

SPEAKING STYLE ADAPTATION IN TEXT-TO-SPEECH SYNTHESIS USING SEQUENCE-TO-SEQUENCE MODELS WITH ATTENTION

Bajibabu Bollepalli Lauri Juvela Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Finland

ABSTRACT

Currently, there are increasing interests in text-to-speech (TTS) synthesis to use sequence-to-sequence models with attention. These models are end-to-end meaning that they learn both co-articulation and duration properties directly from text and speech. Since these models are entirely data-driven, they need large amounts of data to generate synthetic speech with good quality. However, in challenging speaking styles, such as Lombard speech, it is difficult to record sufficiently large speech corpora. Therefore, in this study we propose a transfer learning method to adapt a sequence-to-sequence based TTS system of normal speaking style to Lombard style. Moreover, we experiment with a WaveNet vocoder in synthesis of Lombard speech. We conducted subjective evaluations to assess the performance of the adapted TTS systems. The subjective evaluation results indicated that an adaptation system with the WaveNet vocoder clearly outperformed the conventional deep neural network based TTS system in synthesis of Lombard speech.

Index Terms— TTS, Tacotron, Lombard style, Adaptation

1. INTRODUCTION

For better communication, humans typically change their speaking style depending upon the acoustic and auditory environment. In noisy surroundings, humans adapt to Lombard style [1] in order to improve speech intelligibility. When compared to speech of normal style, Lombard speech differs in many respects both acoustically and phonetically by exhibiting, for example, an increase in vocal intensity and fundamental frequency (F_0), a decrease in spectral tilt, and an increased prominence in the production of vowels compared to consonants [2, 3].

In literature, it has been shown that intelligibility of synthetic normal speech is significantly lower than that of synthetic Lombard speech when evaluated in noisy surroundings [4]. Use of text-to-speech (TTS) systems, however, is becoming increasingly prevalent in real-life noisy environments such as in GPS navigation in busy traffic or public announcement systems in train and bus stations. Thus there is a great need for TTS systems to improve their speech intelligibility in noisy conditions by adapting the speaking style of the synthesis to Lombard. In this article, we investigate a transfer learning technique to adapt a sequence-to-sequence based normal speaking style TTS system to Lombard style¹.

2. RELATED WORK

In TTS, there are many studies on the adaptation of the speaking style, including Lombard, e.g., [5, 6]. These previous studies are

almost exclusively based on statistical parametric speech synthesis (SPSS) due to the technology's benefits in adaptation abilities and flexibility in changing voice characteristics. SPSS systems typically require a moderate amount of high-quality audio data per speaker to generate synthetic speech with good quality (e.g. 5 hours in [7]). However, collecting several hours of speech data from one speaker is difficult, if not impossible, for high vocal effort speaking styles such as shouting and Lombard. To circumvent the data scarcity issue, adaptation techniques are usually employed in SPSS by allowing the use of a smaller amount of data for the particular speaking style to be synthesized. Speaking style adaptation of TTS was first used in hidden Markov model (HMM)-based SPSS systems. In adapting HMM-based systems, initial HMMs, trained on normal speech, were adapted by a small amount of Lombard speech using, for example, a technique called constrained structural maximum *a posteriori* linear regression (CSMAPLR) [5]. In the more recent deep neural network (DNN) based SPSS systems, the adaptation can be done at three levels: 1) input level, 2) model level, and 3) output level [8, 9, 10, 11]. Previous studies have demonstrated that the naturalness of synthetic speech generated with DNN-based SPSS systems is higher than that of HMM-based systems [12, 13], and this applies also to adaptation to Lombard speech [14].

Even though promising results have been obtained in adaptation of synthetic speech using the SPSS framework, this conventional TTS paradigm has limitations that affect the synthesis naturalness. Conventional SPSS systems consists of two separate blocks, 1) front-end and 2) back-end. The front-end processes the input text by producing the numerical, linguistic representation [15]. The back-end first maps the linguistic representation to acoustic features and then passes the mapped acoustic features to a vocoder to render the speech waveform. In this pipeline, both the front-end and back-end are usually constructed independently [15]. Moreover, errors caused in each block can accumulate and degrade the overall performance of the system. Further, each block needs its own expertise to tune the system.

Recently, a more simplified framework using sequence-to-sequence models with attention was proposed for TTS [16, 17, 18]. These models depend heavily on encoder-decoder neural network structures that map a sequence of characters to a sequence of acoustic frames. These models combine the front-end and back-end and learn relations between them from data only. When sequence-to-sequence models are coupled with neural vocoders, they enable generating raw waveforms directly from text [19]. In [20], it was demonstrated that state-of-the-art results in TTS can be achieved with the sequence-to-sequence technology. Despite their success in producing high-quality synthetic speech, sequence-to-sequence systems, however, need a sizable amount of data (i.e. text/audio pairs). In [21], for example, it was concluded that around 10 hours of text/speech pairs are needed to get decent quality in synthetic speech by a sequence-to-sequence model such as Tacotron [17]. From now

¹Samples available at http://tts.org.aalto.fi/lombard_seq2seq/

on, we denote sequence-to-sequence model based TTS as Seq2Seq-TTS.

In Seq2Seq-TTS, only a few investigations have studied methods to synthesize speech of different speakers using a limited amount of data [22, 23]. These studies have employed speaker embeddings, which contain speaker-specific characteristics for multi-speaker speech synthesis. However, extracting the speaker embeddings for unseen speakers in training data may require a huge amount of data to train a separate speaker-encoder network [24]. However, to learn *style-specific* embeddings for challenging speaking styles does not call for having that much data. Thus, in this study we propose a method to effectively leverage an existing large volume of normal speech data to synthesize Lombard speech using a Seq2Seq-TTS system. Our study is to some extent similar to [21] that recently proposed a semi-supervised technique to reduce the data requirements by utilizing freely available data.

The contributions of the paper are twofold. First, we develop a speaking style adaptation system using a Seq2Seq-TTS model. Second, we study the use of a WaveNet vocoder [25] for the application of Lombard speech synthesis. To the best of our knowledge, the current study is the first investigation on speaking style adaptation of speech synthesis using a modern Seq2Seq-TTS system.

3. SEQ2SEQ-TTS SYSTEM

Figure 1 depicts a general block diagram of a TTS system using the sequence-to-sequence model with attention. The model accepts either mono-phonemes or graphemes as inputs and emits acoustic parameters as outputs. It consists of three main components: 1) encoder, 2) attention, and 3) decoder. The encoder takes text sequence \mathbf{x} of length L as input, which represented either in the character or phoneme domain as one-hot vectors. The encoder learns a continuous sequential representation \mathbf{h} using various neural network architectures such as LSTMs [17, 20] and/or CNNs [22].

$$\mathbf{h} = \text{encoder}(\mathbf{x}) \quad (1)$$

At each output time step t , both the attention and decoder modules work together in the following manner:

$$\alpha_t = \text{attention}(s_{t-1}, \alpha_{t-1}, \mathbf{h}) \quad (2)$$

$$c_t = \sum_{j=1}^L \alpha_{t,j} h_j \quad (3)$$

$$y_t = \text{decoder}(s_{t-1}, c_t) \quad (4)$$

where s_{t-1} is the $(t-1)$ -th state of the decoder recurrent neural network and $\alpha_t \in \mathbb{R}^L$ are the attention weights or the alignment and c_t is the context or attention vector. The decoder takes the previous hidden state s_{t-1} and the current context vector c_t as inputs and generates the current output y_t . This process runs until the end of the utterance is reached

In order to synthesize the speech waveform, Seq2Seq-TTS systems use different vocoding approaches. Initial studies predict mel-spectrograms as output, mapping them to linear spectrograms and further to speech waveforms using the Griffin-Lim algorithm [17]. Recent studies, however, generate speech waveforms with the neural WaveNet vocoder, which is conditioned using the predicted mel-spectrograms [20]. In this study, we predict the World vocoder [26] parameters as well as mel-spectrograms as the system outputs, which are later used in conditioning the WaveNet vocoder to generate the final speech waveform.

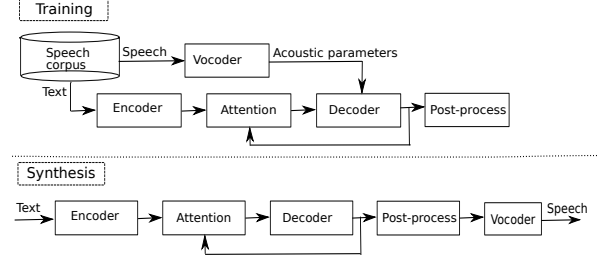


Fig. 1. General block diagram of a sequence-to-sequence based TTS system.

3.1. Adaptation of Seq2Seq models

Previous studies [22, 27] have showed that Seq2Seq models can be adapted to different speakers by adding the speaker embeddings as an input to the model. In addition, some TTS studies [24, 23] have used the fine-tuning technique to adapt the system to new speakers. In this approach, all the parameters of the network are fine-tuned by the new speaker data. In [21], which is the study closest to the current investigation, each component of the model is trained separately with freely available data. Once the encoder and decoder are pre-trained with existing data, a fine-tuning method is used to train the whole model on a small data set. Results of [21] show that the model can generate synthetic speech of good quality using around 30 minutes of data. In the present study, we first train a Seq2Seq-TTS system using a large amount of normal speech of one speaker and then fine-tune the learned model with normal speech of another speaker with limited data. Finally, using Lombard speech of the latter speaker, we fine-tune the model again to synthesize Lombard speech. We predict both mel-spectrograms and the World vocoder parameters as output acoustic frames. To render final speech waveform, we employ both the WaveNet vocoder and the World vocoder.

4. EXPERIMENTS

4.1. Speech material

Our initial Seq2Seq-TTS model was trained on the Blizzard Challenge 2011 speech corpus [28]. The corpus contains around 12000 utterances (which adds up to around 16 hours) read by a US professional female voice talent named Nancy. We employed the Hurricane Challenge speech data [29] for adaptation to Lombard style. The Hurricane Challenge data was spoken by a British male voice professional named Nick. The Nick data consists of both normal and Lombard styles. The normal speech data consists of 2592 utterances (which adds up to 2 hours), and the Lombard speech data consists of 720 utterances (which adds up to 30 minutes). All the speech data was sampled with 16 kHz. The data was partitioned into train, valid and test sets as shown in Table 1.

Table 1. Partition of the data (number of utterances) used in the present study.

Speaker (Style)	Train	Valid	Test
Nancy (normal)	11000	200	800
Nick (normal)	2400	72	120
Nick (Lombard)	500	100	120

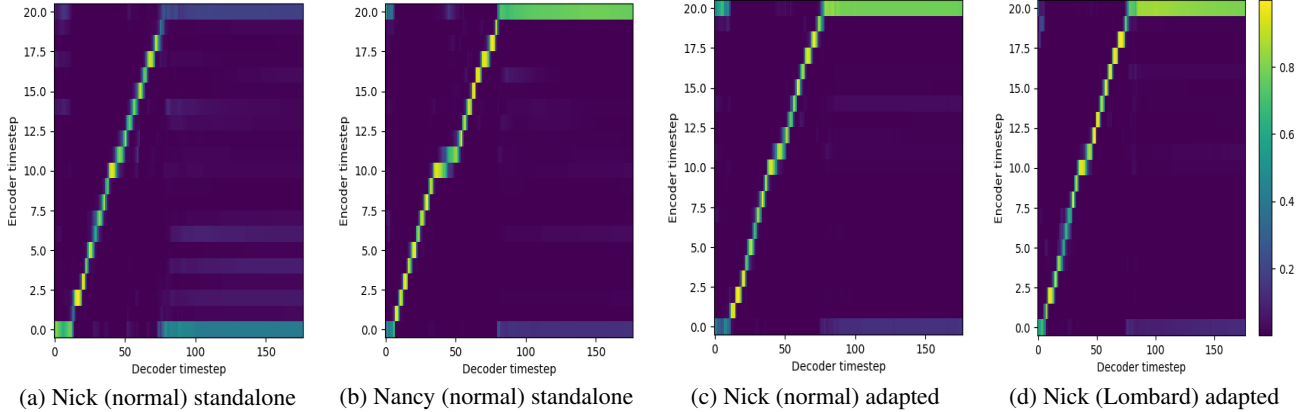


Fig. 2. Alignments obtained by the different systems, the sentence is PAPER WILL DRY OUT WHEN WET. The x-axis and y-axis of each plot correspond to the mel-spectrograms of speech signal and the phonemes of the text, respectively.

4.2. Systems built for comparison

Two types of acoustic parameters were extracted from the speech signals: 1) World vocoder parameters, and 2) mel-spectrograms. The World vocoder parameters consisted of mel-generalized cepstrum (MGC), fundamental frequency (F_0) and band aperiodicity (BAP) with the dimensions 60, 1, and 1, respectively. The mel-spectrograms were extracted using the LibROSA² package and its dimension was 80. The World vocoder parameters were extracted at a 5-ms frame rate whereas the mel-spectrogram features were extracted at a 12.5-ms frame rate. The F_0 values were linearly interpolated in unvoiced regions and transformed into the \log -domain.

We built a total of five systems for comparison as described in Table 2. The systems were different in terms of their output parameter types and the vocoder used. System S1 is the baseline system which uses a LSTM-type of recurrent neural network (RNN)-based TTS system for adaptation, and synthesizes the speech waveform using the World vocoder. System S2 is built using the Seq2Seq-TTS model, and the final waveform is rendered by the World vocoder. Systems S3 and S4 have same architectures as systems S1 and S2, respectively, but they use the WaveNet vocoder for synthesis. System S5 has the same architecture and vocoder as S4, but instead of using the World vocoder parameters, it predicts the mel-spectrogram as the output.

Table 2. Systems developed for experiments.

Sys. ID	TTS model	Ouput	Vocoder
S1	LSTM	MGC+ F_0 +BAP+VUV	World
S2	Seq2Seq	"	"
S3	LSTM	"	WaveNet
S4	Seq2Seq	"	"
S5	"	Mel-spectrogram	"

The baseline S1 system was built as reported in our previous study [14]. We used the fine-tuning method to adapt a LSTM-RNN based TTS system of normal speaking style to Lombard style because this adaptation method showed the best performance. Our previous work used original durations to synthesize Lombard speech. In the current study, however, a separate duration model is built and adapted to Lombard speech. Our Seq2Seq-TTS system is based on the Tacotron-1 architecture [17] with a few modifications such

as predicting the World vocoder parameters instead of the mel-spectrograms as output. Our systems were implemented using an open source repository³. All models were trained on a single GPU with a Nvidia-TITANX 12GB graphics memory card. Before training the model, we processed the data to have even distribution in durations. Because a few utterances are of very long duration and they can slow down training and consume lots of memory, these utterances were removed. The batch size was 32, and 2 acoustic frames were used per each output step. The input linguistic features were mono-phonemes extracted using the Combilex lexicon [30] and represented by one-hot vectors. All acoustic parameters were normalized to have zero mean and unit variance using the standard mean-variance normalization. Linguistic parameters were normalized to lie between 0 and 1 using the min-max normalization.

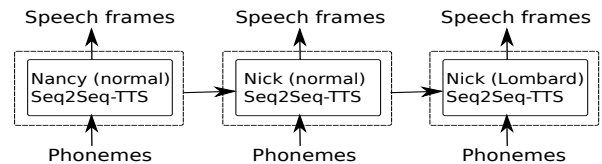


Fig. 3. Flow diagram of the adaptation approach.

Figure 3 illustrates the flow diagram of our adaptation procedure. First, we trained a Seq2Seq-TTS model on the Nancy data of normal style, later that model is adapted to the Nick data of normal style. Then, the Nick normal speech Seq2Seq-TTS model is adapted to Lombard style using the Lombard data of Nick. As seen in Figure 2(a), when we trained the Seq2Seq-TTS model using only the Nick data of normal style (i.e., approx. 2 hours of speech), the alignment between the input phoneme sequences and output acoustic frames is not as clear as in Figure 2(c), which was obtained by adapting the Nick normal speech data using the Nancy Seq2Seq-TTS model. In informal listening tests, pronunciation errors were perceived when we trained the Seq2Seq-TTS model on the Nick data only; this was most likely because the model was unable to learn adequately with little data. Thus we decided to train the initial model using the Nancy data (i.e., approx. 16 hours of speech) to learn a good alignment between input phoneme sequences and output acoustic frames.

²<https://github.com/librosa/librosa>

³<https://github.com/syang1993/gst-tacotron>

4.2.1. WaveNet vocoder

We use a WaveNet configuration similar to [31], i.e., three repetitions of a 10-layer convolution stack with exponentially growing dilations, 64 residual channels, and 128 skip channels. Separate models were trained for World vocoder acoustic features and mel-spectrograms, using 8-bit categorical cross entropy on quantized μ -law companded signals. We found that excluding BAP from the World features improved performance, so the WaveNet vocoder for World only uses MGCs, VUV and $\log F_0$ (interpolated over unvoiced frames). Both the World features and the mel-spectrograms were globally min-max normalized to lie between zero and one.

4.3. Subjective evaluation

Two types of listening tests were conducted: 1) speaking style similarity test and 2) comparison category rating (CCR) test of speech naturalness. The goal of the similarity test is to assess whether the technology developed is capable of generating synthetic speech of different speaking styles (normal vs. Lombard) while the CCR test aims to evaluate how much the naturalness of speech is sacrificed when the speaking style is adapted. We used an evaluation setup similar to [32] for the style similarity test. In this evaluation, each stimulus consists of two utterances, the first being a natural speech signal (either normal or Lombard) and the second one a synthesized signal. The subjects were asked to compare the second utterance to the first one and rate the style similarity on a 4-level scale, ranging from 0 (*Same: Absolutely sure*) to 4 (*Different: Absolutely sure*) [32]. In the CCR test, each stimulus consists of a pair of utterances which were stitched together with a silence of 0.5 seconds between them. Subjects were asked to evaluate the naturalness of the second utterance in comparison to the naturalness of the first utterance on a 7-level scale, ranging from -3 (*First sample sounds much more natural*) to 3 (*Second sample sounds much more natural*).

Both tests were conducted on FigureEight⁴ a crowdsourcing platform (see [31] for more details of conducting the tests). We selected 16 utterances randomly from the test set for each system. Each utterance was evaluated by 50 listeners, and the listeners were screened using natural reference null pairs and artificially corrupted anchor samples.

4.4. Results

The results of the style similarity test are plotted in Figure 4. From the right pane of the figure, it can be observed that synthesized speech by all adapted systems were rated to sound different from natural normal speech with high confidence. When compared to the natural Lombard reference (left pane), system S5 was rated highest, followed by systems S3, S4, S2, and S1. System S5 was built using the Seq2Seq-TTS model and the mel-spectrogram as output. It can be clearly seen that those systems that employed the WaveNet vocoder got higher scores than the ones that used the more traditional World vocoder. Further, system S5, which is based on conditioning the WaveNet vocoder with mel-spectrograms, got a higher score than the ones that used the World vocoder, confirming the findings made in an earlier study [33]. From this results we can conclude that even though we trained the WaveNet vocoder with only 30 minutes of Lombard speech, system S5 generated synthetic speech that was most Lombard-like among the systems compared.

For the CCR test, we included only systems S1, S3 and S5. System S1 can be regarded as the baseline. System S3 was selected

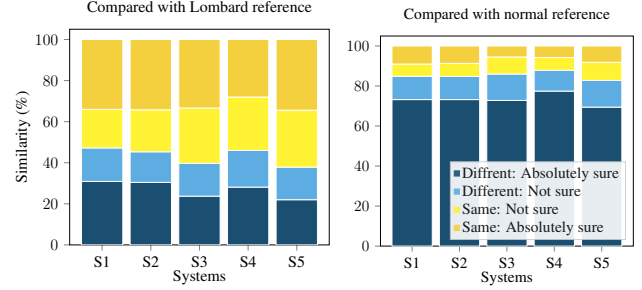


Fig. 4. Results of the style similarity test.

because it was the best system in the similarity test with the LSTM models and the WaveNet vocoder. S5 was selected because it was the best system in overall in the similarity test. The results of the CCR test are shown in Figure 5. The scores were calculated by re-ordering the ratings for each system and pooling together all ratings the system received. Natural Lombard speech was included in the tests as a reference system. The plot shows mean ratings with 95% confidence, corrected for multiple comparisons. As expected, the Lombard reference signal was rated highest followed by S5, S3 and S1. System S5 got a significant better score than the baseline system S1 and the LSTM based system S3.

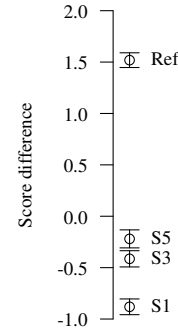


Fig. 5. Combined score differences obtained from the CCR test of naturalness. Error bars are t-statistic based 95% confidence intervals for the mean.

5. CONCLUSIONS

This paper compared different TTS models and vocoders to adapt the speaking style of speech synthesis from normal to Lombard. The study proposes using an adaptation method based on fine-tuning combined with sequence-to-sequence based TTS models and the WaveNet vocoder conditioned using mel-spectrograms. Listening tests show that the proposed method outperformed the previous best method that was developed using a LSTM-RNN based adapted system. Future work includes an extensive subjective evaluations and training both the WaveNet and Seq2Seq-TTS model in a single pipeline.

6. ACKNOWLEDGEMENTS

The study was funded by the Academy of Finland (project 312490). We thank Vassilis Tsiaras for sharing his WaveNet implementation.

⁴<https://www.figure-eight.com/>

7. REFERENCES

- [1] E. Lombard, “Le signe d’élévation de la voix [the sign of the elevation of the voice],” *Annales des maladies de l’oreille et du larynx*, vol. 37, pp. 101–119, 1911.
- [2] J.-C. Junqua, “The Lombard reflex and its role on human listeners and automatic speech recognizers,” *The Journal of the Acoustical Society of America*, vol. 93, no. 1, pp. 510–524, 1993.
- [3] Y. Lu and M. Cooke, “Speech production modifications produced by competing talkers, babble, and stationary noise,” *The Journal of the Acoustical Society of America*, vol. 124, no. 5, pp. 3261–3275, 2008.
- [4] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, “Evaluating the intelligibility benefit of speech modifications in known noise conditions,” *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.
- [5] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm,” *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 17, no. 1, pp. 66–83, 2009.
- [6] T. Raitio, A. Suni, M. Vainio, and P. Alku, “Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise,” *Computer Speech & Language*, vol. 28, no. 2, pp. 648–664, 2014.
- [7] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, “On the training aspects of deep neural network (DNN) for parametric tts synthesis,” in *Proc. ICASSP*, 2014, pp. 3829–3833.
- [8] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, “A study of speaker adaptation for DNN-based speech synthesis,” in *Proc. Interspeech*, 2015.
- [9] H.-T. Luong and J. Yamagishi, “Multimodal speech synthesis architecture for unsupervised speaker adaptation,” in *Proc. Interspeech*, 2018, pp. 2494–2498.
- [10] Z. Huang, H. Lu, M. Lei, and Z. Yan, “Linear networks based speaker adaptation for speech synthesis,” in *Proc. ICASSP*, 2018, pp. 5319–5323.
- [11] X. Wu, L. Sun, S. Kang, S. Liu, Z. Wu, X. Liu, and H. Meng, “Feature based adaptation for speaking style synthesis,” in *Proc. ICASSP*, 2018, pp. 5304–5308.
- [12] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *Proc. ICASSP*, 2013, pp. 7962–7966.
- [13] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, “TTS synthesis with bidirectional LSTM based recurrent neural networks,” in *Proc. Interspeech*, 2014, pp. 1964–1968.
- [14] B. Bollepalli, M. Airaksinen, and P. Alku, “Lombard speech synthesis using long short-term memory recurrent neural networks,” in *Proc. ICASSP*, 2017, pp. 5505–5509.
- [15] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.
- [16] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, “Char2wav: End-to-end speech synthesis,” *Proc. International Conference on Learning Representations (ICLR)*, 2017.
- [17] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly *et al.*, “Tacotron: Towards end-to-end speech synthesis,” *arXiv preprint arXiv:1703.10135*, 2017.
- [18] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang *et al.*, “Deep voice: Real-time neural text-to-speech,” *arXiv preprint arXiv:1702.07825*, 2017.
- [19] W. Ping, K. Peng, and J. Chen, “Clarinet: Parallel wave generation in end-to-end text-to-speech,” *arXiv preprint arXiv:1807.07281*, 2018.
- [20] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang *et al.*, “Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions,” in *Proc. ICASSP*, 2018, pp. 4779–4783.
- [21] Y.-A. Chung, Y. Wang, W.-N. Hsu, Y. Zhang, and R. Skerry-Ryan, “Semi-supervised training for improving data efficiency in end-to-end speech synthesis,” *arXiv preprint arXiv:1808.10128*, 2018.
- [22] 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. International Conference on Learning Representations (ICLR)*, 2018.
- [23] S. O. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, “Neural voice cloning with a few samples,” *arXiv preprint arXiv:1802.06006*, 2018.
- [24] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren *et al.*, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” *arXiv preprint arXiv:1806.04558*, 2018.
- [25] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, “Speaker-dependent WaveNet vocoder,” in *Proc. Interspeech*, 2017, pp. 1118–1122.
- [26] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.
- [27] E. Nachmani, A. Polyak, Y. Taigman, and L. Wolf, “Fitting new speakers based on a short untranscribed sample,” *arXiv preprint arXiv:1802.06984*, 2018.
- [28] S. King and V. Karaiskos, “The Blizzard Challenge 2011.”
- [29] M. Cooke, C. Mayo, and C. Valentini-Botinhao, “Hurricane natural speech corpus,” [sound], 2013.
- [30] K. Richmond, R. Clark, and S. Fitt, “On generating combilex pronunciations via morphological analysis,” in *Proc. Interspeech*, 2010, pp. 1974–1977.
- [31] L. Juvela, V. Tsias, B. Bollepalli, M. Airaksinen, J. Yamagishi, and P. Alku, “Speaker-independent raw waveform model for glottal excitation,” in *Proc. Interspeech*, 2018, pp. 2012–2016.
- [32] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, “The voice conversion challenge 2016,” in *Proc. Interspeech*, 2016, pp. 1632–1636.
- [33] N. Adiga, V. Tsias, and Y. Stylianou, “On the use of WaveNet as a statistical vocoder,” in *Proc. ICASSP*, 2018, pp. 5519–5523.