Effect of data reduction on seq-to-seq acoustic models for speech synthesis

Javier Latorre, Jakub Lachowicz, Jaime Lorenzo-Trueba, Thomas Merritt, Thomas Drugman

Amazon.com

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

Abstract

Recent speech synthesis systems based on sampling from autoregressive neural networks models can generate speech almost undistinguishable from human recordings. To work properly these models required large amounts of data. However, they are more efficient at dealing less homogenous data, which might make possible to compensate the lack of data from one speaker with data from other speakers. This paper evaluates this hypothesis by training several tacotron-like models with different blends of data. The mel-spectrograms generated by these models were converted to audio with a WaveRNN-like neural-vocoder trained on 74 speakers from 17 different languages. Our experiments show that the naturalness of models trained on a blend of 5k utterances from 7 speakers is better than that of speaker dependent (SD) models trained on 15k utterances, and very close to that of SD models trained on 25k utterances. 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 using more than 10 times more utterances from that target speaker data.

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

1. Introduction

Data acquisition is one of the main problem of data-driven text-to-speech (TTS) systems. High-quality unit selection (US) TTS relies on large single speaker databases, usually of more than 15 hours of speech. Classical statistical parametric speech synthesis (SPSS) is more data frugal. Less than one hour of data is enough to train a intelligible speaker dependent (SD) model. More data improves SPSS quality, but from around 4-5 hours of data onwards it tends to saturate [1]. To reduce the dependency on a single speaker, techniques based on an Average Voice Models (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 ($\sim$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 cost over the validation set [4]. Most approaches for multi-speaker model rely on an speaker embedding but they vary on the type of embedding and where to apply it. Whereas some use an external model, e.g. and 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 nobody has evaluated yet practical issues like at which point an SD model becomes better than a multi-speaker one, or whether it's better or worse to use gender-dependent multi-speaker models, or what is the effect of training models with an unbalance mixture of data from the target speaker and other speakers. In this paper we present the results of several experiments aimed at answering these questions. We hope our results will help other developers and researchers to better design their systems and experiments.

The structure of the paper is as follows: section 2 describes the basic structure of the TTS system we have used for the experiments; section 3 describes our experiments; in section 4 we propose some hypotheses to explain the results. Finally, in section 5 conclusion are drawn.

2. System description

Our system architecture follows that of tacotron2 [8]. First, a sequence-to-sequence 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 our acoustic model is described in Fig. 1. The main difference with [8] is that 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 sequence-to-sequence model follows the one proposed in [17] with normalised attention weights [18]. In this attention, 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 and 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 50ms 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 real number between 0 and 1 that reaches the value

of 1 at the end of the sentence. The model was trained with a dropout probability of 0.1.

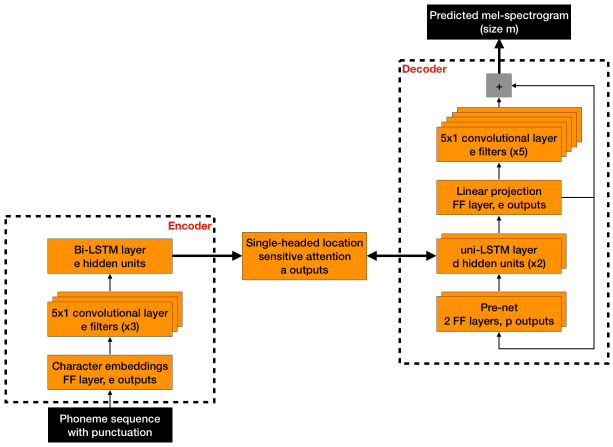

Figure 1: *Acoustic model architecture*

2.2. Neural vocoder

The architecture of the neural vocoder closely follows WaveRNN [19]. The autoregressive part of the network consists of a single forward GRU 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 a two bi-directional LSTMs with a hidden size of 128. The mel-spectrograms for conditioning consisted of 80 coefficients extracted using Librosa library [20] for frequencies from 50 Hz to 12 kHz. The model was trained on data from 75 speakers on 17 different language with between 1k to 2.5k utterances per speaker. Around two third of the speakers were female and the other two third male except for one child. More details about the vocoder architecture and how it was trained can be found in [21].

3. Experiments

The research questions we tried to answer were:

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

To address these questions we ran several MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) tests [22] so that all the systems being evaluated at each experiment could be presented simultaneously in one panel. All the panels include the natural recordings as upper anchor. We ran naturalness and speaker similarity tests. All tests were conducted in Amazon Mechanical Turk. Subjects were people living in the United States who define themselves as native English speakers. To limit the damage of potential careless subjects, there was a limit in the number of panels each subject could evaluate.

Table 1: *Percentage of correctly generated files*

model	training utt	% correct
single speaker	1 x 8.5k	35.4%
	1 x 15k	46.2%
	1 x 25k	69.3%
female only	4x 2.5k	88.3%
	4x 5k	77.33%
	4x 8.5k	77.33%
mix-gender	7 x 2.5k	54.5 %
	7 x 5k	93.5%
	7 x 8.5k	95.6%
mix-gender unbalanced	6 x 5k + 1.25k	91.4%
	6 x 5k + 2.5k	78.9%

The speech data used to train the models came from 7 speakers of our speakers: 2 males, 4 females and one child. The available data for these speaker was >8.5k utterances for four of them, >15k for two and > 25k utterances for one. Out of this, we randomly selected a fixed number of utterances per speaker, as described in the next sections. For each speaker mixed in the model, we used 90% of the utterances for training and 10% for development. For the evaluation we generated utterances for sentences between 5 and 30 words. The significance of the results was analysed with a Wilcoxon rank-test and a standart t-test, both with a Bonferroni-Holm correction.

3.1. Model stability

A problem in sequence-to-sequence 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. The main goal of our experiments was to evaluate speech quality. Therefore, we chose for the subjective tests only those sentences that were correctly generated by all the systems under consideration. However, an analysis of the stability of the models is useful to understand their robustness toward different types of training data.

To measur this, we generated 75 utterances from each speaker on each type of model and discarded those that presented some stability problem. Table 1 shows the proportion of non-discarded utterances for each model together with the number of utterances used to train it. SD models are clearly much more unstable than mixed-speaker one, regardless of whether the mixed speaker ones are gender dependent or gender-mixed. This agrees with the comments on [4] about convergence. Model stability does not seem to be directly linked to 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 mixed-speaker model are more stable than SD ones, despite being trained on less data.

3.2. Naturalness

On the naturalness tests subjects were asked to ”rate the audio samples in terms of their naturalness” with a continuous sliders graded between 0, ”completely unnatural” and 100, ”completely natural”. Each stimuli panel was evaluated by 10 subjects.

In two evaluations we included a guided US among the models to be evaluated. This guided US is a standard one in which the linguistic cost is combined with the acoustic cost computed as the difference between the unit acoustic features and the F0, duration and spectral 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 US data.

3.2.1. *Speaker mixture vs unit selection*

The first experiment evaluates the naturalness of two mixed-speaker models vs unit selection. The mixed-speaker models, 'mx7-5000' and 'mx7-2500', were trained on 5k and 2.5k utterances from each of the seven speakers respectively. We also tried to train SD models on 5k utterances but these models did not converge properly. Speech generated with them were unstable or with much lower segmental quality. We decided not to include this models since having such low anchor would have squeezed the scores of all the other models in the top of the scores As an additional reference point, we included samples resynthesised from the original mel-spectrograms with the neural vocoder, 'nv-resynthesis'.

The evaluation consisted of 27 utterances from each of the 7 speakers making a total of 189 stimuli panels. The order of the panel was randomised and 70 subjects evaluated 27 panels each. The boxplots of the MUSHRA scores can be seen in Fig. 2. All the models were significantly different from each other. As expected the recordings and the 'nv-resynthesis' samples got the higher score followed by the 'mx7-5000' and 'mx7-2500'. The difference between the 'mx7-2500' and 'mx7-5000' is small but statistically significant. The most surprising result was the comparatively low score of the guided US, despite it being built upon more than 99% of all the available data. Obviously, there were differences between speakers, but they don't correlate with the amount of data of the US system. The rank order of the systems was consistent across speakers. Apart from that, the median MUSHRA in the figure show that the 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. The gap due to difference in the amount of training data is comparatively smaller.

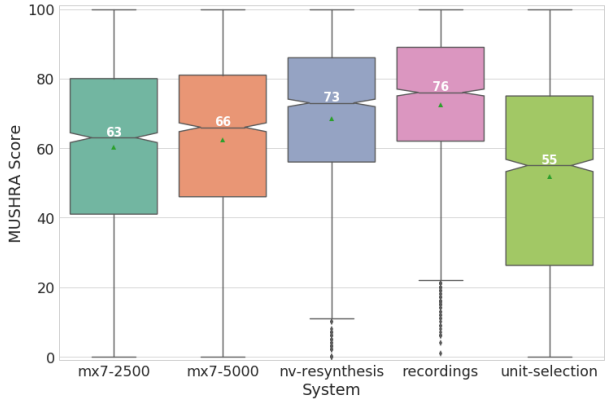

Figure 2: *Mixed models vs. Unit selection*

3.2.2. *Balanced vs unbalanced mixture of speakers*

The second experiment evaluated the naturalness of the models of previous experiments 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 the previous experiment, we added again samples produced by the guided US.

The evaluation procedure was the same as in the previous experiment with a new set of 27 utterances from each of the 7 speakers. Figure 3 shows the results. The ranks of the results of 'unit-selection', 'mx7-2500', 'mx7-5000' and 'recordings' confirm the results of the previous experiment. The 'mx7-2500', 'mx6+1250' and 'mx6+2500' are not significantly different from each other. This indicates that 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 the in terms of quality, 1250 utterance from a target speaker mixed with sufficient data from other speakers can generate better speech quality than a state-of-the-art unit selection system. The only exception to this was one of the speakers 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 rank order between systems was consistent across speakers.

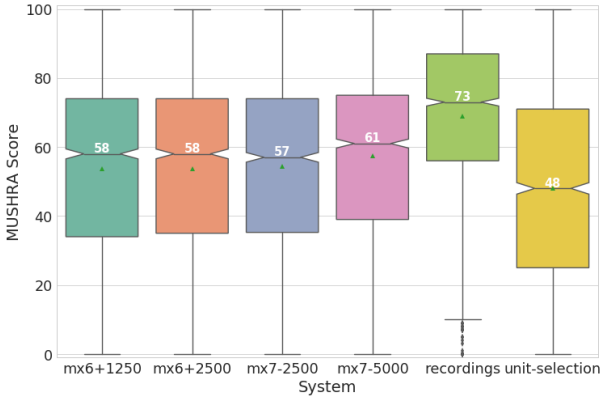

Figure 3: *Mixed models with balanced vs unbalanced data*

3.2.3. *Speaker mixture vs speaker dependent neural TTS*

These set of experiments compared SD models with mixed-speaker models. The mixed-speaker models were trained on 5k and 8.5k utterances from all 7 speakers. We trained SD models on 8.5k utterances for all the speakers, 15k utterances for 3 speakers and on 25k utterances for one. Three separated evaluations were run, for the SD models on 8.5k, 15k and 25k utterances.

Unfortunately, out of the seven 8.5k utterance models only 3 were stable enough to generate samples. To compensate for the lack of data points, in the evaluation of the SD-8500 models we used 42 samples for each of the remaining 3 speaker. 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 generated correctly 24 utterances. To keep the number of utterances per speaker balanced we evaluated these models with 24 utterances/speaker. Finally, for the SD-25000 we generated 45 sentences from that speaker model. To compensate for the lack of data, each stimuli panel in the SD-25000 evaluation was judged by 15 subjects.

Table 2: *Median MUSHRA score*

#utterances in SD	recordings	SD	mx7-8500	mx7-5000
8.5k	71	61	63	62
15k	74	61.5	63	62
25k	77	68	67	66

Table 3: *Average MUSHRA rank (lower is better)*

#utterances in SD	recordings	SD	mx7-8500	mx7-5000
8.5k	1.96	2.78	2.61	2.64
15k	1.91	2.79	2.65	2.65
25k	1.97	2.56	2.73	2.75

Tables 2 and 3 show the median MUSHRA score and average rank of the systems for the 3 MUSHRA experiments. In these three evaluation, the 'mx7-5000' and 'mx-8500' were not significantly different from each other. The SD-8500 and SD-25000 models were significantly different to the mixed models. The SD-15000 models were significantly different to both mixed models according to the t-test but not significantly different to the 'mx7-8500' model according to the wilcoxon rank-test.

This results suggest that, similar as in classical SPSS, SD model trained on sufficient amount of data are better than mixed-speaker models. However, mixed-speaker models outperform SD ones when the ratio of training data is 2.3 times or more.

3.2.4. Gender-dependent vs Gender-independent

The last experiment compared models trained on all the 7 speakers against those trained only on the 4 female speakers. The total amount of data was different but the amount of data per speaker was constant. Figure 4 shows the results. Models trained on different number of utterances per speaker were significantly different. However, for the same amount of data/speaker what type of speaker are mixed does not produce any significant difference. This suggests that the model does indeed some kind of speaker factorisation and not just an averaging of the data from each speaker.

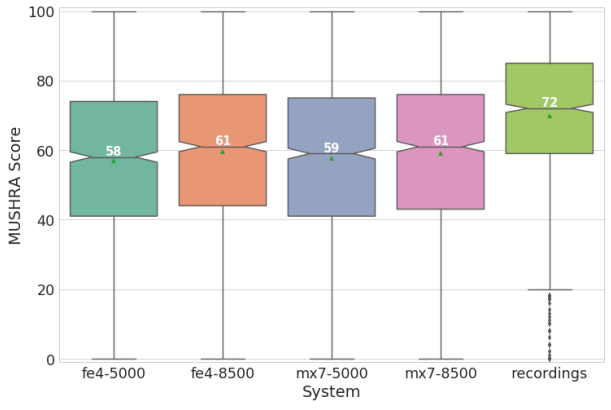

Figure 4: *Mixed of all speakers vs only female speakers*

3.3. Speaker similarity

On the speaker similarity tests subjects were asked to "rate whether the speaker of the reference sounds like the same person as the speakers of the samples." Since we couldn't train SD models for all the speakers, we ran the evaluation independently for each speaker with the best available SD. Subjects were presented with a reference audio from the target speaker and several audio files for a different sentence generated with the different models. The recording of that sentence by the target speaker was also included as an upper anchor. For each speaker we ran an independent MUSHRA test with 10 utterances. Table 4 summarises the results. Each row represents a different speaker. Overall, only the recordings were significantly different from the rest. On a speaker basis there were significant differences between one of the SD model and the rest of models for one of the 8500 utterances and one of the 15000 speakers. However, those difference disappear when both 8500 or 15000 speakers are considered jointly.

Table 4: *Average speaker similarity*

SD data	rec	best SD	mx7			mx6+	
			8.5k	5k	2.5k	1.25k	2.5k
8.5k	78.29	69.48	70.75	70.24	68.77	70.09	70.1
	73.02	68.09	76.29	74.73	76.67	71.86	72.99
	76.47	-	70.67	71.27	71.04	71.35	71.38
	79.33	-	71.54	73.29	74.6	69.09	73.25
15k	68.13	68.44	65.23	68.17	67.45	68.44	64.55
	83.44	77.3	82.33	84.35	84.26	82.29	82.28
25k	74.97	72.15	69.87	71.39	70.75	71.85	70.14
Average	76.01	70.71	72.12	73.10	73.16	71.31	72.20

4. Discussion

Why does mix-speaker models work equal or better than single speaker models? The most probable reason is that by mixing multiple speakers the alignment is more robust against different pronunciations, wrong sentences or different initialisation values. The differences in terms of stability suggest that this is an important factor. However, some of the subjects comments suggest that another potential reason is that by mixing multiple speakers, the prosody becomes more 'generic' which to subjects might be more attractive. If that hypothesis could be confirmed, it would be in line with other studies [23, 24] which found that the closer a voice is to the 'average' the more attractive it is perceived.

5. Conclusions

This paper shows the results of several experiments aimed at reducing the amount of single speaker data needed to train high quality TTS. The results indicate that a mixture of 2.5k utterances 7 speakers can produce better quality than a state-of-the-art US-TTS with a DB ranging between 8.5k and more than 27k utterances. We also show that this is still the case with 1.25k utterances and 5k utterances from another 6 speakers, although in this case the speaker similarity gets degraded. Training on less data but more similar speakers does not affects the speech quality but seems to impact the model stability. Finally, we have shown that for databases up to 15k utterances, mixed-speaker models produce better quality than speaker-dependent ones, and in terms of stability they are always more stable.

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] 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.

[15] S. Pascual and A. Bonafonte, "Multi-output RNN-LSTM for multiple speaker speech synthesis and adaptation." in *24th European Signal Processing Conference EUSIPCO*, 2016.

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

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

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

[19] 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.

[20] 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.

[21] J. Lorenzo-Trueba, T. Drugman, J. Latorre, R. Barra-Chicote, T. Merritt, B. Putrycz, S. Ronanki, and K. , Viacheslav, "Robust universal neural vocoding," in *Submitted to ICASSP 2019*, 2019.

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

[23] J. Yamagishi, O. Watts, S. King, and B. Usabaev, "Roles of the average voice in speaker-adaptive HMM-based speech synthesis," in *Interspeech*, 2010.

[24] L. Bruckert, P. Bestelmeyer, M. Latinus, J. Rouger, I. Charest, G. A. Rousselet, H. Kawahara, and PascalBelin, "Vocal attractiveness increases by averaging," *Current Biology*, vol. 20, no. 2, pp. 116–120, 2010.