



An Overview & Analysis of Sequence-to-Sequence Emotional Voice Conversion

Zijiang Yang¹, Xin Jing¹, Andreas Triantafyllopoulos¹, Meishu Song^{1,2},
Ilhan Aslan³, Björn W. Schuller^{1,4}

¹Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

²Educational Physiology Laboratory, University of Tokyo, Japan

³Device Software Lab, Munich Research Center, Huawei Technologies, Germany

⁴GLAM – Group on Language, Audio, & Music, Imperial College London, UK

zijiang.yang@ieee.org

Abstract

Emotional voice conversion (EVC) focuses on converting a speech utterance from a source to a target emotion; it can thus be a key enabling technology for human-computer interaction applications and beyond. However, EVC remains an unsolved research problem with several challenges. In particular, as speech rate and rhythm are two key factors of emotional conversion, models have to generate output sequences of differing length. Sequence-to-sequence modelling is recently emerging as a competitive paradigm for models that can overcome those challenges. In an attempt to stimulate further research in this promising new direction, recent sequence-to-sequence EVC papers were systematically investigated and reviewed from six perspectives: their motivation, training strategies, model architectures, datasets, model inputs, and evaluation methods. This information is organised to provide the research community with an easily digestible overview of the current state-of-the-art. Finally, we discuss existing challenges of sequence-to-sequence EVC.

Index Terms: affective computing, emotional text-to-speech, emotional voice conversion, sequence-to-sequence

1. Introduction

Speech technology has come a long way in the quest to enable human-like interactions with machines, with research increasingly addressing challenges, originally introduced in the field of affective computing [1]. However, while humans easily convey and react to emotions in interpersonal conversations, today's machines still struggle to synthesise basic emotional speech.

Emotional text-to-speech (ETTS), emotional voice synthesis (EVS), and emotional voice conversion (EVC) would provide user experience designers a powerful new tool to manage and navigate the challenging emotional context in conversational speech interfaces with humans [2]. Ultimately, emotional speech generation and conversion is able to drastically change the value of many products with speech interfaces. ETTS or EVS [3, 4, 5] aim to *directly synthesise* speech with emotional expressivity *from text* – which is a key aspect of text-to-speech (TTS) naturalness that is currently missing. Meanwhile, EVC [6] aims to *convert* the state of a spoken utterance from one emotion to another, while preserving the linguistic information and speaker identity. One could argue that EVC takes a ‘shortcut’ compared to ETTS, since it endows an existing utterance with emotional intonation, rather than synthesising one from the ground up. This allows EVC to be added as an extra step in a traditional TTS pipeline – first synthesise, and then convert to the target emotion. This decomposition into two constituents can reduce computational complexity and dependence on data.

Previous surveys have largely concentrated on ETTS [7, 8]; however, these are mostly outdated as they refer to literature prior to the significant milestones achieved by neural speech synthesis. Moreover, recent surveys that incorporate the deep learning (DL) paradigm have been targeted to ‘standard’ TTS (i. e. without emotional information) [9]. The most relevant survey is that of [6], which provides a comprehensive overview of EVC. However, it devotes little attention to sequence-to-sequence (seq2seq) models, and largely concentrates on generative adversarial networks (GANs) or spectrum and prosody mapping techniques. Thus, it leaves a small –but noteworthy– gap of DL-based approaches for EVC that this contribution attempts to fill.

Frame-to-frame spectral mapping is the mainstream in previous studies [10, 11, 12], however, emotion is inherently suprasegmental and complex with multiple signal attributes concerning both the spectrum and prosody. Thus, frame-based mapping of spectral features of the source and target is insufficient to convert the emotion. Recently, the seq2seq speech synthesis framework raises much interests in EVC.

This research field has recently benefited from the advent of new machine learning techniques such as deep neural networks. Therefore, this paper aims to give an overview of recent developments, pointing out the inherent properties of the various synthesis techniques used, summarising the prosody rules employed, and analysing the evaluation paradigms. Finally, an attempt is made to discuss the existing challenges in EVC.

2. Overview

Seq2seq learning was initially proposed for machine translation by Sutskever *et al.* [13] and has since proved its competitiveness in several natural language processing tasks [14, 15, 16]. Seq2seq models consist of two main modules: the encoder and the decoder. Unlike the decoder taking the output of the encoder like in a standard autoencoder model [17], seq2seq models generate the prediction of the timestep t by using the prediction of the timestep $t - 1$ as the input of the decoder [13]. Therefore, seq2seq models are able to generate outputs with different, variable length. Conventional EVC, from the basic neural network architecture [10] to the recent research with GANs [11, 12, 18], applied frame-to-frame conversion on the spectrum and prosody, which indicates that the converted speech has the same length with the source speech. However, one of the vital characteristic of emotion expression is speech rate [19], which cannot be expressed within the fixed length obtained by using the frame-to-frame EVC [20]. Meanwhile, the dependency between the spectrum and prosody leads to respective conversion mistakes, such as the mismatch when doing a separate study on them [21]. Furthermore, the emotional expression in an utterance often

shows only on part of the utterance (e. g. in only some, but not all, words). Seq2seq models can naturally handle this requirement by adding attention [22], which makes it possible to focus on only those relevant parts [23]. For these reasons, seq2seq models show promising performance compared to conventional methods.

To summarise, there are three main advantages of seq2seq EVC models:

1. Seq2seq models are able to learn feature mapping, alignment, and duration prediction simultaneously;
2. Seq2seq models avoid mistakes caused by the respective mapping of spectrum and prosody;
3. The attention mechanism helps seq2seq models focus on the emotionally emphasised parts in the utterance.

On the other hand, seq2seq training always requires a large-sized dataset [24]. Moreover, the training data should be parallel, which means the same content should be expressed in different emotion categories by the same speaker – a requirement not necessary for other models, like GANs [12, 18]. This represents the major challenge for such models, which has so far prevented the seq2seq paradigm from becoming the dominant one in the field of EVC. As a result, at the time of writing, only 6 papers exist which use seq2seq EVC [20, 21, 23, 25, 26, 27]. These works constitute the background material for this review, and will be comprehensively analysed in the following sections. Table 1 contains the important information in a condensed form.

2.1. Motivation

Robinson *et al.* [26] were the first to introduce a seq2seq model to the EVC task. They presented a model based on converting F0 of the speech. A three-step procedure, including F0 extraction, transformation, and subsequent application of the resulting contour on the signal, is able to convert any neutral speech to the speech with one specific emotional category. In other words, this is a one-to-one neutral-to-emotional EVC model.

Kim *et al.* [25] addressed a major problem plaguing both voice conversion (VC) and EVC tasks – mispronunciation. Instead of text-supervision [28], TTS was introduced to seq2seq EVC for guiding the linguistic information [25]. Furthermore, this was the first many-to-many EVC system based on a seq2seq mechanism, which was facilitated by feeding a reference speech with the target emotion.

A one-to-many seq2seq VC model was presented by Zhao *et al.* [27]. The authors focused on training efficiency and stability by manually balancing the word distribution and increasing the proportion of uncommon words in the dataset. In this way, the size of the training dataset could be also decreased to achieve a similar or even better performance. With the implementation of an emotion encoder, the model is able to convert high-quality emotional speech.

Considering the fact that it is impractical to find a large-sized parallel emotional dataset suitable for seq2seq training, Zhou *et al.* [23] presented in their recent paper a training strategy called ‘two-stage training’, including style initialisation with a TTS dataset and emotion training. This is able to help the many-to-many EVC model improve its performance by using only a small-sized parallel emotional dataset.

Finally, the last two papers focus on emotional intensity control. A key difference is that Choi and Hahn [20] required a parallel multi-speaker emotional dataset with the help of a speaker encoder, while the most recent solution in Zhou *et al.*

[21] only requires a small-sized parallel single-speaker emotional dataset. However, Choi and Hahn [20] used a weight to control the emotional intensity by multiplying it with the emotion embedding, while Zhou *et al.* [21] trained the model with variations of intensity without any annotation on it (cf. Section 2.3).

2.2. Training Strategies

One of the most common training strategies utilised for seq2seq models is called *teacher forcing*. In seq2seq training, the *generated* frame of the previous timestep will be fed into the decoder to generate the frame of the current timestep (using the pre-defined start of the sentence token to generate the frame of the first timestep) [13]. However, this causes a problem when training, because the error will be accumulated during generating. Moreover, the generating takes a long time, since the generating is frame-by-frame. Teacher forcing feeds the *ground truth* frame instead of the generated frame in the training phase. Therefore, it has the ability to help the model learn faster and more accurately, especially at the beginning of the training [15].

However, specific challenges arising in EVC require specialised training regiments. In order to solve mispronunciation and training instability without explicit alignment mechanisms, Kim *et al.* [25] applied multi-task learning by introducing TTS to the EVC task. Besides the content encoder which generates linguistic embedding by using the source speech, a text encoder was implemented to encode the input text to a linguistic embedding. Then, during training, the model was randomly tasked to perform either EVC or TTS – a form of alternating multi-task learning which helped it avoid mispronunciation errors.

Since it is impractical to use a large-sized parallel dataset to fulfil the requirements of seq2seq EVC training, Zhou *et al.* [23] proposed a two-stage training strategy, which begins with a style initialisation phase with a large-sized TTS corpus before doing emotional fine-tuning on a small-sized emotional dataset. Moreover, an emotion classifier was utilised in adversarial fashion to eliminate the emotional information in the linguistic embedding [23]. This adversarial training strategy was aimed to optimise the performance on disentangling the style/emotional information and the linguistic information. Zhao *et al.* [27] presented a similar work by utilising an emotional embedding to the pre-trained VC model.

Zhou *et al.* [21] improved their model by adding emotion supervision training with a pre-trained speech emotion recognition (SER) module. Accounting for the fact that the reconstruction loss between the target speech and the converted speech does not incorporate human emotional perception, a SER module was introduced to compute two perceptual losses: emotion classification loss and emotion embedding similarity loss, thus optimising the emotional perception of the converted speech.

2.3. Model Architectures

As the first to explore seq2seq EVC, Robinson *et al.* [26] used the simplest model architecture, comprising one encoder, one decoder, and one attention module. The encoder accepts the extracted features as the input and generates the context vector, while the decoder uses this vector and the previous frames to generate the converted features frame by frame. The attention mechanism is used to provide an explicit alignment between the input (source) and the output (converted).

Compared to the basic architecture above, Kim *et al.* [25] modified it on the encoder by using three encoders instead of one: a style, a content, and a text one. In the training phase, reference speech was sent to the style encoder for the style embedding,

Table 1: Information of all sequence-to-sequence EVC papers. **Abbreviations:** **ABX:** ABX test, **BWS:** Best-Worst Scaling, **CER:** Character Error Rate, **CS:** Cosine Similarity, **CTC:** Connectionist Temporal Classification, **DDUR:** Differences of Duration, **FFE:** F0 Frame Error, **GPE:** Gross Pitch Error, **MCD:** Mel-cepstral Distortion, **MOS:** Mean Opinion Score, **SSER:** Subjective Speech Emotion Recognition, **VDE:** Voicing Decision Error, **WER:** Word Error Rate.

Paper	Highlights	Feature Set & Vocoder	Emotional Dataset	Language	Emotional Model	Evaluation Methods	Public Code
[26]	First work Syllable-level conversion	F0 SuperVP	~1 100 syllables	French	One-to-one	SSER	✓ ¹
[25]	Multi-task learning	Log Mel-spectrogram Griffin-Lim Algorithm	21 000 utterances	Korean	Many-to-many	WER CS MOS ABX	✓ ²
[27]	Data redundancy reduction CTC leverage EVC fine-tuning	Log Mel-spectrogram HiFi-GAN	6 000 utterances	Chinese	One-to-many	WER CER MOS	✗
[20]	Multi-speaker emo. dataset Context preservation Emotional intensity	Log Mel-spectrogram Parallel WaveGAN	4 000 utterances	Korean	One-to-many	MCD VDE GPE FFE MOS ABX SSER	✗
[23]	Two-stage training Small emo. dataset	Log Mel-spectrogram WaveRNN	350 utterances	English	Many-to-many	MCD DDUR MOS BWS	✓ ³
[21]	Style-pretraining Emotion supervision training Small emo. dataset Emotional intensity	Log Mel-spectrogram Parallel WaveGAN	350 utterances	English	Many-to-Many	MCD DDUR MOS BWS	✓ ⁴

¹ <https://github.com/carl-robinson/voice-emotion-seq2seq>

² <https://github.com/ktho22/vctts>

³ <https://github.com/KunZhou9646/seq2seq-EVC>

⁴ <https://github.com/KunZhou9646/Emovox>

while the contents encoder and the text encoder generated the linguistic embedding from the source speech and the input text, respectively. Then, the style embedding along with the linguistic embedding from either the speech or the text were fed into the decoder to construct the converted emotional speech, in a multi-task way (cf. Section 2.2). In the end, this model has the ability to perform both an EVC task (without the text encoder) and an emotional TTS task (without the content encoder).

Zhou *et al.* [23] also used three encoders: a style/emotion encoder for style/emotion information, seq2seq automatic speech recognition (ASR) utilised for linguistic information from speech features, and a text encoder which is also used to capture for linguistic information but from the input text instead. Furthermore, an emotion classifier was applied to optimise the linguistic embedding obtained from the source speech.

Following up on this, Zhou *et al.* [21] added two extra modules to control the emotional intensity and improve the emotional expressivity of the output speech. Based on the assumption that the emotional intensity can be regarded as the relative difference from the neutral speech (zero intensity) to the emotional speech, relative attributes were applied to train the emotional intensity modelling without any explicit labels. Subsequently, the intensity embedding, which can be derived from the reference speech or given manually, was concatenated with the emotion embedding and the resulting embedding was fed into the decoder to reconstruct the emotional speech with the required intensity. Furthermore, they added a pre-trained SER model and used it to generate two perceptual losses to improve performance. An emotion classification loss was computed by the converted emotional speech being classified by the SER model and compared with the ground truth emotional category, whereas an emotion embedding

similarity loss was computed by using the emotion embedding obtained from the emotion encoder and the SER embedding obtained by sending the converted speech to the SER model. Visualised results showed the perceptual losses helped the emotion encoder discriminate the different emotion categories.

Finally, Choi and Hahn [20] and Zhao *et al.* [27] both applied a speaker encoder, which is used for disentangling the speaker information and makes the use of a multi-speaker emotional dataset possible. Moreover, Choi and Hahn [20] implemented one source decoder and one target decoder in the training phase, to make sure that the content embedding of the source and the target speech preserve the contextual information by comparing the output of these decoders with the source and the target speech, respectively. On the other hand, Zhao *et al.* [27] applied a length regulator module for the length alignment between the encoders and the decoder, and used a connectionist temporal classification (CTC) recogniser [29] after the decoder to guide the alignment between the text and the speech to improve the performance of the EVC model.

2.4. Datasets

To achieve a decent performance by using seq2seq training, a large-sized dataset is very essential [24]. Specifically, seq2seq EVC training requires a large-sized, parallel, one-speaker, emotional dataset. For instance, Kim *et al.* [25] utilised a Korean emotional dataset (mKETTS) including 3 000 utterances per emotional category pronounced by one male speaker, and there are 7 different emotions in total (*neutral, anger, disgust, fear, happiness, sadness, and surprise*). In Robinson *et al.* [26], the dataset used contains only 200 emotional utterances (10 sentences \times 4 emotional category \times 5 levels of intensity, *anger,*

joy, fear, and sadness) recorded by one French actress. However, the researched conversion in this work was on the syllable-level. Thus, the model was trained by using around 1 100 syllable pairs after forced alignment [30].

A speaker encoder allows the use of a multi-speaker dataset – a method utilised by both Zhao *et al.* [27] and Choi and Hahn [20]. Zhao *et al.* [27] used a Chinese emotional dataset including three speakers, three emotional categories (*anger, happiness, and sadness*) and 6 hours of recording to fine-tune their pre-trained VC model. Similarly, the dataset in Choi and Hahn [20] includes 100 sentences in 4 different emotional categories (*neutral, anger, happiness, and sadness*) pronounced by 5 Korean actors and 5 Korean actresses, for a total of 4 000 utterances.

Instead of using a speaker encoder to expand the range of usable emotional datasets, Zhou *et al.* [21, 23] utilised two-stage training on the model to reduce its dependency on the size of the emotional dataset. At the first stage, about 30 hours of recordings recorded by 99 speakers from a multi-speaker corpus called VCTK [31] were applied to pre-train the model. Then, only 350 pairs of emotional speeches from ESD [6] were used to fine-tune, enabling them to improve their performance.

2.5. Model Inputs

Most seq2seq EVC research uses log Mel-spectrograms [20, 21, 23, 25, 27], and the applied vocoders include Parallel WaveGAN [32], WaveRNN [33], and HiFi-GAN [34]. Since Robinson *et al.* [26] focused on F0 conversion on the syllable-level, they estimated the source F0 contour and used the converted F0 contour to synthesise the speech with the SuperVP vocoder.

An important consideration concerns the control of the target emotion. Besides the simple one-to-one EVC [26], there are three different methods to bring the information of the target emotion to the model. The most direct method is feeding an emotion ID to the emotion encoder [27]. In Kim *et al.* [25], an emotion-reference speech was utilised in the inference phase to guide the model. Emotional embeddings were also used [20, 21, 23], where they were calculated by the average of a set of emotional speech embeddings from the same emotional category.

2.6. Evaluation Methods

There are two types of evaluation methods applied in this field: objective and subjective ones. In general, objective evaluation entails the calculation of some measure of difference or correlation between the output and the target. For example, Kim *et al.* [25] used word error rate (WER) for the linguistic consistency and cosine similarity for the performance on emotion conversion. Similarly, both WER and character error rate (CER) were applied in [27]. Zhou *et al.* [21, 23] preferred using Mel-cepstral distortion (MCD) to measure the spectral changes during the conversion, and the differences of duration (DDUR) for the performance on the duration of the converted speech. Besides MCD, Choi and Hahn [20] applied three other objective evaluation methods: voicing decision error (VDE), gross pitch error (GPE), and F0 frame error (FFE), for a more comprehensive analysis of their results.

Moreover, several different subjective evaluation methods were applied: Robinson *et al.* [26] did a subjective emotion recognition survey with 87 participants to measure the performance of the EVC model from a human perspective. Additionally, mean opinion score (MOS) [20, 21, 23, 27], ABX test (identifying whether sample X is from class A or B) [20], and best worst scaling (BWS) [21, 23] on the naturalness, clarity, and similarity were also applied in prior works.

3. Challenges and Conclusion

Although an alternative solution was proposed by Zhou *et al.* [21, 23] to alleviate the requirement on a large-sized parallel emotional dataset, the main challenge of seq2seq EVC task remains the availability of appropriate public datasets. Existing datasets are small or lacking in quality. For example, the newly published dataset ESD [6] is a clean, parallel dataset but includes only 350 utterances per speaker. EmoV-DB [35] has more pairs of samples; however, there are non-speech utterances included, such as laughter and yawn. The collection and release of suitable datasets to the public would foster further research in this promising field and help improve the performance.

Another remaining challenge is the difficulty in comparing different EVC models. Objective evaluation methods are not intuitive enough because they only indicate how ‘close’ or ‘similar’ the converted speech and the target speech are with respect to some metric – however, this is not a guarantee that human perception will also consider them as close. Using evaluation methods based on combinations of pre-trained SER and ASR models, or trying to predict subjective evaluation scores –as is the target of the INTERSPEECH 2022 VoiceMOS challenge¹– are both promising methods of mitigating this challenge. However, until those methods mature enough to enable their usage for practical applications, subjective annotations will remain the gold-standard for EVC evaluations.

Nevertheless, those come with their own drawbacks, for example, when participants between studies are biased (e. g. due to different cultural backgrounds). Using different validation sentences from different datasets is another problem, since context strongly affects the emotion that humans experience [36]. Finally, different evaluation methods on the same criteria can make it more challenging to compare. For example, Choi and Hahn [20] applied MOS while Zhou *et al.* [21] utilised BWS on the speech similarity assessment. With improvements in EVC quality, there will be a need to move away from simple MOS to evaluate utterances which are detached from any context or interaction scenarios. New qualitative evaluation methods will be needed to identify fine grained differences between different EVC solutions, considering conversation contexts and speaker intents in more detail. To this end, inspiration can also be taken from related fields, especially the field of Human-Robot Interaction, where researchers usually evaluate the affect of robots in interactive settings, analysing both users’ immediate reactions and their preferences. For example, Ritschel *et al.* [37] have evaluated the effect of a robot converting non-ironic utterances into ironic utterances in small talks with users in order to be more likeable.

In conclusion, seq2seq EVC is a promising, rapidly maturing research field. While there still remain several challenges (which are not unique to this paradigm), these models have the potential to improve the performance of EVC applications, thus leading to more intelligent human-computer interactions. We presented a short, concise review of recent approaches, which we hope to also fuel novel approaches.

4. Acknowledgements

This research was partially supported by the Affective Computing & HCI Innovation Research Lab between Huawei Technologies and University of Augsburg, and the China Scholarship Council (CSC), Grant # 202006290013.

¹<https://voicemos-challenge-2022.github.io/>

5. References

- [1] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben, E. Marchi, A. Vinciarelli, K. Scherer, *et al.*, “Affective and behavioural computing: Lessons learnt from the first computational paralinguistics challenge,” *Computer Speech & Language*, vol. 53, pp. 156–180, 2019.
- [2] D. Deibel, R. Evanhoe, and K. Vellios, *Conversations with Things: UX Design for Chat and Voice*. Rosenfeld Media, 2021, ISBN: 9781933820262.
- [3] O. Kwon, I. Jang, C. Ahn, and H.-G. Kang, “An effective style token weight control technique for end-to-end emotional speech synthesis,” *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1383–1387, 2019.
- [4] H. Choi, S. Park, J. Park, and M. Hahn, “Multi-speaker emotional acoustic modeling for CNN-based speech synthesis,” in *Proc. ICASSP*, Brighton, UK, 2019, pp. 6950–6954.
- [5] O. Kwon, I. Jang, C. Ahn, and H.-G. Kang, “Emotional speech synthesis based on style embedded Tacotron2 framework,” in *Proc. ITC-CSCC*, Jeju, Korea, 2019, pp. 1–4.
- [6] K. Zhou, B. Sisman, R. Liu, and H. Li, “Emotional voice conversion: Theory, databases and ESD,” *Speech Communication*, vol. 137, pp. 1–18, 2022.
- [7] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, “Emotion representation, analysis and synthesis in continuous space: A survey,” in *Proc. FG*, Santa Barbara, CA, USA, 2011, pp. 827–834.
- [8] M. Schröder, “Emotional speech synthesis: A review,” in *Proc. EUROSPEECH*, Aalborg, Denmark, 2001, pp. 561–564.
- [9] X. Tan, T. Qin, F. Soong, and T.-y. Liu, “A survey on neural speech synthesis,” *arXiv preprint arXiv:2106.15561*, 2021.
- [10] Z. Luo, T. Takiguchi, and Y. Ariki, “Emotional voice conversion using deep neural networks with MCC and F0 features,” in *Proc. ICIS*, Okayama, Japan, 2016, pp. 1–5.
- [11] 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 *Proc. ICASSP*, Barcelona, Spain, 2020, pp. 3502–3506.
- [12] K. Zhou, B. Sisman, M. Zhang, and H. Li, “Converting anyone’s emotion: Towards speaker-independent emotional voice conversion,” in *Proc. INTERSPEECH*, Shanghai, China, 2020, pp. 3416–3420.
- [13] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Proc. NIPS*, Montreal, QC, Canada, 2014, pp. 3104–3112.
- [14] R. Liu, X. Chen, and X. Wen, “Voice conversion with transformer network,” in *Proc. ICASSP*, Barcelona, Spain, 2020, pp. 7759–7763.
- [15] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, *et al.*, “State-of-the-art speech recognition with sequence-to-sequence models,” in *Proc. ICASSP*, Calgary, AB, Canada, 2018, pp. 4774–4778.
- [16] Y. Wang, R. J. 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,” in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 4006–4010.
- [17] M. Elgaar, J. Park, and S. W. Lee, “Multi-speaker and multi-domain emotional voice conversion using factorized hierarchical variational autoencoder,” in *Proc. ICASSP*, Barcelona, Spain, 2020, pp. 7769–7773.
- [18] K. Zhou, B. Sisman, and H. Li, “Transforming spectrum and prosody for emotional voice conversion with non-parallel training data,” in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, Tokyo, Japan, 2020, pp. 230–237.
- [19] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, “The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.
- [20] H. Choi and M. Hahn, “Sequence-to-sequence emotional voice conversion with strength control,” *IEEE Access*, vol. 9, pp. 42 674–42 687, 2021.
- [21] K. Zhou, B. Sisman, R. Rana, B. W. Schuller, and H. Li, “Emotion intensity and its control for emotional voice conversion,” *arXiv preprint arXiv:2201.03967*, 2022.
- [22] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proc. ICLR*, San Diego, CA, USA, 2015, pp. 1–15.
- [23] K. Zhou, B. Sisman, and H. Li, “Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training,” in *Proc. INTERSPEECH*, Brno, Czechia, 2021, pp. 811–815.
- [24] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, “Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining,” in *Proc. INTERSPEECH*, Shanghai, China, 2020, pp. 4676–4680.
- [25] T.-H. Kim, S. Cho, S. Choi, S. Park, and S.-Y. Lee, “Emotional voice conversion using multitask learning with text-to-speech,” in *Proc. ICASSP*, Barcelona, Spain, 2020, pp. 7774–7778.
- [26] C. Robinson, N. Obin, and A. Roebel, “Sequence-to-sequence modelling of F0 for speech emotion conversion,” in *Proc. ICASSP*, Brighton, UK, 2019, pp. 6830–6834.
- [27] Z. Zhao, J. Liang, Z. Zheng, L. Yan, Z. Yang, W. Ding, and D. Huang, “Improving model stability and training efficiency in fast, high quality expressive voice conversion system,” in *Proc. ICMI*, Montreal, QC, Canada, 2021, pp. 75–79.
- [28] 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 *Proc. ICASSP*, Brighton, UK, 2019, pp. 6785–6789.
- [29] S. Kim, T. Hori, and S. Watanabe, “Joint CTC-attention based end-to-end speech recognition using multi-task learning,” in *Proc. ICASSP*, New Orleans, LA, USA, 2017, pp. 4835–4839.
- [30] C. Veaux and X. Rodet, “Intonation conversion from neutral to expressive speech,” in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 2765–2768.
- [31] J. Yamagishi, C. Veaux, and K. MacDonald, “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92),” University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2019.
- [32] R. Yamamoto, E. Song, and J.-M. Kim, “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in *Proc. ICASSP*, Barcelona, Spain, 2020, pp. 6199–6203.
- [33] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van der Oord, S. Dieleman, and K. Kavukcuoglu, “Efficient neural audio synthesis,” in *Proc. ICML*, Stockholm, Sweden, 2018, pp. 3775–3784.
- [34] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Proc. NeurIPS*, Vancouver, BC, Canada, 2020, pp. 17 022–17 033.
- [35] A. Adigwe, N. Tits, K. E. Haddad, S. Ostadabbas, and T. Du-toit, “The emotional voices database: Towards controlling the emotion dimension in voice generation systems,” *arXiv preprint arXiv:1806.09514*, 2018.
- [36] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. W. Schuller, and S. Narayanan, “Context-sensitive learning for enhanced audiovisual emotion classification,” *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.
- [37] H. Ritschel, I. Aslan, D. Sedlbauer, and E. André, “Irony man: Augmenting a social robot with the ability to use irony in multimodal communication with humans,” in *Proc. AAMAS*, Montreal, QC, Canada, 2019, pp. 86–94.