ECAPA-TDNN for Multi-speaker Text-to-speech Synthesis

Jinlong Xue1, Yayue Deng1,2, Yichen Han1, Ya Li1,∗, Jianqing Sun3, Jiaen Liang3

1School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
2Beijing Language and Culture University, Beijing, China
3Unisound AI Technology Co., Ltd, Beijing, China

Abstract

In recent years, neural network based methods for multi-speaker text-to-speech synthesis (TTS) have made significant progress. However, the current speaker encoder models used in these methods still cannot capture enough speaker information. In this paper, we focus on accurate speaker encoder modeling and propose an end-to-end method that can generate better similarity for both seen and unseen speakers. The proposed architecture consists of three separately trained components: a speaker encoder based on the state-of-the-art ECAPA-TDNN model which is derived from speaker verification task, a FastSpeech2 based synthesizer, and a HiFi-GAN vocoder. The comparison among different speaker encoder models shows our proposed method can achieve better speaker similarity. To efficiently evaluate our synthesized speech, we are the first to adopt and evaluate different deep learning based automatic MOS evaluation methods to assess our results, and these methods show great potential in automatic speech quality assessment.

Index Terms: multi-speaker text-to-speech, speaker representation, MOS prediction

1. Introduction

Text-to-speech (TTS) aims to produce natural human speech. In the past few years, deep learning based models have developed rapidly. Recent research shows that the quality and the naturalness of the synthetic voices are comparable with real human speech, such as Tacotron 2 [1], DeepVoice 3 [2], and FastSpeech 2 [3]. Despite the successful achievement of speaker-dependent TTS, how to create expressive and controllable speaking styles in multi-speaker task still needs more research. Besides, the models of voice cloning in unseen speaker circumstances by using a speaker encoder usually tend to synthesize neutral and poor similarity voices compared to the real speaker. Therefore, how to sufficiently extract speaker information from the reference voices becomes significant.

To accomplish the multi-speaker task, a TTS system and a speaker representation are needed. In previous studies, most multi-speaker systems use a speaker encoder to extract speaker embedding to characterize the target speaker’s voice and style. Because models in speaker verification task are designed to extract the text-independent speaker embeddings from the target speaker voices to capture speaker characteristics, they have been widely adopted as the speaker encoder, such as d-vector [4], x-vector [5] and Deep Speaker [6]. Besides, pretrained models are more often used instead of jointly training with the TTS system, because the speaker knowledge for speaker encoder is limited by training dataset in the latter case. Jia et al. [7] investigated the knowledge transfer where the speaker verification model is trained on a dataset with many speakers, like VoxCeleb [8, 9] dataset. Thus, the speaker embedding extracted from the speaker verification model conditioning the TTS system leads to better generalization and performance on the multi-speaker TTS and the voice cloning task.

However, the speaker similarity of the audios synthesized from the current models are less favorable, especially in unseen datasets. The reason is that the ability of the current speaker encoders is not enough to capture enough information of the target speakers in the multi-speaker TTS task. To address these weaknesses, we propose our multi-speaker TTS by adopting the non-autoregressive TTS model FastSpeech 2 and the TDNN-based model ECAPA-TDNN [10] from speaker verification task, which has stronger speaker feature extraction ability and robustness [11]. It introduces multiple enhancements to the basic architecture and outperforms other TDNN based speaker verification models on the VoxCeleb datasets. We compare different speaker encoders, such as x-vector [5], Deep Speaker [6] and SpeakerNet [12], to investigate their generalization ability in two publicly available datasets for both the seen and unseen tests, and our method outstands other methods in speaker similarity.

To better evaluate our methods, we need many subjective evaluations including the mean opinion score (MOS) test and speaker similarity test. However, such measurement requires many humans to be involved, making it time-consuming and expensive. Thanks to the VCC 2016 and VCC 2018 datasets [13, 14], several deep-learning-based automatic speech quality evaluation methods have been proposed, such as MOSNet [15], MBNet [16] and a self-supervised representation based MOS predictor model (denoted as S3PRL) [17]. To our best knowledge, we are the first to evaluate the synthesized speech by using the MOS prediction model to accelerate our research. The results obtained from automatic MOS prediction models are consistent with subjective MOS results, which shows that they have great potential to free us from the burden of MOS tests.

The paper is organized as follows: Section 2 describes related works in terms of speaker representations, and Section 3 illustrates our proposed method with training workflow. Experimental setup and results are shown in Section 4. At last, we conclude our finding in Section 5. Examples of synthesized speech can be found on the project page1.

2. Speaker Representations

Multi-speaker TTS system highly depends on speaker representation for conditioning the acoustic model to clone the voices of

1Audio samples: https://happylittlecat2333.github.io/iscslp2022
target speakers. Many models derived from the speaker verification task have been applied in TTS system because they can extract speaker information from speech. These models are usually trained on a large number of text-independent datasets [8, 9] recorded by many speakers, which provides them with the ability to capture the subtle characteristics and styles from different speakers by using only a short utterance under any circumstances.

In the speaker verification task, deep neural network models have surpassed the classic model i-vector [18]. Among them, d-vector [4], x-vector [5], Deep Speaker [6] and SpeakerNet [12] are representative methods. They show great potential and some have been widely used in multi-speaker TTS [7, 19, 20]. Below, we describe each of these speaker encoder models.

2.1. Deep Speaker

Deep Speaker [6] uses ResCNN and GRU neural network as its backbone for frame-level feature extraction, and adopts pooling and length normalization layers to generate utterance-level speaker embeddings.

2.2. X-vector

In x-vector [5], a time-delay neural network (TDNN) with subsampling is used as the encoder network. An attentive statistics pooling (ASP) [21] layer aggregates frame-level outputs and computes its mean and standard deviation. After sending through segment-level layer, the x-vector speaker embedding is obtained.

2.3. SpeakerNet

SpeakerNet [12] adopts QuartzNet [22] as top-level feature extraction, which is composed of residual blocks with 1D depthwise separable convolutional network. For the pooling layer, it uses x-vector based statistics pooling layer to compute fixed-length embeddings. SpeakerNet is lighter than other models while remaining similar performance.

3. Proposed Method

Inspired by the outstanding performance of ECAPA-TDNN in speaker verification task, we introduce a speaker encoder based on this model to our multi-speaker TTS system. Fig. 1 shows our modified speaker encoder based on ECAPA-TDNN and our proposed method with training and inference workflows.

3.1. Speaker encoder

Based on recent trends in the related field of computer vision and face verification, ECAPA-TDNN uses TDNN as its base architecture and introduces multiple enhancements: using Squeeze-and-Excitation (SE) blocks [23] in encoder modules to explicitly model channel interdependencies; implementing Res2Net with skip connections; aggregating and propagating features of different hierarchical levels in the encoder to capture both the shallow and deep speaker feature maps; improving statistics pooling module with channel- and context-dependent frame attention to focus more on speaker-specific characteristics such as focusing more vowels than consonants. Those improvements endow ECAPA-TDNN with the ability to extract subtle speaker information and outperform other models.

3.2. Acoustic model

We extend the non-autoregressive model FastSpeech 2 [3] architecture to implement our multi-speaker model. FastSpeech 2 is composed of Transformer-based encoder and decoder with a variance adaptor. The encoder generates the hidden embedding from a sequence of phoneme-level inputs. The variance adaptor aims to add variant information to phoneme hidden sequences and it is composed of a duration predictor, a pitch predictor, and an energy predictor. Finally, the decoder generates the mel spectrogram from the hidden sequences expanded by the variance adaptor. Following the module in Tacotron [24], we add a Postnet (Conv1D blocks) module after the decoder to finetune the speech quality.
Table 1: COS, MOS and Sim-MOS with 95% confidence intervals for different speaker encoders.

| EXP.       | COS  | MOS  | SIM-MOS | COS  | MOS  | SIM-MOS | COS  | MOS  | SIM-MOS |
|------------|------|------|---------|------|------|---------|------|------|---------|
| Seen VCTK  |      |      |         | Unseen VCTK |      |      |         | Unseen LibriTTS |      |      |         |
| ground-truth | .826 | 4.10±0.08 | - | .800 | 4.55±0.05 | - | .832 | 4.21±0.10 | - |
| reconstruct | .853 | 4.08±0.08 | 4.65±0.05 | .821 | 4.47±0.06 | 4.61±0.08 | .884 | 4.18±0.10 | 4.83±0.03 |
| baseline    | .806 | 3.72±0.07 | 4.03±0.07 | - | - | - | - | - | - |
| x-vector    | .766 | 3.35±0.06 | 3.68±0.07 | .696 | 3.37±0.06 | 3.62±0.08 | .649 | 3.16±0.09 | 3.24±0.07 |
| deepspeaker | .807 | 3.62±0.06 | 3.94±0.06 | .711 | 3.75±0.08 | 3.83±0.06 | .648 | 3.84±0.07 | 3.50±0.07 |
| speakerNet  | .810 | 3.71±0.06 | 3.94±0.06 | .696 | 3.75±0.06 | 3.77±0.07 | .657 | 3.80±0.08 | 3.52±0.09 |
| ecapa       | .804 | 3.71±0.06 | 3.96±0.07 | .721 | 3.71±0.08 | 3.84±0.07 | .667 | 3.74±0.09 | 3.54±0.09 |

4. Experiments

4.1. Experimental Setup

We use two publicly available English datasets: VCTK [25] and LibriTTS [26]. VCTK corpus includes speech data uttered by 109 English speakers with various accents at 48kHz. Each speaker reads out about 400 sentences and about 44 hours of data in sum. LibriTTS consists of 585 hours of speech data at the 24kHz sampling rate from 2,456 speakers and the corresponding texts. In our experiments, all utterances are downsampled to 22050Hz and are used to extract 80 dimensional mel spectrograms.

We implement pretrained x-vector, Deep Speaker, SpeakerNet and ECAPA-TDNN as speaker encoder. After extraction, the dimensions of the speaker embeddings for x-vector, Deep Speaker, SpeakerNet and ECAPA-TDNN are 512, 512, 256 and 128 respectively.

For acoustic model, we follow the open-source FastSpeech2 implementation. Meanwhile, we use Montreal Forced Aligner [27] to get the ground-truth duration for each phoneme as additional inputs. We use the pretrained HiFi-GAN model as our vocoder to convert the 80 dimensional mel spectrograms to 22050Hz audio files. Our multi-speaker models are all trained for 400K steps with a batch size of 16 on one GeForce RTX 3090.

We use the following methods for evaluation:

1. ground-truth: the real utterances from the datasets.
2. reconstruct: directly convert the ground-truth mel spectrograms back to speech.
3. baseline: FastSpeech2 with look-up table.
4. x-vector: FastSpeech2 with pretrained x-vector.
5. deepspeaker: FastSpeech2 with pretrained DeepSpeaker.
6. speakerNet: FastSpeech2 with pretrained SpeakerNet.
7. ecapa: our proposed method by combining FastSpeech2 with pretrained ECAPA-TDNN.

To comprehensively evaluate our proposed model, we split the VCTK dataset for training and testing: 8 speakers are held as unseen speakers cloning test, and other 101 speakers are used to train and evaluate models for seen speakers. We use LibriTTS for the unseen speaker cloning test.

It should also be noted that the pretrained models will not be finetuned or adapted in unseen speakers in our experiments in order to evaluate the voice cloning ability of our proposed method.

4.2. Objective similarity evaluation

We use a third-party pretrained speaker encoder to evaluate the speaker similarity between the real speech and the synthesized speech. To evaluate how similar synthesized speech and real speech are, we make pairs for each synthesized utterance with a randomly selected real utterance from the same speaker. Then we use the pretrained speaker encoder to extract speaker embeddings for each utterance and compute the average cosine similarity for each pair as our similarity result.

The results of the objective Cosine Speaker Similarity (COS) test are shown in Table 1. It can be seen that in Seen VCTK test, all speaker encoders except x-vector have similar performance and they all match the baseline (usually having the best result in seen speaker). Nevertheless, in the unseen VCTK and unseen LibriTTS test, using pretrained ECAPA-TDNN model as the speaker encoder outperforms other speaker encoders, which suggests ECAPA-TDNN model can capture more speaker information and have more potential to be utilized in unseen voice cloning task.

4.3. Subjective evaluation

We conduct a Mean Opinion Score (MOS) test to evaluate the naturalness of the synthesized speech, and a Similarity Mean Opinion Score (Sim-MOS) test to evaluate the speaker similarity. In our test, we randomly select 20 utterances from each test set. In the MOS test, the listeners are given one synthesized utterance and one real utterance and are asked to give a rating score between 1 to 5 points. In the Sim-MOS test, the listeners are given both a real utterance and a synthesized utterance to evaluate the similarity by scoring between 1 to 5 points. Table 1 shows the MOS and Sim-MOS results.

In MOS test, it can be seen that in Seen VCTK test, all models have similar performance. However, using Deep Speaker or SpeakerNet as the speaker encoder has better naturalness and quality in unseen VCTK and unseen LibriTTS tests. In Sim-MOS test, the results are consistent with the objective similarity evaluation. These results suggest that combining ECAPA-TDNN speaker encoder with TTS synthesizer has the power to gain better speaker similarity in the multi-speaker task at the cost of speech quality.

4.4. Automatic MOS evaluation

To further evaluate the effectiveness of our proposed method, we use several automatic speech quality assessment models to assess our synthesized speech. We use three pretrained MOS prediction models in our experiment: MOSNet, MBNet, and
Table 2: Automatic MOS evaluation results for seen and unseen test sets by using three MOS prediction models.

| Models   | Seen VCTK | Unseen VCTK | Unseen LibriTTS |
|----------|-----------|-------------|-----------------|
| MOSNet   | 3.95      | 4.00        | 3.16            |
| MBNet    | 3.98      | 3.79        | 3.50            |
| S3PRL    | 3.53      | 3.52        | 3.46            |
| ground-truth | 3.95 | 4.00 | 3.16 |
| reconstruct | 3.67 | 3.79 | 3.52 |
| baseline  | 3.67      | 3.79        | 3.52            |
| x-vector  | 3.31      | 3.32        | 3.15            |
| deepspeaker | 3.61 | 3.53        | 3.48            |
| speakernet | 3.66 | 3.50        | 3.49            |
| ecapa     | 3.60      | 3.50        | 3.49            |

Table 3: Cosine similarity between different automatic MOS predictions and human-based subjective MOS results.

| Models   | Seen VCTK | Unseen VCTK | Unseen LibriTTS | ALL  |
|----------|-----------|-------------|-----------------|------|
| MOSNet   | 0.985     | 0.983       | 0.983           | 0.984|
| MBNet    | 0.989     | 0.987       | 0.985           | 0.986|
| S3PRL    | **0.998** | **0.994**   | **0.995**       | **0.996**|

Self-supervised Representation method (S3PRL). These models are pretrained on VCC 2016 [13] and VCC 2018 [14] datasets, which include lots of MOS rating scores evaluated by many participants. It should be noticed that the speech quality in the voice conversion task is less natural than that in speech synthesis task, and the rating scores in the VCC datasets are lower than the usual TTS score. We use the same MOS test set in subjective evaluation and the results of the objective MOS prediction test are shown in Table 2. It can be seen from the results that they are consistent with the results in subjective MOS evaluation.

4.5. Analysis

In order to investigate why our proposed method has higher speaker similarity in the multi-speaker TTS task, we visualize the speaker embeddings in Fig 2. We randomly select the speaker embeddings extracted from 200 utterances from 10 speakers and use t-SNE to reduce them into 2-dimension. It can be seen from the plot that all speaker encoder models can discern the utterance from the same speaker, while the distribution of ECAPA-TDNN is more continuous which suggests that it clusters each speaker but keeps the subtle speaker characteristics from different utterances spoken by the same speaker. This is helpful in multi-speaker synthesis, as its goal is different from speaker verification task. Previous study [28] suggest that the continuous distribution of speaker embeddings has better speaker similarity in the multi-speaker TTS task and our experiment results in similarity tests confirm this study. As a result, using ECAPA-TDNN as a speaker encoder can achieve better speaker similarity.

To analyze the consistency between human and automatic MOS predictors and compare different MOS predictors, we compute the cosine similarity between human subjective MOS results and MOS predictions in different test sets. The results in Table 3 show that S3PRL has the best consistency, and the predictions of these MOS predictors in seen VCTK are more reliable than in unseen scenes. We conclude that the use of these models can free us from the burden of collecting subjective evaluations while they lose their effectiveness in some circumstances. A more comprehensive MOS dataset may increase their robustness and accuracy.

Figure 2: Visualizations of different speaker embeddings. (a) x-vector, (b) Deep Speaker, (c) SpeakerNet, (d) ECAPA-TDNN

5. Conclusions

In order to improve the speaker similarity in multi-speaker text-to-speech synthesis, we propose our end-to-end method by introducing a more powerful speaker encoder based on the ECAPA-TDNN model derived from speaker verification task. We combine the independently pretrained ECAPA-TDNN model with a non-autoregressive acoustic model FastSpeech2. By transferring the knowledge learned from other datasets and applying the SOTA speaker verification model, our proposed model outperforms other methods in speaker similarity. Besides, to lighten the burden of subjective evaluation, we are the first to adopt automatic MOS predictors to assess our testing results and these models show great potential. For future work, we will continue to investigate the multi-speaker speech synthesis.

6. Acknowledgements

This work is supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (202200042) and New Talent Project of Beijing University of Posts and Telecommunications (2021RC37).
7. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Sherry-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.

[2] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, “Deep voice 3: Scaling text-to-speech with convolutional sequence learning,” arXiv preprint arXiv:1710.07654, 2017.

[3] 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,” arXiv preprint arXiv:2006.04558, 2020.

[4] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014, pp. 4052–4056.

[5] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 3329–3333.

[6] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, “Deep speaker: an end-to-end neural speaker embedding system,” arXiv preprint arXiv:1705.02304, 2017.

[7] J. Shen, R. Pang, R. J. Weiss, M. Schuster, 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.

[8] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, L. Lopez Moreno, Y. Wu et al., “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” Advances in neural information processing systems, vol. 31, 2018.

[9] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: a large-scale speaker identification dataset,” in Interspeech, 2017.

[10] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in Interspeech, 2018.

[11] B. Desplanques, J. Thienpondt, and K. Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” arXiv preprint arXiv:2005.07143, 2020.

[12] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, “Ecapa-tdnn embeddings for speaker diarization,” arXiv preprint arXiv:2104.04146, 2021.

[13] N. R. Koluguri, J. Li, V. Lavrukhin, and B. Ginsburg, “Speaker-ernet: 1d depth-wise separable convolutional network for text-independent speaker recognition and verification,” arXiv preprint arXiv:2010.12653, 2020.

[14] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, “The voice conversion challenge 2016.” in Interspeech, 2016, pp. 1632–1636.

[15] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, “The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods,” arXiv preprint arXiv:1804.04262, 2018.

[16] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, “Mositnet: Deep learning based objective assessment for voice conversion,” arXiv preprint arXiv:1904.08352, 2019.

[17] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, “Mbit: Mos prediction for synthesized speech with mean-bias-ness,” in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 391–395.

[18] W.-C. Tseng, C.-y. Huang, W.-T. Kao, Y. Y. Lin, and H.-y. Lee, “Utilizing self-supervised representations for mos prediction,” arXiv preprint arXiv:2104.03017, 2021.

[19] 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). IEEE, 2020, pp. 6184–6188.

[20] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, “Deep voice 2: Multi-speaker neural text-to-speech,” Advances in neural information processing systems, vol. 30, 2017.

[21] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentional statistics pooling for deep speaker embedding,” arXiv preprint arXiv:1803.10963, 2018.

[22] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, “Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6124–6128.

[23] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[24] 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.

[25] J. Yamagishi, C. Veaux, and K. MacDonald, “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit,” 2019.

[26] H. Zeng, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “Libritts: A corpus derived from librispeech for text-to-speech,” arXiv preprint arXiv:1904.02882, 2019.

[27] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal forced aligner: Trainable text-speech alignment using kaldi,” in Interspeech, vol. 2017, 2017, pp. 498–502.

[28] C.-M. Chien, J.-H. Lin, C.-y. Huang, P.-c. Hsu, and H.-y. Lee, “Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech,” in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 8588–8592.