THE HUYA MULTI-SPEAKER AND MULTI-STYLE SPEECH SYNTHESIS SYSTEM FOR M2VOC CHALLENGE 2021

Jie Wang[1,2], *Yuren You*[1], *Feng Liu*[1], *Deyi Tuo*[1], *Shiyin Kang*[1,*], *Zhiyong Wu*[2,3] , *Helen Meng*[2,3]

[1] Huya Inc, Guangzhou, China
[2] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[3] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong SAR, China

jie-wang19@mails.tsinghua.edu.cn, {youyuren, liufeng1, tuodeyi, kangshiyin}@huya.com, {zywu, hmmeng}@se.cuhk.edu.hk

ABSTRACT

Text-to-speech systems now can generate speech that is hard to distinguish from human speech. In this paper, we propose the Huya multi-speaker and multi-style speech synthesis system which is based on DurIAN and HiFi-GAN to generate high-fidelity speech even under low-resource condition. We use the fine-grained linguistic representation which leverages the similarity in pronunciation between different languages and promotes the speech quality of code-switch speech synthesis. Our TTS system uses the HiFi-GAN as the neural vocoder which has higher synthesis stability for unseen speakers and can generate higher quality speech with noisy training data than WaveRNN in the challenge tasks. The model is trained on the datasets released by the organizer as well as CMU-ARCTIC, AIShell-1 and THCHS-30 as the external datasets and the results were evaluated by the organizer. We participated in all four tracks and three of them entered high score lists. The evaluation results show that our system outperforms the majority of all participating teams.

Index Terms— multi-speaker and multi-style TTS, low-resource condition, DurIAN, HiFi-GAN

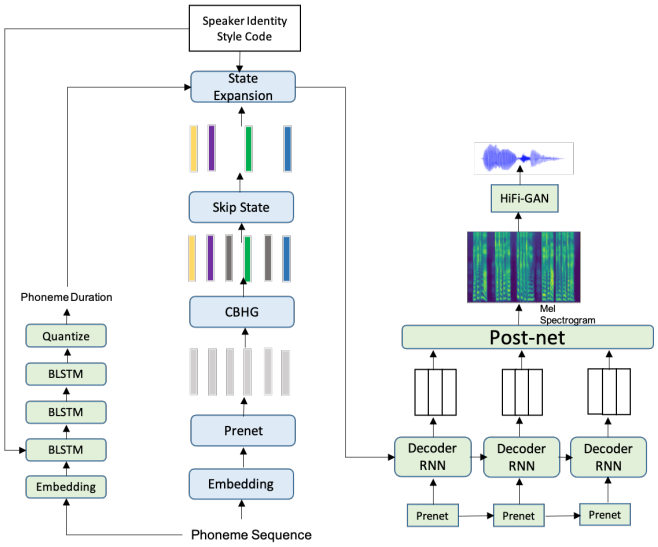


Fig. 1. The Huya TTS system.

1. INTRODUCTION

Text-to-Speech(TTS) aims to convert text to speech which plays an important role in many fields such as voice assistants, audio books and spoken dialog system. Recently, with the rapid developments of acoustic model and neural vocoder technology, TTS systems can generate natural speech for speaker who has a large amount of high quality speech [1, 2, 3]. However, a large amount of speech from single speaker is not always available in low-resource real-world conditions. There is an urgent need to enable stylization and personalization in multi-speaker TTS [4, 5, 6] which can achieve rich prosody control. However, the new speaker data is scarce and the recording condition is usually poor. The production of a multi-speaker corpus is expensive and different speakers may have different recording conditions. The training corpus is difficult to cover abundant speaker distribution. All these limitations restrain the application scenarios of TTS technology.

There have been studies on few-shot or one-shot learning trying to improve the robustness of cloning timbre from unseen speakers with few samples. The voice cloning systems can be roughly divided into two categories according to different ways of mimicking target timbre. The first one is speaker adaptation [7]. It trains a multi-speaker synthesis model conditioned on speaker embedding and the model is fine-tuned with samples from unseen speakers. The fine-tuning can be conducted on the whole synthesis model or only the speaker embedding part [8, 9]. The VoiceLoop [8] fine-tunes the speaker embedding which the speakers are represented by a vector. Global style token(GST) is introduced to represent the target style [9]. The second category is speaker encoding [10, 11]. The core idea is to transfer the knowledge of speaker information from pre-trained speaker verification model [12, 13], which enables the TTS system to learn timbre from different speakers.

In this paper, we proposed the Huya TTS system which can achieve multi-speaker and multi-style speech synthesis under low-resource condition. Contributions of this paper are as follows: Firstly, we model a multi-speaker and multi-style TTS system based on DurIAN [14] and used the fine-grained linguistic representation as the text input to improve the naturalness and intelligibility of speech synthesis. Secondly, we combined the HiFi-GAN vocoder [15] along with DurIAN to generate highly quality and natural audio

* Corresponding author.

which outperforms WaveRNN vocoder [16]. The evaluation results show that our system can effectively clone timbre and style for target speakers. In the final results of the M2VoC Challenge 2021 [17], our system entered the high score lists of track 1B, track 2A and track 2B with ranking 5^{th}, 6^{th} and 4^{th} respectively. Only the top 4$\sim$7 teams can be included in the high score lists.

This paper is presented as follows: Section 2 describes the tasks in M2VoC Challenge 2021. Section 3 introduces our implemented systems for the two tracks and two sub-tracks in detail. Section 4 introduces the experiments and analyses the evaluation results. Section 5 gives the conclusion and future research.

2. THE TASKS IN M2VOC 2021

There are two tracks in the challenge including track 1, few-shots track and track 2, one-shot track. Speakers have different speaking styles and 100 samples are for track 1 and only 5 samples are for track 2. Each track contains two sub-tracks including sub-track A and sub-track B. For the sub-track A, the building of the TTS system is strictly limited to the data released by the organizer. Fot the sub-track B, any data set publicly available can be used to build the system. We participated in all the four tracks.

3. HUYA SPEECH SYNTHESIS SYSTEM

3.1. Data processing

3.1.1. Audio Signal Preprocessing

The challenge organizer released four audio/text datasets including multi-speaker training speech data (MST), target speaker validation speech set (TSV), target speaker testing speech set (TST) and test text set (TT) . The test text set is for final submit and evaluation. The multi-speaker training speech data (MST) contains MST-Originbeat subset and MST-AIShell [18]. The MST-Originbeat subset includes 6.38 hours of high-quality speech from two Mandarin Chinese speakers (one male and one female). Each speaker has 5000 utterances recorded by high-quality microphone in a recording studio. We included the corpus of 174 speakers of the MST-AIShell corpus in the training dataset. The MST-AIShell is nosiy dataset, so we used the RNNoise[1] and Weighted Prediction Error(WPE) [19] to do the denoise and dereverberation respectively. After the signal preprocessing, the voice quality improves which is evaluated by a pretrained MOSNet [20] shown in the Table 1. Besides the dataset released by the organizers, we used the CMU-ARCTIC corpus [21], THCHS [22] and AIShell-1 [23] in track B. The CMU-ARCTIC is an English corpus and others are Chinese corpus.

Table 1. MOSNet scores of original datasets and processed datasets.

	Original	**Processed**
MOS	2.97	2.99

3.1.2. Linguistic representation

We adopted the fine-grained linguistic representation to better utilize the bilingual training corpus and to improve the naturalness of speech synthesis. The fine-grained linguistic representation is composed of pronunciation feature and prosodic structure feature.

[1] https://github.com/xiph/rnnoise

As studied in [24, 25, 26], the phoneme-based TTS models perform significantly better than char- or byte-based variants for Mandarin. Phoneme input eases the learning difficulty of the network to extract linguistic information from the input text. In this research, we adopt the phonemes with tones as the pronunciation feature. The tone symbols are attached to the phoneme symbols. Different from other previous works, we use a smaller dictionary to construct the universal pronunciation space. Our phoneme set can deal with the Chinese and English corpus as it leverages the similarity in pronunciation between different languages. Furthermore, we also takes more characteristics of Mandarin pronunciation into consideration. Phonemes with different initial and final mouth patterns (such as front and back nasals) as well as other unique pronunciation details in Mandarin are both taken into consideration. The articulation coordination is denoted as additional tonal symbols.

The prosodic structure feature is composed of hierarchical prosodic boundary and sentence types. We adopted the common 4-level-based Mandarin prosody standard including prosodic word (PW), prosodic phrase (PP), intonational phrase (IP) and utterance (UTT). We categorize the sentences according to punctuation marks like comma, semicolon, and question mark etc. However, the MST-AIShell and MST-Originbeat corpus are annotated with different prosodic level mark standards. The MST-AIShell train set is annotated with 3-level prosody label while the MST-Originbeat is annotated with 5-level prosody label. Hence, we designed different prosodic mark mapping rules to normalize MST-AIShell and MST-Originbeat corpus. For example, the second and third prosody level symbols in MST-Originbeat are viewed as the same prosodic phrase (PP) in the 4-level prosody structure.

We construct the pronunciation space by the fine-grained linguistic modeling which can reap the full potential rewards of similarities between Chinese and English. The fine-grained linguistic representation boosts the performance of speech synthesis concerning the quality and naturalness while incorporating the English corpus. The results are demonstrated in section 4.1.

3.2. Acoustic model

The overall architecture of Huya TTS system is shown in Figure 1, which is based on DurIAN [14] and incorporates speaker identity information as well as style code to achieve a multi-speaker and multi-style TTS system. We utilized duration informed attention network(DurIAN) as the base acoustic model which is robust to missing or skipping problems. We found that the joint training of the acoustic model and duration prediction model causes the unnatural prosody synthesis results which is adopted by the original DurIAN [14]. In our TTS system, the training process of acoustic model and phoneme duration model are decoupled which means that the two model are trained separately.

We adopted look-up table to model the speaker information. By assigning an ID to each speaker, the corresponding representation for each speaker can be looked up from a trainable embedding table. The style representation is also achieved by a trainable look-up table. The speaker embedding and style code are concatenated to encoder output in all steps which will be decoded later.

For track1 and track2, all audios are sampled to 24kHz with mono-channel. Features are extracted with 50ms window size and 10ms shift size. We will introduce the training details of the acoustic models of track1 and track2 respectively. In track 1, we also consider gender differences besides the speaker identity, and we assigned gender-specific speaker ID to speakers such as 1 denoting male and 2 denoting female. For track 1A, we used the MST-Originbeat to

train the model at first. Then we used the 100 samples of the target speaker (TST) to fine-tune the base-model and the speaker code is assigned according to the target speaker gender. For track 2A, we used MST, TSV and TST to train the multi-speaker acoustic model. Then, the different speaker-IDs were assigned to different speakers respectively. The training procedure of sub-track B is similar to sub-track A. However, we observed that some English words are also contained in the test set. In order to improve the naturalness and intelligibility of code-switch synthesis, we added CMU-ARCTIC database [21] to the joint training of the multi-speaker acoustic model in sub-track 1B and sub-track 2B.

Furthermore, we also tried to mix all the corpus together to train the base-model without distinguishing different speakers. Then we used the samples of the target speaker to fine tune the base-model. Whereas, we found that the base model trained with differentiated speaker-ID is better in speaker similarity and naturalness.

3.3. Neural Vocoder

The noisy datasets and the limited amount samples of the target speakers are the two main challenges for the vocoder training. We compared several vocoders including HiFi-GAN [15] and WaveRNN [16]. Finally, we considered HiFi-GAN as our neural vocoder for the stable synthesis performance in this challenge tasks.

The architecture of WaveRNN is similar to [16], and the architecture of HiFi-GAN is similar to [15]. In order to gain a higher synthesis quality, parameters of the two vocoders were carefully adjusted including decreasing the learning rate, increasing the batch size and increasing the sizes of CNN layers according to the dataset. The long silent segments of the training data were removed using the energy-based VAD which adjusts thresholds according to different audio data. In order to stabilize the training process and improve model performance, we added some noise into the training data. The noise was added to the audio signal or the mel spectrogram. The experimental results showed that the vocoders perform better while directly adding noise in the audio signal in the time domain. In addition, we adopted μ-law to carry out nonlinear transformation of the training audio which can make the model have higher resolution near zero. The above processing methods were applied in the training process of the two vocoders.

For track 1A and track 2A, we used MST-AIShell and MST-Originbeat for vocoder training. For track 1B and track 2B, we additively incorporated THCHS-30 [22] and AIShell-1 [23] in the traning datasets. The TST data was used as the development set to tune the models. We used early stopping to prevent model overfitting. Furthermore, we also tried to add the development set to the training data, but there was no significant improvement due to the small amount of data.

We compared the performance of WaveRNN and HiFi-GAN mentioned above under the same training and testing conditions. For these challenge tasks, the WaveRNN vocoder can generate high expressive speech within the test set. Nevertheless, it is prone to cause bad case as the errors are accumulated by the RNN architecture [16]. By comparison, the HiFi-GAN vocoder performed better while facing the noisy data and can generate high-quality speech stably. In summary, we chose HiFi-GAN as our neural vocoder in all tasks to achieve the high-quality synthesis results.

4. RESULTS

There are 24 participating systems numbered as T01-T24 and two baseline systems denoted as B01 and B02 in the challenge. Our submitted system is annotated as T24. There are three evaluation criteria: speech quality, speaker similarity and style similarity. Our system ranks 5^{th}, 6^{th} and 4^{th} seperately in track 1B, track 2A and track 2B which are included in the high score lists. Below are the detailed evaluation results. We also conduct experiments to demonstrate the effect of the proposed fine-grained linguistic representation.

4.1. The effect of fine-grained linguistic representation

We compared the coverage capacity of the two different pronunciation dictionaries, the traditional one and our proposed. Here, the traditional one refers to the initials and finals representation. The coverage capacity of pronunciation dictionary is defined as the percentage of how many phonemes the target speaker test (TST) contains when one certain pronunciation dictionary is applied. In order to simplify the explanation, we evaluate the coverage inability which is demonstrated as:

$$C = 1 - \frac{n}{N} \tag{1}$$

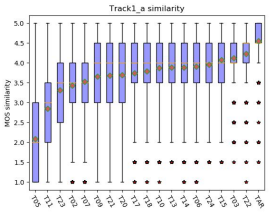
where n is the number of phonemes included in the TST and N is the size of pronunciation dictionary, C reflects the coverage inability of the dictionary.

The size of the traditional pronunciation dictionary is 182 and the size of the proposed is 84. The statistical results are shown in Table 2. The "Chat", "Game" and "Story" in Table 2 denoted the three target speakers in the track 1 and "S3", "S4" and "S5" denoted the other three target speakers in the track 2. As shown in Table 2, the percentage of proposed is small means that the dictionary can model most pronunciation contexts which promise the TTS model a stronger generalization ability. By comparison, the higher percentage of the traditional inputs means that it is quite possible that a lot of audios could not synthesize well when the input is denoted as initials and finals only. The proposed input representation system used a smaller number of phonemes to represent the entire pronunciation space. In the low-resource condition where only limited samples of target speakers are available, the new phoneme system has better coverage.

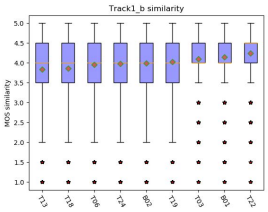
Table 2. The coverage inability C (the lower the better) of two pronunciation dictionaries. "Traditional" refers to the initials and finals linguistic representation. "Proposed" refers to our fine-grained linguistic representation.

	Chat	Game	Story	S3	S4	S5
Traditional	0.11	0.09	0.04	0.64	0.66	0.41
Proposed	0.04	0.04	0.02	0.38	0.37	0.20

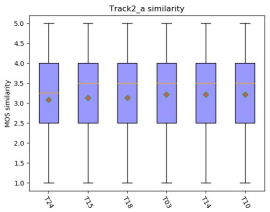
We also explored how the fine-grained linguistic representation endows the TTS systems with stronger modeling ability. In the Table 3, the monolingual corpus refers to the TTS system where only the MST-Originbeat was used as the training dataset and the bilingual corpus refers to that the MST-Originbeat and CMU-ARCTIC were both used as the training dataset. We calculated the L1 loss between the predicted mel spectrogram and ground-truth mel spectrogram. It is clearly shown that compared with the training with only Chinese corpus, the loss is lower when English corpus is incorporated in the training dataset. The linguistic representation can predict the mel spectrogram more accurately with the help of English dataset. We also have done the MOS test to evaluate the intelligibility. For a fair comparison, we synthesized 20 sentences separately for each model but with the same content. There are 9 listeners are involved in the subjective test and the results are shown in Table



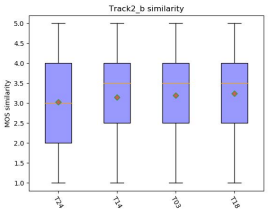
(a) first round track1A similarity



(b) combined track1B similarity

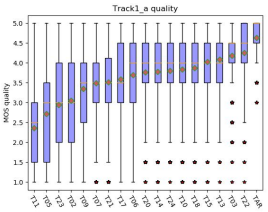


(c) combined track2A similarity

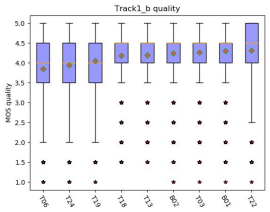


(d) combined track2B similarity

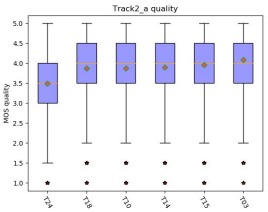
Fig. 2. Boxplot of Speaker Similarity.



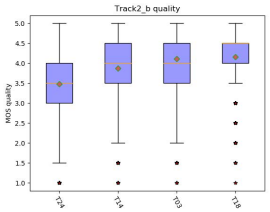
(a) first round track1A quality



(b) combined track1B quality



(c)combined track2A quality



(d)combined track2B quality

Fig. 3. Boxplot of Speech Quality.

3. It shows that our fine-grained linguistic representation can utilize multi-lingual datasets to improve the intelligibility significantly.

Table 3. Effect of fine-grained linguistic representation on mel spectrogram prediction loss and MOS scores.

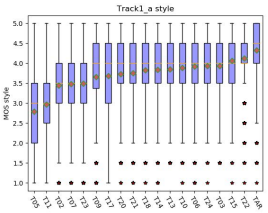
	Monolingual Corpus	Bilingual Corpus
Mel-loss($\times 10^{-2}$)	3.062	2.813
MOS	3.667	3.750

4.2. Speaker similarity

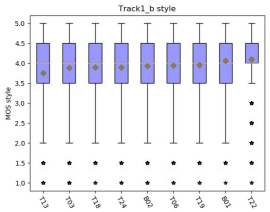
The evaluation results of speaker similarity is shown in Figure 2. The speaker similarity rankings of track2A is 4^{th} showing that our system can synthesize speech which can confuse the human with the groundtruth. For track 2B of one-shot open-set problem, our system ranks 4^{th} which is notable and comparable to the performance of few-shot scenario. We used gender-related speaker ID to alleviate the problem that the data is sparse when only limited numbers of speakers are available. This strategy contributes to the high speaker similarity score. Furthermore, as the speaker information is explicitly provided, the content encoder tends to learn the speaker-independent information which mitigates the speaker information leakage into the linguistic representation.

4.3. Speech quality

The speech quality scores are shown in Figure 3. Multiple systems' median scores are relatively close in the first round as shown in Figure 3(a). The rankings of our systems of track B are higher than that of track A which means that our system can synthesize higher quality speech after we added the external datasets and contributes to the code-switch synthesis. However, the speech quality score is relatively lower than speaker similarity and style similarity. We think a robust vocoder might yield better performance.



(a) first round track1A style



(b) combined track1B style

Fig. 4. Boxplot of Style Similarity.

4.4. Style similarity

As shown in Figure 4, the ranking of our system for track 1 is 4^{th} which is higher than the average level. It demonstrated that our system can clone the target style very similarly which outperforms the other systems. The original DurIAN acoustic model and duration model adopted joint training. We trained the acoustic model and duration model separately for decoupling. The decoupling training ensures a higher accuracy of duration information prediction. Thus, the duration prediction model can estimate the prosody more close to the ground truth and more natural. Besides, we trained one model for one style respectively which means that the TTS model has a higher modeling ability for the certain style.

5. CONCLUSION

This paper presents the details of our submitted systems and the results in the M2VoC Challenge 2021. We built multi-speaker and multi-style voice cloning systems with DurIAN as the acoustic model and HiFi-GAN as the neural vocoder. We used the fine-grained phoneme as the linguistic representation. We also compared the performance of WaveRNN vocoder and HiFi-GAN vocoder in this challenging task. Experimental results showed that our system has good performance and higher robustness which is good for cloning the target style under the low-resource conditions. In the future, we will incorporate a more robust vocoder to improve the performance of speech synthesis.

6. REFERENCES

[1] Stanton D Wang Y, Skerry-Ryan R J et al., "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[2] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[3] Kainan Peng, Wei Ping, Zhao Song, and Kexin Zhao, "Parallel neural text-to-speech," *arXiv preprint arXiv:1905.08459*, 2019.

[4] Y. Fan, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for dnn-based tts synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4475–4479.

[5] Jihyun Park, Kexin Zhao, Kainan Peng, and Wei Ping, "Multi-speaker end-to-end speech synthesis," *arXiv preprint arXiv:1907.04462*, 2019.

[6] David Álvarez, Santiago Pascual, and Antonio Bonafonte, "Problem-agnostic speech embeddings for multi-speaker text-to-speech with samplernn," *arXiv preprint arXiv:1906.00733*, 2019.

[7] Yuchen Fan, Yao Qian, Frank K Soong, and Lei He, "Multi-speaker modeling and speaker adaptation for dnn-based tts synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4475–4479.

[8] Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani, "Voiceloop: Voice fitting and synthesis via a phonological loop," *arXiv preprint arXiv:1707.06588*, 2017.

[9] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif 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.

[10] Sercan O Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou, "Neural voice cloning with a few samples," *arXiv preprint arXiv:1802.06006*, 2018.

[11] Hui Lu, Zhiyong Wu, Dongyang Dai, Runnan Li, Shiyin Kang, Jia Jia, and Helen Meng, "One-shot voice conversion with global speaker embeddings.," in *INTERSPEECH*, 2019, pp. 669–673.

[12] Ye Jia, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *arXiv preprint arXiv:1806.04558*, 2018.

[13] Eliya Nachmani, Adam Polyak, Yaniv Taigman, and Lior Wolf, "Fitting new speakers based on a short untranscribed sample," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3683–3691.

[14] Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei, Dan Su, and Dong Yu, "DurIAN: Duration Informed Attention Network for Speech Synthesis," in *Proc. Interspeech 2020*, pp. 2027–2031.

[15] Kong Jungil, Kim Jaehyeon, and Bae Jaekyoung, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *arXiv preprint arXiv:2010.05646v2*, 2020.

[16] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

[17] Qicong Xie, Xiaohai Tian, Guanghou Liu, Kun Song, Lei Xie, Zhiyong Wu, Hai Li, Song Shi, Haizhou Li, Fen Hong, Hui Bu, and Xin Xu, "The multi-speaker multi-style voice cloning challenge 2021," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

[18] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," *arXiv preprint arXiv:2010.11567*, 2020.

[19] Lukas Drude, Jahn Heymann, Christoph Boeddeker, and Reinhold Haeb-Umbach, "NARA-WPE: A python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *13. ITG Fachtagung Sprachkommunikation (ITG 2018)*, Oct 2018.

[20] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang, "Mosnet: Deep learning based objective assessment for voice conversion," *arXiv preprint arXiv:1904.08352*, 2019.

[21] John Kominek and Alan W Black, "The cmu arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[22] Zhiyong Zhang Dong Wang, Xuewei Zhang, "Thchs-30 : A free chinese speech corpus," 2015.

[23] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[24] Quansheng Duan, Shiyin Kang, Zhiyong Wu, Lianhong Cai, Zhiwei Shuang, and Yong Qin, "Comparison of syllable/phone hmm based mandarin tts," in *2010 20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 4496–4499.

[25] Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," *arXiv preprint arXiv:1907.04448*, 2019.

[26] Shengkui Zhao, Trung Hieu Nguyen, Hao Wang, and Bin Ma, "Towards natural bilingual and code-switched speech synthesis based on mix of monolingual recordings and cross-lingual voice conversion," *arXiv preprint arXiv:2010.08136*, 2020.