*INTERSPEECH 2019*
September 15–19, 2019, Graz, Austria

# Building a mixed-lingual neural TTS system with only monolingual data

*Liumeng Xue*[1^], *Wei Song*[2], *Guanghui Xu*[2], *Lei Xie*[1*], *Zhizheng Wu*[3]

[1]Shaanxi Provincial Key Lab of Speech and Image Information Processing,
School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]JD.com, China
[3]JD.com, USA

{lmxue, lxie}@nwpu-aslp.org, {songwei11, xuguanghui, zhizheng.wu}@jd.com

## Abstract

When deploying a Chinese neural Text-to-Speech (TTS) system, one of the challenges is to synthesize Chinese utterances with English phrases or words embedded. This paper looks into the problem in the encoder-decoder framework when only monolingual data from a target speaker is available. Specifically, we view the problem from two aspects: speaker consistency within an utterance and naturalness. We start the investigation with an average voice model which is built from multi-speaker monolingual data, i.e., Mandarin and English data. On the basis of that, we look into speaker embedding for speaker consistency within an utterance and phoneme embedding for naturalness and intelligibility, and study the choice of data for model training. We report the findings and discuss the challenges to build a mixed-lingual TTS system with only monolingual data.

**Index Terms**: speech synthesis, encoder-decoder, mixed-lingual

## 1. Introduction

When deploying a non-English Text-to-Speech (TTS) system, it is very common that we have to address the mixed-lingual problem. A mixed-lingual TTS system is expected to synthesize utterances with embedded phrases or words from a different language. A straightforward way to build a mixed-lingual TTS system is to use a bilingual speech database recorded by a bilingual speaker. However, it's very hard to find a speaker with excellent multiple language skills and consistent articulation across different languages, and it is not flexible to use pre-recorded data which only has monolingual data. In this work, we look into the mixed-lingual TTS in Mandarin Chinese context with English phrases or words embedded.

### 1.1. Related work

A mixed-lingual TTS system is expected to generate high-quality speech and be perceived as spoken by the same speaker even when switching languages in mixed-lingual utterances. Several studies have been conducted to assess the mixed-lingual problem. In [1], Traber *et al.* builds a mixed-lingual TTS system using a bilingual speech database recorded by a bilingual speaker. In [2], HMM states are shared across languages and speech data from multiple languages are used. In [3, 4, 5], Mandarin and English context-dependent HMM states are shared, and the mapping is learned from a bilingual dataset recorded by a bilingual speaker. In [6], He *et al.* proposes an approach to convert a monolingual TTS into multilingual by employing a bi-linear frequency warping function, and taking into account of cross-language F0 variations and equalizing speaking rate difference between source speaker and reference speaker. Besides, speaker adaptation and voice conversion are also effective ways to mixed-lingual speech synthesis using a set of monolingual or multilingual speech databases. In [7], Ramani *et al.* creates a polyglot corpus using voice conversion on a set of multilingual speech databases including Tamil, Telugu, Malayalam, and Hindi. The HMM-based polyglot TTS built with the polyglot database can synthesize mixed-lingual speech for four languages in target speaker voice. In [8, 9], Sitaram *et al.* presents a code-mixed TTS framework, in which the languages are not written in their native script but borrow the script of the other language. Then, the mapping between the phonemes of both languages is used to synthesize the text using a TTS system trained on a single language. In [10], Chandu *et al.* further extends their code-mixed TTS to a bilingual system using two monolingual speech datasets and a combined phone set for speech synthesis of mixed-language navigation instructions. For deep neural network based speech synthesis, a cross-lingual TTS is built using Kullback-Leibler divergence [11].

Recently, encoder-decoder framework has been successfully applied to TTS system. In [12], Li *et al.* presents two end-to-end models: Audio-to-Byte (A2B) and Byte-to-Audio (B2A), for multilingual speech recognition and synthesis, modeling text using a sequence of Unicode bytes, specifically, the UTF-8 variable length byte sequence for each character. The B2A model is able to synthesize code-switching text and the speech is fluent, but the speaker voice is changed for different language.

### 1.2. The contribution

We conduct investigations based on the encoder-decoder framework, which is proven to generate speech with better prosody. In this study, we attempt to answer the following questions:

- Can the encoder-decoder model learn meaningful phonetic representations in encoder part? Does the encoder interpret Mandarin and English phonemes differently?
- What is the impact of speaker embedding on speaker consistency within mixed-lingual utterances?
- Can phonetic information, i.e., phoneme embedding, be used in attention alignment and context vector and as a result improve naturalness and speaker consistency when switching languages in an utterance?
- Is monolingual data enough to build a mixed-lingual TTS system?

To answer these questions, we conduct analysis on phoneme embeddings to understand what the encoder-decoder

^ Work was done when Liumeng Xue was an intern at JD.com.
* Lei Xie is the corresponding author.
The research work is supported by the National Key Research and Development Program of China (No.2017YFB1002102).



http://dx.doi.org/10.21437/Interspeech.2019-3191

model is learning, and present phoneme-informed attention for the encoder-decoder model. Besides, we compare speaker embedding at different positions to analyze its effects on speaker consistency among mixed-lingual utterances. We also study the way of model training in terms of the choice of training data.

# 2. Mixed-lingual neural TTS system

Even though there are two languages in one utterance, a mixed-lingual TTS system is expected to produce synthesized utterances that sound like one speaker. Hence, speaker consistency within an utterance, intelligibility and naturalness are all important factors for the mixed-lingual TTS.

It will be very challenging if not impossible to learn bilingual phonetic coverage when only monolingual data is available. We start our investigations from building an average voice model (AVM) with multi-speaker monolingual dataset. We note that the Chinese corpus does not have English words and the English corpus does not have Chinese pronunciations. To control speaker consistency, speaker embedding is investigated, and phoneme embedding is also studied to better understand how the encoder-decoder model learns phonetic information.

## 2.1. Multi-speaker voice modeling

There are two ways to build an AVM from multi-speaker data. One way is to mix all the data together and treat the data as from a single speaker. Retrain and adaptation can be performed on top of that. The other way is to assign each speaker a label (e.g., speaker code), and use the label to distinguish data from different speakers . In this work, the second way is used and speaker embedding is applied. We used the first way to analyze phoneme embedding.

## 2.2. Speaker embedding

We use speaker embedding to help the AVM training with multi-speaker monolingual data. The speaker embeddings are assumed to construct a speaker space. There are various approaches proposed for modeling the speaker space [13, 14, 15]. Speaker embedding has been extensively used in multi-speaker speech synthesis to generate the speech of the specific speaker [13, 14, 16, 17, 18]. It has been proved that speaker embedding is an effective way to model a speaker space [19]. In general, speaker embedding can be concatenated with encoder output [15, 20] or fed into decoder as an extra input [17]. In this paper, we use a speaker look-up table to store speaker embeddings, which is trained jointly with the encoder-decoder network. The speaker embeddings are utilized to condition speech synthesis to control speaker voice in both training and inference.

To investigate how to place speaker embedding in the encoder-decoder architecture for better speaker consistency, we consider two different positions: 1) concatenating speaker embedding with the encoder output (SE-ENC) and 2) concatenating speaker embedding with the decoder input (SE-DEC) , i.e., with the LSTM input after prenet. To get the target speaker characteristic, we use two approaches: 1) excluding the target speaker data in the AVM but using it to retrain the decoder, 2) including the target speaker data in the AVM to learn speaker embeddings simultaneously.

## 2.3. Phoneme embedding

Phonetic coverage and the relationship between English and Mandarin phonemes are important to the naturalness and intelligibility for a mixed-lingual TTS system. To this end, we analyze the phoneme embeddings and encoder outputs to understand how the encoder-decoder model learns phonetic representations . Phoneme embeddings are vectorized presentations of discrete phonemes, while encoder outputs are derived from phoneme embeddings, which also have a direct impact on the decoder via attention modeling and the resultant context vector.

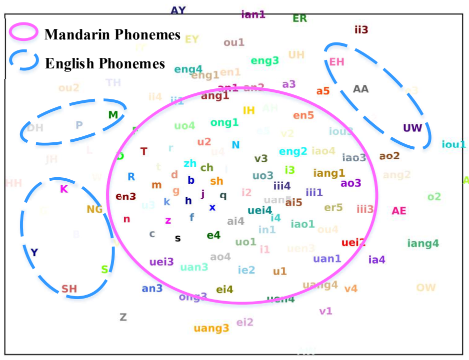


Figure 1: *Phoneme embeddings visualization using t-SNE.*

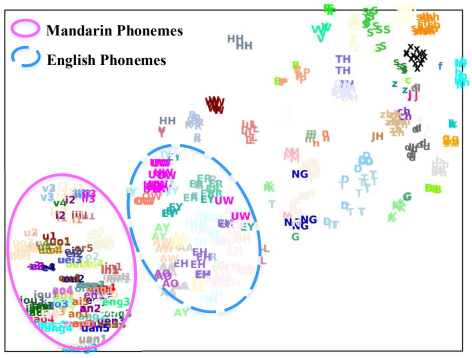


Figure 2: *Encoder outputs visualization using t-SNE.*

Figure 1 and Figure 2 present the t-SNE [21] visualization of phoneme embeddings and encoder outputs, respectively. From the phoneme embedding representations, the English and Mandarin phonemes are separated in some sense but do not have a clear boundary like that in the visualization of encoder outputs. In the visualization of encoder outputs, however, the clustering changes a bit. Mandarin and English phonemes are grouped into two separate clusters. It implies the properties of phoneme embeddings have been changed a bit after several layers of transformations. We suspect that the encoder output is affected more by the audio information which is passed down through back-propagation. We also argue that if the attention alignment is not accurate, and it may also introduce errors into encoder through back-propagation. Hence, it might be useful to have phonetic information when computing the attention alignment or context vector.

## 2.4. Phoneme-informed attention

We investigate the impact of phoneme-informed attention from two directions. One is to calculate an additional phoneme embedding context vector (PECV) by applying attention weights to phoneme embeddings and concatenate it with the attention context vector. The other is to use a residual encoder (RES) architecture by adding the phoneme embeddings to encoder outputs directly. More details about the phoneme-informed attention can be found in the samples webpage[1]. An illustration of the architecture investigated in this study can be found in Figure 3.

[1]Samples can be found at `https://angelkeepmoving.github.io/mixed-lingual-tts/index.html`

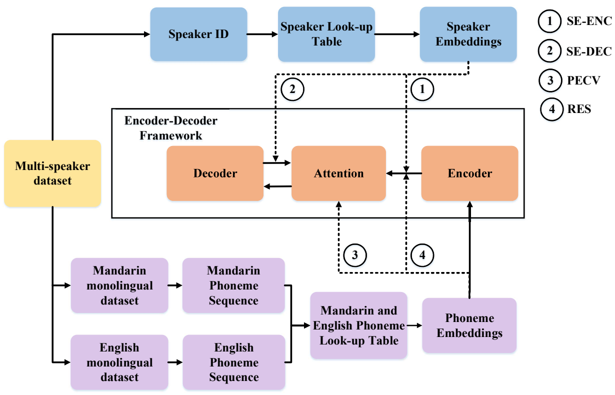


Figure 3: *The architecture investigated in this study.*

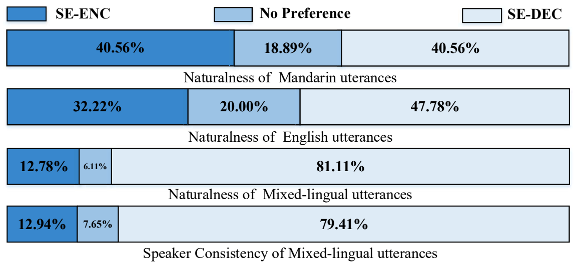


Figure 4: *AB preference results of SE-ENC and SE-DEC.*

# 3. Experimental Results and Analysis

## 3.1. Experimental setup

In this paper, the goal is to study how to build a female Mandarin-English mixed voice using only Mandarin data. We limite the number of training utterances from the target speaker to 500. As the target speaker is a female, only female datasets are utilized to reduce the effect of gender factor. Hence, as we described earlier, to build an AVM, we use an internal Mandarin monolingual dataset from 35 female speakers and an English monolingual dataset from 35 female speakers with American accents, which is a subset of the public available dataset VCTK [22]. Each Mandarin monolingual speaker has around 500 utterances, in total of 17,197 utterances, which is approximately 17 hours of audio. Each English monolingual speaker has varied number of utterances from 200 to 500, in total of 14,464 utterances. English utterances are shorter than Mandarin utterances in duration, so the English dataset is about 8 hours of audio.

All audios are down-sampled to 24kHz. The beginning silence are all trimmed, and the ending silence are trimmed to a fixed length. 80-dimensional mel-spectrograms and 1024-dimensional linear spectrograms are extracted from audios as the model target output. Phoneme sequences are fed to the model as input to predict spectrograms. In this paper, experiments are performed based on the encoder-decoder neural TTS system [23]. Since our work focuses on generating mixed-lingual speech with satisfied intelligibility and a consistent voice, we use the Griffin-Lim [24] algorithm to synthesize waveform from the predicted linear spectrograms like Tacotron-1 [25], instead of using a WaveNet vocoder like Tacotron-2 [23].

## 3.2. Experiments analysis

We perform AB preference tests in terms of naturalness and speaker consistency to assess the performances of different

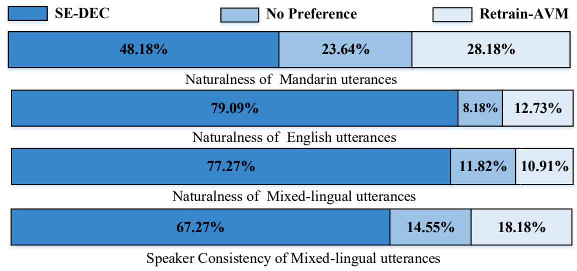


Figure 5: *AB preference results of SE-DEC and Retrain-AVM.*

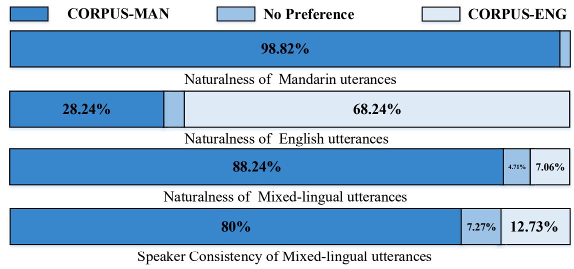


Figure 6: *AB preference results of SE-DEC models trained using CORPUS-MAN and CORPUS-ENG.*

methods. In detail, the speaker consistency AB preference tests are conducted on mixed-lingual sentences, which focus on the voice consistency within mixed sentences. Meanwhile, the naturalness AB preference tests are performed on Mandarin, English and mixed-lingual sentences. For each language, 30 sentences are randomly selected from test set. A group of 18 Mandarin listeners participates in the tests to give their preference choice.

### *3.2.1. Analysis of speaker embedding*

We first analyze the effect of speaker embedding by comparing different positions of speaker embedding. Here, the 500 Mandarin data of target speaker is mixed with data from the speakers for AVM. We compare the effects of speaker embedding at encoder output (SE-ENC) and decoder input (SE-DEC). The results show that SE-DEC brings better performances of speaker similarity and naturalness, as indicated in Figure 4. Placing speaker embedding at decoder input is surprisingly effective for mixed-lingual utterances. We argue that because speaker characteristic is more expressed in speech rather than text and the decoder is more related to speech in TTS task, placing the speaker embedding at the decoder input is more suitable.

### *3.2.2. Including versus excluding the target speaker data in the AVM training*

We then analyze how to use the data of the target speaker to generate mixed-lingual speech properly. The AVM excluding the target speaker data is pre-trained and used to adapt to the target speaker. We retrain decoder based on the AVM (Retrain-AVM) using 500 Mandarin data of the target speaker. Besides, as suggested by the results above, SE-DEC can achieve better performance than SE-ENC. Thus, the AVM including the target speaker data with speaker embedding at decoder (SE-DEC) is built to get the target speaker voice. We compare the performance between Retrain-AVM and SE-DEC. The results in Figure 5 show that the SE-DEC can achieve better performances than Retrain-AVM in terms of naturalness of three languages and speaker consistency. It suggests that including the target

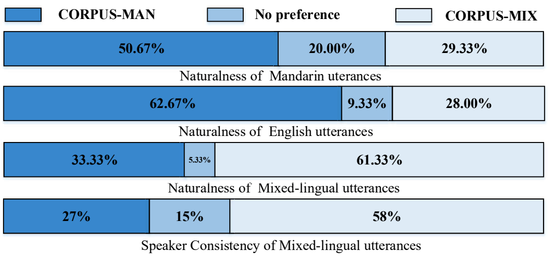


Figure 7: *AB preference results of SE-DEC models trained using CORPUS-MAN and CORPUS-MIX.*

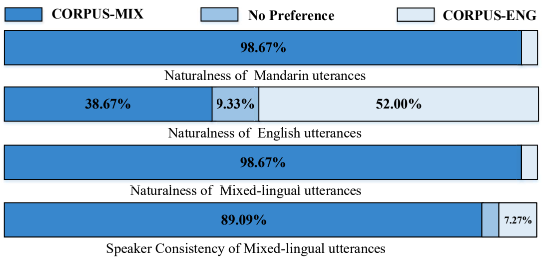


Figure 8: *AB preference results of SE-DEC models trained using CORPUS-MIX and CORPUS-ENG.*

speaker data in AVM is helpful to build the mixed-lingual TTS system with monolingual data.

*3.2.3. The choice of training data*

We then use the SE-DEC structure and study the impact of the choice of training data. Note that the data of target speaker is involved in the AVM training. We use three independent training sets from the same target speaker, 500 Mandarin utterances (CORPUS-MAN), 500 English utterances (CORPUS-ENG) and 500 mixed Mandarin-English utterances (CORPUS-MIX). This is to answer whether monolingual training data can achieve the same performance as that with mixed-lingual training data. The listening test results are presented in Figures 6, 7, and 8. On the mixed-lingual test set, it is always preferred to have mixed-lingual training data. In the case of no mixed-lingual training data, listeners prefer the synthesized audio generated from model trained by Mandarin data. We argue that it may because the primary language of the built TTS system is Mandarin, and better Mandarin synthesis helps in the listening tests. To synthesize monolingual Mandarin, Mandarin training data is always preferred, followed by mixed-lingual. However, to synthesize monolingual English, even though English training data is always preferred, surprisingly Mandarin data is preferred than mixed-lingual data when no English training data is available. We plan to investigate on this aspect further.

*3.2.4. The use of phoneme-informed attention*

We also examine the impact of phoneme-informed attention. On the basis of SE-DEC model, two methods are performed: an additional phoneme embedding context vector (SE-DEC-PECV) and a residual encoder (SE-DEC-RES). Preference comparisons are presented in Figure 9. The results demonstrate that using the residual encoder can achieve better naturalness than the additional context vector for three languages. For speaker consistency, the residual encoder and the additional context vector achieve almost the same preference. Furthermore, we compare the performance of using residual encoder or not. The results shown in Figure 10, demonstrate that using

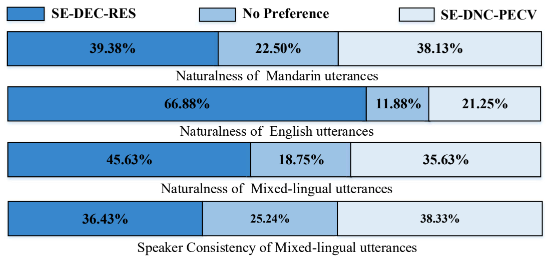


Figure 9: *AB preference results of SE-DEC-RES and SE-DEC-PECV.*

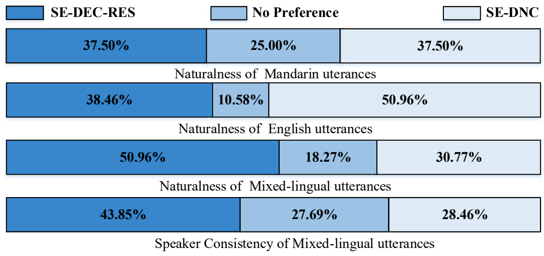


Figure 10: *AB preference results of SE-DEC and SE-DEC-RES.*

residual encoder brings better naturalness for three languages, and also achieves better speaker consistency in mixed-lingual speech.

## 4. Conclusions

In this paper, we investigate how to build a robust Mandarin-English mixed-lingual TTS system when only monolingual data of the target speaker is available. We conduct the study in the encoder-decoder framework. Here are our findings:

- The average voice model built from multi-speaker monolingual data is helpful, and the encoder part can learn phonetic information and the relationship between Mandarin and English phonemes. It also suggests that including the target speaker data in AVM training helps.
- Speaker embedding is useful to control speaker identity and maintain speaker consistency within mixed-lingual utterances. Furthermore, the way to integrate speaker embedding in the encoder-decoder framework is important. Our experimental results show that integrating speaker embedding in the decoder input works better.
- Although monolingual data is able to build a mixed-lingual TTS system, mixed-lingual training data is still preferred to have. It suggests that further investigations should be performed to built a better mixed-lingual TTS system with only monolingual data.
- Experimental results confirm that phoneme-informed attention not only helps with naturalness but also speaker consistency in mixed-lingual utterances.

From informal listening test, we find that when only monolingual data is used to train the model, in some mixed-lingual utterances, the prosody is unnatural, specifically for Mandarin words next to English words, the tones of Mandarin words become inaccurate. In the future work, we plan to integrate our approach with a neural vocoder to produce better audio quality, and have a systematic investigation on phoneme embedding and speaker embedding, and also explore model training with more data.

# 5. References


[1] C. Traber, K. Huber, K. Nedir, B. Pfister, E. Keller, and B. Zellner, "From multilingual to polyglot speech synthesis," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[2] J. Latorre, K. Iwano, and S. Furui, "Polyglot synthesis using a mixture of monolingual corpora," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–1.

[3] H. Liang, Y. Qian, and F. K. Soong, "An HMM-based bilingual (Mandarin-English) TTS," *Proceedings of SSW6*, 2007.

[4] Y. Qian, H. Cao, and F. K. Soong, "HMM-based mixed-language (Mandarin-English) speech synthesis," in *2008 6th International Symposium on Chinese Spoken Language Processing*. IEEE, 2008, pp. 1–4.

[5] Y. Qian, H. Liang, and F. K. Soong, "A cross-language state sharing and mapping approach to bilingual (Mandarin–English) TTS," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1231–1239, 2009.

[6] J. He, Y. Qian, F. K. Soong, and S. Zhao, "Turning a monolingual speaker into multilingual for a mixed-language TTS," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[7] B. Ramani, M. A. Jeeva, P. Vijayalakshmi, and T. Nagarajan, "Voice conversion-based multilingual to polyglot speech synthesizer for Indian languages," in *2013 IEEE International Conference of IEEE Region 10 (TENCON 2013)*. IEEE, 2013, pp. 1–4.

[8] S. Sitaram and A. W. Black, "Speech synthesis of code-mixed text." in *LREC*, 2016.

[9] S. Sitaram, S. K. Rallabandi, S. Rijhwani, and A. W. Black, "Experiments with cross-lingual systems for synthesis of code-mixed text." in *SSW*, 2016, pp. 76–81.

[10] K. R. Chandu, S. K. Rallabandi, S. Sitaram, and A. W. Black, "Speech synthesis for mixed-language navigation instructions." in *INTERSPEECH*, 2017, pp. 57–61.

[11] F.-L. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN approach to cross-lingual TTS," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5515–5519.

[12] B. Li, Y. Zhang, T. Sainath, Y. Wu, and W. Chan, "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes," *arXiv preprint arXiv:1811.09021*, 2018.

[13] 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.

[14] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, "Fitting new speakers based on a short untranscribed sample," *arXiv preprint arXiv:1802.06984*, 2018.

[15] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in Neural Information Processing Systems*, 2018, pp. 4480–4490.

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

[17] 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.

[18] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," *Proc. ICLR*, 2018.

[19] 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.

[20] R. J. Skerry-Ryan, E. Battenberg, X. Ying, Y. Wang, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," 2018.

[21] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[22] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2016.

[23] 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.

[24] 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.

[25] 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.