion:

1) Speaker identity is mainly determined by voice quality;
   -- speaker voice conversion mainly focuses on spectrum conversion
2) Emotion is the interplay of multiple acoustic attributes concerning both spectrum and prosody.
   -- emotional voice conversion focuses on both spectrum and prosody conversion

1. Introduction

1.2 Current Challenges

Our focus

1) **Non-parallel training**

Parallel data is difficult and expensive to collect in real life;

2) **Emotional prosody modelling**

Emotional prosody is hierarchical in nature, and perceived at phoneme, word, and sentence level;

3) **Unseen emotion conversion**

Existing studies use discrete emotion categories, such as one-hot label to remember a fixed set of emotional styles.

Our focus

Outline

1. Introduction
2. Related Work
3. Contributions
4. Proposed Framework
5. Experiments
6. Conclusions

2. Related Work

2.1 Emotion representation

1) Categorical representation

Example: Ekman's six basic emotions theory[1]:
i.e., anger, disgust, fear, happiness, sadness and surprise

2) Dimensional representation

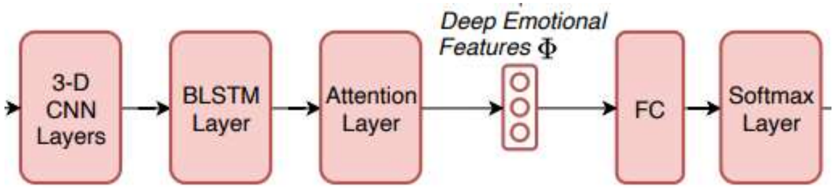
Example: Russell's circumplex model[2]:
i.e., arousal, valence, and dominance

[1] Paul Ekman, "An argument for basic emotions," Cognition & emotion, 1992

[2] James A Russell, "A circumplex model of affect.," Journal of personality and social psychology, vol. 39, no. 6, pp. 1161, 1980.

2. Related Work

2.2 Emotional descriptor

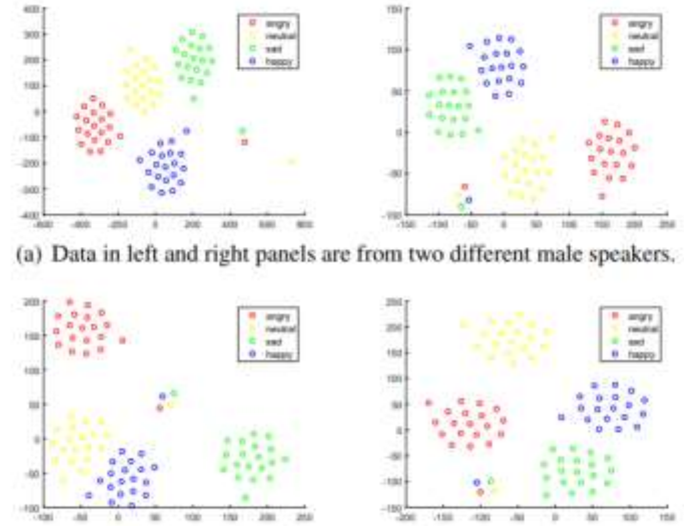


1) One-hot emotion labels

2) Expert-crafted features such as openSMILE [3]

3) Deep emotional features:

-- data-driven, less dependent on human knowledge, more suitable for emotional style transfer…



(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.

[3] Florian Eyben, Martin Wöllmer, and Björn Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, Proceedings of the 18th acm international conference on multimedia, 2010, pp. 1459–1462

2. Related Work

2.3 Variational Auto-encoding Wasserstein Generative Adversarial Network (VAW-GAN)[4]

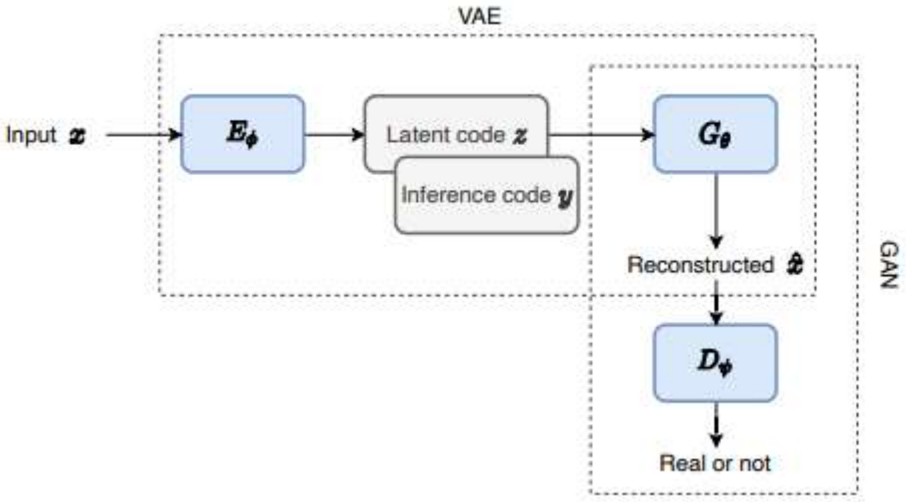


Fig 2. An illustration of VAW-GAN framework.

- Based on variational auto-encoder (VAE)
- Consisting of three components:

Encoder: to learn a latent code from the input;

Generator: to reconstruct the input from the latent code and the inference code;

Discriminator: to judge the reconstructed input whether real or not;

- First introduced in speaker voice conversion in 2017

[4] 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.

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

NUS National University of Singapore

Outline

1. Introduction
2. Related Work
3. Contributions
4. Proposed Framework
5. Experiments
6. Conclusions

3. Contributions

1. One-to-many emotional style transfer framework
2. Does not require parallel training data;
3. Use speech emotion recognition model to describe the emotional style;
4. Publicly available multi-lingual and multi-speaker emotional speech corpus (ESD), that can be used for various speech synthesis tasks;

To our best knowledge, this is the first reported study on emotional style transfer for unseen emotion.

Outline

1. Introduction
2. Related Work
3. Contributions
4. **Proposed Framework**
5. Experiments
6. Conclusions

4. Proposed Framework

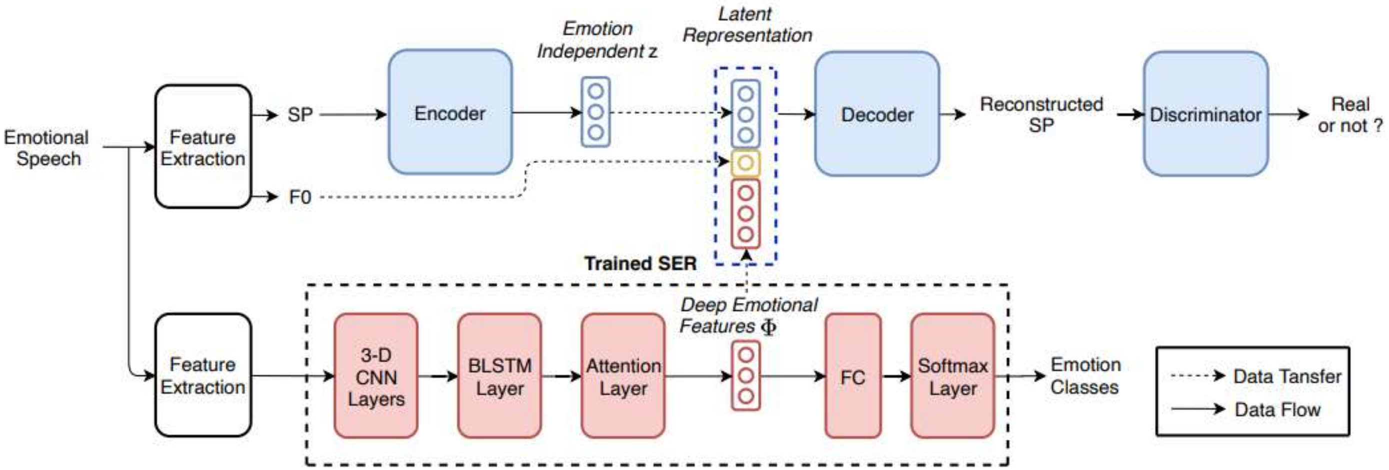


Fig. 3. The training phase of the proposed DeepEST framework. Blue boxes represent the networks that involved in the training, and the red boxes represent the networks that are already trained.

1) Stage I: Emotion Descriptor Training
2) Stage II: Encoder-Decoder Training with VAW-GAN
3) Stage III: Run-time Conversion

4. Proposed Framework

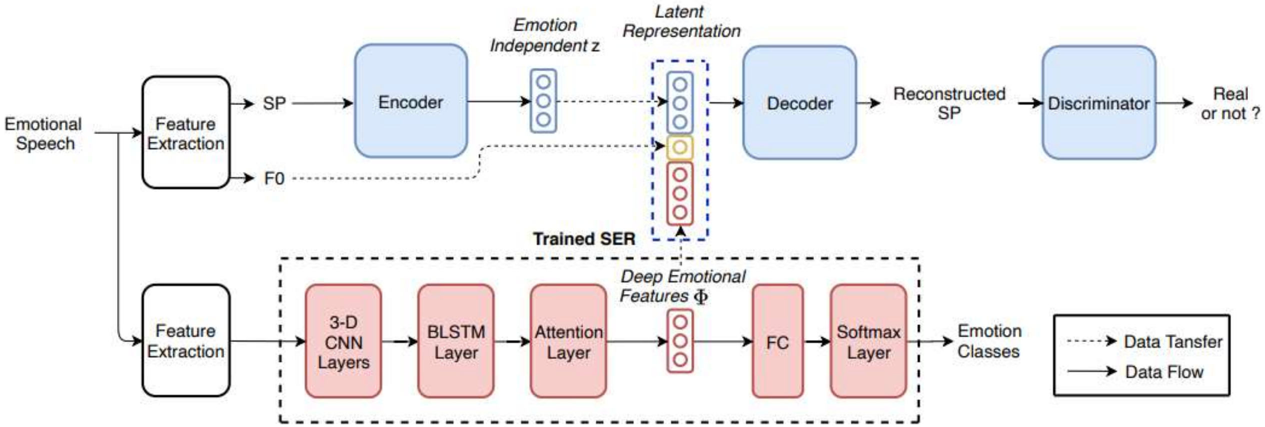


Fig. 3. The training phase of the proposed DeepEST framework. Blue boxes represent the networks that involved in the training, and the red boxes represent the networks that are already trained.

1) Emotion Descriptor Training

- A pre-trained SER model is used to extract deep emotional features from the input utterance.
- The utterance-level deep emotional features are thought to describe different emotions in a continuous space.

4. Proposed Framework

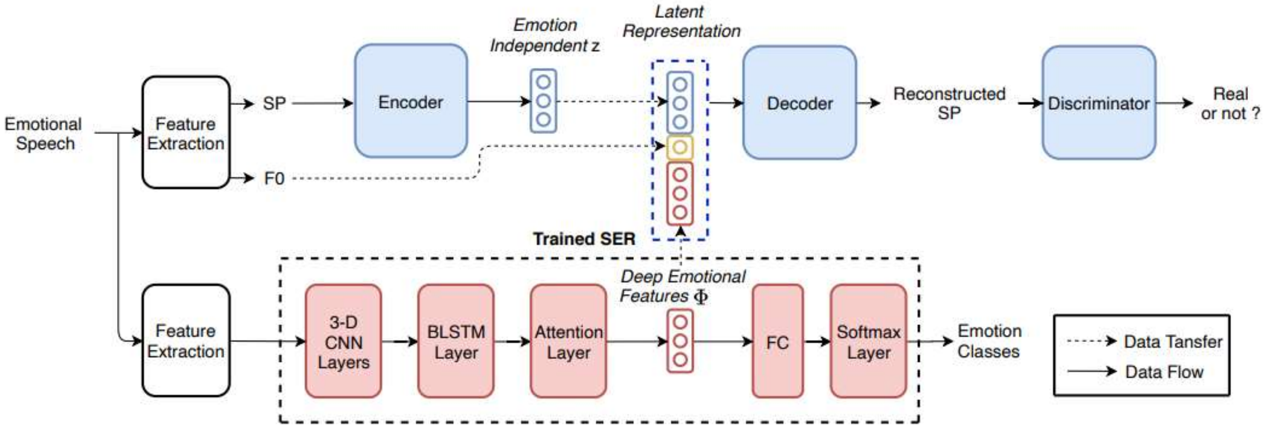


Fig. 3. The training phase of the proposed DeepEST framework. Blue boxes represent the networks that involved in the training, and the red boxes represent the networks that are already trained.

2) Encoder-Decoder Training with VAW-GAN

- The encoder learns to generate emotion-independent latent representation z 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 reconstructed features real or not;

4. Proposed Framework

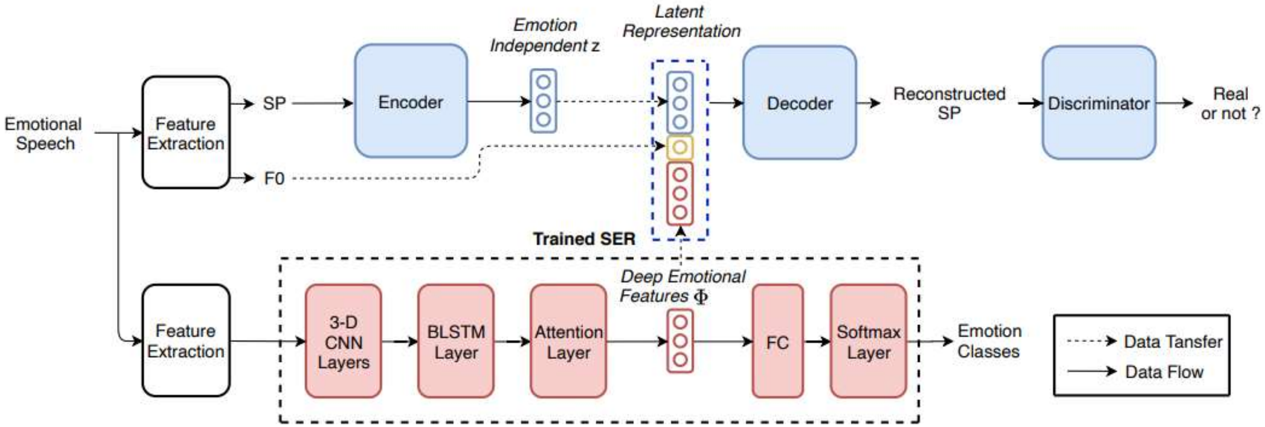


Fig. 3. The training phase of the proposed DeepEST framework. Blue boxes represent the networks that involved in the training, and the red boxes represent the networks that are already trained.

3) Run-time Conversion

- The pre-trained SER generates the deep emotional features from the reference set;
- We then concatenate the deep emotional features with the converted F0 and emotion-independent z as the input to the decoder;

Outline

1. Introduction
2. Related Work
3. Contributions
4. Proposed Framework
5. **Experiments**
6. Conclusions

5. Experiments

5.1 Emotional Speech Dataset -- ESD

1) The dataset consists of 350 parallel utterances with an average duration of 2.9 seconds spoken by 10 native *English* and 10 native *Mandarin* speakers;

2) For each language, the dataset consists of 5 male and 5 female speakers in five emotions:
a) happy, b) sad, c) neutral, d) angry, and e) surprise

3) ESD dataset is publicly available at:

https://github.com/ HLTSingapore/Emotional-Speech-Data

To our best knowledge, this is the first parallel emotional voice conversion dataset in a multi-lingual and multi-speaker setup.

5. Experiments

5.2 Experimental setup

1) 300-30-20 utterances (training-reference-test set);
2) One universal model that takes *neutral*, *happy* and *sad* utterances as input;
3) SER is trained on a subset of IEMOCAP[5], with four emotion types (happy, angry, sad and neutral)
4) We conduct emotion conversion from neutral to seen emotional states (happy, sad) to unseen emotional states (angry).

5) **Baseline:** VAW-GAN-EVC [6]: only capable of one-to-one conversion and generate seen emotions;
Proposed: DeepEST: one-to-many conversion, seen and unseen emotion generation

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” Language resources and evaluation, vol. 42, no. 4, pp. 335, 2008.
[6] Kun Zhou, Berrak Sisman, Mingyang Zhang, and Haizhou Li, “Converting Anyone’s Emotion: Towards Speaker-Independent Emotional Voice Conversion,” in Proc. Interspeech 2020, 2020, pp. 3416–3420.

5. Experiments

5.3 Objective Evaluation

We calculate MCD [dB] to measure the spectral distortion

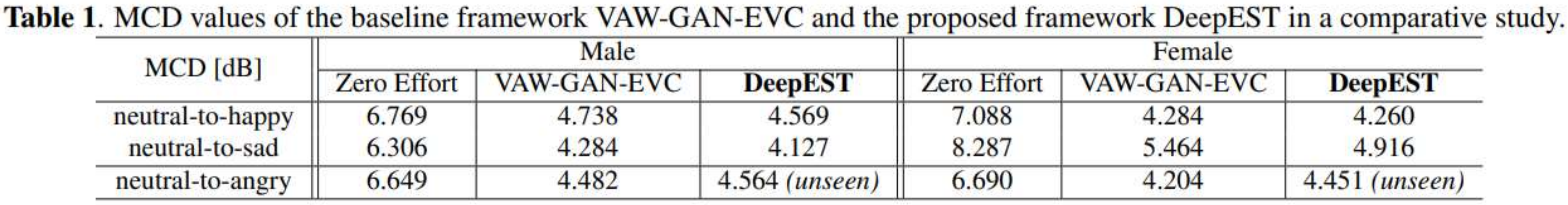


Table 1. MCD values of the baseline framework VAW-GAN-EVC and the proposed framework DeepEST in a comparative study.

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)*

Proposed DeepEST outperforms the baseline framework for seen emotion conversion and achieves comparable results for unseen emotion conversion.

5. Experiments

5.4 Subjective Evaluation

To evaluate speech quality:

1) MOS listening experiments

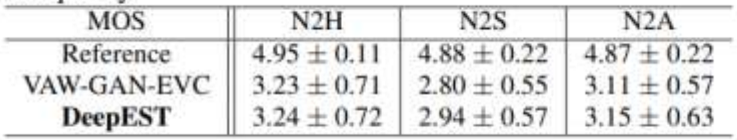
2) AB preference test

To evaluate emotion similarity:

3) XAB preference test

Proposed DeepEST achieves competitive results for both seen and unseen emotions

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

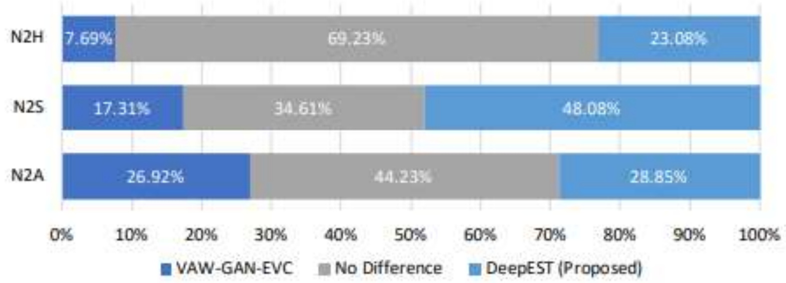


Fig. 3. AB preference test results for the speech quality.

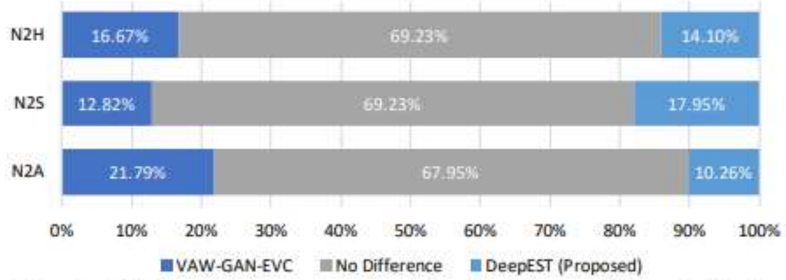


Fig. 4. XAB preference test results for the emotion similarity.

5. Experiments

Codes & Speech Samples

https://kunzhou9646.github.io/controllable-evc/

Outline

1. **Introduction**

2. **Related Work**

3. **Contributions**

4. **Proposed Framework**

5. **Experiments**

6. **Conclusions**

6. Conclusion

1. We propose a one-to-many emotional style transfer framework based on VAW-GAN without the need for parallel data;

2. We propose to leverage deep emotional features from speech emotion recognition to describe the emotional prosody in a continuous space;

3. We condition the decoder with controllable attributes such as deep emotional features and F0 values;

4. We achieve competitive results for both seen and unseen emotions over the baselines;

5. We introduce and release a new emotional speech dataset, ESD, that can be used in speech synthesis and voice conversion.

Thank you for listening!