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

Can Speaker Augmentation Improve Multi-Speaker End-to-End TTS?

Erica Cooper[1⋆], *Cheng-I Lai*[2⋆], *Yusuke Yasuda*[1], *Junichi Yamagishi*[1]

[1]National Institute of Informatics, Japan
[2]Massachusetts Institute of Technology, USA
{ecooper,yasuda,jyamagis}@nii.ac.jp, clai24@mit.edu

Abstract

Previous work on speaker adaptation for end-to-end speech synthesis still falls short in speaker similarity. We investigate an orthogonal approach to the current speaker adaptation paradigms, *speaker augmentation*, by creating artificial speakers and by taking advantage of low-quality data. The base Tacotron2 model is modified to account for the channel and dialect factors inherent in these corpora. In addition, we describe a warm-start training strategy that we adopted for Tacotron2 training. A large-scale listening test is conducted, and a distance metric is adopted to evaluate synthesis of dialects. This is followed by an analysis on synthesis quality, speaker and dialect similarity, and a remark on the effectiveness of our speaker augmentation approach. Audio samples are available online[1].

Index Terms: Speaker augmentation, Speech synthesis, dialect identification, channel modeling, transfer learning

1. Introduction

Recent advances in end-to-end text-to-speech (TTS) synthesis enable the production of synthetic speech of high quality and good speaker similarity [1, 2, 3, 4]. Although the speech quality approaches human naturalness, challenges still remain: first, to model many speakers simultaneously using a common model (termed "multi-speaker TTS") and second, to adapt to voices of arbitrary new speakers while minimizing the amount of data to be collected and requiring little or no additional model training (termed "speaker adaptation").

Previous work on speaker adaptation can be categorized into one of two general approaches. The first approach is simple fine-tuning [5, 6, 7]: the TTS model receives a small amount of additional training with target speaker data, which must be transcribed. The second approach is the use of external speaker embeddings [8, 9], which are extracted from separately trained automatic speaker verification (ASV) models, and the embeddings are input as speaker information to TTS models. This approach does not require transcriptions, and the speaker embedding can be computed from only a few utterances. However, it is reported that speaker similarity of unseen speakers is relatively low [8]. On the other hand, there are a few attempts to use low-quality recordings for TTS. In [10], low-quality recordings were used for fine-tuning based speaker adaptation. Variational autoencoder based clean speech and noise factorisation [11, 12] was also proposed for Tacotron TTS. They conducted speaker adaptation using artificially corrupted speech data [11] or real noisy speech [12] and tried to create speaker-adapted 'clean' TTS voices via the proposed factorisation.

In our previous work [9], we constructed a multi-speaker Tacotron TTS model on the VCTK corpus [13], using speaker embeddings that are transferred from a separately trained ASV model, and performed zero-shot speaker adaptation. The VCTK dataset contains high-quality speech recordings from around a hundred speakers of different English dialects. However, our model was overfitted to seen speakers, and voice characteristics and dialects of unseen speakers were not well reproduced, although the quality of synthetic speech was high [9]. We hypothesized that this number of speakers may be small for our task, and that increasing the number of training speakers can provide better coverage of the speaker space, avoid overfitting to seen speakers, and thus improve similarity and perceived dialects of unseen speakers. However, TTS-quality datasets larger than VCTK are not easily found or created.

A more realistic solution would be ***speaker augmentation***, that is, data augmentation for increasing the number of speakers used for neural network training. This has been investigated for ASV [14], wherein they created "artificial" speakers by simply re-sampling the original audio. They found that this approach improved their speaker models, and also that their system identified the artificial speakers as separate from the original ones. This is known as "vocal tract length perturbation" (VTLP) and it also improved ASR [15]. This could be useful for multi-speaker TTS since by adding more speakers, we can hope that neural networks will be aware of more diverse speaker characteristics and thus avoid overfitting to seen speakers.

In addition to the above artificial speaker augmentation, we also consider another idea for speaker augmentation wherein we use non-ideal TTS data, that is, audio recordings that were collected for purposes other than TTS, and may not meet our usual high-quality recording standards, but have a larger number of speakers. However, carelessly mixing in data from worse recording conditions is expected to degrade the quality of synthesized speech. Furthermore, unlike artificial speaker augmentation, it also increases the number of different dialects included in the training database. We therefore once again borrow ideas from speaker recognition like the neural speaker embeddings in our previous paper, and propose an improved Tacotron speech synthesizer to explicitly handle the two factors, **channel** and **dialect**. Here, the channel is a factor jointly caused by frequency characteristics of recording equipment, noise and reverberation. More precisely, in the proposed synthesizer, neural dialect embedding vectors are used to condition Tacotron's encoder, and channel labels are used to condition Tacotron's postnet.

2. Speaker Augmentation for TTS

Data augmentation has shown to be very effective for speech recognition (e.g. [16, 17]) and speaker recognition [18]. Although data augmentation for speech synthesis was investigated in the past (e.g. [19, 20]), improvements are rather small and the best augmentation strategy for speech synthesis is still unknown. In this paper we consider two speaker augmentation ideas and investigate how such augmentation improves multi-speaker end-to-end TTS.

⋆ Equal contribution.
[1] https://nii-yamagishilab.github.io/samples-multi-speaker-tacotron/augment.html

http://dx.doi.org/10.21437/Interspeech.2020-1229

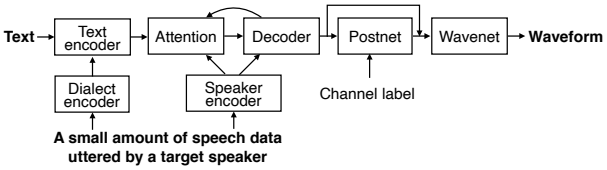

Figure 1: *Diagram of our TTS system for speaker augmentation using low-quality data. This is modified from [9] to accommodate for the additional channel and dialect factors, by adding a channel-aware postnet and a dialect encoder network.*

2.1. Artificial speaker augmentation

The first method of speaker augmentation is the same as [14] wherein we create "artificial" speakers by manipulating the high-quality audio signals. This is a re-sampling of waveforms, and the resulting signals have different fundamental frequency, speaking rate, formants, and spectra. We implemented this augmentation using the SoX [21] 'speed' command, which speeds up or slows down audio by resampling. We created 'x0.9' and 'x1.1' re-sampled versions of each VCTK speaker's speech, and used this augmented dataset to train a speaker-augmented Tacotron model.

2.2. Speaker augmentation using low-quality data

The second method of speaker augmentation is to use low-quality data collected for purposes other than TTS, such as ASR. Such data can represent a diverse range of speakers and dialects, and can be used for the purpose of speaker augmentation for speech synthesis. Our aim is to use the low-quality data for speaker augmentation only, and we assume that target speaker data is limited but recorded in high-quality studios. This is different from previous work, which uses low-quality recordings for speaker adaptation [10, 11, 12] and multi-speaker modeling (e.g. [22]).

However, such ASR data does not meet our high-quality recording standards. It may contain background noise and reverberation unlike typical TTS data recorded in anechoeic chambers. Furthermore, unlike artificial speaker augmentation, it also increases the number of different dialects included in the training database. We therefore modify two parts of our Tacotron TTS that use neural speaker embeddings to explicitly handle the two factors, channel and dialect, brought by low-quality data for speaker augmentation as shown in Figure 1.

Channel-aware postnet: The first revision is to make Tacotron's postnet dependent on channel information. Here, the channel means all of recording equipment, noise, and reverberation. We simply use a one-hot channel label that indicates which dataset each utterance comes from during training. This channel label is input to each convolution layer of the postnet, which controls shaping and enhancement of the spectrum predicted by Tacotron's decoder. Then, at synthesis time, we choose the highest-quality channel setting (VCTK) which will allow the model to produce speech with both a better speaker representation as well as high audio quality. This idea is relevant to [11, 12] wherein the channel factor is used to condition the decoder. In our idea, we view Tacotron as a speech production model and re-interpret its postnet as a channel model.

Dialect encoder network and neural dialect embeddings: The second revision is to make Tacotron's encoder dependent on the dialect of target speakers included in training and adaptation data.[2] We aim to use either a common phone set or character input for all speakers, and factorise Tacotron's encoder based on neural dialect embedding vectors, computed from audio signals. Dialect identification can be considered as a subtask of spoken language recognition, and in general, approaches from speaker recognition tasks can be directly transferred to dialect identification, see [24, 25, 26, 27]. Therefore, similar to our speaker encoder network, we reused the Learnable Dictionary Encoding (LDE) [28] based network architecture for our dialect encoder network. For more details, refer to Section 2 of [9].

[2]Here dialect means English varieties. In traditional TTS, a lexicon and phone set for different dialects are manually prepared [23].

Table 1: *Number of speakers in training, validation and test sets*

Models	Data type	Train	Dev	Test
Baseline	VCTK	100	4	4
Baseline +artificial	VCTK VTLP	100 200	4 8	4 -
Baseline +low-quality	VCTK non-TTS	100 200	4 8	4 -

3. Experiments

3.1. Setup

We use two baseline models, a phoneme-based model which is the same as the best system from our prior work [9], and one with character input. 4 speakers are held out as validation data and 4 speakers are held out as the test set. 80-dimensional mel spectrograms that are output from Tacotron are converted to 16 kHz waveforms using WaveNets [29] that were trained on the same VCTK training set. Details of the setup can be found in Section 3 and 4 of [9], and code is available online[3].

3.2. Artificial speaker augmentation

We created an augmented VCTK dataset by speeding up and slowing down the speech of each original VCTK speaker as described in Section 2.1 and giving them unique speaker identities, resulting in three times as many "speakers" as in the original dataset. Then, we trained both character-based and phone-based models in the same manner as our baselines except using the larger augmented dataset.

3.3. Speaker augmentation using low-quality ASR corpora

We create a large mixed dataset for TTS training using both VCTK and a variety of corpora collected for ASR, which contain a variety of recording conditions and English dialects. While we hold out some portion of each corpus for validation and test, we focus our actual evaluation on VCTK speakers. Once again we train both character-input and phone-input models. We used standard train/validation/test sets where they were defined, as well as predefined adaptation utterances or utterances that were common across speakers for extracting speaker embeddings. We kept the number of training speakers the same as in the artificially-augmented VCTK set. Two speakers were chosen per corpus to add to our development set for the purposes of preliminary model evaluation and selection. Below, we briefly describe the four ASR corpora used in our multi-speaker TTS training (with info about number of speakers in Table 1):

GRID [30]: This corpus consists of 32 English speakers (15 training set speakers) speaking English, Scottish, and Jamaican dialects. Sentences are all of the form "place green at B 4 now." While all sentences are technically unique, they are each very similar (many only varying by one word) and the vocabulary is small. There are 1000 utterances per speaker. Audio is 16 bit

[3]https://github.com/nii-yamagishilab/multi-speaker-tacotron

Table 2: *5 best dialect embeddings (DE) for phone- and character-based TTS. Number of dimensions, mean-only (m) or mean and standard deviation (m,s) pooling, and the number of dictionary components in the pooling layer are shown.*

	Phone			Char		
	dim	pl	dc	dim	pl	dc
DE1	256	m,s	32	128	m,s	32
DE2	256	m	64	256	m	32
DE3	256	m,s	64	32	m,s	64
DE4	32	m,s	64	512	m,s	32
DE5	64	m	64	64	m,s	32

and 50 kHz. Some recordings contain small amounts of background noise such as mouse clicks.

WSJ1 [31]: Wall Street Journal read by speakers of various American English dialects. Audio is 16 bit at a 16kHz sampling rate. We used the first 50 of the 200 'si_tr_s' training set speakers, who each have around 200 utterances.

WSJCAM [32]: Wall Street Journal sentences read by speakers of various British English dialects. Audio is 16 bit at a 16kHz sampling rate. We used 85 of the 96 training speakers, who each read about 110 sentences. Recordings contain loud audible line noise and reverberation.

TIMIT [33]: Speakers of eight American English dialects each read ten phonetically-rich sentences. We picked 50 of the 462 training speakers, balancing for gender and dialect. Audio is 16 bit at a 16kHz sampling rate.

3.4. Modeling channel and dialect factors

Ground-truth channel labels: In addition to training directly by mixing VCTK with the four new ASR corpora, we also trained phone and character models provided with ground-truth channel labels. We used a one-hot encoding indicating which corpus each training utterance comes from[4], and channel labels are input to the Tacotron postnet.

LDE-based neural dialect embeddings: Given our goal of modeling English dialects only, using the standard NIST LRE recipe is not ideal[5]. We opted to use the ATR dialect corpus[6] with six English dialects: Australian, British and various American English. Read and spontaneous speech recordings are sampled such that they are balanced for training. Our dialect encoder network is based on LDE, and we performed a hyper-parameter sweep. Similar to the speaker embeddings in [9], we computed the cosine-similarity scores between dialect embeddings of the synthesized and ground-truth speech[7], and accordingly selected five best embeddings *each* for phone and character models. Details of these embeddings are in Table 2.

Warm-start training strategy[8]: We adopted a warm-start training scheme, in which the full Tacotron training is broken down to four phases (see Figure 2) where the parameters in each phase are initialized from that of previous phase. In Phase 0, a seed single-speaker Tacotron2 is trained on the Nancy dataset from Blizzard 2011 [34]. In Phase 1, we trained a multi-speaker gender-dependent model on 5 corpora (VCTK + ASR),

[4]In addition to the one-hot code, we also tried a binary code simply representing TTS data (VCTK) or not (all other corpora), but this resulted in worse development set alignment error rates.

[5]https://github.com/kaldi-asr/kaldi/tree/master/egs/lre07

[6]https://www.atr-p.com/products/sdb.html

[7]We want to emphasize that our strategy is not optimal, and a strong assumption we imposed here is that the cosine-similarity and speaker/dialect distributions is a one-to-one mapping.

[8]We found this strategy effective, as it produces better synthesis quality and reduces training time.

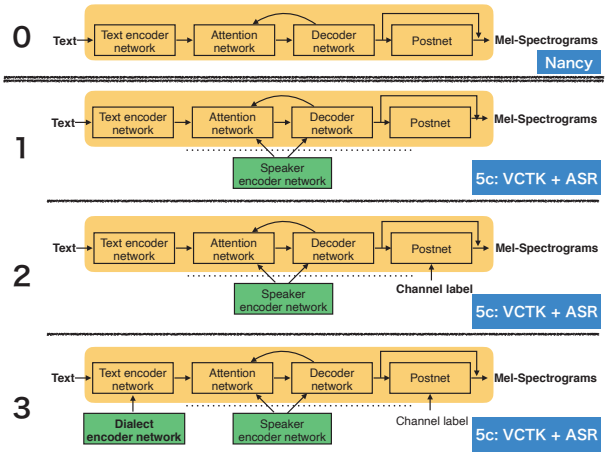

Figure 2: *Illustration of the warmstart training strategy for our Tacotrons (w/o a WaveNet) in this work.* ***Green*** *denotes the pre-trained components,* ***Yellow*** *denotes end-to-end training, and* ***Blue*** *denotes the training data.*

with parameters initialized from the previous step, and included speaker embeddings extracted from a separately-trained LDE model with mean pooling and angular softmax, trained on VoxCeleb [35, 36]. These embeddings are concatenated with the encoder output and input to the attention mechanisms, as well as input to the prenet to the decoder [9]. In Phase 2, we added channel labels. In Phase 3, finally, using one of the top 5 dialect embeddings for phone or character models, we continued training with all five corpora, channel labels, and speaker and dialect embeddings. Each phase is trained until convergence.

3.5. Subjective evaluation setup

We conducted a crowdsourced online listening test with native English listeners. We asked listeners to rate each sample on a mean opinion score (MOS) scale of 1-5 for naturalness and on a differential MOS (DMOS) scale of 1-5 for speaker similarity compared to a ground truth sample from the target speaker. We also asked listeners to provide a categorical opinion about dialect from six choices: American, Canadian, English, Irish, Northern Irish, and Scottish. Since listeners may be unfamiliar with these accents, we also provided reference samples of each accent from VCTK speakers who were not included in the test, on a separate webpage that listeners may optionally refer to. We evaluated 20 different systems: natural speech, vocoded speech using WaveNet, phone and character baselines, VTLP-augmented models, models trained with additional ASR data for a total of 5 training corpora (5c), models with 5c and channel label (CL), and models with 5c + CL + dialect embeddings (DE). For each system, we generated 20 samples using text that was unseen during training from each of 4 VCTK training set (seen) speakers, 4 development set speakers, and 4 test set speakers (completely unseen). We grouped samples into sets of of 40 utterances each, and had 5 different listeners evaluate each set. A total of 60 listeners completed the test, rating 10 sets each.

Metric for evaluating dialect confusion: Since dialect ID can be a challenging task even for native listeners, we evaluated confusion matrices of true vs. guessed accents. We computed Frobenius distance [37, 38] between the confusion matrix for dialects of natural speech and those for each TTS system, based on the idea that if a confusion matrix for TTS is similar to the one for natural speech, then accents are well-represented.

Table 3: *Results: MOS and DMOS on a scale of 1-5 for seen (train) and unseen (dev and test) speakers. Synthesis was done using unseen texts. Here, 5c denotes the 5 training corpora (VCTK + 4 ASR), CL denotes channel label, and DE{1..5} denotes the 5 best dialect embeddings for char and phone models. Significant improvements over the baseline are highlighted in red, and significantly worse systems are in blue.*

	Naturalness			Speaker Similarity		
system	train	dev	test	train	dev	test
natural	4.5	4.4	4.5	4.6	4.5	4.5
vocoded	4.2	4.2	4.3	4.1	3.9	4.0
phone baseline	3.6	3.7	4.0	3.7	2.0	2.7
phone VTLP	3.7	3.7	4.0	3.6	2.1	2.7
phone 5c	3.8	3.5	3.7	3.4	2.1	2.6
phone 5c+CL	3.7	3.7	3.9	3.5	2.0	2.5
phone 5c+CL+DE1	2.4	2.4	2.5	3.2	2.3	2.4
phone 5c+CL+DE2	2.5	2.4	2.5	3.3	2.2	2.6
phone 5c+CL+DE3	3.9	3.7	3.9	3.6	2.1	2.6
phone 5c+CL+DE4	1.1	1.1	1.1	2.8	2.4	2.5
phone 5c+CL+DE5	3.9	3.7	3.8	3.4	2.0	2.6
char baseline	3.7	3.7	3.9	3.6	1.9	2.8
char VTLP	3.8	3.6	3.9	3.5	2.1	2.5
char 5c	3.7	3.5	3.7	3.3	2.0	2.6
char 5c+CL	3.8	3.8	3.9	3.6	2.1	2.6
char 5c+CL+DE1	2.5	2.5	2.5	3.3	2.1	2.5
char 5c+CL+DE2	4.0	3.7	4.0	3.6	2.0	2.5
char 5c+CL+DE3	3.9	3.4	3.8	3.6	2.1	2.4
char 5c+CL+DE4	4.0	3.7	3.9	3.6	2.1	2.5
char 5c+CL+DE5	3.9	3.8	3.9	3.5	2.0	2.6

3.6. Subjective evaluation results and analysis

MOS and DMOS results are presented in Table 3. Statistical significances were measured using the Mann-Whitney U test at a threshold of p=0.01, and systems are compared with their *respective* baselines, i.e. phone or character. Significantly better and worse systems are highlighted in red and blue, respectively.

A) Baseline vs. speaker augmentation: The MOS and DMOS results show an unexpected but interesting tendency. Contrary to our initial expectation, we obtained statistically significant improvements for naturalness of *seen* speakers when the low-quality data, both channel-aware postnet, and dialect-aware encoder are all used for Tacotron training. This is surprising in two ways. First, speaker augmentation contributes to naturalness rather than speaker similarity. Second, adding low-quality data paradoxically resulted in improved quality of synthetic speech for seen speakers. This is somewhat surprising, but this phenomenon has been clearly confirmed for 2 phone-based systems and 4 character-based systems. MOS scores for the phone and character systems have been increased from 3.6 to 3.9 and from 3.7 to 4.0, respectively. Speaker similarity of the development set speakers was also improved from 2.0 to 2.4 for some of the phone-based systems. We may speculate that the addition of dialect modeling and a larger variety of different speakers helps to capture important aspects of speech, but that overfitting to speakers seen during training is still taking place.

B) Artificial vs. low-quality data: Next, we see that VTLP (artificial speaker augmentation) did not improve naturalness or speaker similarity, although it is known that this method works well in other tasks. On the other hand, mixing non-ideal data carelessly does worsen results: we see that simply mixing low-quality data produces significantly worse results in some cases, and that adding the channel-aware postnet only shows improvement when combined with dialect embeddings. This indicates that we need to handle both channel and dialect factors properly.

C) Impacts of dialect encoders: One implication from Table 3 is that the effect of different types of dialect encoders on synthesis is unclear and including them does not consistently improve naturalness and speaker similarity. However, they do appear to be necessary for better dialect modeling (see Table 4).

Table 4: *Frobenius distance results for dialects of seen (train) and unseen (dev and test) speakers, compared to confusion matrices for dialects of natural speech. Distances smaller than baseline are highlighted in red, with best result per category in bold. Distances larger than baseline are highlighted in blue.*

	Dialect confusion		
system	train	dev	test
vocoded	0.06	0.32	0.32
phone baseline	0.20	1.06	1.12
phone VTLP	0.31	0.86	1.20
phone 5c	0.19	0.93	0.93
phone 5c+CL	0.19	**0.84**	0.99
phone 5c+CL+DE1	0.42	0.84	0.88
phone 5c+CL+DE2	0.34	0.95	0.81
phone 5c+CL+DE3	**0.13**	0.93	0.95
phone 5c+CL+DE4	0.44	0.88	0.90
phone 5c+CL+DE5	0.20	0.92	**0.79**
char baseline	0.25	0.96	0.86
char VTLP	**0.12**	1.00	1.17
char 5c	0.25	0.96	0.86
char 5c+CL	0.25	**0.86**	**0.79**
char 5c+CL+DE1	0.41	0.91	0.83
char 5c+CL+DE2	0.17	0.92	1.33
char 5c+CL+DE3	0.18	0.92	1.02
char 5c+CL+DE4	0.21	0.91	1.15
char 5c+CL+DE5	0.22	1.02	1.08

D) Dialect identification and confusion: Frobenius distances representing confusions of perceived dialects are shown in Table 4. The Frobenius distance means how similar confusion matrices of perceived dialects of synthetic speech are compared to those of natural speech. We observed relative improvements for unseen speakers (dev and test). All phone-based systems using the low-quality data have smaller Frobenius distances than the baseline system for unseen speakers. This means adding low-quality data helps our synthesizers generate appropriate phones better and to better match dialects correctly with respect to listeners' perception. It also helps some of the character-based systems to use channel-aware postnet and the dialect-aware encoder. On the other hand, we see that unseen speakers (dev and test) have much larger Frobenius distances compared to seen speakers, even for vocoded speech. This tendency is consistent with speaker similarity judgements in Table 3.

4. Conclusions

In this paper we investigated two realistic speaker augmentation scenarios for multi-speaker end-to-end speech synthesis: artificial augmentation and the use of non-ideal low-quality data. We revised the postnet and encoder of Tacotron to support channel and dialect variations from the low-quality data. Experimental results revealed that using low-quality data with various English accents is an effective data augmentation method for multi-speaker end-to-end speech synthesis. Contrary to our initial expectations, naturalness of seen speakers has been improved and listeners' ratings of perceived dialects are better matched to natural speech for unseen speakers. Our results suggest that improving speaker similarity still remains a challenge, and future work includes the use of large low-quality databases for training an initial seed model and fine-tuning it to a high-quality corpus.

Acknowledgments CI is supported by the Merrill Lynch Fellowship, MIT. This work was partially supported by a JST CREST Grant (JPMJCR18A6, VoicePersonae project), Japan, and by MEXT KAKENHI Grants (16H06302, 18H04112, 18KT0051, 19K24372), Japan. The numerical calculations were carried out on the TSUBAME 3.0 supercomputer at the Tokyo Institute of Technology. We thank Yi Zhao for help with Frobenius distance, Jim Glass and Hao Tang for their discussions.

5. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, Z. Y. Jaitly, Y. Xiao, Z. Chen, S. Bengio, Q. Le *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *INTERSPEECH*, 2017.

[2] W. Ping, K. Peng, and J. Chen, "ClariNet: parallel wave generation in end-to-end text-to-speech," *ICLR*, 2018.

[3] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," *ICASSP*, 2018.

[4] J. Park, K. Zhao, K. Peng, and W. Ping, "Multi-speaker end-to-end speech synthesis," *arXiv preprint arXiv:1907.04462*.

[5] S. Arık, 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.

[6] Z. Kons, S. Shechtman, A. Sorin, C. Rabinovitz, and R. Hoory, "High quality, lightweight and adaptable TTS using LPCNet," *INTERSPEECH*, 2019.

[7] Y. Chen, Y. Assael, B. Shillingford, D. Budden, S. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie *et al.*, "Sample efficient adaptive text-to-speech," *ICLR*, 2019.

[8] 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.

[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," *ICASSP*, 2020.

[10] Q. Hu, E. Marchi, D. Winarsky, Y. Stylianou, D. Naik, and S. Kajarekar, "Neural text-to-speech adaptation from low quality public recordings," *Speech Synthesis Workshop 10*, 2019.

[11] W. Hsu, Y. Zhang, R. J. Weiss, Y. Chung, Y. Wang, Y. Wu, and J. Glass, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5901–5905.

[12] W.-N. Hsu, Y. Zhang, R. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://arxiv.org/pdf/1810.07217

[13] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," University of Edinburgh, The Centre for Speech Technology Research (CSTR), 2017.

[14] H. Yamamoto, K. A. Lee, K. Okabe, and T. Koshinaka, "Speaker augmentation and bandwidth extension for deep speaker embedding," *INTERSPEECH*, 2019.

[15] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013.

[16] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, 2015.

[17] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: a simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[18] 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)*, 2018, pp. 5329–5333.

[19] P.-c. Hsu, C.-h. Wang, A. T. Liu, and H.-y. Lee, "Towards robust neural vocoding for speech generation: A survey," *arXiv preprint arXiv:1912.02461*, 2019.

[20] Y. Hwang, H. Cho, H. Yang, I. Oh, and S.-W. Lee, "Mel-spectrogram augmentation for sequence to sequence voice conversion," *arXiv preprint arXiv:2001.01401*, 2020.

[21] Sox - sound exchange. [Online]. Available: http://sox.sourceforge.net/

[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," in *Proc. Interspeech 2019*, 2019, pp. 1526–1530. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2441

[23] K. Richmond, R. A. J. Clark, and S. Fitt, "On generating Combilex pronunciations via morphological analysis," in *Proc. Interspeech 2010*, 2010, pp. 1974–1977.

[24] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Twelfth annual conference of the international speech communication association*, 2011.

[25] D. Garcia-Romero and A. McCree, "Stacked long-term TDNN for spoken language recognition." in *INTERSPEECH*, 2016, pp. 3226–3230.

[26] W. Cai, Z. Cai, X. Zhang, X. Wang, and M. Li, "A novel learnable dictionary encoding layer for end-to-end language identification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5189–5193.

[27] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, D. Povey, and S. Khudanpur, "Spoken language recognition using x-vectors." in *Odyssey*, 2018, pp. 105–111.

[28] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," *Odyssey 2018, The Speaker and Language Recognition Workshop*, 2018.

[29] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: a generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, 2016.

[30] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[31] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," *Proc. Workshop Speech Natural Lang.*, 1992.

[32] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAM0: a British English speech corpus for large vocabulary continuous speech recognition," *ICASSP*, 1995.

[33] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1," 1993.

[34] S. King and V. Karaiskos, "The Blizzard Challenge 2011," in *Blizzard Challenge Workshop*, 2011.

[35] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[36] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[37] A. Amendola and G. Storti, "Model uncertainty and forecast combination in high-dimensional multivariate volatility prediction," *Journal of Forecasting*, vol. 34, no. 2, pp. 83–91, 2015.

[38] S. Laurent, J. V. Rombouts, and F. Violante, "On the forecasting accuracy of multivariate garch models," *Journal of Applied Econometrics*, vol. 27, no. 6, pp. 934–955, 2012.