Bollepalli, Bajibabu; Juvela, Lauri; Alku, Paavo

Lombard speech synthesis using transfer learning in a Tacotron text-to-speech system

Published in:
Proceedings of Interspeech

DOI:
10.21437/Interspeech.2019-1333

Published: 01/01/2019

Document Version
Publisher's PDF, also known as Version of record

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

Lombard Speech Synthesis using Transfer Learning in a Tacotron Text-to-Speech System

Bajibabu Bollepalli, Lauri Juvela, Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Finland
firstname.lastname@aalto.fi

Abstract

Currently, there is increasing interest to use sequence-to-sequence models in text-to-speech (TTS) synthesis with attention like that in Tacotron models. 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 of good quality. However, in challenging speaking styles, such as Lombard speech, it is difficult to record sufficiently large speech corpora. Therefore, we propose a transfer learning method to adapt a TTS system of normal speaking style to Lombard style. We also experiment with a WaveNet vocoder along with a traditional vocoder (WORLD) in the synthesis of Lombard speech. The subjective and objective evaluation results indicated that the proposed adaptation system coupled with the WaveNet vocoder clearly outperformed the conventional deep neural network based TTS system in the synthesis of Lombard speech.

Index Terms: Text-To-Speech (TTS), Tacotron, Lombard speaking style, Adaptation

1. Introduction

Text-to-speech (TTS) systems are becoming more and more ubiquitous after the proliferation of personal voice assistants such as Amazon Echo, Google Home, and Apple Siri. These devices are usually employed in real-life noisy environments where the intelligibility of synthetic speech can be affected. Humans typically change their speaking style depending upon the acoustic environment for better communication. In noisy surroundings, humans adapt to *Lombard style* [1] in order to improve speech intelligibility. In literature, it has been shown that the intelligibility of synthetic normal speech is significantly lower than that of synthetic Lombard speech when evaluated in noisy surroundings [2]. Thus, TTS systems should be able to be aware of the noisiness of the environment and adapt their speaking style to Lombard style to improve intelligibility.

Speaking style adaptation, including adaptation to Lombard speech, has been studied in TTS [3, 4]. These previous studies are almost exclusively based on hidden Markov model (HMM)-based statistical parametric speech synthesis (SPSS) due to its benefits in adaptation abilities and flexibility in changing voice characteristics. In HMM-based SPSS systems, adaptation can be done by adapting the initial HMMs which are trained on normal speech with a small amount of Lombard speech [3]. In more recent deep neural network (DNN)-based SPSS systems, adaptation can be done at three levels: 1) input level, 2) model level and 3) output level [5–8]. Previous studies have demonstrated that the naturalness of synthetic speech generated with DNN-based SPSS systems is higher than that of HMM-based systems [9, 10], and this also applies to adaptation to Lombard speech [11].

Even though promising results have been obtained in the adaptation of synthetic speech using the SPSS framework, this conventional TTS paradigm has limitations that affect the synthesis's naturalness. Conventional SPSS systems consist of two separate blocks: 1) the front-end and 2) the back-end. In this pipeline, both the front-end and back-end are usually constructed independently [12]. Moreover, errors caused in each block can accumulate and degrade the overall performance of the system. Furthermore, each block needs its own expertise to tune the system.

Recently, a more uniform framework using sequence-to-sequence (Seq2Seq) models with attention was proposed for TTS [13–15]. These models combine the front-end and back-end and only learn the relations between them from data. When Seq2Seq models are coupled with neural vocoders, they enable generating raw waveforms directly from text [16]. In [17], it was demonstrated that state-of-the-art results in TTS can be achieved with the Seq2Seq models. However, despite their success in producing high-quality synthetic speech, Seq2Seq-TTS systems need a sizable amount of data (i.e. text and audio pairs). In [18], it was concluded that around 10 hours of text and speech pairs are needed to get decent quality in synthetic speech using a Seq2Seq model such as Tacotron [14]. 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 the Lombard style. However, to address the data scarcity issue, the *transfer learning* (TL) approach can be used to leverage a large volume of available data.

In [19, 20], a Seq2Seq-TTS system was studied to synthesize the speech of different speakers using a limited amount of data. These studies employed speaker embeddings, which contain speaker-specific characteristics for multispeaker speech synthesis. However, extracting the speaker embeddings for unseen speakers in training data may require a huge amount of data in order to train a separate speaker-encoder network [21]. However, to learn style-specific embeddings for challenging speaking styles does not call for having that much data. Hence, in this study, we propose a method to effectively leverage an existing large volume of normal speech data in order to synthesize Lombard speech using a Seq2Seq-TTS system. The contributions of the paper are twofold. First, we develop a Lombard speaking style adaptation system with a little amount of data by utilizing a TL technique in a Seq2Seq-TTS system. Second, we investigate the effect of using a WaveNet vocoder [22] in Lombard speech synthesis. To the best of our knowledge, the current study is the first investigation of the adaptation of speech synthesis to Lombard style using a modern Seq2Seq-TTS system.

2. The Seq2Seq-TTS system

Seq2Seq models depend heavily on encoder–decoder neural network structures that map a sequence of characters to a se-

http://dx.doi.org/10.21437/Interspeech.2019-1333

quence of acoustic frames. The Seq2Seq-TTS system consists of three main components: 1) the encoder, 2) attention, and 3) the decoder. The encoder takes the text sequence $\boldsymbol{x}$ of length L as an input, which is represented either in the character or phoneme domain as one-hot vectors. The encoder learns a continuous sequential representation $\boldsymbol{h}$ using various neural network architectures such as long short-term memory (LSTM) recurrent neural networks [14, 17] and/or convolutional neural networks (CNNs) [19]:

$$\boldsymbol{h} = \text{encoder}(\boldsymbol{x}). \tag{1}$$

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

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

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

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

where s_{t-1} is the $(t-1)$-th state of the decoder recurrent neural network, $\alpha_t \in \mathbb{R}^L$ 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 [14]. Recent studies, however, generate speech waveforms with the neural WaveNet vocoder, which is conditioned using the predicted mel-spectrograms [17]. In the current study, we predict the WORLD vocoder [23] 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.

3. Adaptation of the Seq2Seq-TTS system by TL

TL is an important and extremely useful framework in machine learning [24]. Let us suppose there are two tasks, a (1) source task and (2) target task, by assuming that they are related. TL can be applied to improve the learning of the target task by utilizing the knowledge learned from the source task. In particular, TL helps the learning significantly when the data of the target task is scarce. In small data conditions, training a new model on a small amount of data might not lead to good generalization. However, the knowledge from the source task which is trained on a large dataset could be very useful. This kind of approach has been widely used in a large number of machine learning tasks [25, 26]. TL can be used in many ways in deep learning. However, two popular approaches are (a) to *fine-tune* the source network for the target task [18, 27, 28] and (b) to learn feature representations using the source network for the target task [29].

In the present study, we propose a method to effectively transfer knowledge from a Seq2Seq-TTS system trained on a large amount of speech of a normal speaking style. As described in Figure 1, the method uses TL with fine-tuning in two steps. We first trained a Seq2Seq-TTS system on the normal speech of a female speaker (called Nancy). Then, we fine-tuned the learned model with the normal speech of a male speaker (called Nick) with limited data. Finally, using Lombard speech of Nick,

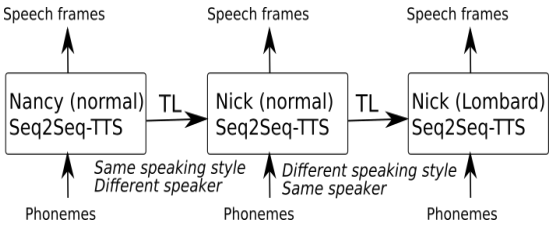


Figure 1: *A flow diagram of the proposed adaptation approach.*

we fine-tuned the model again to generate synthetic Lombard speech. Since success of the TL technique depends on the similarity between the source and target tasks, we used the approach shown in Figure 1 instead of adapting Nancy (normal) directly to Nick (Lombard).

4. Experiments

4.1. Speech material

Our initial Seq2Seq-TTS model was trained on the Blizzard Challenge 2011 speech corpus [30]. The corpus contains around 12,000 utterances (which add up to around 16 hours of speech) read in a normal speaking style by a US professional female voice talent named Nancy. We employed the Hurricane Challenge speech data [31] 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 add up to 2 hours of speech), and the Lombard speech data consists of 720 utterances (which add up to 30 minutes of speech). All the data was sampled 16 kHz. The data was partitioned into train, valid and test sets as shown in Table 1.

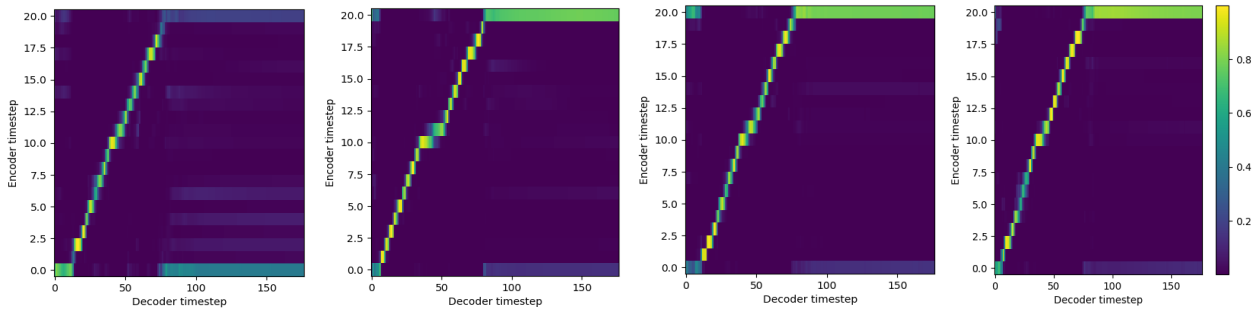
Table 1: *The partition of the data (number of utterances) used in the present study.*

Speaker (Gender)	Style	Train	Valid	Test
Nancy (Female)	Normal	11,000	200	800
Nick (Male)	Normal	2400	72	120
Nick (Male)	Lombard	500	100	120

4.2. Systems built for comparison

A total of five systems were built for comparison as shown in Table 2. The systems were different in terms of their acoustic parameter output types (WORLD vocoder parameters/mel-spectrograms) and the vocoder (WORLD/WaveNet) used. The WORLD vocoder parameters consisted of the mel-generalized cepstrum (MGC), fundamental frequency ($F0$) and band aperiodicity (BAP) with the dimensions 60, 1 and 1 respectively. The mel-spectrograms were extracted with the LibROSA [32] package by using 80 mel bands. 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 $F0$ values were transformed into the log domain and linearly interpolated in unvoiced regions.

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 the same architectures as systems S1 and



(a) Nick (normal) standalone (b) Nancy (normal) standalone (c) Nick (normal) adapted (d) Nick (Lombard) adapted

Figure 2: *An illustration of the alignments in different systems. The sentence "*PAPER WILL DRY OUT WHEN WET*" is taken from the test set. The x-axis and y-axis of each plot correspond, respectively, to the mel-spectrogram (extracted from speech) and phoneme (extracted from text).*

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 uses the mel-spectrogram for acoustic features.

The baseline S1 system was built as reported in our previous study [11]. The input linguistic features of Systems S1 and S3 were full-context labels and extracted using Festival toolkit. 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 oracle 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 [14] 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 repository [33]. All models were trained on a single NVIDIA Titan X GPU. The data is preprocessed in such a way that very long duration utterances are excluded from the training. The batch size was 32, and 2 acoustic frames were used per output step. The input linguistic features were mono-phonemes extracted using the Combilex lexicon [34] 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.

Table 2: *The systems developed for experiments.*

Sys. ID	TTS model	Ouput	Vocoder
S1	LSTM	MGC+$F0$+BAP+VUV	WORLD
S2	Seq2Seq	”	”
S3	LSTM	”	WaveNet
S4	Seq2Seq	”	”
S5	”	Mel-spectrogram	”

For the adaptation, we trained a Seq2Seq-TTS model on the Nancy data of a normal style; later that model was fine-tuned by the Nick data of a normal style. Then, the Nick normal speech Seq2Seq-TTS model was fine-tuned by the Lombard-style data of Nick. The initial Nancy model was trained for 150k steps. The initial learning rate was set to 0.002 and during the training the learning rate was adjusted based on the Noam scheme [35] with 4000 steps as warmup. The pre-trained Nancy model was fine-tuned by Nick date for 10k steps to learn the Nick normal speaking style model, and the initial learning rate was set to 0.00032. The Nick normal style model was further fine-tuned by Nick Lombard data for 10k steps to learn the Nick Lombard speaking style, and the initial learning rate was set to 0.00031. All the parameters of the model were optimized using the Adam optimizer [36].

As seen in Figure 2(a), when we trained the Seq2Seq-TTS model using only the Nick data of a 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 generalize well with little data. Thus we decided to train the initial model using the Nancy data (i.e. approx. 16 hours of speech) in order to learn a good alignment between input phoneme sequences and output acoustic frames.

We used a WaveNet configuration similar to [37], three repetitions of a 10-layer convolution stack with exponentially growing dilations, 64 residual channels and 128 skip channels. Separate models were trained for the 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 used MGCs, VUV and log$F0$ (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 subjective 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 [38] 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)* [38]. In the CCR test, each stimulus consists of a pair of utterances

Table 3: *$SIIB^{Gauss}$ scores measured in bits/s at different SNRs. The higher the score the better the intelligibility.*

Sys. ID	-10 dB	-5 dB	0 dB	5 dB
S1	14.2	25.3	39.7	57.9
S2	16.6	28.3	43.6	65.1
S3	16.0	31.3	50.6	75.5
S4	16.8	32.9	53.1	79.0
S5	**17.8**	**33.0**	**53.7**	**80.3**

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 FigureEight [39], a crowdsource platform (see [37] for more details of conducting the tests). We selected 16 utterances randomly from the test set and used them 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. Instrumental intelligibility evaluation

To measure the effect of Lombard adaptation on speech intelligibility, a recently developed instrumental intelligibility metric called speech intelligibility in bits (SIIB) [40] was used. SIIB measures the mutual information between a clean reference and a noisy signal. The noise signal is created by adding speech shaped noise (SSN) at various SNRs to the clean signal. A modified version of SIIB called as ($SIIB^{Gauss}$) [41] was employed in the present study.

5. Results and discussion

The results of the style similarity test are plotted in Figure 3. From the right pane of the figure, it can be observed that synthesized speech by all adapted systems was rated to sound different from natural normal speech with high confidence. When compared to the natural Lombard reference (the 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 it used the mel-spectrogram as output. It can be clearly seen that the systems that employed the WaveNet vocoder got higher scores than the ones that used the more traditional WORLD vocoder, even though the WaveNet was only trained with 30 minutes of Lombard speech. 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 parameters; a similar finding was observed in an earlier study [42]. From these results we can conclude that the synthetic speech produced by system S5 sounds most Lombard-like among the systems compared[1].

For the CCR test, we only included systems S1, S3 and S5. System S1 can be regarded as the baseline. System S3 was selected 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 overall in the similarity test. The results of the CCR test are shown in Figure 4. The scores were calculated by reordering the ratings for each system

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

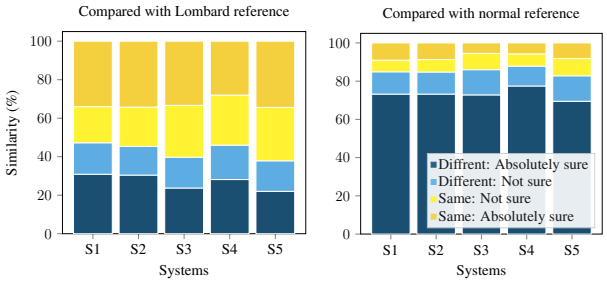


Figure 3: *The results of the style similarity test.*

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 significantly better score than the baseline system S1 and the LSTM based system S3.

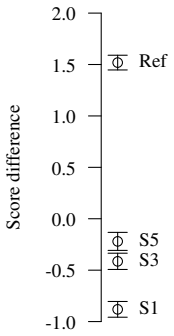


Figure 4: *The combined score differences obtained from the CCR test of naturalness. Error bars are t-statistic–based 95% confidence intervals for the mean.*

Table 3 presents the $SIIB^{Gauss}$ intelligibility scores. It can be observed that the baseline system S1 gave the worst performance and system S5 was best. The systems using the WaveNet vocoder got higher scores than the corresponding systems with the WORLD vocoder, thereby demonstrating how the vocoder choice affects the intelligibility of synthesized Lombard speech. These observations are in line with the results obtained both in the style similarity test and in the speech naturalness test.

6. Conclusions

This article proposed an adaptation approach using TL for the synthesis of Lombard speaking style in modern Seq2Seq TTS systems. Moreover, we also studied the role of a modern neural vocoder, WaveNet, for the synthesis of Lombard speech. Listening tests show that the proposed approach coupled with the WaveNet vocoder outperformed the previous best method that was developed using a LSTM-RNN–based adapted system. Future work includes extensive subjective evaluations and training both the WaveNet and Seq2Seq-TTS models in a single pipeline.

7. Acknowledgements

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

8. 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] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Intelligibility-enhancing speech modifications: the Hurricane challenge." in *Proc. Interspeech*, 2013, pp. 3552–3556.

[3] 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.

[4] 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.

[5] 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, pp. 879–883.

[6] H.-T. Luong and J. Yamagishi, "Multimodal speech synthesis architecture for unsupervised speaker adaptation," in *Proc. Interspeech*, 2018, pp. 2494–2498.

[7] Z. Huang, H. Lu, M. Lei, and Z. Yan, "Linear networks based speaker adaptation for speech synthesis," in *Proc. ICASSP*, 2018, pp. 5319–5323.

[8] 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.

[9] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.

[10] 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.

[11] 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.

[12] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[13] 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.

[14] 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.

[15] 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.

[16] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *arXiv preprint arXiv:1807.07281*, 2018.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, 2017, pp. 1118–1122.

[23] 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.

[24] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[26] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.

[27] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?" *arXiv preprint arXiv:1805.08974*, 2018.

[28] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[29] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.

[30] S. King and V. Karaiskos, "The Blizzard Challenge 2011."

[31] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Hurricane natural speech corpus," [sound], 2013.

[32] LibROSA a python package for music and audio analysis. [Online]. Available: https://github.com/librosa/librosa

[33] GST-Tacotron. [Online]. Available: https://github.com/syang1993/gst-tacotron

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

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] L. Juvela, V. Tsiaras, B. Bollepalli, M. Airaksinen, J. Yamagishi, and P. Alku, "Speaker-independent raw waveform model for glottal excitation," in *Proc. Interspeech*, 2018, pp. 2012–2016.

[38] 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.

[39] Figure-Eight a crowd source platform. [Online]. Available: https://www.figure-eight.com/

[40] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An instrumental intelligibility metric based on information theory," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, 2018.

[41] ——, "An evaluation of intrusive instrumental intelligibility metrics," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2153–2166, 2018.

[42] N. Adiga, V. Tsiaras, and Y. Stylianou, "On the use of WaveNet as a statistical vocoder," in *Proc. ICASSP*, 2018, pp. 5519–5523.