

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

Jinlong Xue<sup>1</sup>, Yayue Deng<sup>1,2</sup>, Yichen Han<sup>1</sup>, Ya Li<sup>1,\*</sup>, Jianqing Sun<sup>3</sup>, Jiaen Liang<sup>3</sup>

<sup>1</sup>School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China

<sup>2</sup>Beijing Language and Culture University, Beijing, China

<sup>3</sup>Unisound AI Technology Co., Ltd, Beijing, China

jinlong\_xue@bupt.edu.cn, dengyayue.stu@gmail.com, adelacvgairo@bupt.edu.cn, yli01@bupt.edu.cn, sunjianqing@unisound.com, liangjiaen@unisound.com

## 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 high-quality speech and 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 naturalness and similarity. To efficiently evaluate our synthesized speech, we are the first to adopt 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, few-shot, 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 synthesized 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 in terms of various speaking styles in multi-speaker task still needs more research. On the other hand, the models of the few-shot voice cloning in unseen speakers circumstance by using a speaker encoder usually tend to synthesize neutral and poor quality 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]. Besides, pretrained models are more of

ten used instead of jointly training with the TTS system, for the speaker knowledge for speaker encoder is limited by training dataset in the latter case. Jia et al. [6] investigated the knowledge transfer where the speaker verification model is trained on a dataset with many speakers, like VoxCeleb [7, 8] 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. Especially the combination of x-vector [5] and TTS system achieves promising results [9].

However, the naturalness and 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 features 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 and investigate their generalization ability in two publicly available datasets for both the seen and unseen tests. Our method outstands other methods in both naturalness and 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 [12, 13], several deep-learning-based automatic speech quality evaluation methods have been proposed, such as MOSNet [14], MBNet [15], a self-supervised representation based MOS predictor model (denoted as S3PRL)[16]. To our best knowledge, we are the first to evaluate the synthesized speech by using the MOS prediction model to accelerate our research. We utilize these methods and compare the MOS score results. 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 page<sup>1</sup>.

\* Ya Li is the corresponding author.

<sup>1</sup>Audio samples: <https://happylittlecat2333.github.io/interspeech2022>

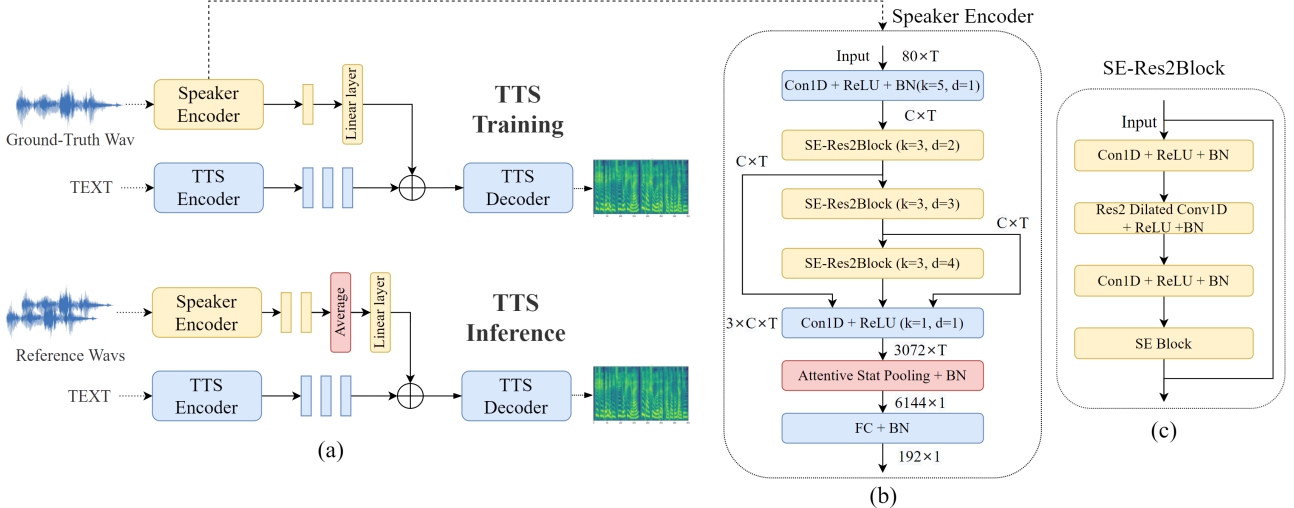


Figure 1: The illustration of our proposed ECAPA-TDNN based multi-speaker speech synthesis model. (a) depicts the training and inference workflows in our experiments. (b) shows the whole network of ECAPA-TDNN.  $K$  stands for kernel size and  $d$  for dilation.  $C$  and  $T$  denote channels and temporal dimensions. (c) is the detailed illustration of SE-Res2Block in the ECAPA-TDNN model.

## 2. Speaker Representations

Multi-speaker TTS system highly depends on speaker representation for conditioning the acoustic model to clone the voices from 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 [7, 8] recorded by many speakers, which provides them the ability to capture the subtle characteristics and styles from different speakers by using only a short utterance under any circumstances.

In speaker verification task, deep neural network models have surpassed the classic model i-vector [17]. Among them, d-vector [4], x-vector [5] are representative methods, and they have been used in multi-speaker TTS [6, 9]. These combinations show great potential and some methods have been widely used [5]. Below, we describe each of these speaker encoder models.

### 2.1. D-vector

The conventional d-vector [4] model uses a DNN architecture as a speaker feature extractor operating at the frame level. After sending the frames through the DNN network, d-vector is obtained by element-wise averaging the frame-level outputs from the last hidden layer of DNN network.

### 2.2. X-vector

In the original x-vector paper [5], a time-delay neural network (TDNN) with sub-sampling is used as the encoder network. An attentive statistics pooling (ASP) [18] layer aggregates all frame-level outputs from the last encoder layer and computes its mean and standard deviation. After sending through segment-level layer (fully-connected layer), the x-vector speaker embedding is obtained. The combination of a TDNN network with dilation has three advantages: reduction of the total number of independent connection parameters, invariance under shifts in times, and larger contextual vision. The use of ASP allows the model to select frames that are indeed relevant to speaker char-

acteristics.

## 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 also uses TDNN as its base architecture but introduces multiple enhancements: using Squeeze-and-Excitation (SE) blocks [19] 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 TDNN based 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 variant adaptor. The encoder generates the hidden embedding from a sequence of phoneme-level inputs. The variant 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 variant adaptor. Following the module in Tacotron [20], we add a Postnet (Conv1D blocks) module after the decoder to finetune the speech quality.

### 3.3. TTS training and inference

At the training stage, we use the speaker encoder model pre-trained on speaker verification task to extract fixed-dimensional embeddings from each utterance for speaker representations. After that, the utterance-level speaker representations are projected to match the dimension of the output from the encoder of acoustic model with one linear layer, and they are expanded and added to the output. Therefore, speaker information is transferred to the synthesizer and the variant adaptor in FastSpeech 2 can be conditioned on speaker information. At the inference stage, we extract a speaker representation for each utterance and compute an average representation on behalf of this speaker. The other process is the same as the training stage.

## 4. Experiments

### 4.1. Experimental Setup

We use two publicly available English datasets: VCTK [21] and LibriTTS [22]. VCTK corpus includes speech data uttered by 109 English speakers with various accents at 48 kHz. 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 24 kHz sampling rate from 2,456 speakers and the corresponding texts. In our experiments, all utterances are down-sampled to 22050 Hz and are used to extract 80 dimensional mel spectrograms.

We implement pretrained x-vector<sup>2</sup>, ECAPA-TDNN<sup>3</sup> as speaker encoder. The x-vector and ECAPA-TDNN are both pre-trained on Voxceleb 1 and Voxceleb 2 [7, 8], and the EER results on Vox1-test are 2.82% and 0.69%. For preprocessing, each utterance is resampled to 16 kHz but converted to 30 dimensional MFCC features for x-vector and 80 for ECAPA-TDNN. After extraction, the dimensions of the speaker embeddings for x-vector and ECAPA-TDNN are 512 and 128 respectively.

For the acoustic model, we follow the open-source FastSpeech2 implementation<sup>4</sup>. We use Montreal Force Aligner [23] to get the ground-truth duration for each phoneme as additional inputs. We use the pretrained HiFi-GAN<sup>5</sup> model as our vocoder to convert the 80 dimensional mel spectrograms to 22050 Hz 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: FastSpeech 2 with look-up table.
4. x-vector: FastSpeech 2 with pretrained x-vector speaker encoder.
5. ecapa: our proposed method by combining FastSpeech 2 with pretrained ECAPA-TDNN speaker encoder.

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.

During testing, we use the average speaker embeddings extracted from all the utterances of the same speaker instead of

from only one utterance, because using the averaged embedding can be more stable and have better similarity in our experiment. It should also be noted that the pretrained models will not be finetuned and be adapted in unseen speakers in our experiments 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 speaker similarity test are shown in Table 1. It can be seen that using the pretrained ECAPA-TDNN model as speaker encoder outperforms x-vector in both seen speaker test and unseen speaker test in VCTK, even in unseen speaker test set from LibriTTS. Moreover, the proposed model matches the baseline (usually having best results in seen speaker) in seen speaker test on VCTK. For unseen speaker test on VCTK and LibriTTS, using a simple lookup table cannot clone unseen speaker voices, and our proposed method has better similarity than the x-vector model, which suggests ECAPA-TDNN model can capture more speaker information and have potential to utilize in multi-speaker task than other speaker encoder models.

### 4.3. Subjective evaluation

We conduct a mean opinion score (MOS) test to evaluate the naturalness and speaker similarity of the synthesized speech. In our test, we randomly select 20 utterances from each test set. In the quality MOS test, the listeners are given one synthesized utterance and are asked to give a rating score between 1 to 5 points for speech quality. In the speaker similarity 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 2 shows the MOS results and the speaker similarity results.

In naturalness test, it can be seen that our proposed method outperforms the x-vector model in all test sets including unseen speaker tests. In similarity test, it shows that the speech synthesized by our model also has better similarity than other methods. The results are consistent with the objective similarity evaluation. These results suggest that combining the ECAPA-TDNN model and acoustic models has the power to gain better speech naturalness and speaker similarity in the multi-speaker task.

### 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 Self-supervised Representation method (S3PRL). These models are pretrained on VCC 2016 [12] and VCC 2018 [13] 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 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 3.

<sup>2</sup>[https://github.com/manojpamk/pytorch\\_xvectors](https://github.com/manojpamk/pytorch_xvectors)

<sup>3</sup><https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb>

<sup>4</sup><https://github.com/ming024/FastSpeech2>

<sup>5</sup><https://github.com/jik876/hifi-gan>

Table 1: Objective speaker similarity for different test sets and different types of speaker encoders.

method	Seen VCTK	Unseen VCTK	Unseen LibriTTS
ground-truth	.979	.975	.986
reconstruct	.976	.972	.984
baseline	<b>.968</b>	-	-
x-vector	.963	.954	.956
ecapa	.967	<b>.959</b>	<b>.959</b>

Table 2: The results of the subjective MOS tests for naturalness and speaker similarity.

Model	method	Seen VCTK	Unseen VCTK	Unseen LibriTTS
MOS	ground-truth	4.19	4.20	4.21
	reconstruct	4.09	4.11	4.10
	baseline	<b>3.70</b>	-	-
	x-vector	3.51	3.51	3.38
	ecapa	3.62	<b>3.62</b>	<b>3.47</b>
	Similarity	reconstruct	4.65	4.69
baseline		3.89	-	-
x-vector		3.71	3.65	3.08
ecapa		<b>3.93</b>	<b>3.66</b>	<b>3.18</b>

It can be seen from the results that the proposed method outperforms the x-vector model and achieves comparable speech quality to the baseline in seen VCTK test, which is consistent with the results in subjective MOS evaluation. Besides, our proposed model also outperforms the x-vector model in unseen VCTK or unseen LibriTTS test.

#### 4.5. Analysis

In order to investigate why our proposed method has better performance 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 both the ECAPA-TDNN model and the x-vector model 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 keeping 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 studies [24] suggest that the continuous distribution of speaker embeddings has better performance in the multi-speaker TTS task. Our experiment results in similarity tests confirm these studies [24]. As a result, using ECAPA-TDNN as a speaker encoder can achieve better speech naturalness and speaker similarity.

By analyzing the results obtained from humans and MOS predictors, we find that assessments from automatic MOS predictors are consistent with evaluations from subjective method, which can both reflect the quality of the synthesized speech in the seen and unseen VCTK test set. The use of these models can free us from the burden of collecting subjective evaluations. Meanwhile, we also see that the currently MOS prediction models lose their effectiveness in some circumstances like

Table 3: Automatic MOS evaluation results for seen and unseen test sets by using three MOS prediction models.

Model	method	Seen VCTK	Unseen VCTK	Unseen LibriTTS
MOSNet	ground-truth	4.16	3.91	3.40
	reconstruct	3.75	3.83	3.34
	baseline	<b>3.32</b>	-	-
	x-vector	3.11	3.44	3.15
	ecapa	3.16	<b>3.52</b>	<b>3.42</b>
MBNet	ground-truth	3.86	3.99	3.05
	reconstruct	3.45	3.81	2.99
	baseline	<b>3.37</b>	-	-
	x-vector	3.07	3.46	3.21
	ecapa	3.35	<b>3.55</b>	<b>3.53</b>
S3PRL	ground-truth	3.53	3.53	3.45
	reconstruct	3.44	3.47	3.37
	baseline	<b>3.45</b>	-	-
	x-vector	3.27	3.42	3.36
	ecapa	3.44	<b>3.52</b>	<b>3.48</b>

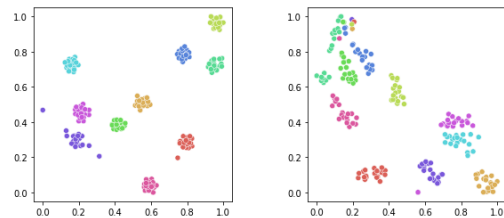


Figure 2: Visualizations of different speaker embeddings. Left: x-vector Right: ECAPA-TDNN

unseen LibriTTS, and a more comprehensive MOS dataset may increase their robustness and accuracy.

## 5. Conclusions

In order to improve the naturalness and 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 both speech naturalness and 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 performance of few-shot 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. 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.
- [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. Variiani, 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. 5329–5333.
- [6] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. 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.
- [7] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: a large-scale speaker identification dataset,” in *Interspeech*, 2017.
- [8] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in *Interspeech*, 2018.
- [9] 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.
- [10] B. Desplanques, J. Thienpondt, and K. Demuyne, “Ecapadnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” *arXiv preprint arXiv:2005.07143*, 2020.
- [11] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, “Ecapadnn embeddings for speaker diarization,” *arXiv preprint arXiv:2104.01466*, 2021.
- [12] 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.
- [13] 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.
- [14] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, “Mosnet: Deep learning based objective assessment for voice conversion,” *arXiv preprint arXiv:1904.08352*, 2019.
- [15] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, “Mbnnet: Mos prediction for synthesized speech with mean-bias network,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 391–395.
- [16] 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.
- [17] N. S. Ibrahim and D. A. Ramli, “I-vector extraction for speaker recognition based on dimensionality reduction,” *Procedia Computer Science*, vol. 126, pp. 1534–1540, 2018.
- [18] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive statistics pooling for deep speaker embedding,” *arXiv preprint arXiv:1803.10963*, 2018.
- [19] 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.
- [20] 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.
- [21] J. Yamagishi, C. Veaux, and K. MacDonald, “Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit,” 2019.
- [22] H. Zen, 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.
- [23] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal forced aligner: Trainable text-speech alignment using kaldii,” in *Interspeech*, vol. 2017, 2017, pp. 498–502.
- [24] 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.