INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

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

Kun Zhou[*], *Berrak Sisman*[†], *Haizhou Li*[*]

[*]Department of ECE, National University of Singapore (NUS), Singapore
[†]Singapore University of Technology and Design (SUTD), Singapore
zhoukun@u.nus.edu, berraksisman@u.nus.edu, haizhou.li@nus.edu.sg

Abstract

Emotional voice conversion (EVC) aims to change the emotional state of an utterance while preserving the linguistic content and speaker identity. In this paper, we propose a novel 2-stage training strategy for sequence-to-sequence emotional voice conversion with a limited amount of emotional speech data. We note that the proposed EVC framework leverages text-to-speech (TTS) as they share a common goal that is to generate high-quality expressive voice. In stage 1, we perform style initialization with a multi-speaker TTS corpus, to disentangle speaking style and linguistic content. In stage 2, we perform emotion training with a limited amount of emotional speech data, to learn how to disentangle emotional style and linguistic information from the speech. The proposed framework can perform both spectrum and prosody conversion and achieves significant improvement over the state-of-the-art baselines in both objective and subjective evaluation.

Index Terms: Emotional voice conversion, sequence-to-sequence, limited data

1. Introduction

Sequence-to-sequence (seq2seq) speech synthesis frameworks, such as Tacotron [1], can generate high-quality synthetic speech. However, such frameworks heavily rely on a large amount of training data. Furthermore, they generally lack emotional variance [2]. Emotional voice conversion aims to convert the emotional state of speech from one to another while preserving the linguistic content and speaker identity. This technique allows us to project a desired emotion into the generated speech, thus bears huge potential in real-world applications, such as expressive text-to-speech [3].

Emotion is inherently supra-segmental and complex with multiple signal attributes concerning both spectrum and prosody [4], thus it is insufficient to convert the emotion only with frame-wise spectral mapping. Prosodic features, such as pitch, energy, and duration, also need to be dealt with for emotional voice conversion. We believe that seq2seq training is a better solution for spectrum and duration conversion in EVC, which will be the focus of this paper.

Emotional voice conversion is a special type of voice conversion [5]. Previous studies are focused on frame-based mapping of spectral features of source and target, including using statistical methods [6, 7] and deep learning methods, such as deep neural network [8], generative adversarial network (GAN) [9] and CycleGAN [10]. Inspired by the success in speaker voice conversion, these methods are adopted to model both spectral and prosodic parameters for emotional voice conversion. Successful attempts include GMM [11], sparse representation [12], deep bi-directional long-short-term memory (BLSTM) network [13], GAN-based [14–17] and auto-encoder-based [18–21] methods. These frameworks model the mapping on a frame-by-frame basis. As emotional prosody is hierarchical in nature [4], frame-based methods are therefore not the best in handling prosody conversion [5].

Recently, seq2seq models with attention mechanism have attracted much interests in speech synthesis [1, 22] and voice conversion such as SCENT [23], AttS2S-VC [24] and ConvS2S-VC [25]. Considering that VC and TTS share a similar motivation in a sense that they both aim to generate speech from internal representations [5], there are studies to leverage TTS systems to further improve seq2seq VC performance, such as adding text supervision [26] or leveraging TTS [27, 28]. Inspired by these studies, seq2seq frameworks have become popular in emotional voice conversion. For example, a seq2seq model is proposed in [29] to jointly model pitch and duration with parallel data. In [30], researchers propose a seq2seq model with multi-task learning for both emotional voice conversion and emotional text-to-speech. We note that these frameworks require tens of hours of emotional speech data to train, which is not practical for real-life scenarios.

In this paper, we propose a 2-stage training strategy for seq2seq emotional voice conversion. In stage 1, we perform style initialization, which aims to disentangle speaking style and the linguistic content with a multi-speaker TTS corpus. In stage 2, we perform emotion training, where all components of the network are trained with limited emotional speech data. By doing this, we obtain *emotion encoder* that learns to disentangle emotional style from the speech, and *emotion classifier* that further eliminates the emotion-related information in the linguistic space. The proposed framework achieves remarkable performance by converting both spectrum and prosody with a limited amount of non-parallel emotional speech data.

The main contributions of this paper include: 1) we propose a seq2seq emotional voice conversion framework leveraging TTS without the need for parallel data, and flexible for many-to-many emotional voice conversion; 2) we propose a novel training strategy that requires a small amount of emotion-labelled data; 3) we significantly improve the performance by modelling the alignment between acoustic and linguistic embedding for emotion styles, which is a departure from frame-based conversion paradigm; 4) we propose emotional fine-tuning for WaveRNN vocoder [31] training with the limited amount of emotional speech data to further improve the final performance.

This paper is organized as follows: In Section 2, we motivate our study through the comparison with existing seq2seq EVC frameworks. In Section 3, we introduce our proposed framework and the proposed training strategy. In Section 4, we report the experiments. Section 5 concludes the study.

Codes & Speech Samples: https://kunzhou9646.github.io/IS21/

http://dx.doi.org/10.21437/Interspeech.2021-781

2. Sequence-to-sequence EVC

The seq2seq model, which was first studied in machine translation [32], was found effective in speech synthesis [1, 22] and voice conversion [23, 25, 27, 28]. In voice conversion, the seq2seq model with attention mechanism has greatly improved the modelling ability by jointly learning the feature mapping and alignment. The seq2seq model marks a departure from the frame-wise modelling [5, 33]. First, the seq2seq model allows for the prediction of the speech duration at the run-time inference which is an essential factor of emotional prosody [11]. Second, emotion labels are usually annotated at the utterance level in speech corpus [33], while emotional prosody is supra-segmental and can be associated with only a few words. The attention mechanism makes it possible for the conversion to focus on emotion-relevant regions, which will be our focus.

There are only a few studies on emotional voice conversion with seq2seq modelling such as jointly modelling pitch and duration with parallel data [29], where the output pitch contour is conditioned on the syllable position and source signal; and multi-task learning where a single system is jointly trained for both emotional voice conversion and text-to-speech [30]. These frameworks perform well but rely on a large emotional speech corpus. In this paper, we would like to study a limited data solution. To the best of our knowledge, this is the first attempt with the seq2seq model that does not need a large amount of emotional speech training data for EVC.

3. Proposed seq2seq EVC model

We propose a seq2seq EVC framework that consists of 5 components, a text encoder, a seq2seq automatic speech recognition (ASR) encoder, a style encoder, a classifier, and a seq2seq decoder. We propose a 2-stage training strategy: 1) Stage I: Style initialization, which disentangles between the speaking style, i.e., speaker style, and the linguistic content with a multi-speaker TTS corpus; 2) Stage II: Emotion training, where all components, initialized by stage I, are further trained with a limited amount of emotional speech data. Finally, during run-time conversion inference, the framework generates the utterance with reference emotion type by combining the source linguistic representation and the reference emotion representation. While the proposed model is trained to perform both EVC and emotional TTS, EVC will be the main focus of this paper.

3.1. Training stage I: Style initialization

At stage I, we adopt a seq2seq VC framework [28], and pre-train it with a publicly available TTS corpus, as shown in Figure 2(a). The framework takes the acoustic features and one-hot phoneme sequences as the inputs. The text encoder and the seq2seq ASR encoder predict the linguistic embeddings from the audio input and the text input respectively. The style encoder embeds the acoustic features into the style embedding. Finally, the seq2seq decoder recovers the acoustic features with the style and linguistic embeddings either from audio or text inputs.

In stage I, the style encoder learns speaker-dependent information, i.e., speaker style, and excludes linguistic information from the acoustic features. To disentangle from speaker style, an adversarial training with a classifier is employed to further eliminate speaker information from the linguistic space. With the text inputs and adversarial training strategy, the framework learns to disentangle the linguistic and style information through a multi-speaker TTS corpus.

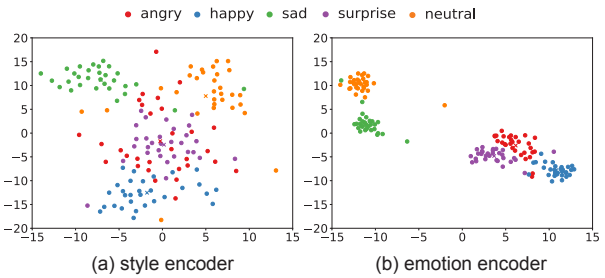


Figure 1: *Visualization of emotion embeddings derived from (a) style encoder and (b) emotion encoder. Each point represents the emotion embedding of a reference utterance.*

However, since the style encoder learns the style information from an emotion-neutral TTS corpus, it does not learn to encode any specific speaking style during stage I, as shown in Figure 1(a). However, the style encoder has rich knowledge about the style and speaker information, we believe it has the potential to learn the emotional style representation given a small amount of emotional speech data. Therefore, we consider stage I as the style initialization, and propose emotion training in stage II, where the style encoder acts as an emotion encoder to learn the emotional style representations.

3.2. Training stage II: Emotion training

We propose to retrain the framework with a limited amount of emotional speech data at stage II, as shown in Figure 2(b). We expect that the network has learnt the basic functions of VC and TTS with a styling mechanism during stage I. The style encoder is then ready to learn the emotional styles from additional emotion-labelled speech data. It acts as an emotion encoder to embed the acoustic features into an emotion vector h^e. Furthermore, the classifier acts as an *emotion classifier* to eliminate the emotion information in the linguistic space. Both the emotion encoder and emotion classifier are trained in a supervised way with a one-hot emotion ID.

3.2.1. Training with limited emotion data

The emotion encoder learns the emotional representations through the loss function L_c as below:

$$L_c = \frac{1}{N}\sum_{n=1}^{N} CE(\phi, \hat{\phi}_n), \tag{1}$$

where $CE(\cdot)$ represents the cross entropy loss function, N represents the length of embedding sequence, ϕ and $\hat{\phi}_n$ denote the one-hot emotion label and the predicted emotion probability respectively. As for the emotion classifier C^s, the adversarial loss L_{adv} is modified as follows:

$$L_{adv} = \frac{1}{N}\sum_{n=1}^{N} \|\alpha - \hat{\phi}_n\|_2^2, \tag{2}$$

where $\alpha = [1/R, ..., 1/R]^T$ is an uniform distribution over the total number of emotion types R. And the emotion classification loss L_{ec} is given as:

$$L_{ec} = CE(\phi, softmax(Vh^e)), \tag{3}$$

where V is the weight matrix of the emotion encoder.

We note that all the components are initialized with the weights learnt at stage I, while the last projection layers of the emotion encoder and the emotion classifier are randomly initialized. At stage II, we update the entire network during training. The training allows the seq2seq ASR encoder and seq2seq decoder to learn a better alignment between acoustic frames and linguistic embedding sequence that particularly characterizes the emotional style of the utterance. Furthermore, we also adapt the speaker style encoder of stage I to an emotion

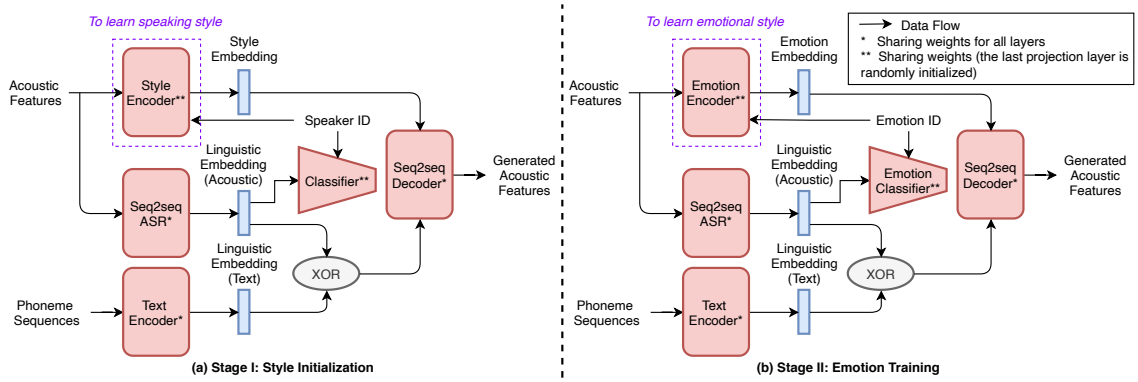


Figure 2: *The proposed 2-stage training strategy for seq2seq emotional voice conversion with limited emotional speech data.*

encoder, the speaker classifier of stage I to an emotion classifier.

Overall, the framework leverages the knowledge of disentanglement between linguistic and style information learnt at stage I, and effectively learns the emotional style disentanglement only with a limited amount of emotional speech data at stage II. The proposed 2-stage training further helps with obtaining better disentangled emotional representation without the support of large emotional speech data.

3.2.2. *Style encoder vs. emotion encoder*

During the emotion training, the style encoder acts as an emotion encoder and takes emotion ID as the input to effectively learn the emotion representation in the speech. To validate our idea, we use t-SNE [34] to visualize the emotion embedding of the reference utterances, which are derived by the style encoder from stage I and the emotion encoder from stage II respectively. To our delight, as shown in Figure 1, 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. From Figure 1, we also observe a significant separation between the emotions with lower values of arousal and valence such as neutral and sad, and those with higher values such as angry, happy and surprise. These observations further validate our 2-stage training for EVC.

3.3. Run-time inference

At run-time, we use the emotion encoder to generate the emotion embeddings from a set of reference utterances belonging to the same emotion category. We use the average emotion embedding to represent the emotion style. Given a source utterance and the intended emotion category, we use a seq2seq ASR encoder to derive the linguistic embedding of the source utterance, and apply the respective emotion embedding to the decoder. The converted acoustic features can be reconstructed by the seq2seq decoder.

3.4. Comparison with related work

The proposed seq2seq EVC framework shares a similar motivation with [28, 30] in terms of leveraging TTS but differs in many aspects. To start with, [28] only focuses on speaker disentanglement, and emotion has not been considered. The proposed 2-stage training strategy allows the network to learn emotion style in training stage II, thus, requires a smaller amount of emotional speech data. Compared with [30] which needs more than *30 hours* of emotional speech data for training, our proposed framework only uses *less than 50 minutes* of emotional speech data. Besides, we further employ adversarial training with an emotion classifier to learn a better emotion disentanglement and use a seq2seq ASR with an explicit loss between the linguistic embedding of EVC and TTS to get a better alignment.

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

4. Experiments

We conduct emotion conversion from neutral to angry, sad, happy and surprise, denoted as *Neu-Ang*, *Neu-Sad*, *Neu-Hap*, and *Neu-Sur* respectively. We first use VCTK corpus [35] for stage I as shown in Figure 2(a), and then use the ESD database [36] for stage II as shown in Figure 2(b). For each emotion pair, we use 300 utterances for training, 30 utterances for reference, and 20 utterances for evaluation. The total duration of emotional speech data used in the stage II is around *50 minutes*, which is small in the context of seq2seq training.

The codes and implementation details of this work are publicly available at: `https://github.com/KunZhou9646/seq2seq-EVC`. We implement two state-of-the-art methods, together with 4 seq2seq EVC systems:

- CycleGAN-EVC [15] (*baseline*): CycleGAN-based emotional voice conversion with WORLD vocoder.
- StarGAN-EVC [16] (*baseline*): StarGAN-based emotional voice conversion with WORLD vocoder.
- Baseline Seq2seq-EVC (*baseline*): Seq2seq-EVC trained directly with limited ESD data without any pre-training, and followed by a Griffin-Lim vocoder [37];
- Seq2seq-EVC-GL (*proposed*): Seq2seq-EVC followed by a Griffin-Lim vocoder;
- Seq2seq-EVC-WA1 (*proposed*): Seq2seq-EVC followed by a WaveRNN vocoder [31] that is pre-trained on VCTK corpus;
- Seq2seq-EVC-WA2 (*proposed*): Seq2seq-EVC followed by a WaveRNN vocoder that is pre-trained on VCTK corpus, and fine-tuned with limited ESD data.

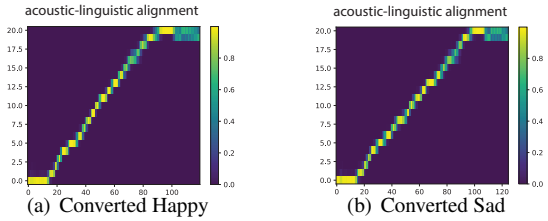
We note that CycleGAN-EVC only can perform the one-to-one conversion, thus we train one CycleGAN-EVC for each emotion pair separately. Both StarGAN-EVC and our proposed Seq2seq-EVC use a unified model for all the emotion pairs.

4.1. Objective Evaluation

We calculate Mel-cepstral distortion (MCD) [5] and the average absolute differences of the utterance duration (DDUR) [28]

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



(a) Converted Happy (b) Converted Sad

Figure 3: *Acoustic-linguistic alignment visualization of utterances that are converted from neutral to (a) happy and (b) sad.*

for the voiced parts to measure the spectral distortion and duration difference respectively. In Seq2seq-EVC models, Mel-spectrograms are adopted as acoustic features, and the Mel-cepstral coefficients (MCEPs) are extracted directly from the waveform to calculate MCD values.

To motivate our proposed 2-stage training, we first conduct experiments with the Baseline Seq2seq-EVC model that is trained directly with limited ESD data without any pre-training procedures. We note that in all experiments, it consistently achieves the worst results in terms of MCD and DDUR values. This observation shows that seq2seq-EVC does not work well with limited training data, which further shows the necessity of our proposed 2-stage training strategy.

As shown in Table 1, our proposed Seq2seq-EVC models with Griffin-Lim, WaveRNN and fine-tuned WaveRNN always outperform baseline CycleGAN-EVC and StarGAN-EVC. We also note that the proposed Seq2seq-EVC-WA1 and Seq2seq-EVC-WA2 consistently achieve the best results of MCD values for all the emotion pairs, which shows the effectiveness of our proposed 2-stage training strategy for limited data EVC.

We note that the attention mechanism allows us to vary the phonetic duration from source to target during the decoding, which is crucial for EVC. Figure 3 shows an example of the acoustic-linguistic alignment for (a) converted happy and (b) converted sad utterances. We note that the source utterance is the same for both conversion mappings, while the converted utterances have different duration. These results further show that our proposed framework is capable of duration manipulation.

We further report DDUR results to evaluate duration conversion performance in Table 2. We observe that the proposed Seq2seq-EVC-WA1 (with WaveRNN) and Seq2seq-EVC-WA2 (with fine-tuned WaveRNN) consistently achieves the best DDUR results for all emotion pairs. We noted that baseline frameworks CycleGAN-EVC and StarGAN-EVC cannot modify speech duration, hence they are not reported in the table. These results show the effectiveness of our proposed Seq2seq-EVC framework in terms of duration conversion.

4.2. Subjective Evaluation

We conduct listening tests to assess the emotion similarity and speech quality. 15 subjects participated in all the experiments and each listened to 128 converted utterances in total.

We first report the emotion similarity results as shown in Figure 4. We use the baseline frameworks CycleGAN-EVC and StarGAN-EVC; and the proposed framework Seq2seq-EVC with Griffin-Lim (Seq2seq-EVC-GL), WaveRNN (Seq2seq-EVC-WA1), and fine-tuned WaveRNN (Seq2seq-EVC-WA2).

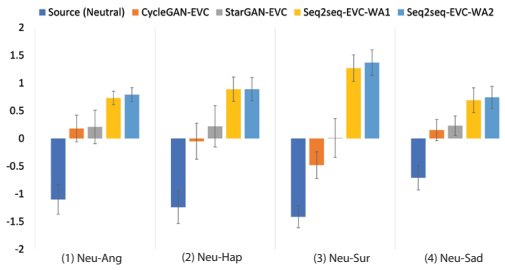


Figure 4: *Emotional similarity results with 95% confidence interval to evaluate emotion similarity with target speech in a scale of -2 to 2 (-2: absolutely different; -1: different; 0: cannot tell; +1: similar; +2: absolutely similar).*

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%**

All participants are asked to listen to the reference target speech first, and then score the speech samples in terms of the emotion similarity to the reference target speech. It is encouraging to see that the proposed Seq2seq-EVC framework with WaveRNN (Seq2seq-EVC-WA1) and fine-tuned WaveRNN (Seq2seq-EVC-WA2) significantly outperform the baselines for all the emotion pairs, especially for Neu-Sur.

We further conduct the best-worst scaling (BWS) [38] test in terms of speech quality of our proposed Seq2seq-EVC framework with 1) Griffin-Lim (Seq2seq-EVC-GL), 2) WaveRNN (Seq2seq-EVC-WA1), and 3) fine-tuned WaveRNN (Seq2seq-EVC-WA2). All participants are asked to choose the best one and the worst one in terms of the overall quality. From Table 3, Seq2seq-EVC-WA2 outperforms the baseline consistently, which proves the effectiveness of our emotional fine-tuning strategy on WaveRNN vocoder.

5. Conclusion

In this paper, we propose a novel training strategy for seq2seq emotional voice conversion leveraging text-to-speech without the need for parallel data. To our best knowledge, this is the first work of seq2seq emotional voice conversion that only needs a limited amount of emotional speech data to train. Moreover, the proposed framework can do many-to-many emotional voice conversion, and conduct spectral and duration mapping at the same time. We also investigate the training strategy of emotion fine-tuning for WaveRNN vocoder training. Experimental results show a significant improvement of the conversion performance over the baselines.

6. Acknowledgment

The research is funded by SUTD Start-up Grant Artificial Intelligence for Human Voice Conversion (SRG ISTD 2020 158) and SUTD AI Grant - Thrust 2 Discovery by AI (SG-PAIRS1821), the National Research Foundation, Singapore under its AI Singapore Programme (Award No: AISG-GC-2019-002) and (Award No: AISG-100E-2018-006), and its National Robotics Programme (Grant No. 192 25 00054), and by RIE2020 Advanced Manufacturing and Engineering Programmatic Grants A1687b0033, and A18A2b0046.

7. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

[2] D. Schuller and B. W. Schuller, "The age of artificial emotional intelligence," *Computer*, vol. 51, no. 9, pp. 38–46, 2018.

[3] R. Liu, B. Sisman, G. Gao, and H. Li, "Expressive tts training with frame and style reconstruction loss," *arXiv preprint arXiv:2008.01490*, 2020.

[4] Y. Xu, "Speech prosody: A methodological review," *Journal of Speech Sciences*, vol. 1, no. 1, pp. 85–115, 2011.

[5] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[6] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[7] B. Sisman, M. Zhang, and H. Li, "Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 1085–1097, 2019.

[8] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[9] B. Sisman, M. Zhang, M. Dong, and H. Li, "On the study of generative adversarial networks for cross-lingual voice conversion," *IEEE ASRU*, 2019.

[10] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv preprint arXiv:1711.11293*, 2017.

[11] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[12] R. Aihara, R. Ueda, T. Takiguchi, and Y. Ariki, "Exemplar-based emotional voice conversion using non-negative matrix factorization," in *APSIPA ASC*. IEEE, 2014.

[13] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion," *Interspeech 2016*, pp. 2453–2457, 2016.

[14] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using dual supervised adversarial networks with continuous wavelet transform f0 features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1535–1548, 2019.

[15] K. Zhou, B. Sisman, and H. Li, "Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 230–237.

[16] 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 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3502–3506.

[17] R. Shankar, J. Sager, and A. Venkataraman, "Non-parallel emotion conversion using a deep-generative hybrid network and an adversarial pair discriminator," *Proc. Interspeech 2020*, pp. 3396–3400, 2020.

[18] J. Gao, D. Chakraborty, H. Tembine, and O. Olaleye, "Nonparallel emotional speech conversion," *Proc. Interspeech 2019*, pp. 2858–2862, 2019.

[19] K. Zhou, B. Sisman, and H. Li, "Vaw-gan for disentanglement and recomposition of emotional elements in speech," *IEEE Spoken Language Workshop (SLT)*, 2020.

[20] 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.

[21] K. Zhou, B. Sisman, M. Zhang, and H. Li, "Converting Anyone's Emotion: Towards Speaker-Independent Emotional Voice Conversion," *Proc. Interspeech 2020*, pp. 3416–3420, 2020.

[22] K. K. J. F. S. Kyle, K. A. C. Y. B. Jose, and S. M. Sotelo, "Char2wav: End-to-end speech synthesis," in *International Conference on Learning Representations, workshop*, 2017.

[23] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, 2019.

[24] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6805–6809.

[25] H. Kameoka, K. Tanaka, D. Kwaśny, T. Kaneko, and N. Hojo, "Convs2s-vc: Fully convolutional sequence-to-sequence voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1849–1863, 2020.

[26] J.-X. Zhang, Z.-H. Ling, Y. Jiang, L.-J. Liu, C. Liang, and L.-R. Dai, "Improving sequence-to-sequence voice conversion by adding text-supervision," in *IEEE ICASSP*, 2019, pp. 6785–6789.

[27] M. Zhang, X. Wang, F. Fang, H. Li, and J. Yamagishi, "Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet," *Proc. Interspeech 2019*, pp. 1298–1302, 2019.

[28] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 540–552, 2019.

[29] C. Robinson, N. Obin, and A. Roebel, "Sequence-to-sequence modelling of f0 for speech emotion conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6830–6834.

[30] T.-H. Kim, S. Cho, S. Choi, S. Park, and S.-Y. Lee, "Emotional voice conversion using multitask learning with text-to-speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7774–7778.

[31] 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 *International Conference on Machine Learning*, 2018, pp. 2410–2419.

[32] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[33] D. M. Schuller and B. W. Schuller, "A review on five recent and near-future developments in computational processing of emotion in the human voice," *Emotion Review*, p. 1754073919898526, 2020.

[34] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[35] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[36] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *arXiv preprint arXiv:2105.14762*, 2021.

[37] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[38] B. Sisman, H. Li, and K. C. Tan, "Sparse representation of phonetic features for voice conversion with and without parallel data," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop*.