INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multi-Lingual Multi-Speaker Text-to-Speech Synthesis for Voice Cloning with Online Speaker Enrollment

Zhaoyu Liu and *Brian Mak*

The Hong Kong University of Science and Technology
Department of Computer Science and Engineering
{zliuar,mak}@cse.ust.hk

Abstract

Recent studies in multi-lingual and multi-speaker text-to-speech synthesis proposed approaches that use proprietary corpora of performing artists and require fine-tuning to enroll new voices. To reduce these costs, we investigate a novel approach for generating high-quality speeches in multiple languages of speakers enrolled in their native language. In our proposed system, we introduce tone/stress embeddings which extend the language embedding to represent tone and stress information. By manipulating the tone/stress embedding input, our system can synthesize speeches in native accent or foreign accent. To support online enrollment of new speakers, we condition the Tacotron-based synthesizer on speaker embeddings derived from a pre-trained x-vector speaker encoder by transfer learning. We introduce a shared phoneme set to encourage more phoneme sharing compared with the IPA. Our MOS results demonstrate that the native speech in all languages is highly intelligible and natural. We also find L2-norm normalization and ZCA-whitening on x-vectors are helpful to improve the system stability and audio quality. We also find that the WaveNet performance is seemingly language-independent: the WaveNet model trained with any of the three supported languages in our system can be used to generate speeches in the other two languages very well.
Index Terms: multi-lingual, multi-speaker, text-to-speech, x-vector, tone/stress embedding

1. Introduction

In traditional text-to-speech (TTS) synthesis methods [1], many system components such as the grapheme-to-phoneme model, phoneme duration model, segmentation model, fundamental frequency estimation model and synthesis model are trained separately, and they require expert domain knowledge to produce high-quality synthesized speech. With the advance of deep learning, they are gradually replaced by neural models. Deep Voice [2] presents a neural TTS system which replaces each separate component with a neural net-based model. Char2wav [3] and Tacotron [4] and its improved version Tacotron2 [5] resort to a totally end-to-end neural model[1] that uses an attention mechanism to convert a sequence of text directly to its corresponding sequence of vocoder features, from which speech audios may be generated using a vocoder. Char2Wav generates WORLD features [6] and uses SampleRNN [7] to generate speech, while Tacotron/Tacotron2 generates linear/mel spectrograms and uses the Griffin-Lim (GL) [8] and WaveNet [9] vocoder, respectively. Tacotron 2 can synthesize natural speech comparable to genuine human speech.

[1]Actually "end-to-end" here only means that both Char2Wav and Tacotron generate vocoder features, not speech audios, from some representation of input texts.

Single-speaker neural TTS systems can be readily extended to support multiple speakers' voices. [10] takes the multi-task learning approach and duplicates the output layer for each of its training speakers so that each speaker is trained with its own speaker-dependent output layer while sharing other hidden layers in the model. Obviously, the model parameters in its output layer grow linearly with the number of training speakers. Multi-speaker Tacotron [11] is introduced by conditioning Tacotron 2's model on pre-trained d-vector speaker embeddings so that new speakers can be enrolled with a few seconds of speech. Similarly, Deep Voice 2 [12] and Deep Voice 3 [13] extends Deep Voice to multi-speaker TTS. Unlike Tacotron 2, Deep Voice 2 and 3 condition each layer of the model with speaker embeddings which is jointly trained with the rest of the TTS system. For example, Deep Voice 3 claims to support 2400 voices. However, enrollment of new speakers in [12] and [13] will require additional training. VoiceLoop [14] uses a fixed-size memory buffer to accommodate speaker-dependent phonological information and facilitates multi-speaker synthesis by buffer shifts. New speaker embeddings can be trained by an optimization procedure while fixing the other model parameters. Neural Voice cloning [15] introduces a similar speaker adaptation method where both model parameters and speaker embeddings are fine-tuned with data from the new speaker. Multi-lingual TTS further extends multi-speaker TTS to support synthesis in more than one language. For example, [16] introduces a cross-lingual TTS system in English and Mandarin trained with IPA without language embedding. It succeeds in synthesizing speech in two languages, however, it can only synthesize native speech but not accented speech. It uses the GL vocoder (instead of WaveNet or other neural-based high fidelity vocoders) resulting in synthesized speech of lower quality.

In this paper, we investigate a multi-lingual and multi-speaker TTS approach to synthesize high-quality speech in three languages and speakers enrol in their own native speech. Our system provides accent control to synthesize accented and native speech when the synthesized language is not the native language of the speaker. [17] proposes a similar approach which shares many ideas in our system. Nonetheless, there are the following notable differences: (a) Our results are reproducible as we used only publicly available training corpora while the system in [17] was trained on proprietary data. (b) [17] aims at synthesizing speech with only training speakers' voices, and their training data consists of few speakers (some are professional voice actors) but each has tens of hours of speech. On the contrary, we trained our system on hundreds of speakers with less than 25 minutes of speech from each speaker. We believe our system is more generalizable to new speakers and we report results on unseen speakers while [17] does not. (c) Both systems employ shared phonemes for inputs and tone/stress em-

http://dx.doi.org/10.21437/Interspeech.2020-1464

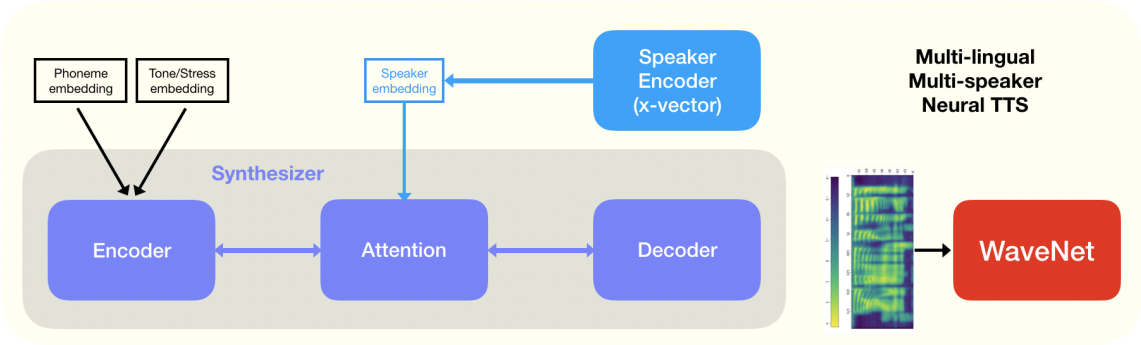

Figure 1: *Multi-lingual multi-speaker TTS system using phoneme embedding, speaker embedding and tone/stress embedding.*

beddings, and speaker embeddings. However, our phoneme set encourages more sharing and is more computationally efficient. Different from their tone/stress embedding, ours combines language information and tone/stress information so that extra language embeddings are unnecessary. And we use the state-of-the-art x-vector for speaker embedding while they use d-vector. We expect our synthesized speech will be better in terms of speaker similarity, especially for unseen test speaker. (d) Our model is simpler with no residual encoding nor adversarial training. Instead, we investigate on the effect of various normalization methods on the speaker embedding vectors for enhancing the intelligibility, naturalness and speaker similarity of the synthesized speech. (e) We also investigate the effect of training the WaveNet vocoder with speech in one language to synthesize speech of all languages in the system.

2. Model Structure

Fig. 1 shows our multi-lingual multi-speaker TTS system.

2.1. Inputs: Phoneme, Tone and Stress Embeddings

Instead of character embedding in [3, 4, 5], we use phoneme embedding which has been shown to generate more natural speech. A shared phoneme set is created by mapping Mandarin pinyin and Cantonese jyupting phonemes to ARPABET, with the exceptions of pinyin phonemes 'j', 'q' and 'x' which are treated as distinct phonemes as no good ARPABET mappings are found. We separately represent 5 Mandarin tones, 6 Cantonese tones and 3 English stresses as 14-D 1-hot embedding and concatenate it to phoneme embedding as shown in Table 1.

2.2. Speaker Encoder

We train a separate speaker encoder using x-vectors described in [18] as speaker embeddings. X-vectors are derived from a TDNN-based speaker discriminative model which is trained to classify the training speakers with a softmax layer. We extract x-vectors from the output of the 6th hidden layer in the TDNN.

We investigate the performance of two normalization techniques: L2-norm normalization and whitening on the generated x-vectors and compared them with unnormalized x-vectors.

2.3. Mel-spectrogram Synthesizer

The mel-spectrogram synthesizer is implemented based on [11]. We input the concatenation of phoneme embeddings and additional tone/stress embeddings to the encoder. Speaker embedding is concatenated with the encoder context output and they are fed into the decoder as in [11]. In our preliminary experiments, we have tried linear addition instead of concatenation but it does not lead to any significant improvement.

Table 1: *One-hot embedding of tones and stresses.*

Index	Tone/stress
0	Mandarin: Neutral tone
1	Mandarin: Tone one
2	Mandarin: Tone two
3	Mandarin: Tone three
4	Mandarin: Tone four
5	English: No stress
6	English: Primary stress
7	English: Secondary stress
8	Cantonese: High level (Tone one)
9	Cantonese: Mid rising (Tone two)
10	Cantonese: Mid level (Tone three)
11	Cantonese: Low falling (Tone four)
12	Cantonese: Low rising (Tone five)
13	Cantonese: Low level (Tone six)

2.4. WaveNet

WaveNet [9] is an auto-regressive sample-by-sample raw audio synthesizer. We construct the WaveNet with 30 layer of dilated causal convolutions and train it with 8-bit mu-law quantization[2] using the CUSENT Cantonese corpus and demonstrate that it can still synthesize high-quality and natural speech in both English and Mandarin. We have also trained another two WaveNet models using English LibriSpeech or Mandarin SurfingTech corpora separately, and the quality of the synthesized speech using any of the three models is similarly good. It seems the WaveNet performance is language-independent.

2.5. Synthesis of Native and Accented Speech

The use of tone/stress embeddings allow us to synthesize native or accented speech in a language X spoken by a speaker whose mother tongue is language Y, where X and Y may be any of the 3 languages supported by our model. To generate native speech, the correct tone or stress is used for each phoneme in the speech. To simulate accented speech by a Cantonese/Mandarin speaker, all phonemes are spoken with Cantonese/Mandarin tone 1, whereas phonemes in an accented speech by an English speaker are spoken with no stress. Table 2 shows which element in the one-hot tone/stress embedding vector will be set to gener-

[2]Training a WaveNet with 8-bit mu-law outputs allows much faster convergence and the output quality is still very good.

Table 2: *Example: Simulation of native and accented English.*

Text	Through out the centuries ...
Phoneme Sequence	TH,R,UW,AW,T,DH,AH,S,EH,N,CH,ER,IY,Z...
Native	5, 5, 5, 6, 5, 5, 5, 5, 6, 5, 5, 5, 5, 5...
Cantonese Accent	8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8...
Mandarin Accent	1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...

ate an English utterance in native English and accented English with Cantonese or Mandarin accent.

3. Experiments and Results

3.1. Training Corpora

We trained our model on 4 datasets in three languages: (1) The "clean" set in Librispeech (LS) [19] consists of 1172/40 training/test English speakers, each with 25 minutes of speech; (2) SurfingTech (ST) [20] is a Mandarin corpus which has 855 speakers, and 10 minutes of speech per speaker. We further randomly select 800 speakers' data for training, others for testing; (3) CUSENT (CU) [21] is a Cantonese corpus with 20 hours of speech from 68/12 training/test speakers; (4) Aishell (AI) [22] is a Mandarin corpus consisting of 150 hours of speech from 340 speakers. We used all data to train the x-vector speaker encoder, and used only CUSENT, Librispeech and SurfingTech data to train the synthesizer. CUSENT was also used to train the WaveNet. Librispeech data were segmented by forced alignment to shorter audios (2s-12s) upon silences that are longer than 0.3s, and denoised by block thresholding [23]. Google Translate was used pinyin transcriptions from the Mandarin texts which were then mapped to ARPABET phonemes. Forced alignment was performed on SurfingTech data to label significant short pauses in its speech.

3.2. Speaker Verification Evaluation

We conducted objective speaker verification (SV) evaluations on x-vector speaker embeddings with an increasing number of training speakers. Enrolment utterances are 3 minutes long and test utterances vary from 5–12s. We first tested 400-D i-vectors, 64/128/512-D x-vectors on Librispeech SV, and the results are shown in Table 3. In our experiments, even though the EERs vary from 1 to 3.25 for different models, the TTS systems using i-vectors or 128-D x-vectors can generate better speech with very similar quality. In contrast, although 64-D x-vectors give the best SV EER, we found that they produce audios of poorer quality in our TTS system. It seems that embeddings that give better SV EER are no guarantee of better synthesized audios. At the end, we chose the 128-D x-vectors that were trained on all corpora for our speaker embeddings, and the SV-EER on LibriSpeech is reduced to 0.75.

Table 3: *Librispeech SV EER (%).*

System	Dim	Train set	Speakers	SV-EER
i-vector	400	LS	1172	3.25
x-vector	64	LS	1172	1.00
x-vector	128	LS	1172	1.50
x-vector	512	LS	1172	1.25
x-vector	128	LS, CU, ST, AI	2380	0.75

3.3. Subjective Evaluation

We conducted two mean opinion score (MOS) tests to subjectively evaluate the performance of our TTS system on speakers who are unseen in both synthesizer and x-vector training. We first constructed three models with the proposed architecture but with unnormalized, L2-norm normalized and whitened 128-D LS+CU+ST+AI x-vectors. We carried out one crowdMOS test using the Amazon Mechanical Turk with the hope of having more raters to rate more synthesized outputs. Another MOS test was conducted with 20 multilingual raters from Guangdong, China where 15 raters are native in both Cantonese and Mandarin; 1 rater is native in Cantonese only and 4 raters are native in Mandarin only; and all are fluent in their non-native languages among the three languages. Both MOS tests use an Absolute Category Rating Scale from 1–5 with 0.5 increments.

3.3.1. CrowdMOS Results

In crowdMOS test conducted in the US, we randomly selected 40 ground truth utterances in total from 10 test speakers unseen in both synthesizer and x-vector training for each language and synthesized 40 unseen utterances from them using each model[3]. We asked all raters to rate speaker similarity but we asked only those raters who understand the target language (verified by a transcription question) to rate the naturalness of the synthesized speech. The number of distinct raters (given in parentheses in Table 4) is smaller than expected because some raters may rate multiple assignments. We found that there are much fewer raters who understand Cantonese and Mandarin. Additionally, the original crowdMOS results were very noisy. The naturalness MOS of ground truth English speech was only 3.65. Thus, we added a qualification question in each MOS task to ask the raters to rate a ground truth utterance, and filtered out those responses that rated the ground truth utterances lower than 3.5. The results are shown in Table 4.

Table 4a shows the naturalness and speaker similarity MOS of the ground truth utterances. It seems difficult for English raters to rate the similarity of Chinese speakers, especially Mandarin speakers. Table 4b shows the speaker similarity MOS of same-language voice cloning of Cantonese and English speakers using models trained with different x-vector normalization techniques. Both whitening and L2-norm normalization help improve the performance. However, we found that the synthesis of some texts failed to stop: Out of our 600 syntheses, 1 failed to stop with L2-norm normalization; 24 failed to stop with whitening, and 9 failed to stop with no normalization. It shows that L2-norm normalization helps improve the model stability. Nonetheless, the quality of synthesized speech produced by whitening is slightly better and whitening normalization was used in all the remaining experiments. Table 4c gives the naturalness MOS of voice cloning of speakers of different mother tongues to speak native/accented English. The synthesized native speech is more natural than accented speech as expected. Same-language voice cloning performs slightly better than cross-lingual voice cloning. Table 4d and 4e show the speaker similarity MOS of voice cloning to native and accented speech, respectively. Interestingly, it seems that speaker similarity is not highly correlated to either the target or the source language, or the accents.

[3]Only performance on unseen speakers is reported as this is the more difficult task. The performance on seen speakers are generally better.

Table 4: *CrowdMOS results with 95% confidence interval (Student's t-distribution). The figures in ( ) are the numbers of distinct raters after filtering in each case. (TL: target language of synthesized speech; SL: source language or mother tongue)*

Language	Naturalness	Speaker Similarity
Cantonese	4.53±0.15 (4)	3.95±0.18 (10)
English	4.15±0.12 (12)	4.22±0.22 (11)
Mandarin	4.31±0.18 (3)	3.59±0.19 (11)

(a) *MOS of ground truth speech.*

Language\Normalization	whitening	L2-norm	none
Cantonese (23)	3.15±0.16	3.14±0.15	3.08±0.15
English (30)	3.33±0.15	3.31±0.14	2.93±0.15

(b) *Effect of x-vector normalization on speaker similarity MOS of same-language voice cloning.*

TL\SL	Cantonese	English	Mandarin
Native English	3.67±0.15 (18)	3.83±0.16 (30)	3.64±0.15 (22)
Accented English	3.27±0.17 (18)	-	3.01±0.18 (22)

(c) *Naturalness MOS of voice-cloning for speakers of different mother tongues to speak native/accented English.*

TL\SL	Cantonese	English	Mandarin
Cantonese	3.15±0.15 (23)	3.35±0.11 (20)	3.34±0.10 (17)
English	3.27±0.12 (18)	3.33±0.15 (30)	3.19±0.12 (22)
Mandarin	3.29±0.11 (21)	3.33±0.10 (17)	3.06±0.15 (25)

(d) *Speaker similarity MOS of cross-lingual voice-cloning for a speaker of mother tongue SL to speak like a native speaker of TL.*

TL\SL	Cantonese	English	Mandarin
Cantonese	-	3.18±0.11 (20)	3.25±0.10 (17)
English	3.28±0.12 (18)	-	3.27±0.12 (22)
Mandarin	3.40±0.11 (21)	3.49±0.10 (17)	-

(e) *Speaker similarity MOS of cross-lingual voice-cloning for a speaker of mother tongue SL to speak TL with his/her own SL accent.*

3.3.2. MOS Results from Multilingual Raters

Table 5 shows the MOS results from 20 multilingual raters. We randomly selected 2 ground truth utterances from 2 unseen speakers per language, and synthesized 1 unseen utterance per language for selected speakers. Each rater had to rate all ground truth and synthesized utterances. Table 5a shows the ground truth MOS. The speaker similarity is particularly high for Mandarin speech, and a probable reason is that the raters almost all raters can speak native Mandarin.

Table 5b/5c gives the intelligibility MOS of native/accent speech, respectively, whereas 5d/5e gives their naturalness MOS. As expected, same-language voice cloning performs significantly better than cross-lingual voice cloning in terms of both intelligibility and naturalness. Cross-lingual voice cloning of native speech between Cantonese and Mandarin speakers performs better than their voice cloning to native English probably because of the similarity between Cantonese and Mandarin. However, the foreign accent simulated by wrong input tones in the synthesized accented Cantonese and Mandarin speech results in worse MOS than accented English. Interestingly, using our proposed model, Cantonese speakers can speak English better than Mandarin speakers while English speakers can speak better Mandarin than Cantonese. The results seem to indicate that the x-vector speaker embedding contains language information of the speakers. Table 5f and 5g give the speaker similarity MOS. Different from the crowdMOS results, this group of multilingual raters gave higher similarity MOS for same-language voice cloning than cross-lingual voice cloning.

Table 5: *Multilingual raters' MOS with 95% confidence interval (Student's t-distribution).*

Language	Intelligibility	Naturalness	Spkr Similarity
Cantonese	4.45±0.08	4.05±0.11	4.55±0.06
English	4.34±0.08	4.45±0.07	4.45±0.06
Mandarin	4.60±0.07	4.03±0.13	4.98±0.01

(a) *Ground truth speech.*

TL\SL	Cantonese	English	Mandarin
Cantonese	4.50±0.11	3.28±0.23	3.84±0.20
English	4.07±0.15	4.54±0.07	3.78±0.14
Mandarin	4.43±0.09	4.17±0.15	4.42±0.15

(b) *Intelligibility MOS of native speech.*

TL\SL	Cantonese	English	Mandarin
Cantonese	-	2.05±0.27	1.70±0.15
English	3.09±0.14	-	2.14±0.17
Mandarin	2.26±0.21	2.57±0.20	-

(c) *Intelligibility MOS of accented speech.*

TL\SL	Cantonese	English	Mandarin
Cantonese	4.28±0.18	2.92±0.23	3.51±0.21
English	3.79±0.20	4.30±0.13	3.45±0.20
Mandarin	4.24±0.12	3.70±0.15	4.32±0.18

(d) *Naturalness MOS of native speech.*

TL\SL	Cantonese	English	Mandarin
Cantonese	-	2.13±0.17	1.54±0.09
English	2.91±0.20	-	2.05±0.17
Mandarin	2.12±0.35	2.30±0.19	-

(e) *Naturalness MOS of accented speech.*

TL\SL	Cantonese	English	Mandarin
Cantonese	3.82±0.37	2.91±0.38	3.00±0.51
English	3.57±0.32	3.84±0.37	3.30±0.40
Mandarin	3.28±0.36	3.20±0.33	4.49±0.18

(f) *Speaker similarity MOS of native speech.*

TL\SL	Cantonese	English	Mandarin
Cantonese	-	3.07±0.39	2.96±0.69
English	2.53±0.39	-	3.01±0.53
Mandarin	2.45±0.44	3.39±0.35	-

(g) *Speaker similarity MOS of accented speech.*

4. Conclusion

This paper presents a novel multi-lingual multi-speaker TTS model that can enroll new speakers without fine-tuning and synthesize speech in languages different from the speakers' mother tongue. The model can clone a voice to speak intelligibly and naturally in its own language or another language as if he/she is a native speaker of the other language. It can also generate accented speech in another language with an accent due to the speakers' mother tongue. We further find that the WaveNet could be trained in any of the supported languages in this paper and then used to synthesize speech in the other languages well.

5. Acknowledgements

This work was supported by grants from the Research Grants Council of the Hong Kong SAR, China (Project Nos. HKUST16200118, HKUST16215816 and T45-407/19N-1).

6. References

[1] T. Dutoit, *An Introduction to Text to Speech Synthesis*. Kluwer Academic Publishers, 1997.

[2] S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep Voice: Real-time neural text-to-speech," *CoRR*, vol. abs/1702.07825, 2017.

[3] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. C. Courville, and Y. Bengio, "Char2wav: End-to-end speech synthesis," in *the 5th International Conference on Learning Representations, ICLR 2017. Workshop Track Proceedings*, 2017.

[4] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech*, 2017, pp. 4006–4010.

[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 4779–4783.

[6] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions*, vol. 99-D, no. 7, pp. 1877–1884, 2016.

[7] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," *arXiv preprint arXiv:1612.07837*, 2016.

[8] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[9] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[10] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4475–4479.

[11] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in neural information processing systems*, 2018, pp. 4480–4490.

[12] 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," in *Advances in neural information processing systems*, 2017, pp. 2962–2970.

[13] 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.

[14] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "Voiceloop: Voice fitting and synthesis via a phonological loop," *arXiv preprint arXiv:1707.06588*, 2017.

[15] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 019–10 029.

[16] M. Chen, M. Chen, S. Liang, J. Ma, L. Chen, S. Wang, and J. Xiao, "Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding," *Proc. Interspeech 2019*, pp. 2105–2109, 2019.

[17] 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 foreign language: Multilingual speech synthesis and cross-language voice cloning," in *Interspeech*, 2019.

[18] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 5329–5333.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5206–5210.

[20] "ST-CMDS-20170001_1, Free ST Chinese Mandarin Corpus."

[21] T. Lee, W. K. Lo, P. C. Ching, and H. Meng, "Spoken language resources for Cantonese speech processing," *Speech Communication*, vol. 36, no. 3-4, pp. 327–342, 2002.

[22] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *the 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.

[23] G. Yu, S. Mallat, and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Transactions on Signal processing*, vol. 56, no. 5, pp. 1830–1839, 2008.