EFFECT OF DATA REDUCTION ON SEQUENCE-TO-SEQUENCE NEURAL TTS

Javier Latorre, Jakub Lachowicz, Jaime Lorenzo-Trueba, Thomas Merritt, Thomas Drugman,
Srikanth Ronanki, Viacheslav Klimkov

Amazon.com

{jlatorre, lachj, truebaj, thommer, drugman, ronanks, vklimkov }@amazon.com

Abstract

Recent speech synthesis systems based on sampling from autoregressive neural network models can generate speech almost indistinguishable from human recordings. However, these models require large amounts of data. This paper shows that the lack of data from one speaker can be compensated with data from other speakers. The naturalness of Tacotron2-like models trained on a blend of 5k utterances from 7 speakers is better than or equivalent to that of speaker dependent models trained on 15k utterances. Additionally, in terms of stability multi-speaker models are always more stable. We also demonstrate that models mixing only 1250 utterances from a target speaker with 5k utterances from another 6 speakers can produce significantly better quality than state-of-the-art DNN-guided unit selection systems trained on more than 10 times the data from the target speaker.

Index Terms: statistical parametric speech synthesis, autoregressive, neural vocoder, generative models, sequence-to-sequence

1. Introduction

Data acquisition is one of the main problems of data-driven text-to-speech (TTS) systems. High-quality unit selection TTS relies on large single-speaker databases, usually of tens of hours of speech. Classical statistical parametric speech synthesis (SPSS) is more data frugal. Less than one hour of data is enough to train an intelligible speaker dependent (SD) model. More data improves SPSS quality, but from 4-5 hours of data onwards the quality tends to saturate [1]. To reduce the dependency on a single speaker, techniques based on mixing data from multiple speakers into an Average Voice Model (AVM) were developed. These techniques produce reasonable quality with as little as 3 minutes of target speaker data [2]. However, when the available target speaker data is above 2 hours (∼2k utterances), Speaker-Dependent (SD) models were better [3].

The change of paradigm introduced by auto-regressive models [4, 5, 6, 7, 8], has produced synthetic speech of unprecedented quality. These new models require much more data than traditional TTS but they are also more efficient at integrating diverse data [9, 10, 11]. Several studies have reported that it is easy to train multi-speaker models [9, 12] and that adding more speakers improves the loss function over the validation set [4]. Most approaches for multi-speaker models rely on a speaker embedding but they vary on the type of embedding and where to apply it. Whereas some use an external model, e.g. speaker classification, to provide the embeddings [13, 12] others train the speaker embedding together with the model out of a one-hot speaker ID vector [4, 9, 5]. Some approaches use the embedding at the input only as a global conditioning [4], whereas others apply it at different levels within the model [9, 5].

Despite all the recent attention to multi-speaker models, to the best of our knowledge, there have been no published studies yet on practical issues such as; 1) at which point an SD model becomes better than a multi-speaker one, 2) whether it is better or worse to use gender-dependent multi-speaker models, 3) what is the effect of training models with an unbalanced mixture of data from the target speaker and other speakers. This paper presents the results of several experiments aimed at answering these questions. We hope our results will help other developers and researchers in designing their systems and experiments.

The structure of the paper is as follows: Section 2 describes the basic structure of our TTS system; Sections 3 and 4 describe the experimental protocol and results respectively. Finally, in Section 5 conclusions are drawn.

2. System description

Our system architecture follows that of Tacotron2 [8]. First, a sequence-to-sequence (S2S) acoustic model predicts the mel-spectrograms from a sequence of linguistic inputs. Then a neural vocoder converts the mel-spectrograms into a waveform.

2.1. Acoustic model

The architecture of the acoustic model is described in Figure 1. It is a S2S model with attention mechanism similar to [8]. However, instead of using raw graphemes as inputs, our system first converts the graphemes into phonemes which are then encoded with a one-hot vector. For the vowels, we use 3 different symbols depending on their level of stress (0,1,2). The punctuation after each word, including blanks, is treated as if it were another phoneme.

The attention mechanism for the S2S model follows the one proposed in [14], with normalised attention weights [15]. In this mechanism the attention weights for the current frame depend both on the previous output of the decoder and on the attention weights of the previous frame. The speaker conditioning is similar to [4], with a one-hot speaker ID global conditioning.

The output of the model are blocks of 5 frames of mel-spectrograms, each consisting of an 80-dimensional vector spanning frequencies between 50 Hz and 12 kHz. Each frame is computed over 50 ms and shifted every 12.5 ms. The last frame of the previous block is passed as input to both the attention model and the decoder to generate the next 5-frame block. During training, this recursive input is randomly switched between real spectrograms and self-generated ones (scheduled sampling). The probability for taking real spectrograms is 0.9. In addition to the mel-spectrograms, the model also predicts a stop token to mark the end of the utterance. The stop token is encoded as a linear function with 0 at the beginning and 1 at the end of the sentence. The model was trained with a dropout probability of 0.1 for both the decoder and the auto-regression, but without dropout for the encoder. The dropout was also applied at inference time.

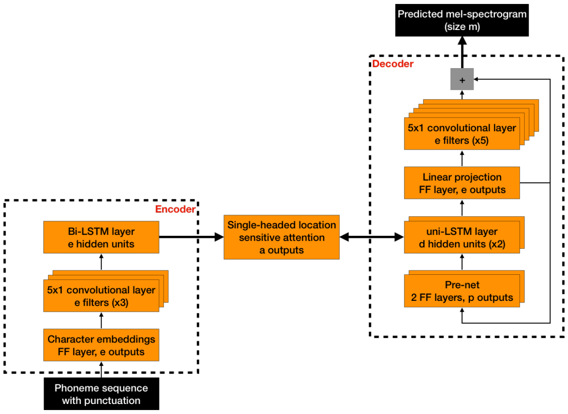

Figure 1: *Acoustic model architecture*

2.2. Neural vocoder

The architecture of the neural vocoder closely follows WaveRNN [16]. The autoregressive part of the network consists of a single forward Gated Recurrent Unit with a hidden size of 896 and a pair of affine layers followed by a softmax layer with 1024 outputs to predict the 10-bit mu-law samples with 24 kHz sampling rate. The conditioning network consists of two bi-directional Long Short-Term Memory (LSTM) layers with a hidden size of 128. The mel-spectrograms for conditioning consisted of 80 coefficients extracted using Librosa library [17] for frequencies from 50 Hz to 12 kHz. The model was trained on data from 74 speakers on 17 different languages with between 1k to 2.5k utterances per speaker. Around two thirds of the speakers were female and the other third male, except for one child. More details about the vocoder architecture and how it was trained can be found in [18].

3. Experimental protocol

The research questions we attempted to answer were:

1. Can a multi-speaker model with limited data per speaker achieve similar quality to a SPSS-guided unit selection with a large database?
2. Can we train multi-speaker models with less data for the target speaker than for the supporting speakers?
3. How much data is needed for a SD model to be better than a multi-speaker one?
4. Is it better to combine all the available speakers or only the most similar ones, e.g., only female speakers?
5. Does mixing speakers affect the speaker similarity?

The results of our experiments to answer these questions are presented in Section 4.

3.1. Training data and model stability

The speech data used to train the models came from 7 internal speakers: 2 males, 4 females and one child. The available data was 8.5k utterances for four speakers (2 females, one male and the child), 15k for two (one male and one female) and 25k utterances for one female speaker. Out of this, we randomly selected a fixed number of utterances per speaker depending on the model. For each speaker in the model, we used 90% of the utterances for training and 10% for development. The first three columns of Table 1 shows the speaker blends and amount of data used to train each type of model in our experiments.

Table 1: *Percentage of correctly generated files*

model	model name	#training utt	% stable
single speaker	sd-8500	8.5k	35.4%
(1 spkrs)	sd-15000	15k	46.2%
	sd-25000	25k	69.3%
female only	fe4-2500	4×2.5k	88.3%
(4 spkrs)	fe4-5000	4×5k	77.33%
	fe4-8500	4×8.5k	77.33%
mixed-gender	mx7-2500	7×2.5k	54.5 %
(7 spkrs)	mx7-5000	7×5k	93.5%
	mx7-8500	7×8.5k	95.6%
mixed-gender	mx6+1250	6×5k + 1.25k	91.4%
unbalanced (7 spkrs)	mx6+2500	6×5k + 2.5k	78.9%

A problem in S2S models is that the attention sometimes gets lost at inference time. This produces errors such as skipping one or more phones, repeating part of the sentence, getting stuck in silences, etc. An analysis of the stability of the models is useful to understand their robustness toward different blends of training data. To measure this, we generated 75 utterances from each speaker on each type of model and marked those that, after listening, presented any of the above mentioned stability problems. The last column of Table 1 shows the proportion of stable utterances for each model. Multi-speaker models are clearly much more stable than SD models, regardless of whether they are female-only or mixed-gender. This result agrees with the comments in [4] about convergence of multi-speaker models. Model stability does not seem to be directly linked to the amount of training data. The female-only model trained on 2.5k utterances/speaker was more stable than the female-only models trained on more data. Also, some multi-speaker models are more stable than SD ones, despite being trained on less data. The type of problems seems to depend on the speaker, even in the multi-speaker models. All these suggest that stability depends on the characteristics of the data itself. However, we could not find any clear pattern for it.

3.2. Subjective evaluation

To address our research questions, we ran several MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) tests [19]. The advantage of MUSHRA over Mean Opinion Score (MOS) is that for each sentence, all the systems being evaluated are presented simultaneously in one panel. On every test the positions of the systems on the panel were randomised. All the panels included the natural recordings as an upper anchor but, similar to [20], subjects were not forced to assign the top score to any of the systems. All tests were conducted on Amazon Mechanical Turk. Subjects were people living in the United States who defined themselves as native English speakers. For each evaluation, we selected sentences ranging between 5 and 30 words. For all the tests the significance of the results was analysed with a Wilcoxon signed-rank test and a standard t-test, both corrected with Holm-Bonferroni [21]. The main goal of the subjective tests was to evaluate speech quality and speaker similarity. Therefore, for each of the MUSHRA tests we chose only those utterances for which none of the systems under consideration presented any of the attention stability problems described in 3.1.

3.2.1. Naturalness

For tests addressing questions 1-4, subjects were asked to *"Rate the audio samples in terms of their naturalness"* with a continuous slider between *"Completely unnatural"* (0) and *"Completely natural"* (100). Each stimuli panel was evaluated by 10 subjects. The set of sentences used on each experiment were slightly different since due to stability problems, not all the systems on each test were able to synthesise all the utterances. In the tests for questions 1 and 2, a guided unit selection was included among the systems to be evaluated. This guided unit selection was a standard system in which a rule-based linguistic cost [22] is combined with an acoustic cost, computed as the distance between the F0, duration and spectrum of the units and those predicted by a state-level DNN model. The models for the acoustic cost were speaker-dependent and trained with all the available data for each speaker. At synthesis time, the evaluation sentences were blacklisted so that their units could not be selected. This blacklisting removed less than 0.5% of the unit selection data. Since these tests proved that the guided unit selection was worse than the other systems, we didn't include it in subsequent experiments.

3.2.2. Speaker similarity

For the test on question 5, subjects were asked to *"Rate whether the speaker of the reference sounds like the same person as the speakers of the samples."* between *"Definitely a different person"* (0) and *"Definitely the same person"* (100). Subjects were presented with a reference audio from the target speaker (sentence1) and speech audio samples for a different sentence (sentence2) generated by the evaluated models. The recording of sentence2 by the target speaker was also included as an upper anchor. For each of the seven speakers, we ran an independent MUSHRA test with 10 utterances and the best available SD model for that speaker.

4. Results

4.1. Multi-speaker vs unit selection

The first experiment evaluates the naturalness of two multi-speaker models, 'mx7-5000' and 'mx7-2500' (see Table 1) vs the guided unit selection. As an additional reference point, we included samples re-synthesised from the original mel-spectrograms with the neural vocoder, 'nv-resynthesis'.

The evaluation consisted of 27 utterances from each of the 7 speakers resulting in a total of 189 stimuli panels. A total of 70 subjects evaluated 27 panels each. The boxplots of the MUSHRA scores can be seen in Figure 2. All the models were significantly different from each other at $p < 0.05$. As expected, the recordings and the 'nv-resynthesis' samples achieved the higher score followed by 'mx7-5000' and 'mx7-2500'. The difference between 'mx7-2500' and 'mx7-5000' is small but statistically significant. The most surprising result was the comparatively low score of the guided unit selection, despite it being built upon more than 99% of all the available data. Obviously, there were differences between speakers, but they do not correlate with the amount of data of the unit selection system. The rank order of the systems was consistent across speakers. The median MUSHRA in Figure 2 also shows that the gap between 'nv-resynthesis' and the recordings is very small, despite the vocoder being a generic one trained on multiple speakers in different languages. The main gap is between the models and the 'nv-resynthesis', i.e. in the modelling of the mel-spectrograms. Comparatively, the gap due to differences in the amount of training data is smaller.

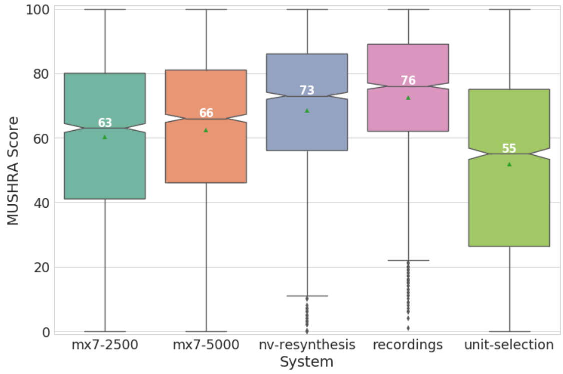

Figure 2: *Multi-speaker models vs. Unit selection*

4.2. Balanced vs unbalanced mixture of speakers

The second experiment evaluated the naturalness of the models from Section 4.1 vs models trained with 5k utterances from six speakers plus 2.5k or 1.25k utterances from a target speaker, 'mx6+2500' and 'mx6+1250' respectively. We train one 'mx6+...' model for each speaker and used them only to generate speech with the voice of that speaker. To keep the lower anchor of Section 4.1, we added again samples produced by the guided unit selection system.

The evaluation consisted of 27 utterances from each of the 7 speakers. They were evaluated by a total of 70 subjects. Figure 3 shows the results. The rank order of the results for 'unit-selection', 'mx7-2500', 'mx7-5000' and 'recordings' confirms the results of Section 4.1. Models 'mx7-2500', 'mx6+1250' and 'mx6+2500' are not significantly different from each other. This indicates that the benefit of using 5k utterances instead of 2.5k for the non-target speakers is not in terms of quality, but in terms of stability as was shown in Section 3.1. A second interesting result was that in terms of quality, 'mx6+1250' models were significantly better than a state-of-the-art unit selection system. The only exception to this was for the unit selection of the male speaker on >15k utterances. There were some other minor differences between speakers, especially in the relative ranking of the two 'mx6' models. However, with the above mentioned exception, the results were fairly consistent across speakers.

4.3. Multi-speaker vs speaker dependent

This set of experiments compared SD models with multi-speaker models 'mx7-5000' and 'mx7-8500' (see Table 1). We trained 'sd-8500' models for all seven speakers, 'sd-15000' for 3 speakers and 'sd-25000' for one speaker, depending on the amount of data available for the speakers. Three separate evaluations were conducted, for the SD models on 8.5k, 15k and 25k utterances.

Unfortunately, out of the seven 'sd-8500' models only 3 (two female and one male) were stable enough to generate samples. To compensate for the lack of data points in the evalua-

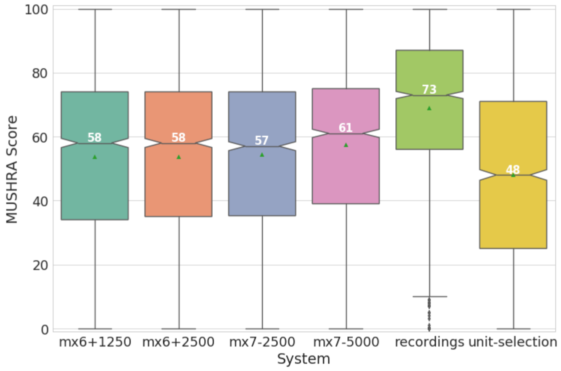

Figure 3: *Mixed models with balanced vs unbalanced data*

tion of the ‘sd-8500’ models we used 42 samples for each of the remaining 3 speakers. For the evaluation of the ‘sd-15000’ models we only have 3 speakers with enough data. As shown in Table 1 these models were also very unstable, especially for one of the speakers which only correctly generated 24 utterances. To keep the number of utterances per speaker balanced we evaluated these models with 24 utterances/speaker. Finally, for ‘sd-25000’ we generated 45 sentences from that speaker model. To compensate for the lack of data, each MUSHRA panel of the ‘sd-25000’ evaluation was judged by 15 subjects.

Table 2 shows the median MUSHRA score for the three tests and the average rank of the systems. In the three evaluations, the differences between ‘mx7-5000’ and ‘mx7-8500’ were not statistically significant, as can be seen by the small differences in their averaged rank order. Both multi-speaker models were better than ‘sd-8500’ models and worse than the ‘sd-25000’ model. The ‘mx7-8500’ model was better than the ’sd-15000’ model. These differences were statistically significant. The differences between ‘sd-15000’ model and ‘mx7-5000’ were not significant with the Wilcoxon signed-rank test.

These results suggest that, similar to classical SPSS, a SD model can sound more natural than a multi-speaker model when trained on sufficient amounts of data. However, multi-speaker models are better than SD models when they are trained on more than 2.3 times the amount of data or, alternatively, when the SD model is trained on less than 15k utterances. Further work is needed to clarify this last point.

Table 2: *Median score and average rank (in parentheses) of the tests comparing multi-speaker vs speaker dependent models*

Evaluated SD model	Recordings	Models		
		SD	mx7-8500	mx7-5000
sd-8500	71 (1.96)	61 (2.78)	63 (2.61)	62 (2.64)
sd-15000	74 (1.91)	61.5 (2.79)	63 (2.65)	62 (2.65)
sd-25000	77 (1.97)	68 (2.56)	67 (2.73)	66 (2.75)

4.4. Female only vs mixed gender

The last naturalness experiment compared models trained on all 7 speakers against those trained only on 4 female speakers. The total amount of data was different but the amount of data per speaker was constant. Figure 4 shows the results. The differences between models trained on different number of utterances per speaker were statistically significant, but the differences between models trained on the same data per speaker were not.

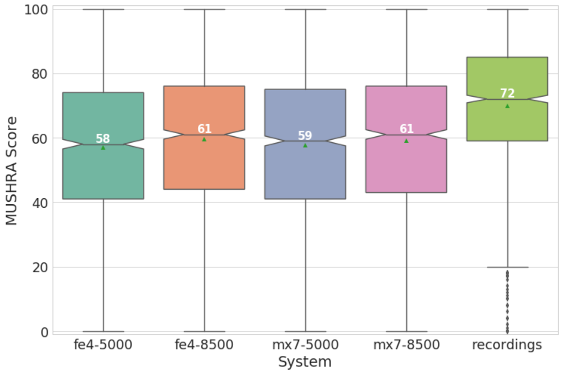

Figure 4: *Mixed of all speakers vs only female speakers*

4.5. Speaker similarity

Table 3 summarises the results. Each row represents a different speaker. On average, only the differences between recordings and the other systems were statistically significant at $p < 0.05$. On a per-speaker basis, the differences for two speakers (in bold type face in Table 3) and the rest of the models were statistically significant. However, that significance disappears when both ‘sd-8500’ or ‘sd-15000’ speakers are considered jointly.

Table 3: *Average speaker similarity*

Evaluated SD model	Record-ings	Models					
		best	mx7-...			mx6+...	
		SD	8500	5000	2500	1250	2500
sd-8500	78.3	69.5	70.7	70.2	68.8	70.0	70.1
	73.0	**68.1**	**76.3**	**74.7**	**76.7**	**71.9**	**73.0**
	76.5	-	70.7	71.3	71.0	71.3	71.4
	79.3	-	71.5	73.3	74.6	69.1	73.2
sd-15000	68.1	68.4	65.2	68.2	67.4	68.4	64.5
	83.4	**77.3**	**82.3**	**84.3**	**84.2**	**82.3**	**82.3**
sd-25000	75.0	72.1	69.9	71.4	70.7	71.8	70.1
Average	76.0	70.7	72.1	73.1	73.2	71.3	72.2

5. Conclusions

This paper presents several experiments aimed at reducing the amount of single speaker data required to train high-quality S2S TTS systems. The results show that models trained on a mixture of speakers can produce better quality than a state-of-the-art guided unit selection TTS system with an inventory of units ranging between 8.5k and 25k utterances. We show that this is true for S2S models trained on 2.5k utterances from 7 speakers and also for S2S models trained on a mixture of 1.25k utterances from the target speaker and 5k utterances from 6 other speakers. Our results also show that for databases with up to 15k utterances, multi-speaker models produce better quality than speaker-dependent ones. SD models with more data can produce marginally better quality but in terms of stability SD models are always less stable. The most probable reason for this is that by mixing multiple speakers, the alignment is more robust against different pronunciations, incorrectly labelled sentences or different initialisation values. This seems to also be the case when training on less data but more similar speakers. The different speaker blends do not seem to affect the speech quality but lower variability of speakers seems to negatively impact the model’s stability.

6. References

[1] J. Yamagishi, L. Zhenhua, and S. King, “Robustness of HMM-based Speech Synthesis,” *Proc. INTERSPEECH*, 2008.

[2] J. Yamagishi, K. Takao, Y. Nakano, K. Ogata, and J. Isogai, “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm,” *IEEE Transaction on Speech, Audio & Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[3] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, Y. Guan, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo, “Thousands of voices for HMM-based speech synthesis,” *Proc. INTERSPEECH*, pp. 420–423, 2009.

[4] A. v. d. 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 *arXiv:1609.03499*, 2016.

[5] 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,” in *arXiv:1710.07654*, 2017.

[6] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta, and M. Shoeybi, “Deep voice: Real-time neural text-to-speech,” in *arXiv:1702.07825v2*, 2017.

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

[8] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, “Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions,” *arXiv preprint arXiv: 1712.05884*, 2017.

[9] S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, “Deep voice 2: Multi-speaker neural text-to-speech,” in *Advances in Neural Information Processing Systems*, 2017, pp. 2966–2974.

[10] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, “Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis,” *Speech Communication*, vol. 99, pp. 135–143, 2018.

[11] M. Podsiadlo and V. Ungureanu, “Experiments with training corpora for statistical text-to-speech systems.” in *Proc. Interspeech 2018*, 2018, pp. 2002–2006. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-2400

[12] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez-Moreno, and Y. Wu, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” *arXiv preprint arXiv:1806.04558v1*, 2018.

[13] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, “Voiceloop: Voice fitting and synthesis via a phonological loop.” in *International Conference on Learning Representations*, 2018.

[14] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv: 1409.0473*, 2014.

[15] T. Salimans and D. P. Kingma, “Weight normalization: A simple reparameterization to accelerate training of deep neural networks,” *arXiv preprint arXiv: 1602.07868*, 2016.

[16] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. v. d. Oord, S. Dieleman, and K. Kavukcuoglu, “Efficient neural audio synthesis,” *arXiv preprint arXiv:1802.08435*, 2018.

[17] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in python,” in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[18] J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, and R. Barra-Chicote, “Robust universal neural vocoding,” *arXiv preprint arXiv: 1811.06292*, 2019.

[19] ITUR Recommendation, “Bs. 1534-1. method for the subjective assessment of intermediate sound quality (mushra),” *International Telecommunications Union, Geneva*, 2001.

[20] T. Merritt, B. Putrycz, A. Nadolski, T. Ye, D. Korzekwa, W. Dolecki, T. Drugman, V. Klimkov, A. Moinet, A. Breen, R. Kuklinski, and N. S. andRoberto Barra-Chicote, “Comprehensive evaluation of statistical speech waveform synthesis,” in *Proc. SLT*, 2018.

[21] R. A. J. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, “Statistical analysis of the blizzard challenge 2007 listening test results,” in *Proc. Blizzard 2007*, 2007.

[22] S. King, “Measuring a decade of progress in text-to-speech,” *Loquens*, vol. 1, no. 1, p. 006, 2014.