30 August – 3 September 2021
INTERSPEECH 2021 BRNO | CZECHIA
Speech everywhere!

Limited Data Emotional Voice Conversion Leveraging Text-to-Speech: Two-Stage Sequence-to-Sequence Training

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

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

NUS National University of Singapore
SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Introduction

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

- Sequence-to-sequence (seq2seq) methods:
- allows for the duraion prediction;
- jointly model spectrum and prosody;
- focus on emotion-relevant regions through attention;
- But always require a large amount of training data!

Our contributions:

- without the need of parallel data, and flexible for many-to-many emotional voice conversion;

- only needs limited amount of emotional

- The first work of seq2seq emotional voice conversion that only needs a limited amount of emotional speech! data to train!

Proposed Framework

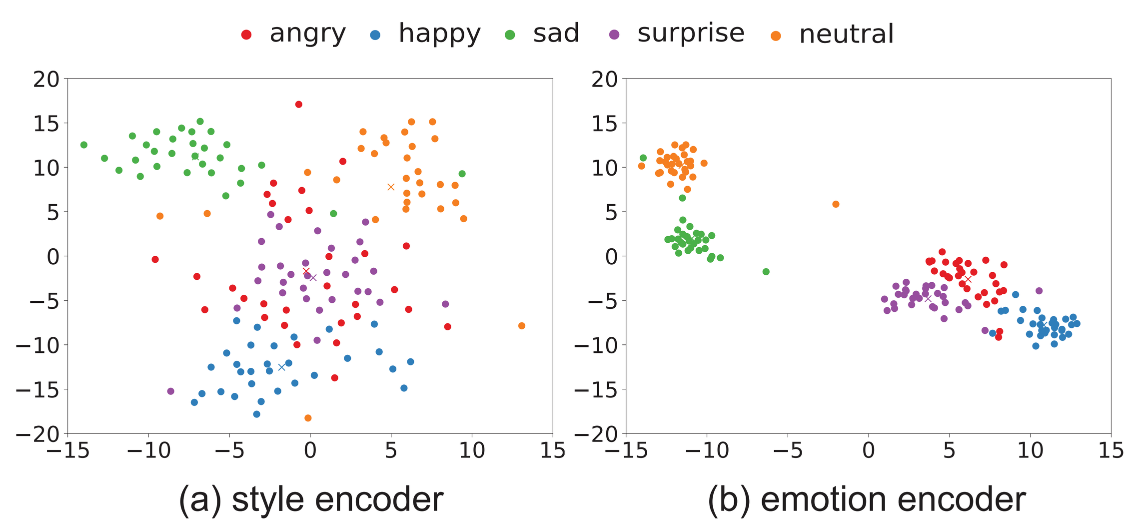


Figure 1: Visualization of emotion embedding derived from (a) style encoder and (b) emotion encoder.

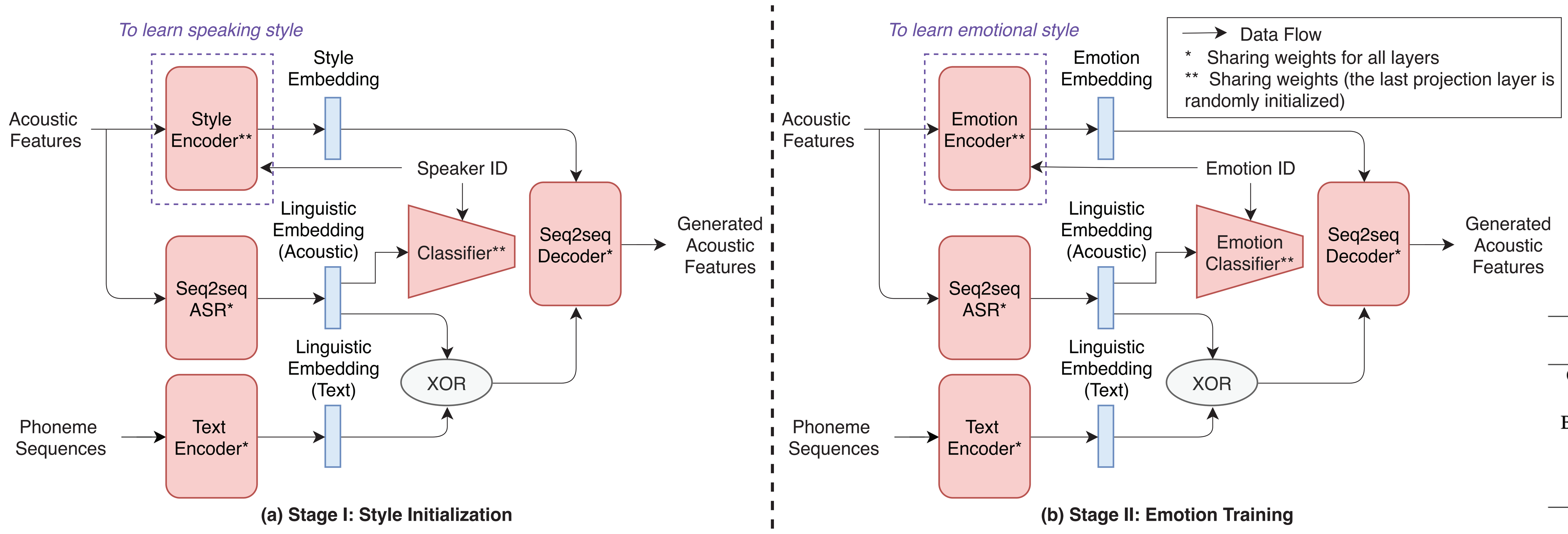


Figure 2: The proposed 2-stage training strategy for seq2seq emotional voice conversion with limited emotional speech data.

1) Training Stage I: Style Initialization

- We adopt a seq2seq VC framework [2], and pretrain it with a publicly available TTS corpus, as shown in Figure 2(a);
- Style encoder learns speaker-dependent information, i.e., speaker style, and excludes linguistic information from the acoustic features;

2) Training Stage II: Emotion Training

- Style encoder acts as emotion encoder to learn the emotional styles from additional emotion-labelled speech data;
- Classifier acts as an emotion classifier to eliminate the emotion information in the linguistic space;
- Style encoder vs. Emotion enocder: We visualize the emotion embedding of the reference utterances, as shown in Fig. 1;

Findings: the emotion embeddings derived by the emotion encoder form separate groups for each emotion type, while those from the style encoder fail to provide a clear pattern!
-- Validate our idea of 2-stage training!

Codes & Speech Samples:

For any inquiries:
Please email: zhoukun@u.nus.edu

Experiments

- Database: VCTK corpus [3] for stage I, ESD database [4] for stage II.
- Baseline: 1) CycleGAN-EVC[5]; 2) StarGAN-EVC[6]; 3) Baseline Seq2seq-EVC;
- Proposed: 1) Seq2seq-EVC-GL; 2) Seq2seq-EVC-WA1; 3) Seq2seq-EVC-WA2
- Objective Evaluation

1) MCD:

Table 1: *A comparison of MCD [dB] values.*

Framework	MCD [dB]			
	Neu-Ang	Neu-Sad	Neu-Hap	Neu-Sur
CycleGAN-EVC [15]	4.57	4.32	4.46	4.68
StarGAN-EVC [16]	4.51	4.31	4.24	4.39
Baseline Seq2seq-EVC	5.14	5.27	5.04	5.40
Seq2seq-EVC-GL	3.98	3.83	3.92	3.94
Seq2seq-EVC-WA1	3.72	3.73	3.71	3.83
Seq2seq-EVC-WA2	3.73	3.73	3.70	3.80

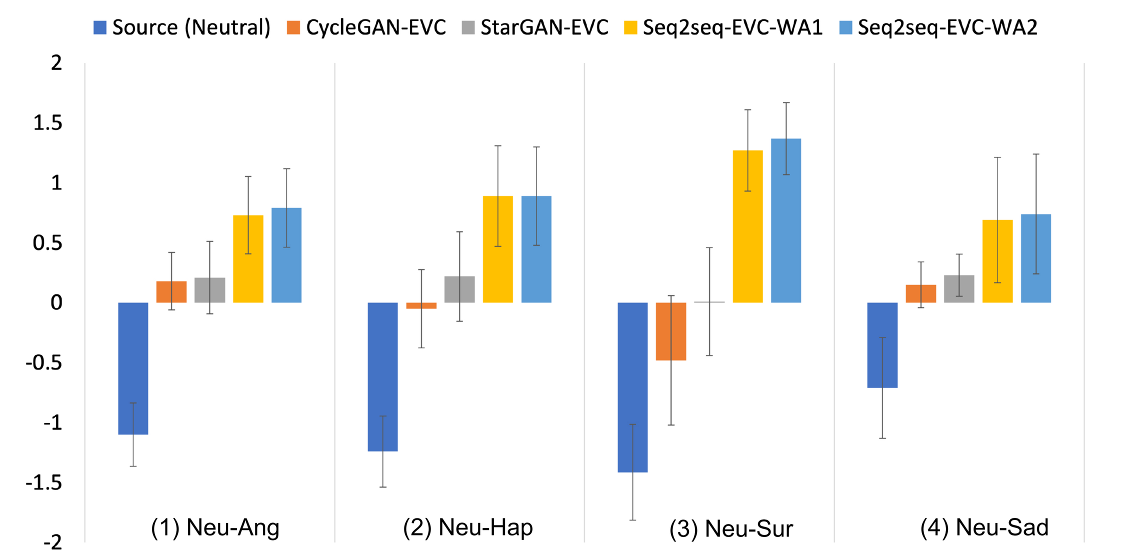
2) DDUR:

Table 2: *A comparison of DDUR [s] values for the voiced parts.*

Framework	DDUR [s]			
	Neu-Ang	Neu-Sad	Neu-Hap	Neu-Sur
Source-Target	0.36	0.46	0.26	0.44
Baseline Seq2seq-EVC	0.65	0.91	0.69	0.54
Seq2seq-EVC-GL	0.38	0.41	0.26	0.33
Seq2seq-EVC-WA1	0.39	0.39	0.27	0.33
Seq2seq-EVC-WA2	0.34	0.40	0.24	0.32

- Subjective Evaluation

1) MOS for emotion similarity



2) BWS for overall speech quality

Table 3: *Best Worst Scaling (BWS) listening experiments to evaluate the overall speech quality.*

Systems		Seq2seq-EVC-GL	Seq2seq-EVC-WA1	Seq2seq-EVC-WA2
Neu-Ang	Best	0%	19%	**81%**
	Worst	94%	6%	**0%**
Neu-Hap	Best	0%	32%	**68%**
	Worst	97%	3%	**0%**
Neu-Sur	Best	6%	25%	**69%**
	Worst	94%	**3%**	**3%**
Neu-Sad	Best	0%	10%	**90%**
	Worst	94%	6%	**0%**

Conclusions

- A novel training strategy for limited data seq2seq emotional voice conversion leveraging text-to-speech without the need for parallel data;
- Can do many-to-many emotional voice conversion, and conduct spectral and duration mapping at the same time;
- Investigate different training strategies for WaveRNN vocoder training;
- Experimental results show a significant improvement of the performance.

References

[1] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in ICML 2018;
[2] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel sequenceto-sequence voice conversion with disentangled linguistic and speaker representations," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019
[3] C. Veaux, J. Yamagishi, K. MacDonald et al., "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.
[4] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," IEEE ICASSP, 2021.
[5] K. Zhou, B. Sisman, and H. Li, "Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data," in Proc. Odyssey 2020 ;
[6] G. Rizos, A. Baird, M. Elliott, and B. Schuller, "Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in ICASSP 2020;