Proceedings of the 9th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'23)
Brunel University, London, United Kingdom - August 03-05, 2023
Paper No. MHCI 111
DOI: 10.11159/mhci23.111

GAN-Based Fine-Grained Feature Modeling For Zero-Shot Voice Cloning

Zhongcai Lyu[1], Jie Zhu[2]
[1]Shanghai Jiao Tong University
800 Dongchuan RD. Minhang District, Shanghai, China
lzcsjtu@sjtu.edu.cn
[2]Shanghai Jiao Tong University
800 Dongchuan RD. Minhang District, Shanghai, China
zhujie@sjtu.edu.cn

Abstract - With the continuous development of deep learning and speech signal processing, speech synthesis technology has greatly improved in naturalness and comprehensibility, and many application technologies such as artificial intelligence voice assistant and personalized navigation have been widely used in real life, and the demand for personalized speech synthesis is increasing. Personalized speech synthesis requires models that can achieve speech timbre migration, also known as speech reproduction, with only a small number of target speaker speech samples. However, since human speech is highly expressive and contains rich information, including speaker identity information, prosody, rhythm, emotion and other factors, the limited speech data will lead to poor similarity and rhythmic performance of the model-generated speech, and the model needs to be fine-tuned to improve the quality of the synthesized speech. Therefore, personalized speech synthesis with few samples is a very challenging task. To achieve the goal of speech cloning, this paper proposes a personalized speech synthesis method based on FastSpeech2. By using fine-grained feature modeling module containing prosody extractor and prosody predictor, and a training strategy based on Generative adversarial network (GAN) and meta-learning, it is realized that personalized speech with high similarity and naturalness can be generated with a very short reference audio. The subjective and objective experiments also demonstrate that the model proposed in this paper can achieve high quality speech replication without fine-tuning the model under a few or even a single reference audio of the target speaker.
Keywords**:** Voice cloning, zero-shot, feature modelling, fine-grained, GAN

1. Introduction

Text-to-speech (TTS) technology aims to generate intelligible and natural-sounding speech from a given text, enabling computers and smart devices to communicate with humans. With the increasing demand for personalized speech synthesis, TTS models are expected to generate high-quality speech for any user, even with limited speech data samples. Thus, personalized speech synthesis has become an important research direction for TTS technology, which requires TTS models to not only generate high-quality speech but also to be able to capture well with only a small number of speech data samples the speech of a given speaker. To achieve this goal, the TTS model should be able to generate the speech of multiple speakers and adapt well to the voices of unseen speakers.

Personalized speech synthesis has been a popular research topic and it has different terms in academia and industry, such as voice reproduction, voice adaptation [1], voice cloning [2], custom speech synthesis [3], etc. With the increasing demand for personalized speech generation, the adaptation of TTS models to new speakers has been extensively investigated by domestic and international researchers. For example, many statistical parametric speech synthesis has been studied for speech adaptation [4], and the recent speech cloning challenge has attracted many participants [5-7].

The research directions of personalized speech synthesis TTS can be mainly divided into the following.

(1) Firstly, train a multi-speaker speech synthesis system on a multi-speaker dataset, and then use samples of target speakers with less data to fine-tune the whole model [1-2] or some components of the model [8-10]. This approach works well but requires too many adaptive parameters.

(2) Adaptation for multi-speaker model generalization, including generalization of the TTS model and adaptation to different domains to enhance the expressiveness of model synthesis. To allow the model to achieve fine-grained control and for modeling and decoupling of speech feature learning, this information can be analyzed and combined [11] from coarse to

fine granularity when modeling. Recent studies have used this approach for speaker modeling in TTS, such as obtaining language ID, speaker ID, style, and rhyme [12-13] from tag data. Rhyme information can be tagged according to some annotation patterns, such as AuToBI [14], INTSINT [15], etc. There is also the extraction of pitch and energy information from speech and duration from paired text and speech data[16-18]. AdaSpeech3 [19] designs specific of padding pause adaptation, rhythm adaptation, and timbre adaptation to make the personalized speech synthesis model more adaptive.

(3) Personalized speech synthesis is achieved by modeling the speaker identity and then using the modeled speaker embedding vector in combination with a multi-speaker speech synthesis model. The most mainstream approach is still speaker reference encoder [20-24] and variational auto-encoder [25-27].

In addition, some adversarial generative models have been applied to speaker modeling, such as using advanced generative models to implicitly learn speaker feature information, which can better model multimodal distributions and solve the one-to-many mapping problem [28].

In this paper, we focus on the research of personalized speech synthesis with zero learning under few samples. The main research contents and innovation points of this paper can be summarized as follows.

(1) To address the problems of large training data volume, high sound quality requirements, and weak generalization ability of current speech synthesis systems, we propose a fine-grained feature modeling module containing a prosody extractor and a prosody predictor to model the implicit prosody features in speech and realize the improvement of naturalness for the generated speech.

(2) To address the problem that the model has weak generalization ability under the short reference speech of unknown speakers, we propose a discriminator module including speaker embedding discriminator and phoneme discriminator, and speaker prototype vector based on the training strategy of GAN and meta-learning [29], which enhances the adaptive ability of the model to unknown speakers and improves the sound quality and generalization ability of the model.

(3) Through multiple comparative experiments, including MOS, CMOS evaluation, AB preference testing, our model's effectiveness is demonstrated, effectively meeting the needs of personalized speech synthesis.

The remaining part of this paper proceeds as follows. The overview and components of the proposed model are described in section 2. Experiments and results are reported in section 3. Lastly, the conclusion and future work are covered in section 4.

2. Methodology

The model is mainly based on the FastSpeech2 model and its overall structure is mainly divided into a fine-grained modeling module and a GAN-based meta-learning module. The fine-grained modeling module contains a speaker encoder, a prosody extractor, and a predictor with the SALM structure, collectively referred to as the Hybrid TTS Model (HTM) structure. The GAN-based meta-learning module is divided into a generator module and a discriminator module. The overall structure of the model FFM-GAN is shown in the following diagram.

3.1. Fine-Grained Feature Modeling Module

The overall structure of the Hybrid TTS Model (HTM) includes: a speaker encoder with the SALM structure (which is introduced in our previous work [30] and fine-grained modeling module. By modeling the speaker identity, the model can synthesize speech with specific speaker characteristics without the need for model fine-tuning.

We design a fine-grained feature modeling module based on implicit variables, including a prosody extractor based on the reference speaker's mel-spectrogram and a prosody predictor based on phoneme features and speaker information to extract acoustic features, which can model the speed, energy, pitch, and pauses of speech. The target speaker's prosody and tone can be combined for speech synthesis, ultimately improving the naturalness and robustness of the adaptive speech synthesis model.

The structure of the personalized speech synthesis model with added fine-grained modeling module is shown in Figure 2.

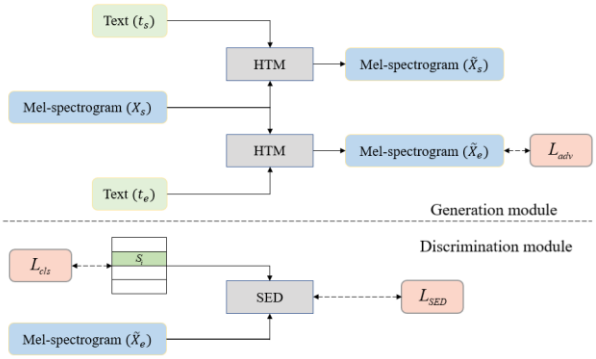

Fig. 1: The overall structure of our model FFM-GAN.

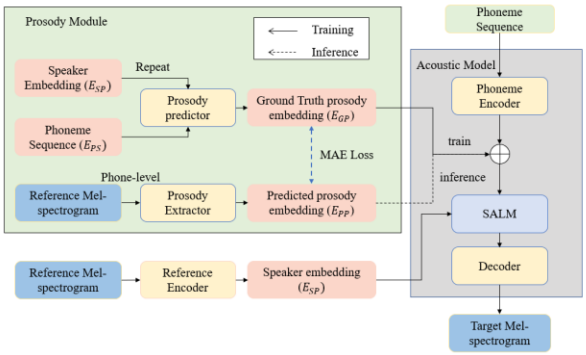

Fig. 2: The structure of the personalized speech synthesis model with added fine-grained modeling module.

(1) Prosody Extractor

During training, the fine-grained implicit prosodic labels are extracted from the real Mel spectrogram, and each prosodic label contains the prosodic information of a very short speech segment (ie, a word or a phoneme). Is a low-dimensional vector or scalar, they are mapped to the prosodic embedding vector, whose dimension is equal to the dimension of the phoneme-level feature, and then the prosodic embedding vector is added to the phoneme feature sequence, mixed together as the input of the subsequent decoder to predict the final Mel Spectrogram. This module is jointly learned with the TTS model.

(2) Prosody Predictor

During the inference stage, there is typically no corresponding reference audio of the target speaker available to serve as a guide for synthesized speech. Therefore, prosody predictor is designed to predict the speaker's prosody embedding vector. Prosody features are modeled using both the original phoneme sequence and predicted phoneme-level intonation vectors obtained from the reference Mel spectrogram. This approach enables effective learning of the target speaker's pronunciation habits, gender information, and other factors, resulting in improved similarity and naturalness of the synthesized speech in terms of fine-grained feature.

After fine-grained feature modeling, the final phoneme embedding vector E_{FP} is calculated using the following equation:

$$E_{FP} = \begin{cases} E_{PS} + E_{GP} = E_{PS} + E(X) & \text{When training} \\ E_{PS} + E_{PP} = E_{PS} + P(E_{PS} + E_{SP}) & \text{When infering} \end{cases} \quad (1)$$

Here, E represents the function represented by the intonation extractor and P represents the function represented by the intonation predictor. During the training phase, the true intonation vector E_{GP} is extracted from the intonation extractor using the reference Mel spectrogram X. During the inference process, the predicted intonation vector E_{PP} is obtained from the phoneme sequence vector E_{PS} and the speaker embedding vector E_{SP}.

3.2. GAN-Based Meta-learning Module

Given the lack of adaptability of the model to unknown speakers, the model is based on the GAN meta-learning to realize the adaptation of the speech synthesis model with few samples and zero samples, and the TTS model trained the meta-learning algorithm is used for unknown speakers to adapt efficiently. The model is mainly divided into two a generation module and a discrimination module. The main model of the generation module is a hybrid TTS model (HTM) structure with a fine-grained feature modeling module, and the discrimination module is composed of a speaker embedding discriminator (SED).

In addition to randomly sampling a sample (t_s, X_s) containing text and corresponding speech from the corpus of the target speaker, an additional text information t_e is introduced during training. In the training process, in addition to using (t_s, X_s) to generate the predicted speech $\tilde{X}_s$ through the HTM model, and calculate the reconstruction loss Lrecon with the speech signal X_s as the training target, an additional The text t_e and the speaker embedding vector E_{SP} obtained from the target speaker's reference speech X_s using the reference encoder generates an additional speech signal $\tilde{X}_e$, and use the subsequent discriminator to calculate the adversarial loss to achieve better personalization Generalization ability of speech synthesis models to unknown speakers.

The speaker embedding discriminator (SED) has a similar architecture to the speaker reference encoder, the difference is that the speaker embedding discriminator uses a one-dimensional convolution instead of the original gated convolution (GLU), which is faster the training process of the accelerated model, the generated speaker embedding vector E_{sd} and the speaker embedding vector E_{SP} generated by the reference encoder have the same dimension.

In the speaker embedding discriminator, a set of speaker prototype embedding vectors S is also introduced as a reference, where K represents the number of speakers in the training set, and S_i represents the i-th speaker's prototype embedding vector. This prototype embedding vector Directly uses the corresponding speaker ID in the training set to obtain it through embedding layer transformation, which has the same dimension as the speaker embedding vector E_{sd} generated by the discriminator. During training, the model constraints E_{sd} and prototype vectors using a classification loss function.

Calculate the dot product by calculating the speaker embedding vector E_{sd} generated by the speaker embedding discriminator and the prototype vector S of all speakers in the training set, and then achieve the purpose of multi-classification by calculating the cross-entropy loss, the speaker the calculation formula of the prototype vector loss function L_{cls} is as follows:

$$L_{cls} = -\log \frac{\exp\left(E_{sd}^{T} s_i\right)}{\sum_{i'} \exp\left(E_{sd}^{T} s_{i'}\right)} \tag{2}$$

Where E_{sd}^{T} means that E_{sd} has been transposed.

3. Experiments

3.1. Dataset and Settings

In the preprocessing stage, the speech signal is obtained from the Mel spectrogram after windowing, framing, short-time Fourier transform, and Mel filtering. The 80-dimensional Mel spectrogram is selected as the target output and input for training. A Hanning window with a window length size of 1024 is used in the short-time Fourier transform, with a frameshift set to 256 sampling points, FFT size set to 1024, and frequency range set to [0, 8000]. The speech signal is downsampled using Librosa before and after mute cropping, with all audio files downsampled to 22.05 kHz. The model employs four FFT blocks for both the phoneme encoder and the Mel spectrogram decoder, based on FastSpeech2. Additional preprocessing networks are set up for the encoder and decoder, comprising two convolutional layers and a linear layer with residual connections in the encoder and two linear layers in the decoder, to increase the model's robustness. All hidden vectors, including the phoneme embedding vector, speaker embedding vector, and input and output of FFT blocks, are set to 256 dimensions. All inputs are padded to maximum length during training.

The HiFi-GAN introduced in Chapter 2 is utilized as the vocoder to convert the generated Mel spectrogram into audio waveform files, exhibiting strong speaker adaptability. The Adam optimizer is used for optimization in the model training, with the initial learning rate set to 0.0002. The learning rate decay strategy is implemented by multiplying by 0.9 after every 10,000 iterations. The batch size is set to 16, and both the model and the comparison models are trained on two GPU 3080Ti RTX machines for 200,000 iterations.

3.2. Subjective Evaluations

To ensure the overall performance of the model and evaluate the impact of individual modules, we have designed several baseline models for comparison:

(1) GT: Target speaker reference speech audio as ground truth; (2) GT_HiFi, which generates synthetic audio by passing the Mel spectrogram obtained from the target speaker's reference speech audio through HiFi-GAN; (3) Base: Use the multi-speaker FastSpeech2 model to synthesize the Mel spectrogram, and then uses HiFi-GAN to obtain the synthesized audio; (4) Base-FFM: Based on the model (3), a fine-grained feature modeling module is added; (5) Base-GAN: Based on the model (3), a GAN-based meta-learning module is introduced; (6) FFM-GAN: Our proposed final model with a fine-grained feature modeling module and a GAN-based meta-learning module.

In our experiments, we verified the adaptability of the model to unknown speakers. Specifically, we randomly selected an audio sample from the speaker in the test set as a reference and used the reference speech and the given text to generate speech. We evaluated the performance of the model using the MOS (mean opinion score) and SMOS (standard deviation of MOS) indicators, which were manually rated on a scale of 1 to 5 by 10 judges for each audio sample. The evaluation results with 95% confidence intervals are presented in Table 1.

Table 1: Subjective evaluation results of comparative experimental verification.

Model	MOS	SMOS
GT	4.42±0.13	4.77±0.14
GT-HiFi	4.12±0.15	4.62±0.14
Base	3.45±0.12	2.97±0.13
Base-FFM	3.52±0.15	3.43±0.16
Base-GAN	3.58±0.13	3.59±0.13
FFM-GAN	3.67±0.11	3.70±0.14

We conducted a naturalness test using AB test, where listeners were presented with pairs of synthesized samples with identical text and asked to select the sample that sounded more natural. The experimental results, as shown in Figure 3, indicate the level of naturalness for each synthesized sample.

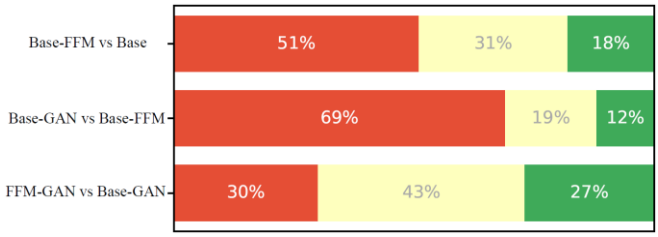

Fig. 3: The results of AB tests.

It is evident from the results that the fine-grained feature modeling module enables speaker representation learning for the target speaker, resulting in higher SMOS values and fluent synthesized speech. Moreover, using the reference speech of a single target speaker can achieve better-personalized speech synthesis in terms of sound quality and similarity. The GAN-

based meta-learning module further enhances the similarity and naturalness of synthesized speech, leading to a 0.2-point improvement in SMOS compared to other models. Additionally, the sound quality of the synthesized speech is closest natural speech. The proposed final model, FFM-GAN, exhibits clear advantages over other models.

3.3. Objective Evaluations

In addition to subjective evaluation, we also use many objective evaluation indicators to assist in the analysis of the synthesized speech to prove the effectiveness of the prosodic module, including:

(1) Mel Cepstrum Distortion (MCD): MCD is one of the objective indicators for evaluating model-synthesized speech, and it is a measure of the numerical difference between the Mel cepstrum sequences of two speeches. (2) Gross Pitch Error (GPE): Measures the pitch similarity between the reference audio and the predicted audio. The lower the error rate, the more accurate the prosody of the speech. (3) Voicing Decision Error (VDE): Measures the difference in voice decision between reference audio and predicted audio. (4) $F0$ Frame Error (FFE): Combining the two errors of GPE and VDE, count the percentage of frames with both errors at the same time.

The comparison results are shown in Table 2.

Table 2: Objective evaluation results of comparative experimental verification.

Model	GPE↓	VDE↓	FFE↓	MCD↓
Base	0.47±0.17	0.33±0.16	0.49±0.21	4.72±0.20
Base-FFM	0.41±0.13	0.29±0.15	0.45±0.14	4.30±0.18
Base-GAN	0.45±0.16	0.37±0.14	0.54±0.23	4.83±0.22
FFM-GAN	0.39±0.18	0.28±0.19	0.43±0.17	4.28±0.19

The table shows that the model with fine-grained feature modeling module consistently outperforms other models. Adding a prosodic predictor with speaker embeddings allows sequential control of prosody, and a GAN-based meta-learning module improves the robustness of synthesized speech. Together, they can help synthesize high-quality and natural speech.

3.4. Ablation Experiment

We further conduct ablation experiments to verify the effectiveness of various components in the GAN-based meta-learning module of the personalized speech synthesis model, including the verification of the role of speaker embedding discriminator and speaker prototype vector. In the ablation study, we use three metrics: MCD, MOS and SMOS, and the results are shown in the Table 3.

Table 3: Validation results of ablation experiments.

Model	MOS↑	SMOS↑	MCD↓
FFM-GAN	3.76±0.14	3.83±0.13	4.28±0.18
w/o SED	3.75±0.13	3.59±0.16	4.52±0.21
w/o S	3.72±0.13	3.68±0.14	4.31±0.13

The ablation study results show that the speaker embedding discriminator and speaker prototype vector are effective components in the GAN-based meta-learning module. Without the speaker embedding discriminator, the SMOS score drops significantly, indicating that the model cannot effectively learn the speaker representation. Similarly, removing the speaker prototype vector also leads to a decrease in SMOS score, indicating that the speaker prototype vector plays a crucial role in controlling the speaker similarity of the synthesized speech. Overall, these results demonstrate the importance of these components for achieving high-quality and personalized speech synthesis.

3.5. Model Generalization Capability Verification

Finally, we used the speech audio of the target speaker with different lengths as a reference, and selected four different different lengths: <1s, 1-3s, 3-10s and 10-20s, we scored MOS and SMOS for subjective evaluation of the synthesis effect effect of each model under the above reference audio. The scoring results are shown in the Table 4.

Table 4: Evaluation results of synthesized audio with different lengths of reference audio.

Model	MOS				SMOS			
	<1s	1-3s	3-10s	10-20s	<1s	1-3s	3-10s	10-20s
Base	3.32±0.12	3.36±0.10	3.42±0.10	3.47±0.14	2.33±0.12	2.41±0.13	2.65±0.13	2.67±0.10
Base-FFM	3.45±0.10	3.48±0.11	3.51±0.13	3.53±0.13	2.18±0.17	3.37±0.10	3.39±0.13	3.41±0.12
Base-GAN	3.53±0.13	3.58±0.12	3.62±0.12	3.59±0.12	3.27±0.13	3.46±0.16	3.66±0.12	3.56±0.14
FFM-GAN	3.54±0.12	3.57±0.14	3.66±0.10	3.64±0.12	3.32±0.16	3.63±0.14	3.71±0.10	3.68±0.12

A longer reference audio can help the model to better learn the characteristics of the target speaker. Our model performance has also reached the best under various lengths of reference audio, which directly illustrates the fine-grained feature modeling module and based on Effectiveness of GAN-based meta-learning module.

4. Conclusion

As a core technology in speech interaction, speech synthesis has advanced significantly with the fusion of deep learning and speech signal processing, allowing for high naturalness, clarity, and intelligibility in speech synthesis. However, there are still numerous challenges to be addressed, such as adapting speech synthesis to low-sample and low-quantity training data, and achieving expressive emotional speech synthesis.

This paper presents a personalized speech synthesis model based on the FastSpeech2 architecture that uses minimal training data and requires no fine-tuning to achieve high similarity and naturalness in replicating a target speaker's voice based on a single reference speech. Additionally, the model will be improved to allow for the combination and control of different styles and tones, as well as the ability to generate target speaker speech with different rhythms and styles, thus expanding its application in real-world scenarios.

References

[1] 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 adap- tive text-to-speech," arXiv preprint arXiv:1809.10460, 2018.

[2] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," Advances in neural information processing systems, vol. 31, 2018.

[3] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T.-Y. Liu, "Adaspeech: Adaptive text to speech for custom voice," arXiv preprint arXiv:2103.00993, 2021.

[4] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for dnn-based tts synthesis," in 2015 IEEE interna- tional conference on acoustics, speech and signal processing (ICASSP). IEEE, 2015, pp. 4475-4479.

[5] Q. Xie, X. Tian, G. Liu, K. Song, L. Xie, Z. Wu, H. Li, S. Shi, H. Li, F. Hong et al., "The multi-speaker multi-style voice cloning challenge 2021," in ICASSP 2021-2021 IEEE International Conference on Acous- tics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 8613– 8617.

[6] C.-H. Hu, Y.-C. Wu, W.-C. Huang, Y.-H. Peng, Y.-W. Chen, P.-J. Ku, T. Toda, Y. Tsao, and H.-M. Wang, "The as-nu system for the m2voc challenge," arXiv preprint arXiv:2104.03009, 2021.

[7] D. Tan, H. Huang, G. Zhang, and T. Lee, “Cuhk-ee voice cloning sys- tem for icassp 2021 m2voc challenge,” arXiv preprint arXiv:2103.04699, 2021.
[8] H. B. Moss, V. Aggarwal, N. Prateek, J. Gonz´alez, and R. Barra-Chicote, “Boffin tts: Few-shot speaker adaptation by bayesian optimization,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 7639-7643.
[9] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, A. Rosenberg, B. Ram- abhadran, and Y. Wu, “Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and autoregressive prosody prior,” in ICASSP 2020-2020 IEEE International Conference on Acous- tics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6699–6703.
[10] C. Zhang, Y. Ren, X. Tan, J. Liu, K. Zhang, T. Qin, S. Zhao, and T.-Y. Liu, “Denoispeech: Denoising text to speech with frame-level noise model- ing,” in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 7063-7067.
[11] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, “A survey on neural speech syn- thesis,” arXiv preprint arXiv:2106.15561, 2021.
[12] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, “Learning to speak fluently in a for- eign language: Multilingual speech synthesis and cross-language voice cloning,” arXiv preprint arXiv:1907.04448, 2019.
[13] T. Nekvinda and O. Du˘sek, “One model, many languages: Meta-learning for multilingual text-to-speech,” arXiv preprint arXiv:2008.00768, 2020.
[14] A. Rosenberg, “Autobi-a tool for automatic tobi annotation,” in Eleventh annual conference of the international speech communication associa- tion, 2010.
[15] D. Hirst, “Automatic analysis of prosody for multilingual speech cor- pora,” Improvements in speech synthesis, pp. 320-327, 2001.
[16] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fast- speech: Fast, robust and controllable text to speech,” Advances in neural information processing systems, vol. 32, 2019.
[17] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Fast- speech 2: Fast and high-quality end-to-end text to speech,” arXiv preprint arXiv:2006.04558, 2020.
[18] Y. Ren, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, “Almost unsuper- vised text to speech and automatic speech recognition,” in International conference on machine learning. PMLR, 2019, pp. 5410-5419.
[19] Y. Yan, X. Tan, B. Li, G. Zhang, T. Qin, S. Zhao, Y. Shen, W.-Q. Zhang, and T.-Y. Liu, “Adaspeech 3: Adaptive text to speech for spontaneous style,” arXiv preprint arXiv:2107.02530, 2021.
[20] 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.
[21] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, “Deep voice 3: 2000-speaker neural text-to- speech,” proc. ICLR, pp. 214-217, 2018.
[22] S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, “Deep voice 2: Multi-speaker neural text-to-speech,” arXiv preprint arXiv:1705.08947, 2017.
[23] Y. Yan, X. Tan, B. Li, T. Qin, S. Zhao, Y. Shen, and T.-Y. Liu, “Adaspeech 2: Adaptive text to speech with untranscribed data,” in ICASSP 2021- 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 6613-6617.
[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. PMLR, 2018, pp. 5180– 5189.
[25] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, “Learning latent represen- tations 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). IEEE, 2019, pp. 6945-6949.

[26] N. Dilokthanakul, P. A. Mediano, M. Garnelo, M. C. Lee, H. Salim- beni, K. Arulkumaran, and M. Shanahan, “Deep unsupervised clus- tering with gaussian mixture variational autoencoders,” arXiv preprint arXiv:1611.02648, 2016.
[27] Y. Lee, J. Shin, and K. Jung, “Bidirectional variational inference for non- autoregressive text-to-speech,” in International conference on learning representations, 2022.
[28] S. Ma, D. Mcduff, and Y. Song, “Neural tts stylization with adversarial and collaborative games,” in International conference on learning repre- sentations, 2018.
[29] R. Vilalta and Y. Drissi, “A perspective view and survey of meta- learning,” Artificial intelligence review, vol. 18, pp. 77-95, 2002.
[30] Z. Lyu and J. Zhu, “Enriching style transfer in multi-scale control based personalized end-to-end speech synthesis,” in 2022 12th International Conference on Information Science and Technology (ICIST). IEEE, 2022, pp. 114-119.