Juvela, Lauri; Bollepalli, Bajibabu; Yamagishi, Junichi; Alku, Paavo

Reducing mismatch in training of DNN-based glottal excitation models in a statistical parametric text-to-speech system

Published in:
Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH

DOI:
10.21437/Interspeech.2017-848

Published: 01/08/2017

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

Please cite the original version:
Juvela, L., Bollepalli, B., Yamagishi, J., & Alku, P. (2017). Reducing mismatch in training of DNN-based glottal excitation models in a statistical parametric text-to-speech system. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH (Vol. 2017-August, pp. 1368-1372). (Interspeech: Annual Conference of the International Speech Communication Association). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2017-848

INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Reducing mismatch in training of DNN-based glottal excitation models in a statistical parametric text-to-speech system

Lauri Juvela[1], *Bajibabu Bollepalli*[1], *Junichi Yamagishi*[2,3], *Paavo Alku*[1]

[1]Aalto University, Department of Signal Processing and Acoustics, Finland
[2] National Institute of Informatics, Japan
[3]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom
{lauri.juvela,bajibabu.bollepalli,paavo.alku}@aalto.fi, jyamagis@nii.ac.jp

Abstract

Neural network-based models that generate glottal excitation waveforms from acoustic features have been found to give improved quality in statistical parametric speech synthesis. Until now, however, these models have been trained separately from the acoustic model. This creates mismatch between training and synthesis, as the synthesized acoustic features used for the excitation model input differ from the original inputs, with which the model was trained on. Furthermore, due to the errors in predicting the vocal tract filter, the original excitation waveforms do not provide perfect reconstruction of the speech waveform even if predicted without error. To address these issues and to make the excitation model more robust against errors in acoustic modeling, this paper proposes two modifications to the excitation model training scheme. First, the excitation model is trained in a connected manner, with inputs generated by the acoustic model. Second, the target glottal waveforms are re-estimated by performing glottal inverse filtering with the predicted vocal tract filters. The results show that both of these modifications improve performance measured in MSE and MFCC distortion, and slightly improve the subjective quality of the synthetic speech.

Index Terms: statistical parametric speech synthesis, excitation modeling

1. Introduction

Statistical parametric speech synthesis (SPSS) [1] has gained popularity due to its several favorable properties, such as good generalization to unseen text, flexible model adaptation to new speakers with relatively small amount of data, and small runtime resource requirements compared to unit-selection synthesis [2]. The emergence of the neural network-based acoustic models has improved the quality of SPSS systems [3, 4], and the use of sequence models, such as Long Short-Term Memory (LSTM) networks [5, 6] has become widely adopted in SPSS.

Traditionally, the back-end of a SPSS system consists of two separate parts: speech parametrization and waveform generation are done with a vocoder, such as STRAIGHT [7], and an acoustic model is trained to map linguistic features onto the speech parameters. In conjunction, the two major issues in SPSS, over-smoothness of the generated acoustic parameters, and "buzzy" synthetic sound quality have been attributed to the acoustic model and vocoder, respectively. Recent efforts have improved the acoustic model performance resulting in more natural synthetic parameter trajectories using, for example, autoregressive mixture density networks [8], or generative adversarial network-based post-filtering [9]. However, the performance of these systems is still upper-bounded by the analysis-synthesis quality of the vocoder.

Recently, there has been growing interest in joint optimization for the acoustic model and speech parametrization. One such approach was proposed in [10, 11], where a network mapping text features to a cepstrum was trained directly on speech waveforms. Another viable approach to direct speech modeling operates in the spectrogram domain: deep auto encoders for spectral envelope prediction were proposed in [12] and full spectrogram prediction with phase recovery-based synthesis was presented in [13]. The emergence of advanced neural net-based waveform generation methods [14, 15] also seems to lead towards a closer integration between acoustic models and waveform synthesis. Indeed, success in training end-to-end speech synthesis systems with these methods has been reported [16, 17]. Although connected end-to-end, these systems are still initially trained to map text to pre-estimated low-dimensional acoustic features, i.e. mel-generalized cepstrum (MGC) [16] and filter-bank mel-frequency cepstral coefficients (MFCC) [17]. Overall, these two proposed end-to-end systems retain the general structure of parametric TTS systems, while all the system parts are now interconnected neural nets.

Training end-to-end neural TTS systems can be difficult, and they require considerable amounts of training data and computational resources. Furthermore, when a system is fine-tuned end-to-end, any initial intermediate representations, such as filter parameters, lose their interpretability and can not be used directly in signal processing. From the perspective of human voice production, the source-filter model is still relevant, and the explicit use of auto-regressive filters to model the vocal tract is powerful and widespread in many speech applications. The glottal source, the excitation of voiced speech, is a more elementary signal than the speech waveform itself because the glottal source is generated at the level of vocal folds and therefore does not include resonances of the vocal tract. Hence, the glottal excitation is an attractive domain for generative waveform modeling. Indeed, glottal excitation generation with neural nets has been applied successfully to TTS in e.g. [18, 19], while still using relatively lightweight networks for the task. With this approach, glottal pulse excitation waveforms can be generated from acoustic features with a simple feedforward net.

Until now, however, the models for generating glottal excitation waveforms have been trained separately from the acoustic model, and it is therefore not known whether combining the two into an end-to-end framework would benefit the synthesis quality. Specifically, the separate training of excitation and acoustic models leads to mismatch between training and synthesis at both the excitation model input and the desired output. This paper addresses the mismatch by proposing connected training of the system: first the acoustic model is trained normally and

http://dx.doi.org/10.21437/Interspeech.2017-848

fixed, after which the excitation model training (both inputs and outputs) are changed to compensate for the errors in the acoustic model, as detailed in Section 2. Synthesis systems are trained with the suggested modifications for female and male voices in Section 3, and evaluated with objective measures and listening tests. Finally, conclusions are drawn in Section 4.

2. Synthesis system

This paper uses the same text features and the same acoustic model as in [20], while training of the glottal excitation model is modified. For details on using glottal pulses in DNN training, see [18]. Fig. 1 shows an overview of the synthesis system. There are three variants in training the excitation model (right side in figure): (1) the baseline training scheme corresponds to [20], whereas (2) uses generated acoustic inputs for training, and (3) additionally performs inverse filtering with generated vocal tract filters, as elaborated on in Sections 2.4 and 2.5, respectively. The acoustic model part (left side in figure) is shared in all the variants, and the variations marked with (1–3) only affect how the excitation model is trained.

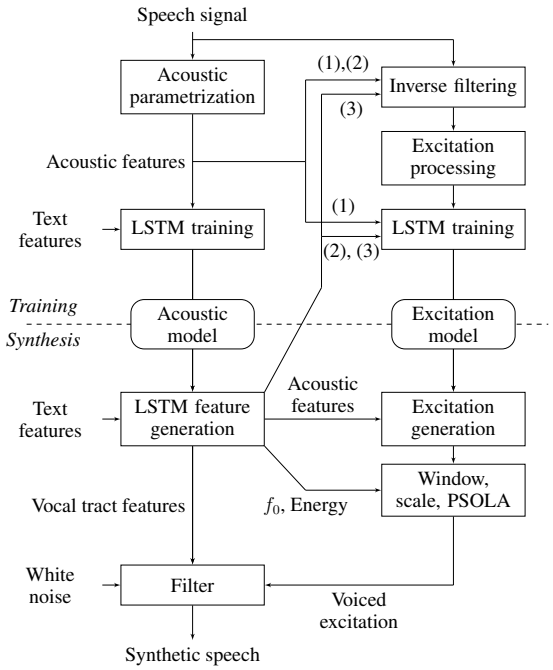

Figure 1: *Overview of the speech synthesis system. The baseline approach (1) uses the original estimated vocal tract filter for both inverse filtering and training the excitation model, whereas systems (2) and (3) are trained with features generated by the acoustic model. Furthermore, the system (3) re-estimates the target waveforms by inverse filtering with the generated vocal tract filter.*

2.1. Text features

In this work, text features refer to the acoustic model neural network inputs, which are derived from full-context linguistic labels extracted from text. The full-context labels include phoneme, syllable, word, phrase, and sentence level information and are created using the Flite [21] speech synthesis frontend and the Combilex [22] lexicon. The text and acoustic features are aligned with the HMM-based speech synthesis system (HTS) [23] and the resulting durations are combined with the labels to create 396-dimensional (per time-frame) text features.

2.2. Acoustic model

In the training phase, acoustic features are estimated from speech signal at a 5-ms frame rate similarly to [20]. The vocal tract filter is estimated with the Quasi Closed Phase (QCP) [24] algorithm, and parametrized with line spectral frequencies (LSFs). Fundamental frequency ($\log f_0$) and voicing decisions (VUV) are extracted with RAPT [25], and Reaper [26] is used for glottal closure instant (GCI) detection. Additionally, glottal source spectral envelope and harmonic-to-noise ratio (HNR) are estimated similarly to [27].

For training the acoustic model, deltas and delta-deltas of the acoustic features are included. The model is a bidirectional LSTM network that maps the frame-rate text features (see Section 2.1) to the dynamic acoustic features. The network configuration is given in Table 1. At synthesis stage, acoustic features are generated on given text inputs, after which the maximum likelihood parameter generation (MLPG) is applied to produce smooth feature trajectories. Additionally, the generated vocal tract parameters are post-filtered with the formant enhancement method presented in [28]. Voiced excitation signal is obtained from the excitation model by conditioning it on the generated acoustic features, while white noise excitation is used for unvoiced speech.

2.3. DNN-based excitation model

Glottal excitation pulses are processed by first performing inverse filtering, and then extracting a two pitch-period segment with one GCI in the middle and two at the both ends (see left side of Fig. 2 for illustration). These pulses are then cosine-windowed and zero-padded to a constant length of 400 samples. Pulses are associated with acoustic features by assigning the pulse nearest to the mid-point of the 5-ms frame, and unvoiced frames are excluded from the training sequences to accommodate LSTM. The network structure used for the excitation models is listed in Table 1. At synthesis time, generated acoustic features are fed into the excitation model, after which the generated pulses are windowed, scaled to the desired energy, and joined with pitch-synchronous overlap-add (PSOLA) [29]. In the baseline excitation model training scheme [18, 20], the original estimated acoustic features are used as input, and the corresponding original glottal pulses are set as the target outputs.

2.4. Training with generated acoustic features

The first issue to be observed in the excitation model training is the difference between how the model is trained and how it is used in synthesis: if the model is trained with the original acoustic features as the input, a mismatch occurs because only synthetic acoustic features will be available at test time. This problem is straightforward to fix by using the acoustic model to generate the training set acoustic features from the text input, and further feed them into the excitation model inputs at training time. This modification is denoted in Fig. 1 by number (2).

Since we want to use the acoustic model outputs for signal processing, we fix the acoustic model after training, and only alter the excitation model to retain connectivity at the input. The acoustic model will inevitably produce some errors, and the synthetic acoustic features will be corrupted versions of the

original ones. As a result, this mismatch may cause the waveform model to overfit the natural acoustic features, while generalizing poorly with the synthetic acoustic input. This behavior is illustrated in Fig. 3 (training details in section 3.2). The use of generated acoustic features for excitation model input can also been seen as a form of regularization, where training with corrupted input helps to prevent overfitting.

2.5. Re-estimating glottal excitation waveforms

The second mismatch issue is related to the output of the excitation model: since the acoustic model is fixed after training, synthesized vocal tract filters will differ from the original (see Fig. 2), and can no longer give perfect reconstruction even with the original excitation waveform. However, we can re-estimate the excitation model target waveforms by performing inverse filtering with the predicted filter—this aims to compensate for the errors made in acoustic modeling. A similar idea has been adopted in speech coding, where the encoder uses quantized filter parameters (that are always utilized by the decoder), rather than using the more accurate non-quantized filter parameters, e.g. [30]. This change is denoted in Fig. 1 by number (3).

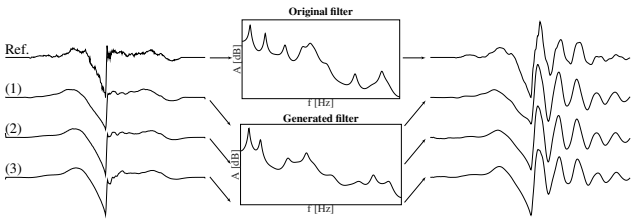

Figure 2: *Since the excitation waveforms are re-estimated for system (3), comparing generated excitations (left) directly with the reference excitation is not meaningful. Instead, objective metrics are calculated in speech signal domain (right) by filtering the excitations with either the true vocal tract filter (top mid) or the generated one (bottom mid).*

3. Experiments

3.1. Speech material

Two speaker-specific systems were trained for the experiments. Both speakers (one male, one female) are professional British English voice talents. The dataset for the male speaker "Nick" comprises 2542 utterances, totaling 1.8 hours, and the dataset for the female speaker "Jenny" comprises 4314 utterances, totaling 4.8 hours. For both speakers, 100 utterances were randomly selected for both validation and testing, while the rest were used for training the systems. The material was downsampled to 16 kHz sample rate from the original 48 kHz rate.

3.2. Training

The acoustic model network consist of two feedforward (FF) layers with logistic activation function, with two bidirectional LSTM layers stacked on top of them. The model was trained with stochastic gradient descent and early stopping was applied after 5 epochs of no improvement on validation set. Dynamic features were included for the acoustic model outputs.

The excitation models were trained similarly, but now only static acoustic features were used at the model input. This is because we want to exactly match the domain of the acoustic features being used at waveform synthesis, which for generated features includes MLPG and the vocal tract formant enhancement post-filter. To illustrate the mismatch issue, Fig. 3 shows the excitation model training and test set errors as a function of epochs. Direct comparison is applicable between the two systems which attempt to predict the original glottal waveform, either from the original (AC-GL) or the generated (GEN-AC-GL) acoustic features. In the first case, poor generalization on test set is evident for both voices: excitation model fits tightly to the original acoustic features of the training set, while the test set performance does not improve. On the other hand, training with the generated acoustic features does not reach as low training error, but performs considerably better on the test set. The third system using re-estimated target waveforms was trained with network structure and generated acoustic inputs similar to GEN-AC-GL. The CURRENNT toolkit [31] was utilized for training all the networks.

Table 1: *Two types of networks are used in the system: acoustic model maps text features (TXT) to acoustic (AC) delta-features, while excitation model maps acoustic features (excluding voicing decision) to glottal pulse waveforms (GL).*

Network layer	Acoustic model	Excitation model
Input (size)	TXT (396)	AC (47)
Hidden (size)	FF (512)	LSTM (128)
	FF (512)	FF (512)
	LSTM (256)	FF (512)
	LSTM (256)	FF (512)
Output (size)	AC (142)	GL (400)

3.3. Objective evaluation

For objective evaluation, we aim to measure the difference between the target speech signal and the synthesised speech after filtering the generated excitation. Since the systems were trained with the MSE criterion, the same metric should be used for the evaluation. However, measuring point-wise errors directly on the final synthetic speech waveform is difficult, since synchronism is broken in overlap-add due to the original and generated pitch being different. Nevertheless, the final result can be closely approximated by individually filtering the generated pulses frame-by-frame, as illustrated in Fig. 2. The target waveform is created by filtering the original estimated glottal pulse with the original vocal tract filter, whereas the generated pulses from various systems are filtered with the filter given by the acoustic model.

In addition to MSE, MFCC distortion is calculated, as this measure is commonly used in speech applications and roughly correlates with perceptual differences. The MFCCs were calculated with a filterbank size of 24, 13 cepstrum coefficients were used, and the distortion based on squared MFCC error is given in decibels. Fig. 4 shows the average MSE and MFCC distortions over the voiced frames in the test set. The trend is clear for both voices: training with generated acoustic features reduces the objective errors, and using re-estimated target waveforms further improves performance.

3.4. Listening experiment

For the listening test samples, text features with forced alignment durations were used as the acoustic model input in order to focus on the combined performance of the acoustic and excitation generation models. An A-B preference test was per-

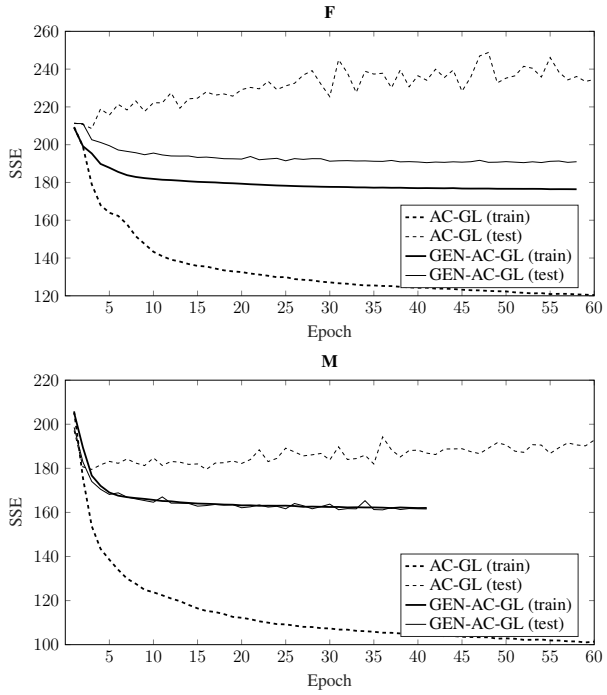

Figure 3: *Errors on the training (thick lines) and test (thin lines) sets for female (top) and male (bottom) voices illustrate the issue with disconnected training. Training set acoustic feature inputs are either the original (AC-GL) or generated (GEN-AC-GL) from the acoustic model, while the test set inputs are always generated. Due to the mismatch effect, AC-GL does not generalize well on the test set.*

formed on the three DNN excitation models, where the listeners were asked to evaluate the overall quality of the two samples, and indicate which one they preferred. Additionally, an option to give no preference was allowed. The listening test was implemented on the CrowdFlower [32] crowd-sourcing platform CrowdFlower. The test material consisted of 20 samples for both voices and each method pair, presented in random order. Additionally, 10 samples degraded with added noise in the excitation and over-smoothed acoustic feature trajectories were included for post-screening of subjects. Participants who answered less than 70% of the screening questions correctly were excluded. The test was made available in English-speaking countries, in addition with the top four countries in EF English Proficiency Index [33]. 50 listeners participated in the test, resulting in a total 3084 of pair comparisons after screening. Results are listed in Table 2. The p-values were calculated with a binomial test between A and B, excluding the no-preference (neutral) answers from the calculation.

The results show a small but statistically significant difference in favor of the proposed training schemes (2) and (3) using generated acoustic features over the baseline approach (1). Training on the re-estimated glottal waveforms (3) also seems to the improve performance slightly.

4. Conclusions

This study addresses two sources of mismatch that occur when glottal excitation models and acoustical models are trained

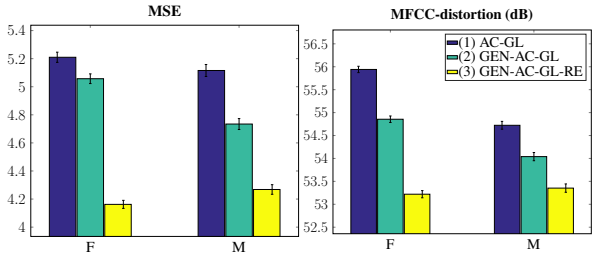

Figure 4: *Average test set MSE and MFCC distortion with 95% confidence intervals for the male (M) and the female (F) voice. The baseline training scheme (1) consistently gives higher errors than the ones (2–3) trained with generated acoustic features. Using re-estimated excitation waveforms in training (3) further reduces the errors.*

Table 2: *A-B preference test scores (system numbering as above). Systems trained with generated acoustic features were slightly preferred over the baseline training scheme. Given p-values were calculated with a binomial test between A-B, while excluding the no-preference ratings from the calculation.*

speaker	(1)	(2)	(3)	neutral	p-value
Female	7.98	**14.79**		77.2	0.0008
	9.30		**12.98**	77.7	0.0464
		5.81	**12.79**	81.4	0.0002
Male	6.30	**12.60**		81.1	0.0007
	9.34		11.09	79.6	0.2176
		7.36	**12.02**	80.6	0.0105

and used for parameter generation in TTS. The results show that training the excitation model with inputs generated by the acoustic model improved generalization to test set, reduced the objective error metrics and slightly improved the perceived quality of synthetic speech. Additionally, training with excitation waveforms re-estimated by glottal inverse filtering with the generated vocal tract filters further improved both objective and subjective performance.

Apart from the proposed training scheme, another way to reduce mismatch between original and generated acoustic features is to improve the acoustic model. Nevertheless, the ideas presented here still apply while the acoustic features are used to control generative neural models. The shortcomings in the acoustic model can be addressed in the future with the use of more advanced models, such as mixture density network LSTMs [8].

From a waveform generation perspective, the current DNN-based excitation models trained with the squared error criterion are inherently limited to generating conditional averages. As such, they cannot reproduce the stochastic properties desirable in excitation signals. Potential extensions to our excitation modeling framework include the use of more powerful generative neural network methods, such as generative adversarial networks, or some form of the recent powerful sample-by-sample generative methods, e.g. [15, 14].

5. Acknowledgements

This work was supported by the Academy of Finland (proj. no. 284671) and MEXT KAKENHI Grant Numbers (26280066, 15H01686, 15K12071, 16H06302).

6. References

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

[2] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

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

[4] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, and L. Deng, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *Signal Processing Magazine, IEEE*, vol. 32, no. 3, pp. 35–52, May 2015.

[5] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks." in *Interspeech*, 2014, pp. 1964–1968.

[6] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4470–4474.

[7] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *MAVEBA*, 2001.

[8] X. Wang, S. Takaki, and J. Yamagishi, "An autoregressive recurrent mixture density network for parametric speech synthesis," in *Proc. of ICASSP*. IEEE, 2017, pp. 4895–4899.

[9] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *Proc. of ICASSP*, March 2017, pp. 4910–4914.

[10] K. Tokuda and H. Zen, "Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis," in *Proc. of ICASSP*, April 2015, pp. 4215–4219.

[11] ——, "Directly modeling voiced and unvoiced components in speech waveforms by neural networks," in *Proc. of ICASSP*, March 2016, pp. 5640–5644.

[12] S. Takaki and J. Yamagishi, "A deep auto-encoder based low-dimensional feature extraction from FFT spectral envelopes for statistical parametric speech synthesis," in *Proc. of ICASSP*, March 2016, pp. 5535–5539.

[13] S. Takaki, H. Kameoka, and J. Yamagishi, "Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis," in *Interspeech (submitted)*, 2017.

[14] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *Pre-print*, 2016, https://arxiv.org/pdf/1609.03499.pdf.

[15] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. C. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," in *ICLR 2017 (submission)*, 2017. [Online]. Available: http://arxiv.org/abs/1612.07837

[16] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2Waw: End-to-end speech synthesis," in *ICLR 2017 workshop (submission)*, 2017, https://openreview.net/pdf?id=B1VWyySKx.

[17] 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 *ICML 2017 (submission)*, 2017, https://arxiv.org/pdf/1702.07825.pdf.

[18] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," in *"Proc. of ICASSP"*, Mar. 2016, pp. 5120–5124.

[19] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku, "GlottDNN—a full-band glottal vocoder for statistical parametric speech synthesis," in *Proc. of Interspeech*, 2016.

[20] L. Juvela, X. Wang, S. Takaki, M. Airaksinen, J. Yamagishi, and P. Alku, "Using text and acoustic features in predicting glottal excitation waveforms for parametric speech synthesis with recurrent neural networks," in *"Proc. of Interspeech"*, Sep. 2016, pp. 2283–2287.

[21] A. W. Black and K. A. Lenzo, "Flite: a small fast run-time synthesis engine," in *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.

[22] K. Richmond, R. A. Clark, and S. Fitt, "Robust LTS rules with the Combilex speech technology lexicon," in *Proc. of Interspeech*, Brighton, September 2009, pp. 1295–1298.

[23] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system version 2.0," in *Proc. of ISCA SSW6*, Bonn, Germany, August 2007, pp. 294–299.

[24] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 3, pp. 596–607, March 2014.

[25] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[26] ——, "REAPER: Robust Epoch And Pitch EstimatoR," https://github.com/google/REAPER, 2015.

[27] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 153–165, January 2011.

[28] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Comparison of formant enhancement methods for HMM-based speech synthesis." in *SSW*, 2010, pp. 334–339.

[29] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech communication*, vol. 16, no. 2, pp. 175–205, 1995.

[30] "ETSI TS 126 090: Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding functions" ," European Telecommunications Standards Institute, Technical specification, 2016.

[31] F. Weninger, "Introducing CURRENNT: The Munich open-source CUDA recurrent neural network toolkit," *Journal of Machine Learning Research*, vol. 16, pp. 547–551, 2015.

[32] CrowdFlower Inc. (2017) Crowd-sourcing platform. [Online]. Available: https://www.crowdflower.com/

[33] (2017) EF English proficiency index. [Online]. Available: http://www.ef.com/epi/