Reinforcement Learning for Emotional Text-to-Speech Synthesis with Improved Emotion Discriminability

Rui Liu[1,2], *Berrak Sisman*[1], *Haizhou Li*[2,3]

[1] Singapore University of Technology and Design (SUTD), Singapore
[2] ECE Department, National University of Singapore (NUS), Singapore
[3] Machine Listening Lab, University of Bremen, Germany

`liurui_imu@163.com, berrak_sisman@sutd.edu.sg, haizhou.li@nus.edu.sg`

Abstract

Emotional text-to-speech synthesis (ETTS) has seen much progress in recent years. However, the generated voice is often not perceptually identifiable by its intended emotion category. To address this problem, we propose a new interactive training paradigm for ETTS, denoted as *i-ETTS*, which seeks to directly improve the emotion discriminability by interacting with a speech emotion recognition (SER) model. Moreover, we formulate an iterative training strategy with reinforcement learning to ensure the quality of *i-ETTS* optimization. Experimental results demonstrate that the proposed *i-ETTS* outperforms the state-of-the-art baselines by rendering speech with more accurate emotion style. To our best knowledge, this is the first study of reinforcement learning in emotional text-to-speech synthesis.

Index Terms: Reinforcement Learning, Emotional Text-to-Speech Synthesis, Speech Emotion Recognition.

1. Introduction

Emotional text-to-speech (ETTS) seeks to synthesize human-like natural-sounding voice for a given input text with desired emotional expression. The recent advances have enabled many applications such as virtual assistants, call centers, dubbing of movies and games, audiobook narration, and online education.

The early studies of ETTS are based on hidden Markov models [1–3]. For example, we can synthesize speech with a desired emotion through model interpolation [2] or by incorporating unsupervised expression cluster during training [3]. Recently, deep learning opens up many possibilities for ETTS [4, 5], where emotion codes can be used as control vectors to change text-to-speech (TTS) output in different ways. Successful attempts include global style tokens (GST) that have been used to control the expressiveness in emotional TTS [6–8]. Such approaches typically use style embedding to indicate the emotion rendering and can learn speech variations in an unsupervised manner. We note that the latent style embeddings has no explicit meaning, hence lacks interpretability. Therefore, GST-Tacotron studies often face emotion confusion problem [9], i.e., the projected emotion in the synthetic speech is not rendered accurately.

Recently, some approaches [10–12] were proposed to enhance the interpretability of the style embedding. These approaches add an additional emotion recognition loss [10, 11] or perceptual loss [12] to force the latent style embedding to pay more attention to the emotion rendering. These works have made a great contribution to the development of ETTS. However, they mostly focus on the hidden features [13], not the explicit output; and optimize the output acoustic features

Speech samples: https://ttslr.github.io/i-ETTS.

with mean square error (MSE) loss, which calculates the distance between the synthesized emotional speech and natural reference. We note that such MSE-based objective function does not focus directly on emotion discriminability, hence often face emotion confusion problem in generated speech.

We are motivated by the fact that the interaction of speech emotion recognition model with ETTS can improve the emotion discriminability of synthesized speech, hence overcome emotion confusion problem. We note that ETTS can be trained with SER classification result in a supervised manner. However, this may require a large amount of emotion-labeled speech data, that limits the scope of applications. Therefore, this paper studies the use of reinforcement learning for ETTS in an interactive manner for improved emotion discriminability.

The reinforcement learning (RL) algorithm [14] learns how to achieve a complex goal in an interactive manner. Specifically, RL involves agents to learn their behavior by trial and error [14]. RL agents aim to learn decision-making by successfully interacting with the environment where they operate. It has enabled speech processing systems [15] through a well designed feedback that reflects appropriate perceptual metrics in speech enhancement [16,17], speech recognition [18] and speaker recognition [19]. We note that the use of interactive paradigm with RL algorithm in emotional TTS remains to be explored, which will be the focus of this paper.

In this paper, we propose a novel ETTS framework with reinforcement learning, denoted as *i-ETTS*. The proposed approach aims to overcome emotion confusion problem of synthesized speech by optimizing the ETTS model through an interaction with a SER model. The proposed idea is expected to reduce the requirement of emotion-labeled training data size. Experimental results demonstrate that i-ETTS consistently outperforms the state-of-the-art baselines, by achieving improved emotion discriminability of synthesized speech.

The main contributions of this paper are summarized as follows: 1) we introduce a novel interactive emotional text-to-speech synthesis paradigm that overcomes emotion confusion problem; 2) we formulate the training problem with RL, and optimize the ETTS model with policy gradient and a reward function correlated with the SER accuracy; 3) we utilize an iterative training strategy for a stable training procedure. To our best knowledge, this is the first study of reinforcement learning in emotional speech synthesis.

This paper is organized as follows. We motivate our study through the comparison with existing RL and ETTS frameworks in Section 2. We formulate the proposed paradigm in Section 3. We report the results of a systematic evaluation and comparison in Sections 4. Finally, Section 5 concludes the study.
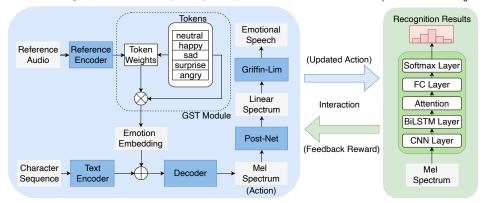
Figure 1: *Block diagram of the proposed interactive paradigm for emotional text-to-speech synthesis with reinforcement learning.*

2. Related Work

The reinforcement learning (RL) algorithm takes an action in an environment in order to maximize the cumulative reward [14]. It has been successfully applied to various tasks, such as computer game [20], natural language processing [21], and speech processing [15–19]. For example, RL has been used to optimize speech enhancement system based on character recognition error. Moreover, it has also been used for DNN-based source enhancement [16] where the objective sound quality assessment scores are given as the reward. Last but not least, RL-based methods have also been proposed for speech recognition [18] and speaker recognition [19]. We note that these approaches utilize limited or unlabeled data and optimize the target model by interacting with the feedback, which correlates with an evaluation metric from a third party directly.

With the advent of deep learning, end-to-end TTS systems, such as Tacotron [22], Tacotron2 [23] and their variants [24, 25] greatly improve the voice quality of the synthesized speech. However, it remains a challenge as to how we generate speech with an intended emotion due to the fact that it is hard to interpret the learned style embedding.

Some recent studies make use of SER loss [10, 11] and perceptual loss [12] to assign specific emotional information to the style embedding. However, the commonly used MSE loss for output feature does not directly improve the perceptual discriminability of the synthesized speech. In this paper, we devise a training strategy under reinforcement learning, which employs an interactive game between the ETTS system and a SER to directly improve emotion discriminability.

3. Interactive ETTS: Methodology

We propose a novel interactive training paradigm for emotional TTS under the reinforcement learning framework, denoted as *i-ETTS*. Our method is a combination of ETTS and RL, hence we first present the main components of proposed i-ETTS that includes *agent*, *policy*, *action* and *reward*. We then introduce the action update method, denoted as policy gradient. Finally, we describe the iterative training algorithm.

3.1. i-ETTS: Interactive ETTS with RL

As in Fig. 1, the ETTS model can be viewed as an *agent* under the reinforcement learning framework. The parameters of this agent define a *policy*. The policy aims to predict the emotional acoustic features at each time step. The emotional acoustic features define the *action*. After the emotional acoustic feature prediction is finished, the pre-trained SER starts to feedback the emotion recognition accuracy, which is denoted as *reward*. The policy gradient strategy was used to perform back-propagation

smoothly, and optimize the ETTS model to achieves maximum reward.

3.1.1. Agent: ETTS

As shown in the left panel of Fig. 1, our emotional text-to-speech synthesis model is built on the GST-based Tacotron network [24]. It consists of text encoder, reference encoder with a GST module attached, attention-based decoder and Post-Net. We use Griffin-Lim algorithm [26] to reconstruct the speech waveform.

Text encoder shares a similar architecture with that of Tacotron2 [23], which takes the character embeddings as input and generates high-level encoder output embeddings. The reference encoder and GST module were defined to encode the emotion of reference audio into a fixed-length emotion embedding. Their architectures are also similar to that of [24], except for the definition of style tokens. To force the GST module to pay more attention to learn the emotion-related style, we set the number of tokens to the number of emotion categories in the corpus. The attention-based decoder takes the encoder output and emotion embedding as input and predicts the acoustic features frame-by-frame. Note that we also adopt a similar decoder structure to Tacotron2 [24], except that the post-net as in [22] was added to convert the mel spectrum to linear spectrum. The GMM attention [27] was utilized to learn the <text, wav> alignment.

3.1.2. Reward: SER Accuracy

In order to achieve emotion expressiveness and discriminability, we introduce a reward function correlated with the SER evaluation metric.

As shown in the right panel of Fig. 1, the SER network includes a CNN layer, a BiLSTM layer, an attention layer, a fully connected (FC) layer and a softmax layer. The above architecture is proven effective to learn discriminative features for SER [28]. The CNN first reads a mel spectrum from input utterance and then outputs a fixed size latent representation. The BiLSTM summarizes the temporal information into another latent representation. The attention layer learns the weights for each frame. The final softmax layer outputs the probability of five emotion types, i.e., happy, angry, neutral, sad and surprise.

For ETTS training, given an input text x_i and reference audio $\hat{y}_i$ with emotion label $l_{\hat{y}_i}$, the goal is to generate an emotional speech y_i' that is not only natural, but also in accordance with the desired emotion $l_{\hat{y}_i}$. One way to employ the pre-trained SER as an expert system in place of a human judge is to evaluate the emotion category of the generated speech y_i'. We then use the recognition accuracy as the reward to update the ETTS model.

Specifically, upon generating the end-of-sequence (EOS) token, the pre-trained SER is used to evaluate how well the generated mel spectrum feature y_i' matches the reference emotion label $l_{\hat{y}_i}$. Mathematically, the recognition probability p_i about the target emotion $l_{\hat{y}_i}$ of y_i' is formulated as:

$$p_i = \text{SER}\left(l_{\hat{y}_i} \mid y_i'; \varphi\right) \qquad (1)$$

where φ represents all the parameters of SER, that are pre-trained before the training of ETTS. Note that the value of p_i ranges from 0 to 1.

During ETTS training, we sample a fixed number of predicted mel spectrum features y' from a batch of data to calculate the SER reward, which is formulated as:

$$R = \frac{N}{K} = \frac{\sum_{i=1}^{K} \mathbf{1}(p_i > \lambda)}{K} \qquad (2)$$

where R is the reward of the sampled mel spectrum features y' from ETTS model. Note that the value of reward R also ranges from 0 to 1. K denotes the sample size. λ is a threshold that is set to 0.5 in this work. N represents the number of samples of which the probability p_i of the target emotion exceeds the threshold λ. $\mathbf{1}(\cdot)$ is an indicator function of a set $\{0, 1\}$.

3.1.3. Action update: Policy Gradient

We use policy gradient algorithm [29] to estimate the gradient that leads to a larger expected reward $\mathbb{E}[R]$ of the generated acoustic features y' for input text x. Note that the gradient of the ETTS model $P(\cdot)$ w.r.t. the model parameters θ is estimated by sampling as follows:

$$\begin{aligned}
\nabla_{\theta}\mathbb{E}[R] &= \nabla_{\theta} P\left(y' \mid x; \theta\right) R \\
&= P\left(y' \mid x; \theta\right) R \nabla_{\theta} \log\left(P\left(y' \mid x; \theta\right)\right) \qquad (3) \\
&\simeq R \nabla_{\theta} \log\left(P\left(y' \mid x; \theta\right)\right)
\end{aligned}$$

Unlike the traditional gradient descent algorithm, the policy gradient algorithm estimates the weights of an optimal policy through gradient ascent. Note that it can assign an explicit emotion-aware supervised signal to the gradient during the model training, resulting in emotional speech with accurate emotion category.

3.2. Iterative Training

In practice, it is hard to train the whole network from scratch with policy gradient since the ETTS model may find an unexpected way to achieve a high reward but fail to guarantee the naturalness or robustness of the synthesized speech [21]. Therefore, we formulate an iterative training algorithm, reported in Algorithm 1, that contains two main steps: 1) pre-training, and 2) iterative-training. First, we use *MSE Loss* to pre-train the ETTS model with text-wav pairs from the training set. After the pre-training phase, iterative-training phase aims to optimize the ETTS model with reward and the MSE loss alternatively.

4. Experiments

We report emotional TTS experiments on ESD database [30], which is a new publicly available emotional speech dataset for emotional speech synthesis. ESD is a multi-lingual dataset and has 350 parallel utterances spoken by 10 native English and 10 native Mandarin speakers. We use the English corpora with a total of nearly 13 hours of speech by 5 male and 5 female speakers in five emotions, namely happy, angry, neutral, sad and surprise. In ESD, scripts are provided, speech data are sampled

---

Algorithm 1: Iterative Training Algorithm.

Input: Training set: $D = \{x, y, l\}$
    x: character sequence
    y: acoustic feature sequence
    l: emotion label
Output: ETTS model: θ
Begin
0. Pre-train SER model φ with (y, l)
▷ **Pre-training**
1: Pre-train ETTS model θ using *MSE Loss* with (x, y, l)
▷ **Iterative-training**
3: **for** epoch = 1,2,...N **do**
4.   Sample a batch B in D
5.   Sample (x, y, l) from B
6.   Generate mel spectrum features y' via θ
7.   Compute reward R based on Eq. 2
8.   Update θ using R based on Eq. 3
9.   Update θ using *MSE Loss* with (x, y, l)
10: **end for**
End

---

at 16 kHz and coded in 16 bits. We note that ESD is considered to be large enough for various voice conversion tasks [31], while it provides limited data for TTS and ETTS tasks.

4.1. Comparative Study

We implement two baseline systems together with the proposed *i-ETTS*, as summarized next.

- MTL-ETTS [11]: An emotional TTS model that jointly trains an auxiliary SER task with the TTS model;

- CET-ETTS [12]: An emotional TTS model that uses two reference encoders with SER module and perceptual loss to enhance the emotion-discriminative ability;

- i-ETTS: the proposed *i-ETTS* that optimizes the ETTS model with a reward function correlated with the SER accuracy.

For a fair comparison, all frameworks use the same SER module, as also illustrated in Section 3.1.2, and Griffin-Lim algorithm [26] is used for waveform generation. We note that the use of neural vocoder will further improve the speech quality [32], which is not the main focus of this paper.

4.2. Experimental Setup

For all systems, the encoder takes the character sequence as input. The 80-channel mel spectrum and 1025-channel linear-spectrum are extracted with 12.5ms frame shift and 50ms frame length. The decoder predicts 5 output frames at each decoding step. We randomly split the audio set of each emotion in each speaker into training/validation/test set with the number of 300/30/20, resulting in a total of 15,000/1,500/1,000 training/validation/test proportion for whole dataset. We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ to optimize the model. The learning rate is set to 10^{-3} before 100k steps, then exponentially decays to 10^{-5} after 100k steps. We set the batch size to 32. The sample size in Eq. 3 is set to 20. The MTL-ETTS and CET-ETTS systems are trained with 300k steps.

For our i-ETTS, we first pre-train a SER [28], that reports a classification accuracy of 90.4% for all emotions on the test set. During the iterative learning, i-ETTS model is first pre-trained
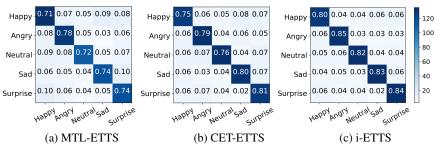
Figure 2: *Confusion matrices of synthesized speech from MTL-ETTS, CET-ETTS and the proposed i-ETTS. The X-axis and Y-axis of subfigures represent predicted and truth emotion label, respectively.*
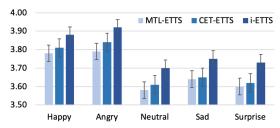

Figure 3: *The MOS test results of MTL-ETTS, CET-ETTS and i-ETTS for five emotion categories, with 95% confidence interval.*

Figure 4: *The preference test results between 1) MTL-ETTS vs. i-ETTS, and 2) CET-ETTS vs. i-ETTS for five emotion categories, with 95% confidence interval.*

200k steps via MLE, then the iterative training runs until the performance on validation set converges. Note that we set the number of tokens in GST module to 5 to be consistent with the number of emotion categories in ESD dataset. The GST module produces a 256-dimensional emotion embedding. At run-time, the task is to generate emotional speech by selecting a reference audio with the desired style. In this paper, we follow [33] and manually specify the weight of style tokens by averaging the style token weights of the test set for each emotion. All systems use Griffin-Lim algorithm [26] with 64 iterations.

4.3. Experiment Results

4.3.1. Objective Evaluation

We first evaluate the emotion-discriminative ability of the synthesized speech with SER accuracy among the baselines MTL-ETTS, CET-ETTS and the proposed i-ETTS. As reported in Fig. 2, the proposed *i-ETTS* consistently outperforms two baselines by achieving average accuracy of 82.8%, which is much higher than that of the MTL-ETTS (73.8%) and CET-ETTS (78.2%). The results show the effectiveness of our proposed model in terms of emotion expressiveness and discriminability.

4.3.2. Subjective Evaluation

We first conduct mean opinion score (MOS) listening experiments to compare overall emotion expression among different frameworks. The MOS values are calculated by taking the arithmetic average of all scores provided by the subjects. We keep the linguistic content same among different models to exclude other interference factors. We invite 15 subjects to participate in these experiments. Each listener listened to 100 synthesized speech samples. As shown in Fig. 3, we observe that the proposed *i-ETTS* achieves remarkable results by outperforming both MTL-ETTS and CET-ETTS baseline systems for all emotion categories.

We also conduct A/B preference experiments as reported in 4 to evaluate the emotion expressiveness among MTL-ETTS, CET-ETTS and *i-ETTS* systems. We ask listeners to listen to all speech samples from different systems, and then choose
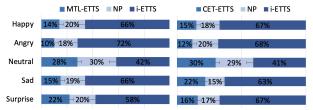
the preferred one in terms of emotion expression. We also invite 15 subjects to this experiment. Each listener listened to 100 synthesized speech samples. Consistent with the previous experiments, the preference test results in Fig. 4 show that our proposed *i-ETTS* system can generate more expressive emotional speech, and always achieves better results than MTL-ETTS and CET-ETTS systems in all five emotion categories. All the above observations validate the effectiveness of our proposed *i-ETTS* system in terms of emotion expressiveness and discriminability.

5. Conclusion

We have studied an interactive training paradigm for emotional TTS, denoted as *i-ETTS*, to synthesize emotional speech with accurate emotion category. In doing so, we devise a training strategy under reinforcement learning, which employs an interactive game between ETTS and SER. We formulate a policy gradient strategy and a reward function correlated with the SER accuracy. A series of experiments were conducted to evaluate the emotion expression. The proposed *i-ETTS* achieves remarkable performance by consistently outperforming the ETTS baseline systems in terms of voice quality and emotion discriminability. Future work includes improving the performance by investigating more effective ways to optimize the ETTS model.

6. Acknowledgements

7. References

[1] K. Tokuda, H. Zen, and A. W. Black, "An hmm-based speech synthesis system applied to english," in *IEEE Speech Synthesis Workshop*, 2002, pp. 227–230.

[2] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for hmm-based speech synthesis," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[3] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. J. Gales, and K. Knill, "Unsupervised clustering of emotion and voice styles for expressive tts," in *ICASSP 2012-2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4009–4012.

[4] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, "Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis," *Speech Communication*, 2018.

[5] H. Choi, S. Park, J. Park, and M. Hahn, "Multi-speaker emotional acoustic modeling for cnn-based speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6950–6954.

[6] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen *et al.*, "Hierarchical generative modeling for controllable speech synthesis," in *International Conference on Learning Representations*, 2018.

[7] Y. Zhang, S. Pan, L. He, and Z. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6945–6949.

[8] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7254–7258.

[9] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[10] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai, "End-to-end emotional speech synthesis using style tokens and semi-supervised training," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 623–627.

[11] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1–5.

[12] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[13] R. Liu, B. Sisman, G. Gao, and H. Li, "Expressive tts training with frame and style reconstruction loss," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1806–1818, 2021.

[14] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[15] S. Latif, H. Cuayáhuitl, F. Pervez, F. Shamshad, H. S. Ali, and E. Cambria, "A survey on deep reinforcement learning for audio-based applications," *arXiv preprint arXiv:2101.00240*, 2021.

[16] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "Dnn-based source enhancement to increase objective sound quality assessment score," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1780–1792, 2018.

[17] Y. Shen, C. Huang, S. Wang, Y. Tsao, H. Wang, and T. Chi, "Reinforcement learning based speech enhancement for robust speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6750–6754.

[18] T. Kala and T. Shinozaki, "Reinforcement learning of speech recognition system based on policy gradient and hypothesis selection," in *ICASSP 2018-2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5759–5763.

[19] M. Seurin, F. Strub, P. Preux, and O. Pietquin, "A machine of few words: Interactive speaker recognition with reinforcement learning," *Proc. Interspeech 2020*, pp. 4323–4327, 2020.

[20] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[21] F. Luo, P. Li, J. Zhou, P. Yang, B. Chang, Z. Sui, and X. Sun, "A dual reinforcement learning framework for unsupervised text style transfer," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019*, 2019.

[22] 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.

[23] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP 2018-2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[24] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*, 2018, pp. 5180–5189.

[25] R. Liu, B. Sisman, J. Li, F. Bao, G. Gao, and H. Li, "Teacher-student training for robust tacotron-based TTS," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6274–6278.

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

[27] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[28] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[29] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[30] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," *arXiv preprint arXiv:2010.14794*, 2020.

[31] B. Sisman, S. King, J. Yamagishi, 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*, vol. 29, pp. 132–157, 2021.

[32] Y. Zhou, X. Tian, and H. Li, "Multi-task wavernn with an integrated architecture for cross-lingual voice conversion," *IEEE Signal Processing Letters*, vol. 27, pp. 1310–1314, 2020.

[33] 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.