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

Improving Performance of Seen and Unseen Speech Style Transfer in End-to-end Neural TTS

Xiaochun An[1†], *Frank K. Soong*[2], *Lei Xie*[1⋆]

[1]Audio, Speech and Language Processing Group (ASLP@NPU),
School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Microsoft China, China

xiaochunan@npu-aslp.org, frankkps@microsoft.com, lxie@nwpu.edu.cn

Abstract

End-to-end neural TTS training has shown improved performance in speech style transfer. However, the improvement is still limited by the training data in both target styles and speakers. Inadequate style transfer performance occurs when the trained TTS tries to transfer the speech to a target style from a new speaker with an unknown, arbitrary style. In this paper, we propose a new approach to style transfer for both seen and unseen styles, with disjoint, multi-style datasets, i.e., datasets of different styles are recorded, each individual style is by one speaker with multiple utterances. To encode the style information, we adopt an inverse autoregressive flow (IAF) structure to improve the variational inference. The whole system is optimized to minimize a weighed sum of four different loss functions: 1) a reconstruction loss to measure the distortions in both source and target reconstructions; 2) an adversarial loss to "fool" a well-trained discriminator; 3) a style distortion loss to measure the expected style loss after the transfer; 4) a cycle consistency loss to preserve the speaker identity of the source after the transfer. Experiments demonstrate, both objectively and subjectively, the effectiveness of the proposed approach for seen and unseen style transfer tasks. The performance of the new approach is better and more robust than those of four baseline systems of the prior art.

Index Terms: neural TTS, style transfer, style distortion, cycle consistency, disjoint datasets

1. Introduction

Recent advancement of end-to-end neural TTS has demonstrated that it can synthesize very natural, human-like speech [1, 2, 3, 4, 5]. The trained neural TTS models usually consist of an encoder-decoder neural network [6, 7] which can map a text sequence to a sequence of speech frames. Extensions of these models have shown that speech styles (e.g., speaker identity, emotion and prosody), which are essential for expressive and diverse voice generation, can be also modelled and controlled [8, 9, 10, 11, 12, 13]. As applicable scenarios of speech synthesis have rapidly developed, such as the audio reading scenario, there is a growing demands for the single-speaker, multi-style synthesis, where a person can simultaneously speak multiple styles, and yet research in this area is still in its infancy.

Currently most neural TTS systems [14, 15, 16, 17, 18] are modelled by using a corpus of a single expressive style. Acquiring and annotating a large set of single-speaker speech data with multiple styles for training a neural TTS is usually expensive and time consuming. It is an effective solution to use transfer learning to perform speech style transfer, which allows a speaker to learn the desired style from the data with this style of other speakers without the data of a certain style, and keeps his own timbre consistency. Recently, neural TTS model with global style tokens (GST) [19, 20] or a variational autoencoder (VAE) [21] has received interests for controlling and transferring speech styles. Theoretically, these models can model any complex styles in a continuous latent space, so that one can control and transfer style by manipulating the latent variables or variational inference from a reference audio.

However, these researches model all speech styles into one style representation, which contains too much interfering information to be robust and interpretable, and lacks the ability to control a specific speech feature independently. When conducting style transfer, one has to transfer all styles whether desired or not, which may not fit the contexts thus hurts generalization. When conducting style control, one can hardly confirm the relationship between the styles and the coefficients of each dimension of the style representations.

Recently, Bian et al. [22] introduce a multi-reference encoder to GST [19] and adopt an intercross training scheme, which together ensure that each sub-encoder of the multi-reference encoder independently disentangles and controls a specific style. They show successful style transfer on a multi-style data scenario. However, their intercross training scheme does not guarantee each combination of style classes is seen during training, causing a missed opportunity to learn disentangled representations of styles and sub-optimal results on disjoint, multi-style datasets.

In order to improve style transfer for the combined style that is underrepresented in the dataset, [23] proposes an adversarial cycle consistency training scheme with paired and unpaired triplets to ensure the use of information from all style classes. Unlike intercross training, the scheme sweeps across all combinations of style classes via paired and unpaired triplets. This provides disentanglement of multiple style classes, enabling the model to transfer style in a more faithful manner than existing methods.

Though [23] improves performance of style transfer, it suffers a limitation, similar to [22], that can only transfer the style seen during training, and is inadequate to transfer the speech to a target style from a new speaker with an unknown, arbitrary style, thus narrowing down the applicable scenarios of neural TTS systems. In addition, recording training samples for new style (e.g., customer-service style and poetry style) is challenging and labor-intensive, transferring style from one dataset to another (i.e., disjoint, multi-style datasets) is an appealing feature for TTS systems. Therefore, unseen style transfer on disjoint, multi-style datasets needs to be improved.

[†] Work partially done during internship at Microsoft.
[⋆] Corresponding author.

http://dx.doi.org/10.21437/Interspeech.2021-1407

In this paper, we propose a new approach to style transfer for both seen and unseen styles. As a result, it tackles the single-speaker, multi-style synthesis in a more flexible and convenient manner, and further meets the needs of audio reading scenario. The main contributions of this paper are summarized as follows:

- To facilitate seen and unseen style transfer in end-to-end neural TTS, we first adopt an inverse autoregressive flow (IAF) structure [24] to improve the style representation, and then propose four different loss functions to together make sure the seen and unseen style transfer: 1) using a reconstruction loss to measure the distortions in both source and target reconstructions; 2) injecting an adversarial loss to "fool" a well-trained discriminator; 3) introducing a style distortion loss to measure the expected style loss after the transfer; 4) incorporating a cycle consistency loss to preserve the speaker identity of the source after the transfer. With the proposed approach, we can transfer the speech to a target style from a new speaker with an unknown, arbitrary style, which does not even need to be seen during training.
- The proposed seen and unseen style transfer scheme is used as a data augmentation method to generate a single-speaker, multi-style speech data, which is significant for various speech tasks, such as multi-style TTS and voice conversion.
- Our approach outperforms the four prior art baselines and the improvement is confirmed in both subjective and objective tests. The resultant performance of seen and unseen style transfer is better and more robust than the counterpart in the prior art.

2. Proposed approach

Fig. 1 illustrates our proposed framework to style transfer for both seen and unseen styles, where we adopt Tacotron 2 [3] as the decoder and employ the Mel LPCNet vocoder [25] to reconstruct the waveforms from Mel-spectrogram.

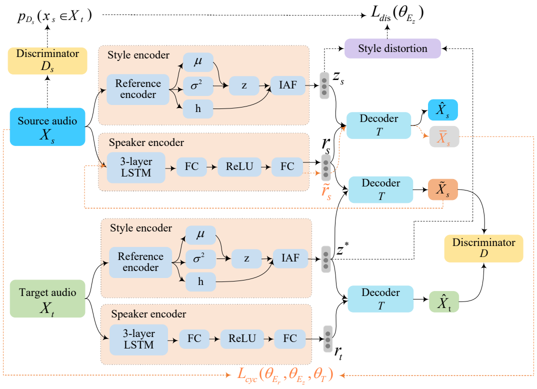

Figure 1: *Proposed approach for seen and unseen style transfer.*

2.1. Preliminary

Our approach dealing with seen and unseen style transfer is built on the encoder-decoder framework. Below we briefly describe the encoder-decoder based style transfer framework.

In the encoder, we assume that each speech utterance x can be decomposed into the style representation, $z \in Z$, and the speaker representation, $r \in R$. Each source utterance, $x_s^{(i)} \in X_s = \{x_s^{(1)}, \dots, x_s^{(n)}\}$, has its individual style, $z_s^{(i)}$, while the target utterances, $x_t^{(i)} \in X_t = \{x_t^{(1)}, \dots, x_t^{(m)}\}$, share the same style, z^*. We use two encoding functions, $E_z(x)$ and $E_r(x)$, to calculate the z and r of an utterance x respectively: $z_s^{(i)} = E_z(x_s^{(i)})$, $r_s^{(i)} = E_r(x_s^{(i)})$, $z^* = E_z(x_t^{(j)})$, $r_t^{(j)} = E_r(x_t^{(j)})$.

For the decoder T, we leverage Tacotron 2 [3] as the decoding function and enforce that for a sample x_s, its decoded sequence using T given its speaker representation r and target style representation z^*, should be in the target domain X_t. A reconstruction loss is used for maintaining reconstruction fidelity of the utterance in the decoder T as:

$$\begin{aligned} &\mathcal{L}_{rec}(\theta_{E_r}, \theta_{E_z}, \theta_T) \\ &= \mathbb{E}_{x_s \sim X_s}[-\log p_T(x_s|z_s, r_s)] \\ &+ \mathbb{E}_{x_t \sim X_t}[-\log p_T(x_t|z^*, r_t)] \end{aligned} \tag{1}$$

Following GAN [26, 27, 28], we introduce an adversarial loss to be minimized in decoding and adopt a discriminator D, as shown in Fig. 1, to distinguish between $T(r_s, z^*)$ and $T(r_t, z^*)$, while the task of decoder is to fool the discriminator, as shown in Eq. 2. The details of each individual component in our model will be described in Section 3.1.

$$\begin{aligned} &\mathcal{L}_{adv}(\theta_{E_r}, \theta_{E_z}, \theta_T, \theta_D) \\ &= \mathbb{E}_{x_s \sim X_s}[-\log(1 - D(T(r_s, z^*)))] \\ &+ \mathbb{E}_{x_t \sim X_t}[-\log D(T(r_t, z^*))] \end{aligned} \tag{2}$$

2.2. Seen and unseen style transfer

For a sample $x_s \in X_s$, z_s can be an arbitrary value that minimizes the above reconstruction loss and adversarial loss, which may not necessarily capture the utterance style. This will affect the speaker representation, making it not fully represent the speaker identity, which should be invariant with the style.

To address the issue, this paper introduces a style distortion loss into the above framework to constrain that style representation of an utterance should be closer to the target style representation. As shown in Fig. 1, a discriminator, D_s, is first trained to predict whether a given utterance x has the target style with an output probability, $p_{Ds}(x \in X_t)$. When learning the style representation z_s, we then enforce that the distortion between this style representation z_s and target style representation z^* should be consistent with the output probability of D_s. Here, we use the L_2 norm to measure the style distortion, $d(z_s, z^*) = \|z_s - z^*\|_2$, and want to have the style distortion positively correlated with $1 - p_{Ds}(x_s \in X_t)$. To incorporate this idea into our model, we adopt a probability density function, which is modelled with a standard normal distribution in our experiments, to evaluate the style distortion loss. Intuitively, when an utterance x_s have a large output probability $p_{Ds}(x_s \in X_t)$, our model encourages a large probability density and a small style distortion. This means z_s will be closer to z^*, and the style distortion loss can finally be written as:

$$\mathcal{L}_{dis}(\theta_{E_z}) = \mathbb{E}_{x_s \sim X_s}[p_{Ds}(x_s \in X_t)d(z_s, z^*)^2] \tag{3}$$

where D_s is a pre-trained model trained with a portion of the training data. Here, if we integrate D_s into our training, we may start with a D_s with a low accuracy, and then our model is inclined to optimize a wrong style distortion loss for many epochs and gets stuck into a poor local optimum.

In addition, the discriminator D_s can only constrain the generated utterance to be aligned with the target style, but cannot guarantee to keep the speaker of the source utterance intact.

To address the problem, we further introduce a cycle consistency loss [29, 30] to our model shown in Fig. 1, which requires that a transferred utterance should preserve the speaker identity of its source utterance, and thus it is enable to recover the source utterance in a cyclic manner. The cycle consistency loss is shown as follows:

$$\begin{aligned}&\mathcal{L}_{cyc}(\theta_{E_r}, \theta_{E_z}, \theta_T) \\ &= \mathbb{E}_{x_s \sim X_s}[-\log p_T(x_s|E_r(\widetilde{x}_s), z_s)] \\ &+ \mathbb{E}_{x_t \sim X_t}[-\log p_T(x_t|E_r(\widetilde{x}_t), z^*)]\end{aligned} \tag{4}$$

where $\widetilde{x}_s$ is the transferred utterance from a source sample x_s and has the target style z^*. We encode the $\widetilde{x}_s$ with the speaker encoder $E_r(\widetilde{x}_s)$ to obtain its speaker representation $\widetilde{r}_s$, which is combined with its source style z_s for decoding. Here, we expect that the source utterance can be generated with a high probability. For a target sample x_t, although we do not aim to change its style in our model, similar to x_s, we still calculate its cycle consistency loss for the purpose of additional regularization and hope that the target utterance x_t should be generated.To summarize, the final form of our loss function is:

$$\begin{aligned}&\mathcal{L}(\theta_{E_r}, \theta_{E_z}, \theta_T, \theta_D) \\ &= \alpha\mathcal{L}_{rec} + \beta\mathcal{L}_{adv} + \gamma\mathcal{L}_{dis} + \lambda\mathcal{L}_{cyc}\end{aligned} \tag{5}$$

where $\alpha = \beta = \lambda = 1.0$ and $\gamma = 5.0$ are preset weights for balancing the different loss terms. In our experiments, we find the results are insensitive to these parameters through cross-validation.

3. Experiments

This paper focuses on disjoint, multi-style datasets and an internal Chinese corpus is used in experiments: source data contains examples of four styles (i. e., reading style, broadcasting style, talking style and story style) from four different speakers whereas target data contains samples of two styles (i. e., customer-service style and poetry style) from two different speakers. This represents a minimalistic scenario of the disjoint, multi-style datasets: a single model must be able to properly transfer an arbitrary and unknown style to target style where there is no variation of speaker identity. The corpus contains 20,698 samples ($\sim$ 23.3 hours) and each style contains 4,000 samples except for poetry style. There are 698 samples in poetry style. We remove long silence ($>$ 0.1 sec) at the beginning and ending of each utterance. Mel spectrum is extracted as target speech representations with a Hanning window of 50 ms and 12.5 ms frame shift. Phoneme sequences are used as the text input. For all different systems in our experiments, we train $\sim$ 220k steps with a single Nvidia Tesla P40 GPU. The models on which we conduct experiments include:

- GST: similar to [19], we introduce "global style tokens" (GSTs) to Tacotron 2 [3] to perform various style control and transfer tasks, and make a fair comparison.
- VAE: we incorporate VAE [21] into Tacotron 2 [3] to learn the latent representation of speaking styles to guide the style in synthesizing speech.
- MRF-IT: we augment a multi-reference encoder structure into GST-Tacotron 2 [22] and adopt intercross training approach to control and transfer desired speech styles.
- MRF-ACC: we adopt an adversarial cycle consistency training scheme for multi-reference neural TTS stylization [23] to ensure the use of information from all style classes and achieve style transfer.
- Proposed model: we adopt an IAF structure [24] to improve style representation, and introduce four different loss functions, including: reconstruction loss, adversarial loss, style distortion loss and cycle consistency loss, to together make sure the seen and unseen style transfer on disjoint, multi-style datasets.

We conduct mean opinion score (MOS) and preference listening tests (ABX) of speech quality to evaluate the reconstruction performance of different experimental systems. An ABX test of style similarity is conducted to assess the style conversion performance, where subjects are asked to choose the speech samples which sound closer to the target style in terms of style expression. We further conduct a comparative mean opinion score (CMOS) test of speaker similarity to evaluate how well the transferred speech matches that of the source speaker. For each system, we randomly select the reading style as the seen style and make two experiments: from reading style to customer-service style (R2C) and reading style to poetry style (R2P). We randomly choose an unique Taiwanese-reading style from a new female speaker as the unseen style, and conduct two tests from Taiwanese-reading style to customer-service style (TR2C) and Taiwanese-reading style to poetry style (TR2P) to assess the performance of unseen style transfer.

3.1. Model details

The style encoder contains a reference encoder and an IAF [24] flow to extract underlying stylistic properties, which form a more discriminative and expressive latent style representation. Similar to [21], the reference encoder consists of a stack of six 2-D convolutional layers cascaded with one unidirectional 128-unit GRU layer, and the architecture of the IAF is the same as that in [31]. The output of reference encoder network is then used to estimate initial mean μ, initial variance δ, and hidden output h. Afterward, the initial latent variable z along with hidden output h is provided to k steps of IAF transformation to obtain flexible posterior probability distribution with latent variable z_s. For the speaker encoder, we use a 3-layer LSTM with the projection operations shown in Fig. 1. In our model, we adopt Tacotron 2 [3] as the decoder, which takes the concatenation of the speaker and style representations as the initial hidden state. As for the discriminator D, we follow the architecture of the discriminator in [28]. The pre-trained discriminator D_s used in the style distortion loss has the same structure as the style encoder followed by a sigmoid output layer.

3.2. Experimental results

Speech quality We use the seen and unseen test sets, which contain 20 sentences of each style, respectively, to compare the performance of all models in speech quality and naturalness with the MOS and ABX listening tests[1]. The 15 subjects need to mark a sentence unintelligible when any part of it is unintelligible in listening. Table 1 and Fig. 2 show the results of these two subjective evaluations. The proposed model outperforms the baseline models on both seen and unseen style transfer tasks. The performance of the proposed model on unseen style transfer is much better than other models. The results show a better generalization of the proposed model on the unseen style transfer. These observations validate the effectiveness of our proposed model in terms of speech quality. Seen style transfer performs better speech quality than the unseen style transfer.

[1] Samples can be found at `https://xiaochunan.github.io/transfer/index.html`

Table 1: *MOS results with 95 % confidence interval for speech quality.*

Models	Seen style transfer		Unseen style transfer	
	R2C	R2P	TR2C	TR2P
GST	3.33±0.06	3.27±0.08	2.91±0.10	2.82±0.06
VAE	3.42±0.04	3.38±0.03	2.95±0.05	2.87±0.05
MRF-IT	3.54±0.07	3.46±0.07	3.07±0.06	2.98±0.12
MRF-ACC	3.75±0.11	3.69±0.09	3.57±0.04	3.48±0.07
Proposed	**3.96±0.03**	**3.89±0.12**	**3.80±0.05**	**3.73±0.02**

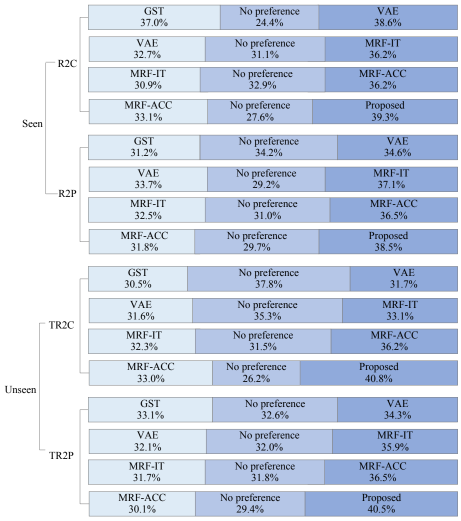

Figure 2: *ABX preference results for speech quality.*

The difference is partially due to the smaller speech data set of the poetry style, hence its MOS score is lower than that of the customer-service style.

Style similarity Adopting the same test sets, we conduct an ABX test of style similarity to assess the style conversion performance. The same 15 listeners are asked to choose which speech sample sounds closer to the target style. The results are shown in Fig. 3, where the listeners give preference to the proposed system, showing the proposed method improves performance of style transfer. For the unseen style, we find that GST, VAE and MRF-IT models, in most cases, fail to transfer unseen style of the Taiwanese-reading style to the target style of customer service or poetry style. The MRF-ACC system, the best of all baseline system, is still significantly inferior to the proposed model in the style similarity test. The results demonstrate the effectiveness of our proposed approach for both seen and unseen style transfer.

Speaker similarity To evaluate how well the transferred speech matches that of the source speaker's timbre, we conduct CMOS tests between the proposed model and each baseline by using the same test sets. The same 15 listeners are asked to select audio that represents a closer speaker to source audio. Table 2 reports CMOS results for speaker similarity, where score of the proposed model is fixed to 0. Comparisons between the proposed model and the baseline models show that our approach delivers better speaker similarity performance than all baseline models, on both seen and unseen style transfer tasks. We further adopt the cosine distance to calculate the similarity between the speaker embedding of a transferred sample and the speaker embedding of a randomly selected ground truth utterance from the same speaker to objectively measure the speaker conversion performance. The results are shown in Table 3 where we observe that the proposed model delivers a higher similarity than the other models. On the unseen style transfer of TR2C and TR2P, the GST, VAE and MRF-IT, MRF-ACC models are not capable of keeping the speaker's timbre, resulting in low similarity scores. The best baseline, MRF-ACC system, still performs significantly worse than the proposed model.

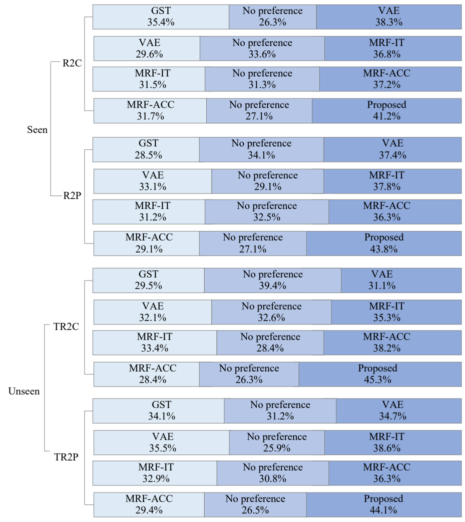

Figure 3: *ABX preference results for style similarity.*

Table 2: *CMOS results for speaker similarity.*

Models	Seen style transfer		Unseen style transfer	
	R2C	R2P	TR2C	TR2P
Proposed	0	0	0	0
GST	-0.78	-0.80	-0.90	-0.92
VAE	-0.55	-0.56	-0.65	-0.68
MRF-IT	-0.34	-0.35	-0.40	-0.42
MRF-ACC	-0.16	-0.20	-0.26	-0.28

Table 3: *Cosine distance results for speaker similarity.*

Models	Seen style transfer		Unseen style transfer	
	R2C	R2P	TR2C	TR2P
GST	0.29	0.28	0.19	0.17
VAE	0.35	0.34	0.21	0.20
MRF-IT	0.45	0.43	0.34	0.32
MRF-ACC	0.57	0.56	0.42	0.40
Proposed	**0.69**	**0.68**	**0.65**	**0.64**

4. Conclusions

This paper investigates how to train an end-to-end neural TTS for both seen and unseen speech style transfer with disjoint training datasets. We adopt an IAF structure to encode the style information and optimize a weighed sum of four, reconstruction, adversarial, style and cycle consistency, loss functions. The weighted total loss is minimized to optimize style transfer performance. The source speaker's identity or the voice timbre is well preserved by the additional cycle consistency loss. Experiments demonstrate that the proposed approach, for both seen and unseen style transfer, can outperform four other systems of the prior art, both objectively and subjectively.

5. Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2020AAA0108600).

6. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, 2017, pp. 4006–4010.

[2] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," in *Proc. ICLR*, 2018.

[3] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[4] F. Yang, S. Yang, P. Zhu, P. Yan, and L. Xie, "Improving mandarin end-to-end speech synthesis by self-attention and learnable gaussian bias," in *Proc. ASRU*, 2019, pp. 208–213.

[5] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, A. Rosenberg, B. Ramabhadran, and Y. Wu, "Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and autoregressive prosody prior," in *Proc. ICASSP*, 2020.

[6] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, 2014, pp. 3104–3112.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.

[8] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. NIPS*, 2018, pp. 4485–4495.

[9] 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 *Proc. APSIPA ASC*, 2019.

[10] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *Proc. SLT*, 2018, pp. 595–602.

[11] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," in *Proc. ICML*, 2018, pp. 4693–4702.

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

[13] Y. Lei, S. Yang, and L. Xie, "Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis," in *Proc. SLT*, 2021.

[14] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *Proc. INTERSPEECH*, 2018, pp. 3067–3071.

[15] X. An, Y. Wang, S. Yang, Z. Ma, and L. Xie, "Learning hierarchical representations for expressive speaking style in end-to-end speech synthesis," in *Proc. ASRU*, 2019, pp. 184–191.

[16] W. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," in *Proc. ICLR*, 2019.

[17] R. Habib, S. Mariooryad, M. Shannon, E. Battenberg, R. Skerry-Ryan, D. Stanton, D. T. H. Kao, and T. Bagby, "Semi-supervised generative modeling for controllable speech synthesis," in *Proc. ICLR*, 2020.

[18] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchical fine-grained prosody modelling for interpretable speech synthesis," in *Proc. ICASSP*, 2020.

[19] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, 2018, pp. 5180–5189.

[20] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," in *Proc. ISCSLP*, 2021, pp. 1–5.

[21] Y. Zhang, S. Pan, L. He, and Z. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *Proc. ICASSP*, 2019, pp. 6945–6949.

[22] Y. Bian, C. Chen, Y. Kang, and Z. Pan, "Multi-reference tacotron by intercross training for style disentangling, transfer and control in speech synthesis," in *Proc. INTERSPEECH*, 2019.

[23] M. Whitehill, S. Ma, D. McDuff, and Y. Song, "Multi-reference neural TTS stylization with adversarial cycle consistency," in *Proc. INTERSPEECH*, 2020, pp. 4442–4446.

[24] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improving variational inference with inverse autoregressive flow," in *Proc. NIPS*, 2016, pp. 4743–4751.

[25] J. Valin and J. Skoglund, "LPCNET: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, 2019, pp. 5891–5895.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[27] M. Mathieu, J. Zhao, P. Sprechmann, A. Ramesh, and Y. LeCun, "Disentangling factors of variation in deep representations using adversarial training," in *Proc. NIPS*, 2016.

[28] H. Guo, F. K. Soong, L. He, and L. Xie, "A new GAN-based end-to-end TTS training algorithm," in *Proc. INTERSPEECH*, 2019.

[29] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017, pp. 2223–2232.

[30] L. Xue, S. Pan, L. He, L. Xie, and F. K. Soong, "Cycle consistent network for end-to-end style transfer TTS training," *Neural Networks*, vol. 140, pp. 223–236, 2021.

[31] V. Aggarwal, M. Cotescu, N. Prateek, J. Lorenzo-Trueba, and R. Barra-Chicote, "Using VAEs and normalizing flows for one-shot text-to-speech synthesis of expressive speech," in *Proc. ICASSP*, 2020, pp. 6179–6183.