Disentangling Speaker Representations from Intuitive Prosodic Features for Speaker-Adaptative and Prosody-Controllable Speech Synthesis

Pengyu Cheng*, Zhenhua Ling†, Meng Meng*, Yujun Wang*
* Xiaomi Inc., Beijing, China
E-mail: {chengpengyu, mengmeng, wangyujun}@xiaomi.com
† NERC-SLIP, University of Science and Technology of China, Hefei, China
E-mail: zhling@ustc.edu.cn

***Abstract*—In this paper, we propose a method of speaker-adaptative and prosody-controllable speech synthesis with a disentanglement between intuitive prosodic features and speaker representations. In this method, the intuitive prosodic features include utterance-level pitch, pitch range, speak rate and energy. A residual speaker information encoder with a set of adversarial classifiers is designed to extract the speaker characteristics that can't be described by these intuitive prosodic features. Further, the outputs of the residual speaker information encoder are concatenated with intuitive prosodic features to obtain complete speaker representations for acoustic feature prediction. Experimental results have demonstrated that our proposed method can synthesize speech with better naturalness and higher prosody controllability than its counterpart without the disentanglement.**

***Index Terms*—speech synthesis, speaker adaptation, speaker representation, intuitive prosodic features, disentanglement**

I. Introduction

Currently, state-of-the-art text-to-speech (TTS) systems adopt sequence-to-sequence (seq2seq) acoustic models [1]–[3] to convert input text sequences into Mel-scale spectrograms, and then utilize neural network vocoders [4], [5] to reconstruct audio waveforms from the generated Mel-spectra. On the basis of speaker-dependent model training which usually requires a large amount of training data from the target speaker, many studies have paid their attentions on the speaker adaptation task, which aims to synthesize high-quality voice when the data amount of the target speaker is limited. Most of these studies build a pre-trained TTS model with a multi-speaker dataset at first and then use the recordings from the target speaker to update the whole or part of pre-trained model's parameters. According to the speaker representation method used in acoustic modeling, these speaker adaptation methods can be roughly divided into two main categories, i.e., speaker-encoding-based ones [6]–[11] and speaker-embedding-based ones [12]–[16].

In speaker-encoding-based adaptation methods [6]–[11], a neural speaker encoder is designed to model each sentence in the corpus. Such speaker encoder is usually pre-trained on a large-scale multi-speaker corpus through a speaker verification task and is fixed while training the TTS model. In speaker-embedding-based methods [12]–[16], a speaker embedding vector is designed to represent the global voice characteristics of each speaker. At the training stage, these speaker embedding vectors are usually optimized together with TTS model, and the most widely used strategy is the look-up table. Considering prosody characteristics influence the subjective perception of synthetic speech significantly, our previous work [17] proposed a method of multi-speaker training and speaker adaptation with intuitive prosodic features for seq2seq speech synthesis. The intuitive prosodic features consisted of utterance-level pitch, pitch range, speak rate, and energy. Experimental results demonstrated that this method can effectively improve the subjective similarity of speaker adaptation under both speaker-encoding-based and speaker-embedding-based.

Another advantage of utilizing intuitive prosodic features is that it enables us to control the prosody characteristics of synthetic voice conveniently. This is useful for improving the expressiveness of speech synthesis and creating new voices for avatars. For example, Tuomo et al. [18] proposed a method of prosody modeling and controlling with intuitive prosodic features for single speaker TTS. Morrison et al. [19] presented a deep autoregressive model that supported controllable, context-aware fundamental frequency (F_0) generation. Valle et al. [20] designed a multi-speaker TTS framework combining explicit prosodic variables (F_0 and voicing decision) and latent space modeling. Comparing with the latent prosody representations learned from speech spectra [21], [22], intuitive prosodic features exclude segmental characteristics and have better interpretability at each dimension. One issue with current approaches to integrate intuitive prosodic features into multi-speaker acoustic modeling [17], [20] is the entanglement between speaker representations and intuitive prosodic features. The speaker representations either derived by a speaker encoder or learned as speaker embeddings may contain the prosody characteristics of reference utterances. Thus, they may conflict with the modified intuitive prosodic features at the synthesis stage, which constrains the performance of prosody control.

Therefore, this paper proposes a method of disentangling speaker representations from intuitive prosodic features for

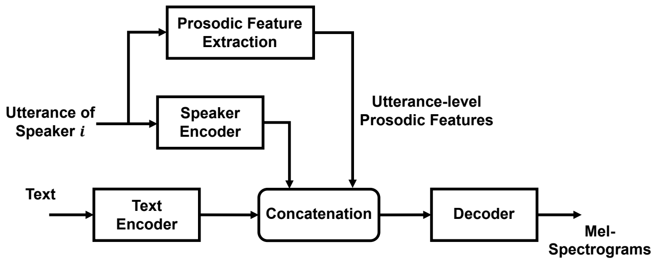

Fig. 1. The architecture of our baseline model [17].

speaker-adaptative and prosody-controllable speech synthesis. The four-dimensional utterance-level intuitive prosodic features used in our previous study [17] are adopted here. A residual speaker information encoder is designed to extract the speaker characteristics that are not represented by intuitive prosodic features. A set of prosody classifiers with adversarial losses were applied to achieve the disentanglement. The outputs of the residual speaker information encoder are concatenated with intuitive prosodic features to obtain complete speaker representations, which are constrained by a speaker classifier and are sent into the decoder for acoustic feature prediction. Our experimental results on the AISHELL-3 [23] dataset demonstrated that the proposed method achieved higher naturalness of synthetic speech and better prosody controllability over all prosodic variables than the baseline model [17].

II. Baseline with Intuitive Prosodic Features

The baseline model used in this paper follows our previous work [17]. Its architecture is shown in Fig. 1. It adopts Tacotron2 [1] as its backbone and consists of four parts: a prosodic feature extraction module, a speaker encoder, a text encoder and a decoder. In this model, the extracted prosodic features are all at utterance-level, and the detailed procedure of prosodic feature extraction can be found in [17]. The speaker encoder has a ResNet-based architecture, following the structure proposed in [24]. The parameters of the speaker encoder are pre-trained by a speaker verification (SV) task on a large-scale multi-speaker corpus and are fixed while training the speech synthesis model. The text encoder and the decoder are pre-trained with a multi-speaker dataset and are then adapted to the target speaker. For each training utterance, the speaker representations extracted using the speaker encoder are concatenated with the utterance-level intuitive prosodic features and are then sent into the decoder for predicting the Mel-spectrograms of this utterance. At the synthesis stage, the average speaker representations and intuitive prosodic features of all adaptation utterances from the target speaker are employed. In our previous experiments, this baseline model outperformed its counterpart without intuitive prosodic features (named *Baseline w/o IPF* in this paper) on both similarity and naturalness after speaker adaptation [17].

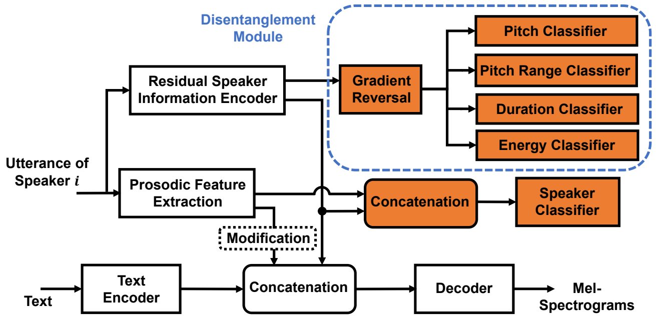

Fig. 2. The architecture of our proposed model, the orange parts are only conducted at the training stage.

III. Proposed Methods

A. Overall Structure

The overall structure of our proposed model is illustrated in Fig. 2 which consists of five parts: a prosodic feature extraction module, a disentanglement module, a residual speaker information encoder, a text encoder and a decoder .

The prosodic feature extraction module, the text encoder and the decoder have exactly the same structure as the ones in the baseline model [17]. The residual speaker information encoder receives the 80-dim Mel-spectra of the input utterance and output a disentangled speaker vector which is expected to be irrelevant to the extracted intuitive prosodic features of this utterance. To achieve this goal, a set of prosody classifiers with a gradient reversal layer [25] are designed in the disentanglement module in Fig. 2. To guarantee the completeness of speaker representation, the disentangled speaker vector and ground-truth (GT) intuitive prosodic features are concatenated and then sent into a speaker classifier with a cross-entropy (CE) loss. Like the baseline model, the concatenated speaker representation together with the output of the text encoder are fed into the decoder for predicting corresponding Mel-spectra.

At the inference stage, an utterance randomly selected from the target speaker's adaptation dataset is employed as the input of the residual speaker information encoder, while the average intuitive prosodic features of the speaker's adaptation data is adopted, which can be further modified to control the prosody characteristics of synthetic speech.

B. Residual Speaker Information Encoder

We first tried the ResNet-based [24] structure commonly used in SV tasks to build the residual speaker information encoder, but it did not work well. This may be due to that the structure designed for SV tasks learns to extract prosody-related representations which can help to distinguish between different speakers, resulting in that the output speaker vectors can't be successfully disentangled from intuitive prosodic features by adversarial training.

Therefore, a simpler architecture is adopted in our proposed residual speaker information encoder, which consists of 6 convolutional layers and a bidirectional long short-term memory (BiLSTM) layer. Each convolutional layer is followed by a batch normalization operation. A post linear layer is attached to the BiLSTM layer as the output layer. Finally a L2-norm activation is used for the output hidden variables to accelerate the convergence of model training.

C. Disentanglement Module

In order to eliminate the information of intuitive prosodic features from speaker representations, a disentanglement module based on domain adversarial training [25] is applied in this paper. Four prosodic classifiers corresponding to the four intuitive prosodic features used in this paper are designed as shown in Fig. 2. A gradient reversal layer is placed between the residual speaker information encoder and the four prosodic classifiers to invert the gradients calculated by the prosody classification loss.

The four prosodic classifiers share the same structure, including a dropout operation with pre-dense and post-dense layers. Since the previously extracted intuitive prosodic features are continuous ones, they are discretized first to get classification labels. For each prosodic feature, its minimum and maximum values in the training set are normalized to $[0, 1]$ respectively. The normalized $[0, 1]$ range is then divided into 256 intervals equally. Thus, each continuous value of prosodic features can be quantized into a categorical label for training the disentanglement module.

D. Model Training Strategy

Three groups of loss functions are used for model training. The first group is the conventional decoder loss, which consists of the original Tacotron2 [1] loss function of Mel-spectrum prediction and a guided attention loss [26] to help model learn alignment faster. The second group contains the adversarial losses given by the prosodic classifiers in the disentanglement module. A cross-entropy (CE) loss is calculated for each classifier since the prosodic features are quantized to discrete labels. The third group is the cross-entropy (CE) loss calculated by the speaker classifier in order to guarantee the completeness of speaker representation.

The model parameters are first pre-trained with a multi-speaker dataset and then adapted to the target speaker. At the pre-training stage, all three groups of losses are applied with equal weights to train all model parameters simultaneously. At the adaptation stage, the text encoder is fixed. The CE loss of the speaker classifier is removed since only the data of the target speaker is used at this stage. The other two groups of losses are utilized to fine-tune other model parameters.

IV. Experiments

A. Datasets

The AISHELL-3 [23] corpus was adopted in our experiments, which contains a total of 88,305 utterances spoken by 218 Mandarin Chinese speakers. Considering that some utterances in this dataset had no phrase boundaries in their corresponding transcripts, these samples were abandoned. The remaining part had a total of 63,263 utterances from 174 speakers, including 31 male and 143 female speakers. 4 male and 4 female speakers were randomly selected as adaptation target speakers. For each target speaker, 80, 10 and 20 utterances were randomly selected as the training set, the development set and the test set for speaker adaptation. The recordings of the remaining 166 speakers were used as the multi-speaker dataset for pre-training.

TABLE I
Objective evaluation results of three systems on the test sets of female and male speakers.

Models	MCD (dB)		F_0 RMSE (Hz)	
	Female	Male	Female	Male
Baseline w/o IPF	3.85	5.36	47.08	27.79
Baseline	3.76	**5.10**	44.07	26.98
Proposed	**3.57**	5.21	**42.23**	**26.14**

B. Experimental Configurations

The original recordings were down-sampled to 16kHz for training the TTS model. The 80-dimensional Mel-spectrograms were computed with 50 ms frame length and 12.5 ms shift.

To verify the effectiveness of our proposed method, two models were constructed for comparison in addition to the proposed model. One was the baseline model introduced in Section II. Another was the *Baseline w/o IPF* model mentioned at the end of Section II. We trained each model on the multi-speaker dataset for 140k steps with a batch size of 40. The pre-trained model was then fine-tuned for fixed 800 steps for each unseen speaker with a batch size of 20. All the three models shared a same Parallel WaveGAN [27] vocoder to generate speech from predicted Mel-spectra. We simply trained the vocoder on the complete AISHELL-3 dataset since this paper focuses on acoustic modeling instead of vocoding techniques.

C. Evaluation on Speaker Adaptation

Objective and subjective evaluations were conducted to compare the performance of different models on speaker adaptation.

We used 25-dimensional Mel-cepstral coefficients (MCCs) and F_0 to calculate Mel-cepstrum distortion (MCD) and root mean square error of F_0 (F_0 RMSE) as objective evaluation metrics. Dynamic time warping (DTW) [28] was conducted based on MCCs to align the sequences of MCCs and F_0 extracted from synthetic speech with the ones from natural recordings for calculating distortions. Objective metrics were calculated on the test sets of 8 target speakers with 20 utterances per speaker.

The evaluation results are reported in Table I. We can see that the overall objective performances of the baseline model and our proposed model were similar. Our proposed method outperformed baseline on female speakers. For male speakers, our proposed method achieved lower F_0 RMSE than baseline, but higher MCD. Both baseline and proposed

TABLE II
RESULTS OF ABX PREFERENCE TESTS (%) BETWEEN TWO SYSTEMS ON THE TEST SETS OF FOUR SPEAKERS, WHERE N/P DENOTES "NO PREFERENCE" AND p MEANS THE p-VALUE OF PAIRED T-TEST.

	Baseline	Proposed	N/P	p
naturalness	24.77	**46.36**	28.87	0.0002
similarity	19.09	22.05	58.86	0.23

TABLE III
MEAN OPINION SCORES ON NATURALNESS (NAT.) AND SIMILARITY (SIM.) WITH 95% CONFIDENCE INTERVALS OF SPEECH SYNTHESIZED BY TWO MODELS WITH PROSODY CONTROL.

Prosody Variables		Baseline	Proposed
Pitch	Nat.	3.115 ± 0.136	**3.475 ± 0.114**
	Sim.	3.920 ± 0.107	3.955 ± 0.100
Pitch Range	Nat.	3.300 ± 0.118	**3.645 ± 0.105**
	Sim.	3.765 ± 0.104	**3.975 ± 0.088**
Speak Rate	Nat.	3.005 ± 0.130	**3.665 ± 0.100**
	Sim.	3.575 ± 0.116	**3.995 ± 0.094**
Energy	Nat.	3.270 ± 0.125	**3.780 ± 0.100**
	Sim.	3.825 ± 0.086	**4.025 ± 0.087**

models outperformed Baseline w/o IPF, which demonstrated the effectiveness of integrating intuitive prosodic features.

The subjective evaluation was conducted following the experimental configurations in our previous work [17]. 4 target speakers (2 female and 2 male) were randomly selected to control the scale of listening tests. For each target speaker, an ABX preference test was conducted to compare the baseline model with our proposed model. Each test had 10 pairs of utterances synthesized by these two models respectively. At least 11 native listeners participated in each test, and the listeners were asked to give their preference opinion for each pair (which one in the pair was better or there was no preference) on both naturalness and similarity. Paired t-test was also carried out to examine the significance of the preference between two models.

Table II shows the results of ABX preference tests. We can see that the proposed model achieved significantly ($p = 0.0002$) better naturalness of synthetic speech than the baseline. This demonstrates the effectiveness of our proposed disentanglement module, which helped to eliminated the mismatch between speaker representation and intuitive prosodic features at the inference stage. Comparing the similarity of speech synthesized by these two models, the preference difference was insignificant ($p = 0.23$)[1].

D. Evaluation on Prosody Control

To test whether our proposed model can effectively disentangle prosodic information from speaker representations, a prosody control experiment was carried out. Two speakers (one male and one female) were randomly selected from the 8 target speakers and we conducted speaker adaptation for both speakers following the baseline and proposed methods respectively. For each intuitive prosodic feature, its minimum and maximum values in the multi-speaker pre-training dataset were calculated after discarding the highest and lowest 10% samples. Then, we mapped the range between the minimum and maximum values to [-1, 1] to get normalized prosody values. In order to control each intuitive prosodic feature at the

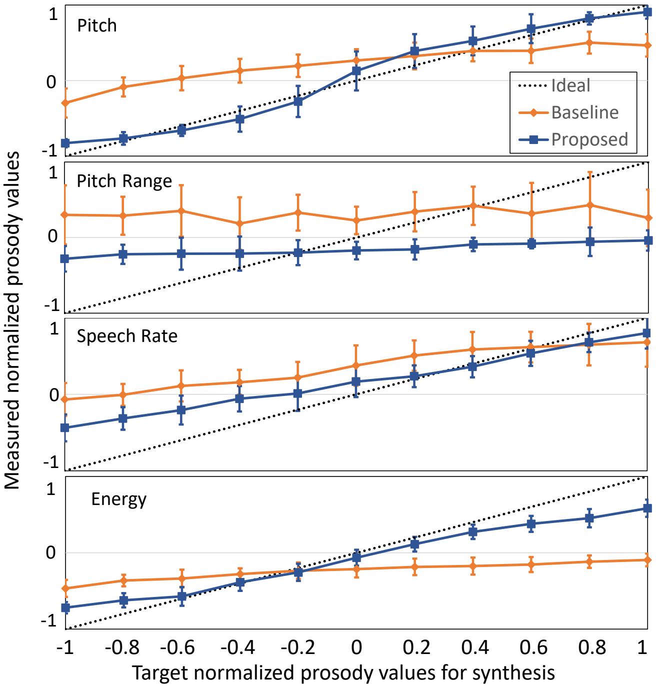

Fig. 3. The mapping curves between measured and target prosody values for different intuitive prosodic features and models on the female target speaker.

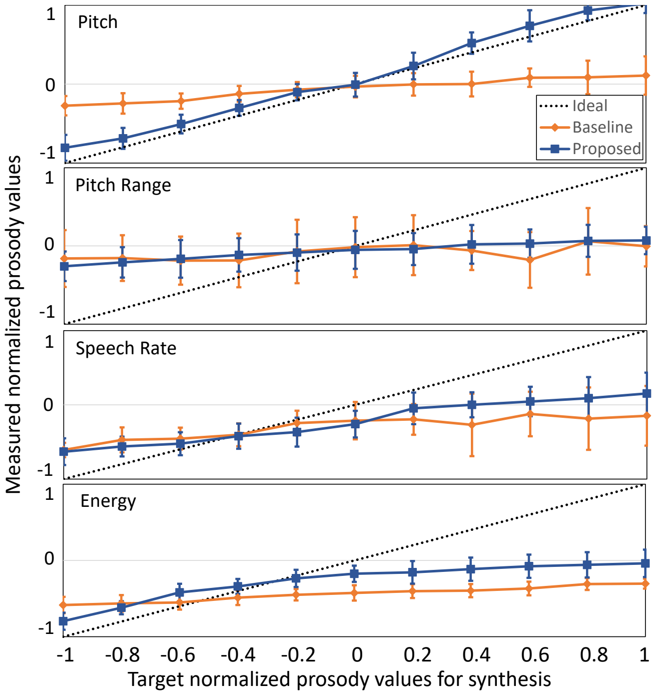

Fig. 4. The mapping curves between measured and target prosody values for different intuitive prosodic features and models on the male target speaker.

[1] Samples can be found at https://chengpy22.github.io/DSRIPF_demo/.

synthesis stage, a group of target normalized prosody values were sampled with intervals of 0.2 in [-1, 1] and were then denormalized to the original values of each prosodic feature for decoding the utterances in the test set. For each target normalized prosody value, intuitive prosodic features were extracted from synthetic utterances, and were then normalized to get the measured normalized prosody values. The mean and standard deviation of the measured prosody values were calculated from all test utterances. Finally, the mapping curves between measured and target prosody values were drawn to compare the prosody controllability of different models as shown in Fig. 3 and Fig. 4. The ideal curve should be a diagonal one as shown by the dotted lines in these figures.

The results show that our proposed method can better reflect the target prosody values in synthetic speech than the baseline method in the prosody control experiment. For pitch, the proposed method presented almost ideal correlation between target and measured prosody values on both speakers. The performance of controlling pitch range was the worst among the four prosodic features for both models. The reason may be that the reading style of most speakers in the AISHELL-3 dataset was rather bland, and the models failed to learn the correlations between pitch ranges and Mel-spectra. This issue is worth further investigation in the future.

We also evaluated subjective performance of prosody control for these two models. For each intuitive prosodic feature, we selected ten pairs of synthetic utterances with the same measured prosody values from the samples used to draw Fig. 3 and Fig. 4. The two utterances in each pair had the same contents and were synthesized by the two models respectively. Then, the mean opinion scores (MOS) on naturalness and similarity of these selected samples were evaluated by a listening test. 11 native listeners took part in the test to give a 5-scale opinion score (5:excellent, 4: good, 3: fair, 2: poor, 1: bad) on both naturalness and similarity for each sample.

The subjective evaluation results are shown in Table III. We can see that for each prosodic feature, the proposed model outperformed the baseline model on both naturalness and similarity, except the similarity when controlling pitch. This further confirms the effectiveness of our proposed method on prosody control by eliminating the possible conflict between speaker representations and modified prosodic features.

V. Conclusions

In this paper, we have proposed a multi-speaker TTS and speaker adaptation framework with disentanglement between intuitive prosodic features and speaker representations. Experiments on speaker adaptation demonstrated that our method can achieve better naturalness of synthetic speech than the baseline model with intuitive prosodic features but without disentanglement. The objective and subjective evaluation results of the prosody control experiment further proved that our method outperformed the baseline model on prosody controllability. Our future work includes integrating more intuitive acoustic features, such as voice quality features, into current framework for adaptive and controllable speech synthesis.

References

[1] 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. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 2018.

[2] 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.

[3] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6706–6713, 2019.

[4] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[5] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.

[6] A. Tjandra, S. Sakti, and S. Nakamura, "Machine speech chain with one-shot speaker adaptation," *arXiv preprint arXiv:1803.10525*, 2018.

[7] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting new speakers based on a short untranscribed sample," in *International Conference on Machine Learning*, pp. 3683–3691, PMLR, 2018.

[8] Q. Hu, E. Marchi, D. Winarsky, Y. Stylianou, D. Naik, and S. Kajarekar, "Neural text-to-speech adaptation from low quality public recordings," in *Speech Synthesis Workshop*, vol. 10, 2019.

[9] Z. Kons, S. Shechtman, A. Sorin, C. Rabinovitz, and R. Hoory, "High quality, lightweight and adaptable TTS using LPCNet," in *Interspeech*, pp. 176–180, 2019.

[10] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6184–6188, IEEE, 2020.

[11] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in Neural Information Processing Systems 31*, pp. 4485–4495, 2018.

[12] S. O. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," *arXiv preprint arXiv:1802.06006*, 2018.

[13] Y. Chen, Y. Assael, B. Shillingford, D. Budden, S. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie, *et al.*, "Sample efficient adaptive text-to-speech," *arXiv preprint arXiv:1809.10460*, 2018.

[14] Y. Huang, L. He, W. Wei, W. Gale, J. Li, and Y. Gong, "Using Personalized Speech Synthesis and Neural Language Generator for Rapid Speaker Adaptation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7399–7403, 2020.

[15] Y. Deng, L. He, and F. Soong, "Modeling multi-speaker latent space to improve neural tts: Quick enrolling new speaker and enhancing premium voice," *arXiv preprint arXiv:1812.05253*, 2018.

[16] Y. Zheng, X. Li, and L. Lu, "Investigation of fast and efficient methods for multi-speaker modeling and speaker adaptation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6618–6622, IEEE, 2021.

[17] P. Cheng and Z. Ling, "Speaker adaption with intuitive prosodic features for statistical parametric speech synthesis," in *Proceedings of the 6th International Conference on Digital Signal Processing*, pp. 187–193, 2022.

[18] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," in *Interspeech*, 2020.

[19] M. Morrison, Z. Jin, J. Salamon, N. J. Bryan, and G. J. Mysore, "Controllable neural prosody synthesis," *Proc. Interspeech 2020*, pp. 4437–4441, 2020.

[20] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6189–6193, IEEE, 2020.

[21] 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*, pp. 5180–5189, PMLR, 2018.

[22] C. Lu, X. Wen, R. Liu, and X. Chen, "Multi-speaker emotional speech synthesis with fine-grained prosody modeling," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5729–5733, 2021.

[23] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus," *Proc. Interspeech 2021*, pp. 2756–2760, 2021.

[24] W. Cai, J. Chen, J. Zhang, and M. Li, "On-the-Fly Data Loader and Utterance-Level Aggregation for Speaker and Language Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1038–1051, 2020.

[25] 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.

[26] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4784–4788, 2018.

[27] 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 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6199–6203, IEEE, 2020.

[28] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, vol. 1, pp. 125–128, IEEE, 1993.