REFEREE: TOWARDS REFERENCE-FREE CROSS-SPEAKER STYLE TRANSFER WITH LOW-QUALITY DATA FOR EXPRESSIVE SPEECH SYNTHESIS

Songxiang Liu, Shan Yang, Dan Su, Dong Yu

Tencent AI Lab

ABSTRACT

Cross-speaker style transfer (CSST) in text-to-speech (TTS) synthesis aims at transferring a speaking style to the synthesised speech in a target speaker's voice. Most previous CSST approaches rely on expensive high-quality data carrying desired speaking style during training and require a reference utterance to obtain speaking style descriptors as conditioning on the generation of a new sentence. This work presents Referee, a robust reference-free CSST approach for expressive TTS, which fully leverages low-quality data to learn speaking styles from text. Referee is built by cascading a text-to-style (T2S) model with a style-to-wave (S2W) model. Phonetic PosteriorGram (PPG), phoneme-level pitch and energy contours are adopted as fine-grained speaking style descriptors, which are predicted from text using the T2S model. A novel pretrain-refinement method is adopted to learn a robust T2S model by only using readily accessible low-quality data. The S2W model is trained with high-quality target data, which is adopted to effectively aggregate style descriptors and generate high-fidelity speech in the target speaker's voice. Experimental results are presented, showing that Referee outperforms a global-style-token (GST)-based baseline approach in CSST.

Index Terms— style transfer, low-quality data, neural speech synthesis

1. INTRODUCTION

Neural text-to-speech (TTS) synthesis has been greatly improved in terms of quality and robustness of synthesized speech in recent years [1–4]. Nowadays, how to improve the expressiveness of TTS systems for an even better listening experience has attracted more attention and research efforts. Cross-speaker style transfer (CSST) is a promising technology for expressive speech synthesis, which aims at transferring a speaking style from a source speaker to the synthesized speech in a target speaker's voice.

Most previous CSST approaches require a reference utterance conveying the desired speaking style to obtained style descriptors, either in utterance level [5–7] or in more fine-grained levels [8–10], during the run-time inference phase. This hinders their practical use since a reference utterance is not universally applicable for different textual input and should be carefully selected for the desired speaking style. Hence, reference-free CSST approach for expressive TTS is more applicable in real-world scenarios.

Reference-free CSST presents several challenges. First, representing speaking style in a computable form is challenging, since a speaking style is the confluence of several factors including pronunciation patterns, intonation, intensity, etc. Due to the fact that all kinds of speech information are entangled within an acoustic signal, for robust CSST it is of great importance to decompose style representations from speaker identity. Second, it is difficult to predict style descriptors from the text. Sufficient amount of training data conveying the desired speaking style is required to learn a robust text-to-style model. Furthermore, it is challenging for a synthesis module to effectively and efficiently aggregate style descriptors and generate high-fidelity speech in a desired target speaker's voice.

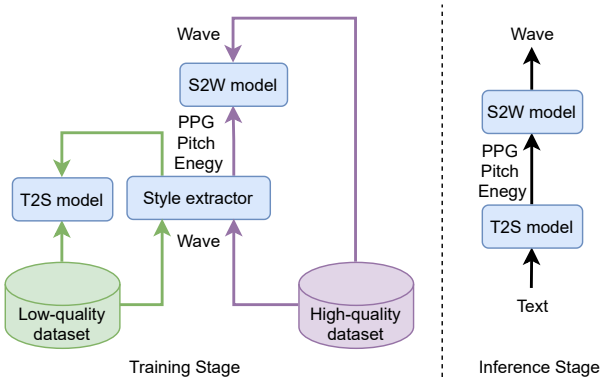

Fig. 1. Training and inference procedures of the proposed CSST approach for TTS.

Few work focuses on reference-free CSST approach. In [11], an auto-regressive multi-speaker and multi-style Transformer-based TTS model is augmented with a phoneme-level prosody module, and is able to conduct reference-free CSST. However, a sufficient amount of high-quality style data from non-target speakers are required during the training phase.

This work attempts to address the aforementioned challenges, and proposes Referee, a robust **refer**ence-f**ree** CSST approach. An overview of the proposed approach is illustrated in Fig. 1. We regard pronunciation, duration, pitch, and intensity as basic information characterizing a speaking style. Phonetic PosteriorGram (PPG) has been widely used to represent pronunciation in speech synthesis tasks due to its speaker-independent property, such as cross-lingual TTS [12], code-switched TTS [13], accent conversion [14], one-shot voice conversion [15], just to name a few. A PPG is obtained from the acoustic model in a speaker-independent automatic speech recognition (SI-ASR) system, by stacking a sequence of phonetic posterior probabilistic vectors. Since PPG features have the same frame-rate as their corresponding acoustic features, they contain not only pronunciation information but also duration information of a speech signal. Moreover, PPGs computed from an SI-ASR system, which is trained on large-scale speech corpus containing various levels environmental noise, reverberation and any other kinds of acoustic distortions, are robust representations of pronunciational and durational information for low-quality data. Therefore, this work uses PPGs as pronunciation and duration descriptors. Besides, phoneme-

level pitch and energy contours are utilized to represent pitch and intensity of a speech signal, respectively. In Referee, a style extractor is adopted to obtain these features from wave signals.

For reference-free CSST, this work trains a text-to-style (T2S) model to predict PPGs, phoneme-level pitch, and energy contours from textual input. Large-scale high-quality dataset of a desired speaking style is expensive if not impossible to collect. In the meantime, we observe that a large number of low-quality speech data with rich speaking styles can be easily obtained in the era of data explosion. Based on this, we first pretrain a base T2S model with a multi-style low-quality dataset, and then refine it with the low-quality samples carrying the desired style we want to transfer with a novel meta-learning mechanism, as illustrated by the green arrows in Fig. 1. Finally, we train a style-to-wave (S2W) model with high-quality data from the target speaker to aggregate PPGs, pitch and energy (see purple arrows in Fig. 1) and use it to generate speech carrying desired speaking style in the target speaker's voice.

The contributions of this work are summarized as follows:

- We achieve robust reference-free CSST for expressive speech synthesis by cascading a T2S model with a powerful S2W model.
- We present a novel pretrain-refinement method to fully exploit readily accessible low-quality data.

2. PROPOSED APPROACH

As shown in the left part of Fig. 1, the proposed method contains a style extractor, a T2S model and an S2W model. The style extractor aims at extracting PPGs, phoneme-level pitch and energy contours from speech, while the T2S model learns to map the input texts into those features with low-quality data. Besides, the S2W model focuses on reconstructing high-quality speech in target voice with the predicted features from the T2S model.

2.1. Style Extractor

The goal of the style extractor is to obtain robust speaker-agnostic style descriptor from low-quality speech data. To achieve this goal, we utilize an ASR model [16] trained with large-scale corpus to compute PPGs from speech waveform, which contain rich pronunciational and durational information of phonemes. Frame-level F0 and energy contours are extracted from speech waveform, representing the pitch and intensity information, which is further mean-variance normalized and converted into phoneme-level according to text-audio alignment information obtained using a hidden-Markov-model (HMM)-based forced aligner.

2.2. Text-to-Style Model

Fig. 2 shows the architecture of the T2S model, which replaces layer normalization in the FastSpeech 2 [17] model with conditional layer normalization (CLN) introduced in AdaSpeech [18]. The T2S takes phoneme sequence as input and predicts style descriptors (i.e., PPGs, phoneme-level pitch and energy contours) conditioned on style IDs. Since learning to predict style descriptors purely from textual input requires considerable amount of training data, we use a low-quality multi-style audio-book corpus (each speaker has only one style) to train a base T2S model and then conduct refinement to obtain target style with limited target style data.

During base model training, we incorporate style IDs into both the phoneme encoder and the PPG decoder via CLN, since we hypothesize that the phoneme encoder output contains not only content information but also pronunciation patterns. Specifically, one-hot style IDs are first transformed into embedding vectors via the style look-up table (LUT); and then consumed by layer-dependent linear layers to generate scale and bias vectors, which are used to modulate the hidden features after each self-attention and feed-forward block. Mean squared error (MSE) criterion is adopted to compute PPG, pitch, energy and duration (in log-scale) losses.

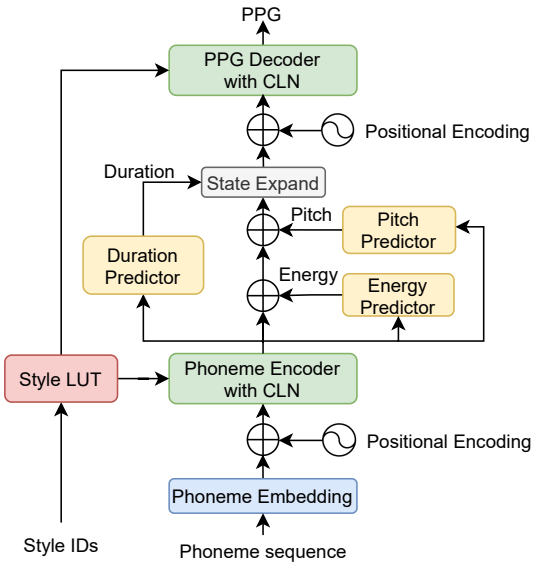

Fig. 2. Text-to-Style (T2S) model. "LUT" is abbreviation of "look-up table". "CLN" is abbreviation of "conditional layer normalization".

With limited target style data and a pre-trained base T2S model, we further propose a novel refinement procedure using a meta-learning mechanism to model the target style. In general, one challenge of cross-speaker style transfer is that the transfer performance will significantly degrade using out-of-domain (OD) textual inputs, since the text and style are somewhat entangled due to data sparsity in the training set. To mitigate this issue, we make the T2S model simulate the scenario of conducting style transfer for OD texts using episodic meta-learning [19], as illustrated in Fig. 3.

Define the ground-truth set of {PPG, pitch, energy} as S and the predicted set by the T2S model as $\hat{S}$, we simulate one-shot learning for OD text input by assuming there is only a single text sample for each speaking style available. In each episode, we regard the target style as the support ID z_s and sample a support text T_s in the target style; besides, a query text T_q is sampled from another style. Since there is no ground-truth S for the query text in the target style, we set a style discriminator D_s and a phoneme discriminator D_t as shown in Fig. 3 to provide learning signals to the T2S model. During refinement, we only update parameters of the style embedding, CLN linear layers, pitch predictor, energy predictor and duration predictor in the T2S model, aiming at mitigating possible over-fitting issue. The adversarial training process is based on the least squares GAN (LS-GAN) [20]:

$$\mathcal{L}_{D_s} = \mathbb{E}[(D_s(S) - 1)^2 + D_s(T2S(T_q, z_s))], \quad (1)$$

$$\mathcal{L}_{D_t} = \mathbb{E}[(D_t(S, E_s) - 1)^2 + D_t(T2S(T_q, z_s), E_q)]. \quad (2)$$

The adversarial loss of the generator (T2S model) is the sum of the

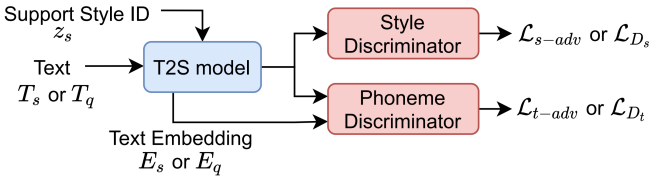

Fig. 3. Refinement process of the T2S model. The reconstruction loss is not shown for simplicity.

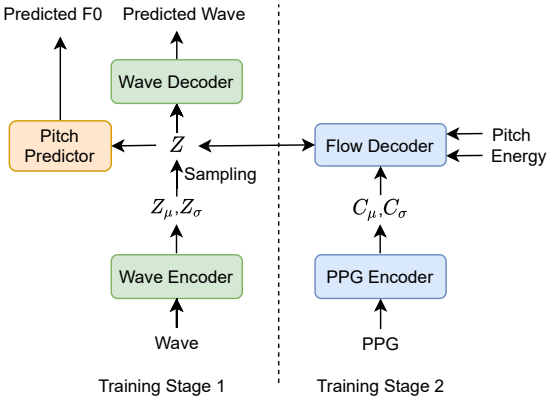

Fig. 4. Training procedure of the style-to-wave (S2W) model.

adversarial signals from D_s and D_t:

$$\begin{aligned}\mathcal{L}_{adv} &= \mathcal{L}_{s-adv} + \mathcal{L}_{t-adv} \\ &= \mathbb{E}[(D_s(T2S(T_q, z_s)) - 1)^2] \\ &+ \mathbb{E}[(D_t(T2S(T_q, z_s), E_q) - 1)^2].\end{aligned} \tag{3}$$

To stabilize the refinement process, we also use the L1 reconstruction objective for the T2S model as:

$$\mathcal{L}_{recon} = \mathbb{E}[||T2S(T_s, z_s) - S||_2^2]. \tag{4}$$

Therefore, the ultimate losses of the T2S model and the discriminators are:

$$\mathcal{L}_{T2S} = \alpha\mathcal{L}_{recon} + \mathcal{L}_{adv}, \tag{5}$$

$$\mathcal{L}_D = \mathcal{L}_{D_s} + \mathcal{L}_{D_t}, \tag{6}$$

where we empirically set $\alpha = 10$ to balance scale of each loss, and alternatively update the T2S model and the discriminators.

2.3. Style-to-Wave Model

To transfer the learned style from T2S model to a target voice, we build a style-to-wave (S2W) model to generate target voice with the style features, whose training procedure is shown in Fig. 4. The S2W modifies the skeleton of the powerful Glow-WaveGAN model [21] for high-quality speech reconstruction, which contains a VAE-based WaveGAN model to extract latent distribution $p(z) \sim \mathcal{N}(\mu, \sigma)$ from speech and a flow-based acoustic model to predict $p(z)$ from input style features. Specifically, the duration module in Glow-WaveGAN is no longer needed since the style features are already aligned to the target $p(z)$. Besides, we locally condition the Flow-based decoder with frame-wise pitch and energy features, obtained by expanding their phoneme-level counterparts according to phoneme durations.

2.4. Inference

The inference procedure for CSST is illustrated in the right part of Fig. 1. Text with arbitrary content is fed into the T2S model, which is refined on the target style, to predict PPG, phoneme-level pitch, energy and duration. The phoneme-level pitch and energy are then expanded to frame-level according to the predicted duration. Finally, the S2W model takes the predicted PPG and frame-level pitch and energy as input and generates waveform in the target speaker's voice conveying the target style, thus accomplishing the reference-free CSST.

3. EXPERIMENTS

3.1. Setups

Two internal Chinese corpora are used in our experiments. The low-quality multi-style audio-book corpus contains 38 speakers and 28.8 hours speech data in total, where each speaker only has one speaking style. Sampling rate of some audio samples is 16 KHz. We resample all audio to 24 KHz. We randomly partition the low-quality corpus into training, validation and test sets according to an 86%-7%-7% scheme. In the experiments, we choose a strong northeast Chinese accent as our target style, whose corresponding speaker is a male speaker. The high-quality corpus is recorded for developing TTS systems by a female target speaker in reading style, which contains 14 hours of speech data at 24 KHz. The PPG feature has dimensionality of 218 with a 10ms frame-shift.

The T2S model in the proposed approach uses 4 feed-forward Transformer (FFT) blocks in both the phoneme encoder and the PPG decoder. The hidden size, number of attention heads, kernel size and filer size of the 1D convolution in the FFT block are set as 256, 2, 9 and 1024, respectively. The dimention of the style LUT is 128. The style discriminator and phoneme discriminator adopt the same network hyper-parameter settings as in [19]. The S2W model follows the basic setting as that used in [21]. As for the style features, we firstly transform them through three 1D-convolutional layers with kernel size 3, and then add them into each affine coupling layer flow decoder as local conditions. The Adam optimizer with $\beta_1 = 0.9$, $\beta_1 = 0.98$ and $\epsilon = 10^{-9}$ is used to train all the models. The T2S model is pre-trained on the multi-style low-quality data with a batch size of 64 for 50K steps. Then we meta-refine the base T2S model on the target style for 5K steps using a fixed learning rate of 10^{-5} and batch size of 16.

3.2. Comparisons

To evaluate the performance of style transfer with low-quality data, we build four systems in our experiments:

GST-DurIAN (baseline) We first build a baseline DurIAN model [3] extended with a global-style-token (GST)-based style encoder [6], which converts the text into mel-spectrogram domain conditioned on a reference audio. We utilize the style embedding from the GST encoder and speaker IDs as auxiliary inputs to both the acoustic model and duration model. To disentangle style and speaker, we adopt the domain adversarial training mechanism [22]. GST-DurIAN is first pre-trained on the combination of both the low-quality multi-style corpus and the high-quality target speaker corpus, and then refined to the low-quality target style and the high-quality target voice. During cross-speaker style transfer at inference, a reference audio conveying the target style is used as the input to the GST encoder to provide target style. A HifiGAN vocoder [23] is used to generate waveform from mel-spectrogram.

Table 1. Mean opinion score (MOS) and similarity MOS (SMOS) results with 95% confidence interval.

Setting	In-domain text		Out-of-domain text	
	MOS (↑)	SMOS(↑)	MOS (↑)	SMOS(↑)
Recordings	4.60±0.09	-	-	-
GST-DurIAN	3.38±0.14	3.53±0.19	3.27±0.21	3.25±0.18
Style-GWG	**4.04±0.17**	**3.67±0.22**	-	-
Prop-1	3.85±0.17	3.55±0.20	3.33±0.20	3.23±0.19
Referee	3.86±0.19	3.53±0.21	**3.35±0.21**	**3.33±0.22**

Style-GWG (topline) This model (**Style-G**low-**W**ave**G**AN) is the topline of the proposed approach. We directly feed ground-truth PPG, pitch and energy obtained from a reference utterance to the S2W model.
Prop-1 (proposed) This model is the proposed approach with a traditional refinement procedure. During the refinement, we use only the target style data to refine the T2S model. To avoid over-fitting, we also only refine the parameters of the style embedding, CLN linear layers, and each style predictor (pitch, energy and duration).
Referee (proposed) The proposed approach introduced in Section 2 with episodic meta-learning.

3.3. Evaluations

We conduct the mean opinion score (MOS) tests and similarity MOS (SMOS) tests to evaluate the naturalness and voice similarity of the generated audio samples. To evaluate the style transfer performance, ABX style perceptual preference tests are conducted between compared approaches. 10 samples are generated from hold-out texts in each compared approach. We invite 10 native Mandarin Chinese speakers as the raters for each listening test, and the audio samples can be found online[1].

3.3.1. CSST for in-domain text

For the in-domain evaluation, texts from the test set of the target style are used. The MOS and SMOS results are presented in the left part of Table 1. We can see that the topline approach Style-GWG achieves the best MOS and SMOS results among the four compared approaches. Prop-1 and Referee have similar performance in naturalness and voice similarity of the generated samples, both outperforming the baseline approach GST-DurIAN in naturalness. Since low-quality data is incorporated in the training process of the GST-DurIAN model, the style vector computed from a low-quality utterance by the GST encoder also encodes noise and channel information, leading to quality degradation in generated speech.

The ABX test results for evaluating style transfer performance are shown in Fig. 5. We can see that Referee is significantly better than Prop-1 (p-value ≈ 0.016) and the baseline GST-DurIAN (p-value $= 2.0 \times 10^{-17}$) in terms of style transfer performance. The topline Style-GWG beats Referee with a large margin (p-value = 0.00023). Aggregating the observations mentioned above, we can safely say that Referee achieves significantly better performance than GST-DurIAN in terms of naturalness and cross-speaker style transfer, and on-par performance in terms of voice similarity. And adopting the proposed pretrain-refinement scheme introduced in section 2.2 boosts Referee in naturalness and style transfer, compared with Prop-1. However, there is still a gap between Referee and the topline Style-GWG. One possible reason is that there still exists mismatch between the generated style features by the T2S model and those directly extracted from waveform. Fine-tuning the S2W model with generated style features should be the first attempt to address this issue.

[1][Online] `https://liusongxiang.github.io/Referee/`

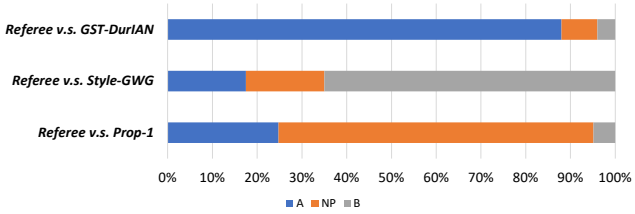

Fig. 5. In-domain Style similarity ABX test results.

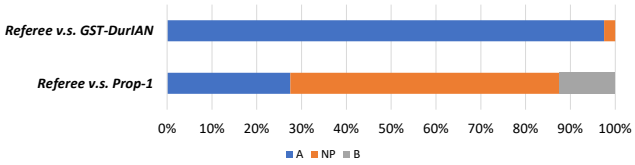

Fig. 6. Out-domain Style similarity ABX test results.

3.3.2. CSST for out-of-domain text

This subsection presents the out-domain evaluation, where texts for generating samples are from an unseen reading-style corpus. Since Style-GWG cannot conduct style transfer when there is no speech sample accessible for the target style, evaluations on it are not conducted. The MOS and SMOS results are shown in the right part of Table 1. The results have a significant drop compared to their in-domain counterparts, showing that the models overfit the text in the training set in some degree. Enlarging the training set size could be a good solution to this issue. Referee achieves the best results in naturalness and voice similarity. The ABX test results are shown in Fig. 6, where we see that Referee is significantly better than GST-DurIAN (p-value $= 3.7 \times 10^{-33}$). And the proposed episodic meta-learning refinement also helps Referee achieve slightly better style transfer performance compared with Prop-1 (p-value $= 0.068$).

4. CONCLUSION

This paper has introduced Referee, which is a robust cross-speaker style transfer approach and does not require a reference stylish utterance at inference. PPG, phoneme-level pitch and energy are used as style descriptors and are predicted from text using a T2S model. Referee adopts a novel pretrain-refinement scheme through meta-learning to fully exploit easily accessible low-quality data. Our experiments show that Referee outperforms a GST-based baseline system in CSST performance for both in-domain text and out-of-domain text. Future work includes extending Referee for supporting robust many-to-many cross-speaker style transfer.

References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[2] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[3] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei, et al., "Durian: Duration informed attention network for multimodal synthesis," *arXiv preprint arXiv:1909.01700*, 2019.

[4] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *arXiv preprint arXiv:1905.09263*, 2019.

[5] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *international conference on machine learning*. PMLR, 2018, pp. 4693–4702.

[6] 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*. PMLR, 2018, pp. 5180–5189.

[7] W.-N. Hsu, Y. Zhang, R. Weiss, H. Zen, Y. Wu, Y. Cao, and Y. Wang, "Hierarchical generative modeling for controllable speech synthesis," in *International Conference on Learning Representations*, 2019.

[8] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5911–5915.

[9] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, "Fine-grained robust prosody transfer for single-speaker neural text-to-speech," *arXiv preprint arXiv:1907.02479*, 2019.

[10] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman, "CopyCat: Many-to-Many Fine-Grained Prosody Transfer for Neural Text-to-Speech," in *Proc. Interspeech 2020*, 2020, pp. 4387–4391.

[11] S. Pan and L. He, "Cross-speaker style transfer with prosody bottleneck in neural speech synthesis," *arXiv preprint arXiv:2107.12562*, 2021.

[12] L. Sun, H. Wang, S. Kang, K. Li, and H. M. Meng, "Personalized, cross-lingual tts using phonetic posteriorgrams.," in *INTERSPEECH*, 2016, pp. 322–326.

[13] Y. Cao, S. Liu, X. Wu, S. Kang, P. Liu, Z. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "Code-switched speech synthesis using bilingual phonetic posteriorgram with only monolingual corpora," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7619–7623.

[14] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent conversion using phonetic posteriorgrams," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5314–5318.

[15] S. Liu, J. Zhong, L. Sun, X. Wu, X. Liu, and H. Meng, "Voice conversion across arbitrary speakers based on a single target-speaker utterance.," in *INTERSPEECH*, 2018, pp. 496–500.

[16] Z. You, D. Su, J. Chen, C. Weng, and D. Yu, "Dfsmn-san with persistent memory model for automatic speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7704–7708.

[17] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.

[18] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, sheng zhao, and T.-Y. Liu, "Adaspeech: Adaptive text to speech for custom voice," in *International Conference on Learning Representations*, 2021.

[19] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," *arXiv preprint arXiv:2106.03153*, 2021.

[20] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[21] J. Cong, S. Yang, L. Xie, and D. Su, "Glow-wavegan: Learning speech representations from gan-based variational auto-encoder for high fidelity flow-based speech synthesis," *arXiv preprint arXiv:2106.10831*, 2021.

[22] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[23] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 17022–17033, Curran Associates, Inc.