Bollepalli, Bajibabu; Juvela, Lauri; Alku, Paavo

Generative adversarial network-based glottal waveform model for statistical parametric speech synthesis

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

DOI:
10.21437/Interspeech.2017-1288

Published: 01/08/2017

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

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

Generative adversarial network-based glottal waveform model for statistical parametric speech synthesis

Bajibabu Bollepalli, Lauri Juvela, Paavo Alku

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

Abstract

Recent studies have shown that text-to-speech synthesis quality can be improved by using glottal vocoding. This refers to vocoders that parameterize speech into two parts, the glottal excitation and vocal tract, that occur in the human speech production apparatus. Current glottal vocoders generate the glottal excitation waveform by using deep neural networks (DNNs). However, the squared error-based training of the present glottal excitation models is limited to generating conditional average waveforms, which fails to capture the stochastic variation of the waveforms. As a result, shaped noise is added as post-processing. In this study, we propose a new method for predicting glottal waveforms by generative adversarial networks (GANs). GANs are generative models that aim to embed the data distribution in a latent space, enabling generation of new instances very similar to the original by randomly sampling the latent distribution. The glottal pulses generated by GANs show a stochastic component similar to natural glottal pulses. In our experiments, we compare synthetic speech generated using glottal waveforms produced by both DNNs and GANs. The results show that the newly proposed GANs achieve synthesis quality comparable to that of widely-used DNNs, without using an additive noise component.

Index Terms: Glottal souce modelling, GAN, TTS, DNN

1. Introduction

Statistical parametric speech synthesis (SPSS) and concatenative synthesis are the two predominant paradigms in text-to-speech technology. SPSS systems have several advantages over concatenative synthesis, such as their flexibility to transform the synthesis to different voice characteristics, speaking styles and emotions, as well as their small memory footprint and robustness to unseen text prompts [1]. The main drawback of SPSS is that the quality of synthetic speech is worse than that of concatenative synthesis. There are three major factors behind this: quality of vocoders, acoustic modeling accuracy, and over-smoothing [2].

Recent use of neural network-based acoustic models [3], especially sequence models, such as long short-term memory (LSTM) networks [4, 5], have addressed primarily the acoustic modelling accuracy and to some extent also the over-smoothing problem [6, 7]. Although the progress in acoustic modelling has improved the synthesis, the quality achieved by the best SPSS systems is still limited by the copy-synthesis quality of the vocoder. Thus in this study we focus on improving the quality of vocoders.

In SPSS systems, vocoders are used for speech parametrization and waveform generation. Vocoders used in SPSS can be grouped into three main categories: mixed/impulse excited vocoders (e.g. STRAIGHT [8, 9]), glottal vocoders (e.g. GlottHMM [10] and GlottDNN [11]), and sinusoidal vocoders (e.g. quasi harmonic model [12, 13]). The first two categories are based on the source-filter model of speech production and they differ mainly on the interpretation of voiced excitation signal. In glottal vocoding, the excitation is assumed to correspond to the time-derivative of the true airflow generated at the vocal folds (consisting of the combined effects of the glottal volume velocity and lip radiation [14]), and the filter corresponds to a transfer function that is created by the physiology of the human vocal tract. Recent studies have shown that glottal vocoding can improve the synthesis quality [10, 15, 16].

The first glottal vocoder [10] used a single glottal pulse to create the voiced excitation waveform—the pulse was modified according to the estimated acoustic parameters to build the entire excitation signal. This straightforward use of a single glottal pulse was replaced in later studies [17, 15] with deep neural networks (DNNs) to predict the glottal pulse waveforms from acoustic features, where the actual estimated glottal pulses were set as optimization targets. However, the mean squared error-based training of the DNN-based glottal excitation models is only able to generate conditional average waveforms, which fails to capture the stochastic variation of the waveforms. In order to tackle this drawback, the excitation was post-processed by adding shaped noise to the waveform. In this study, we propose an alternative method for predicting glottal excitation waveforms for SPSS, by using a new training strategy with generative adversarial networks (GANs) [18].

As the research in deep learning progress, new advanced neural networks capable of generating raw signal waveforms directly from linguistic features are proposed in text-to-speech (e.g WaveNet [19], Deep Voice [20], and [21]). Although these systems produce high-quality synthetic speech they are not yet applicable in real-time speech synthesis due to heavy computational requirements. In contrast, the widely-used source-filter model is an applicable means to express speech in parametric forms as shown by its widespread applications [22, 23]. In addition, the excitation of the source-filter model, the glottal flow, is an elementary time-domain signal (particularly when compared to the speech pressure signal) because it is produced at the level of glottis in the human larynx in the absence of vocal tract resonances. Hence, the glottal excitation is an attractive domain for generative waveform modeling.

Recently, GANs have started to emerge in TTS applications. In [24], a GAN was employed as a post-filter to address the over-smoothing problem in predicted acoustic parameters in SPSS. Results of [24] showed that GANs are capable of producing detailed speech spectra, including also the modulation spectrum, resulting in increased synthesis quality. In [25], adversarial type of training was used to take into account an anti-spoofing verification as an additional constraint in the acoustic model training. The current study is the first investigation to use GANs to model the glottal waveform as an excitation waveform in SPSS.

http://dx.doi.org/10.21437/Interspeech.2017-1288

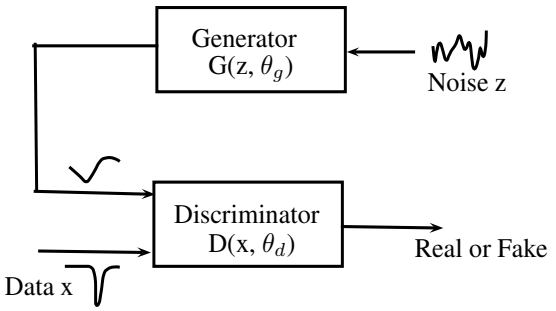


Figure 1: *General block diagram of generative adversarial networks (GANs).*

2. Generative adversarial networks

Generative adversarial networks (GANs) are generative models that have shown a huge success in unsupervised learning [26]. In GANs a new type of training procedure is employed by an adversarial process where two models, generator G and discriminator D, compete with each other. Figure 1 illustrates the block diagram of a GAN. During training, G starts from sampling input variables $\boldsymbol{z}$ from a uniform or Gaussian distribution $p_{\boldsymbol{z}}(\boldsymbol{z})$, then maps the input latent variables $\boldsymbol{z}$ to data space $G(\boldsymbol{z};\theta_g)$ through a differentiable network. D is a classifier $D(\boldsymbol{x};\theta_d)$ that aims to discriminate whether a sample is a real one from the training data or a fake generated by G. In this framework, D and G play a two-player minmax game with the following binary cross entropy:

$$\min_G \max_D V_{\text{GAN}}(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]. \quad (1)$$

In training, updates are alternated between G and D, but the error gradient always propagates through the classifier D. The main theoretical advantage of this framework is that the parameters θ_g and θ_d can be learned through back propagation without making any assumptions on the data distribution [18].

3. GAN-based glottal waveform model

The regular or vanilla GAN [18] framework is modified in the following manner to model glottal waveforms.

3.1. Conditional generative adversarial networks (CGAN)

Generator G in regular GANs has no control on modes of data it generates. In [27] it was shown that by conditioning the model on additional information it is possible to direct the data generation process. Since the goal of the current study is to generate glottal pulses based on acoustic parameters, we conditioned both the generator and discriminator by the acoustic parameters $\boldsymbol{y}$. The objective function in Eq. 1 can be rewritten with conditional variable $\boldsymbol{y}$ as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x}|\boldsymbol{y})] + \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z}|\boldsymbol{y})))]. \quad (2)$$

3.2. Convolutional architecture

In regular GANs, both discriminator and generator employ a simple feed-forward neural network for learning. However, numerous studies have shown (e.g. [26, 28] that convolutional architectures yield better generated outputs than simple feed-forward networks. Our generator network consists of only convolutional layers and hence the local temporal characteristics in glottal waveforms can be effectively preserved with a relatively small number of weights.

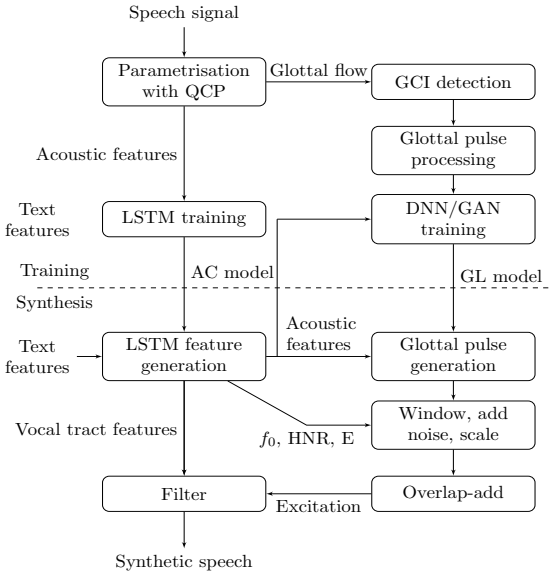


Figure 2: *Block diagram of the LSTM-based speech synthesis system using the GlottDNN vocoder.*

3.3. Least Squares Generative Adversarial Networks

In the regular GAN, the discriminator is a classifier and uses binary cross-entropy as a loss function. In [29] it was shown that this kind of loss function can lead to problems due to vanishing gradients when updating the parameters of the generator. Thus the loss function of discriminator in Eq. 2 is modified to the least square function:

$$\min_D V_{\text{LSGAN}}(D) = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}\left[(D(\boldsymbol{x}|\boldsymbol{y}) - 1)^2\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}\left[(D(G(\boldsymbol{z}|\boldsymbol{y})))^2\right] \quad (3)$$

4. Experiments

4.1. Speech material

We employed data of one female speaker recorded by a professional British English voice talent, labeled as "Jenny". The speech data consisted of 4314 utterances summing to 7 hours and 51 minutes. A total of 100 utterances were randomly selected for validation and testing, and the rest were used for training the systems. The sampling frequency of the corpus was 16 kHz.

4.2. Feature extraction

4.2.1. Linguistic features

Figure 2 illustrates the block diagram of our LSTM-based speech synthesis system used in this study. The text files of the utterances were provided along with corpus. The full contextual labels were obtained by the Flite [30] speech synthesis front end and the Combilex [31] lexicon. To align the labels and acoustic features at phoneme level, the HMM-based force alignment was used. The full-context labels were represented

Table 1: *Acoustic features used in training the LSTM-based AC model, the DNN and the GAN-based GL model.*

Feature	Type/Unit	Dimension
Vocal tract spectrum	LSF	30
Energy	dB	1
Fundamental frequency	$\log F_0$	1
Harmonic-to-noise ratio	dB/ERB	5
Voice source spectrum	LSF	10
Total	-	47

into binary and numerical features by the question file used in the HMM-based speech synthesis system. These features convey information about the phoneme identity, syllable location, part-of-speech, number of words in an utterance, and number of phrases in an utterance. In total, the input feature vector was 396 in dimension (per time-frame) including the extra numerical values which provide information about the frame position in a given phoneme.

4.2.2. Acoustic features

Both the vocal tract and voice source parameters, shown in Table 1, were extracted using the GlottDNN vocoder [11]. The acoustic parameters were extracted at a 5-ms frame rate. The $\log F0$ was linearly interpolated to fill unvoiced regions and an extra binary V/UV feature was added to code the voiced/unvoiced information. The output parameters included both static and dynamic (with delta, and delta-delta) features. Thus, in total, the output feature was 142 dimensional. The input features were normalized to the range of [0.1, 0.99] by using the min-max method, while output features were normalized using the mean-variance normalization method. The development and evaluation set were normalized by the values derived from the training data. At synthesis time, the maximum likelihood parameter generation (MLPG) [32] algorithm was applied on predicted acoustic parameters using the global variances to generate smooth parameter trajectories.

4.3. Acoustic (AC) model

The acoustic model network consisted of four hidden layers which were followed by a linear layer at the output. The four hidden layers consisted of two feed-forward layers at bottom and two bidirectional LSTM layers on top. The bottom feed-forward layers were intended to act as feature extraction layers, with 512 hidden units using logistic activation function in each layer. The top two layers had 256 bidirectional LSTM blocks in each layer. The stochastic gradient descent algorithm was used to learn the parameters and early stopping criterion was adopted to reduce overfitting.

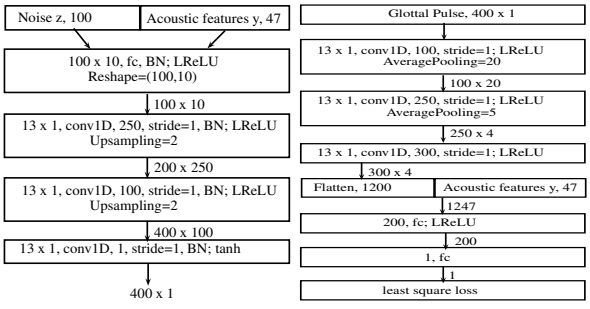
4.4. Glottal excitation (GL) model

4.4.1. DNN-based glottal excitation model

The GL model using DNNs was developed as described in [15]. The input features to the network were the same acoustic features as described in Table 1 (i.e. 47 in dimension) and the outputs were two pitch-period windowed glottal flow waveforms centered and zero-padded to 400 time-domain samples. The acoustic parameters predicted by the AC model were employed to train the GL model instead of the original acoustic features, in contrast to [15]. The main motivation for this change is to reduce the mismatch in the acoustic feature inputs between training and testing time – we provide a detailed analysis on this issue in [33]. The DNN architecture consisted of three hidden layers each with 512 units. The logistic and linear activations were used for hidden and output layers, respectively.

4.4.2. GAN-based glottal excitation model

Four types of GL models were developed using GANs [1]. The first model was a vanilla GAN, denoted as "GAN", where the DNNs were employed for both the generator and discriminator. The generator consisted of three hidden layers followed by an output layer. The discriminator consisted four hidden layers followed by an output layer. Each hidden layer had 1024 units and the activation function was a leaky rectified linear unit (LReLU). The tangent hyperbolic and sigmoid activation functions were employed in the output layer of the generator and discriminator, respectively. The batch-normalization was employed on the generator network [34]. The second model was a conditional GAN, denoted as "CGAN", same as the vanilla GAN except that it was conditioned by acoustic features in both the generator and discriminator. The third model, denoted as "CGAN+CNN", was a conditional GAN with deep convolutional neural networks in both the generator and discriminator. The fourth model, denoted as "CGAN+CNN+LS", was similar to the third model except that the least square loss was used in the discriminator. Architectures of the third and fourth models were illustrated in Figure 3. The noise vector $\boldsymbol{z}$ had a dimension of 100 and was sampled from Gaussian distribution $\mathcal{N}(0, 0.5)$.



(a) Generator (b) Discriminator

Figure 3: *Architectures of models CGAN+CNN and CGAN+CNN+LS. BN: batch normalization, fc: fully connected layer, conv1D: 1D convolution.*

4.5. Objective evaluation

The main drawback of GANs is the lack of an explicit objective score to measure the performance of the generator [18]. Therefore, visual inspection is typically adopted [26]. In the current study, simple objective scores, the mean square error (MSE) and Pearson correlation coefficient (PCC) computed between the actual and generated glottal pulses, were used. The obtained objective scores, computed as an average over 100 utterances (with 43746 glottal pulses), are presented in Table 2.

The MSE value of the baseline DNN system was lower than that of the other systems, but this was expected since the DNN system was trained to minimize the MSE cost function. Among the proposed models, deep convolutional neural network (CNN) -based GAN models outperformed the DNN-based GAN models in both MSE and PCC. Figure 4 shows a few example pulses generated by the proposed methods. It can be seen that the glottal pulses generated by the CNN-based GAN models are visu-

[1] Code is available at https://github.com/bajibabu/GlottGAN

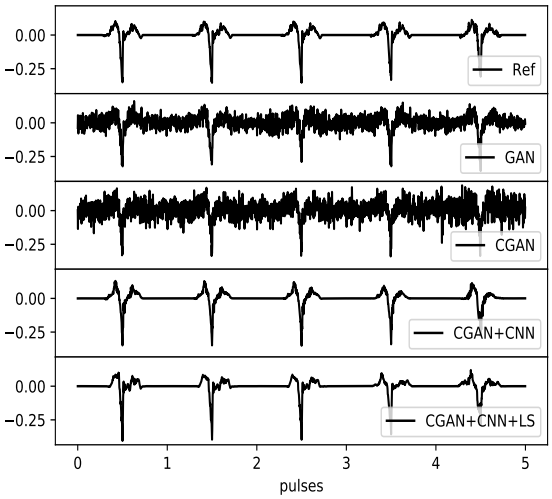


Figure 4: *Glottal pulses generated by different generative adversarial networks (GANs). Ref: natural reference, GAN: vanilla GANs, CGAN: conditional GAN, CGAN+CNN: conditional GAN with deep convolutional neural networks, and CGAN+CNN+LS: same as CGAN+CNN but least square loss is used by the discriminator.*

Table 2: *The objective scores of GANs and DNN. MSE: mean square error, PCC: Pearson correlation coefficient.*

Model	MSE	PCC
DNN	0.2458	0.86
GAN	0.66	0.68
CGAN	1.9135	0.54
CGAN+CNN	0.4469	0.76
CGAN+CNN+LS	0.4644	0.76

ally much closer to the reference pulses than the corresponding pulses generated by the DNN-based GAN.

Figure 5 shows the voiced souce excitation signal after pitch-synchronous ovelap-add (PSOLA) [35]. The excitation signal generated by the baseline DNN is smooth and without a noise component, and therefore shaped noise is added to the signal to match the predicted HNR values. The GAN-based model, however, is able to generate a noise component similar to the reference waveform without using any HNR-based post-processing.

4.6. Subjective evaluation

Subjective evaluation was conducted with the comparison category rating (CCR) test [36] between three systems: the baseline DNN (denoted "DNN"), the baseline DNN with HNR (denoted "DNN+HNR") and the GAN-based glottal generation (denoted "GAN"). Among the GAN-based glottal generation models, we selected the CGAN+CNN+LS system since it performed better than the other systems in informal listening tests. A total of 11 utterances from the test set were randomly selected for the listening test.

A crowd sourcing platform, CrowdFlower [37], was employed for the subjective evaluation and followed the same setup as in [16]. A set of 13 utterances were used as control utterances that included null pairs and anchor samples [36]. Listeners who performed with at least 75 % accuracy were allowed to participate in the actual listening test. The tests were made available to the English speaking countries, and top four countries in EF English Proficiency Index rankings [38]. A total of

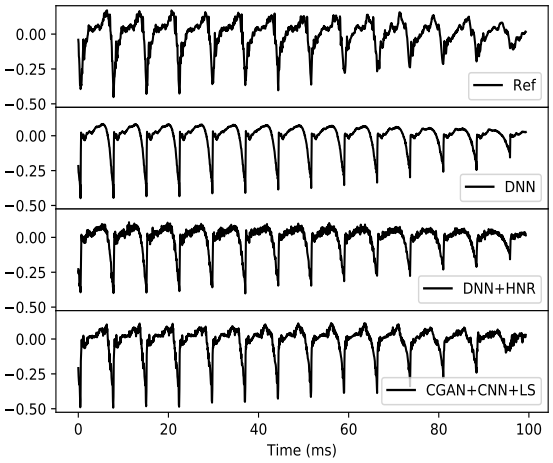


Figure 5: *Glottal excitation signals after PSOLA. Ref: Reference excitation signal. DNN: excitation generated using the baseline DNN model. DNN+HNR: baseline DNN with additive shaped noise. CGAN+CNN+LS: convolutive LS-GAN conditioned with acoustic features.*

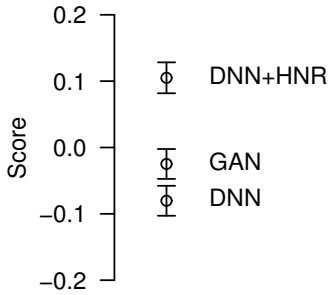


Figure 6: *Subjective listening test results (CCR test) with their 95% confidence intervals on synthesis quality.*

3850 judgments were made by 50 listeners.

The results of the listening test are shown in Figure 6. The DNN+HNR method performed better than other two methods, indicating the perceptual relevance of a stochastic component in excitation. Moreover, in the comparison between GAN and the DNN without HNR, the former was rated slightly higher. This is a likely related to GANs ability to generate stochastic variability, rather than producing smooth glottal waveforms as done by the DNN.

5. Conclusions

This study proposed a new method to model glottal excitation waveforms in statistical parametric speech synthesis using generative adversarial networks (GANs). We modified the vanilla GAN in various forms comparing the system performance in generation of glottal pulses. In our experiments, the deep convolutional neural networks -based GANs outperformed the DNN-based GANs. We also compared glottal pulses generated by the GANs with DNNs. The subjective evaluation gave encouraging evidence showing that GANs are more able to reproduce the stochastic component in the glottal excitations than DNNs. The GANs are still relatively new and definitely require more research to understand their full potential in SPSS.

6. References

[1] 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.

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

[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] 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.

[5] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. of ICASSP*. IEEE, 2015, pp. 4470–4474.

[6] Z. Wu and S. King, "Improving trajectory modelling for DNN-based speech synthesis by using stacked bottleneck features and minimum generation error training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1255–1265, 2016.

[7] 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.

[8] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[9] 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.

[10] 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.

[11] 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.

[12] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[13] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 184–194, April 2014.

[14] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2–3, pp. 109–118, 1992, Eurospeech '91.

[15] 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.

[16] M. Airaksinen, B. Bollepalli, J. Pohjalainen, and P. Alku, "Glottal vocoding with frequency-warped time-weighted linear prediction," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 446–450, April 2017.

[17] T. Raitio, A. Suni, L. Juvela, M. Vainio, and P. Alku, "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort," in *Proc. of Interspeech*, Singapore, September 2014, pp. 1969–1973.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[19] 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.

[20] 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.

[21] H. Zen, "Generative model-based text-to-speech synthesis," 2017, invited talk given at CBMM workshop on speech representation, perception and recognition.

[22] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Computer Speech & Language*, vol. 28, no. 5, pp. 1117–1138, 2014.

[23] P. Alku, "Glottal inverse filtering analysis of human voice production – a review of estimation and parameterization methods of the glottal excitation and their applications. (invited article)," *Sadhana – Academy Proceedings in Engineering Sciences*, vol. 36, no. 5, pp. 623–650, 2011.

[24] 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.

[25] Y. Saito, S. Takamichi, and H. Saruwatari, "Training algorithm to deceive anti-spoofing verification for dnn-based speech synthesis," in *ICASSP, New Orleans, USA*, 2017, pp. 4900–4904.

[26] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[27] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[28] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," *arXiv preprint arXiv:1703.02205*, 2017.

[29] X. Mao, Q. Li, H. Xie, R. Y. Lau, and Z. Wang, "Least squares generative adversarial networks," *arXiv preprint arXiv:1611.04076v2*, 2017.

[30] 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.

[31] 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.

[32] H. Zen, K. Tokuda, and T. Kitamura, "An introduction of trajectory model into HMM-based speech synthesis," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[33] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "Reducing mismatch in training of DNN-based glottal excitation models in a statistical parametric text-to-speech system," in *"Submitted to Interspeech"*, 2017.

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[35] 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.

[36] Recommendation ITUTP, "800, methods for subjective determination of transmission quality," *International Telecommunication Union*, 1996.

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

[38] (2017) EF English Proficiency Index. [Online]. Available: http://www.ef.com/epi/