Waveform Generation for Text-to-speech Synthesis Using Pitch-synchronous Multi-scale Generative Adversarial Networks

WAVEFORM GENERATION FOR TEXT-TO-SPEECH SYNTHESIS USING PITCH-SYNCHRONOUS MULTI-SCALE GENERATIVE ADVERSARIAL NETWORKS

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

[1]Aalto University, Finland [2]National Institute of Informatics, Japan
[3]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

ABSTRACT

The state-of-the-art in text-to-speech (TTS) synthesis has recently improved considerably due to novel neural waveform generation methods, such as WaveNet. However, these methods suffer from their slow sequential inference process, while their parallel versions are difficult to train and even more computationally expensive. Meanwhile, generative adversarial networks (GANs) have achieved impressive results in image generation and are making their way into audio applications; parallel inference is among their lucrative properties. By adopting recent advances in GAN training techniques, this investigation studies waveform generation for TTS in two domains (speech signal and glottal excitation). Listening test results show that while direct waveform generation with GAN is still far behind WaveNet, a GAN-based glottal excitation model can achieve quality and voice similarity on par with a WaveNet vocoder.

Index Terms— Neural vocoding, text-to-speech, GAN, glottal excitation model

1. INTRODUCTION

Recent advances in deep learning have led to text-to-speech (TTS) systems achieving near-human naturalness [1]. This is partially due to neural sequence-to-sequence mapping methods that can learn to align and map between input text and output acoustic feature sequences [2]. Another major progress in TTS is the introduction of waveform generation methods, such as WaveNet [3], that have been adopted to use as "neural vocoders" [4]. Generally, a neural vocoder is a neural network that, given some input acoustic features, outputs (speech) waveforms. This approach effectively decouples the acoustic mapping and waveform generation models from each other. Indeed, WaveNet-based neural vocoders have been successfully applied in TTS using either sequence-to-sequence techniques [1] or more traditional statistical parametric speech synthesis (SPSS) acoustic models [5]. Although WaveNets and other autoregressive models produce impressive results, their inference process is inherently slow, and requires heavy optimization for real time applications [6, 7]. Alternative approaches capable of parallel inference have been proposed [8, 9], but these models are increasingly complex and difficult to train.

Meanwhile, generative adversarial networks (GANs) have achieved impressive results in image synthesis [10], and they are currently of increasing interest in speech applications. Recent advances are powered in part by more appropriate architectures, including residual connections [11] progressive upsampling [10, 12] and multi-scale processing [13]. Similar mechanisms can be seen in the WaveNet architecture (cf. exponentially growing dilations and residual connections). On the other hand, improved training techniques (e.g. [14, 11]) have also resulted in better stability and convergence properties for GANs. Nevertheless, while some speech-related GAN applications apply direct waveform-to-waveform transformations [15, 16], most TTS applications have so far used GANs on improving acoustic models instead of directly generating waveforms [17, 18].

Another approach to facilitate neural waveform generation for speech synthesis is the pitch-synchronous utilization of glottal waveforms (i.e. signals representing the true air flow excitation of speech generated by the vocal folds). In this approach, glottal inverse filtering [19] is applied to the speech signal in order to remove the vocal tract resonances and to obtain a more elementary signal that is easier to model. Early approaches used a point-wise least squares loss in time-domain [20, 21], and while this captures the gross waveshape well, the produced output is essentially a conditional average and lacks in stochastic high frequency contents (due to the averaging). The missing stochastic component can be recreated using signal processing techniques for aperiodicity modification, resulting in high quality synthetic speech [22, 23], but this involves making signal model assumptions that may not hold generally. Further efforts have been made to model the stochastic part directly using GANs [24] or WaveNet ("GlotNet") [21]. However, the former approach suffers from the unstable GAN training techniques available at the time, and for the latter WaveNet inference limitations still apply. A similar excitation modeling approach can be extended to generating residual excitation signals for waveform synthesis from MFCCs, as MFCCs can be readily interpreted as spectral envelopes and cancelled from the signal via inverse filtering [25].

This paper proposes a novel multi-scale GAN architecture for pitch-synchronous waveform generation. The proposed generator performs progressive upsampling of feature maps and outputs waveforms at multiple timescales, while the discriminator ensures that the waveforms remain valid at each timescale. The model is trained using a Wasserstein GAN [14] with modified gradient penalties [11] and an FFT-based auxiliary loss, leading to stable training behavior for both direct speech waveform and glottal excitation signals. The proposed model is evaluated as a neural vocoder for an SPSS system and compared with the GlottDNN and WaveNet vocoders. The results show that the proposed "GlotGAN" can achieve similar performance to the WaveNet vocoder, while the direct waveform "WaveGAN" still falls short in performance compared to the other methods. Computational benefits of the proposed model include parallel inference (due to GAN) and explicit fundamental frequency control (due to pitch-synchronous processing).

2. SPEECH SYNTHESIS SYSTEM

The focus of this paper is on neural vocoders, and we use a conventional SPSS pipeline for the text-to-acoustic-features mapping. First,

text is converted to linguistic features using the Flite speech synthesis front-end [26] and the Combilex lexicon [27]. Alignments between the linguistic and acoustic features are found using the HMM-based speech synthesis system (HTS) [28] and we use a bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) for the acoustic model. For system details, see [29].

This paper uses a common acoustic feature set of glottal vocoder features [22] for all neural vocoders: 30 vocal tract filter line spectral frequencies (LSFs), 10 glottal source spectral envelope LSFs, 5 harmonic-to-noise ratio (HNR) parameters, fundamental frequency value on mel-scale (interpolated over unvoiced frames) and a binary voicing flag.

2.1. Speech material

In the experiments, we use speech data from two speakers (one male, one female), who both are professional British English voice talents. The dataset for the male speaker "Nick" comprises 2 542 utterances, totaling 1.8 hours, and the dataset for the female speaker "Jenny" comprises 4 314 utterances, totaling 4.8 hours. For both speakers, 100-utterance test and validation sets were randomly selected from the data, while the remaining utterances were used for training. The material is used at a 16 kHz sample rate.

3. PROPOSED WAVEFORM GENERATION MODEL

3.1. Waveform representation

In this work, we further simplify the waveform representation from [30] by removing the pitch-adaptive cosine windowing and merely phase-lock the window midpoint to a glottal closure instant (GCI). See Fig. 1 for illustration. We use REAPER for GCI and pitch estimation [31]. Similar phase-locked representations have been successfully applied not only in our previous work ([30, 24, 25]), but also in [23, 32]. Furthermore, using GCIs to center waveforms in a window can be seen analogous to using facial landmarks to center images, as done in the highly successful CelebA-HQ dataset [10].

The target waveform can be the glottal excitation or the speech waveform directly. Our primary interest in this paper is the glottal excitation, as it appears relatively simpler to model [21], but we also include experiments on modeling the speech waveform directly. At synthesis time, the generated waveforms are assembled using pitch-synchronous overlap-add (PSOLA) [33], but in principle, the non-tapered windowing enables using other concatenation techniques, including waveform similarity overlap-add (WSOLA) [34]. When modeling glottal excitation signals, the assembled excitation signal is further filtered with the vocal tract filter to produce speech.

3.2. Model architecture

A general view of our GAN architecture is shown in Fig. 2. Progressive upsampling aims for the lower resolution layers to learn the low frequency global structure, while the high resolution layers can focus on the high frequency stochastic signal components. By construction, the model allocates more capacity on the perceptually more relevant low frequencies.

The generator G consists of gated convolutional residual blocks with progressive feature map upsampling at each block. Fig. 3 shows a schematic of the generator block. Upsampling is performed along the sample dimension using linear interpolation to avoid checkerboard artefacts commonly arising from transposed convolution upsampling [35]. In total, the generator contains five blocks that perform progressive upsampling from 32 points to 512 points.

The discriminator D consists of similar gated convolution blocks (without residual connections). Each block downsamples the input by a factor of two, using strided convolutions, until input width of one is reached. The first five discriminator blocks concatenate the generator output or real image and the conditioning to the hidden activations at their respective timescale before applying the convolution. Both the generator and the discriminator use 128-channel 2D-convolutions with filter width three across frames and seven across samples.

The conditioning model C encodes the 200 Hz frame rate acoustic sequence with a non-causal WaveNet-like dilated convolution structure (similar to [36]). The model consists of eight residual blocks with the dilation pattern $\{1, 2, 4, 8\}$ repeated twice. The dilation stack is attached via skip connections to a projection module that outputs conditioning at each time-scale. The residual blocks use 64 channels and a filter width three.

3.3. GAN training

Denote the full-resolution signal segment by $\boldsymbol{x}_n$, and a collection $\boldsymbol{x}$ containing the segment $\boldsymbol{x}_n$ at all time resolutions by $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=0,\ldots,n}$, i.e. $\boldsymbol{x}_i$ is $\boldsymbol{x}_n$ downsampled by a factor of 2^i. Similarly, the generator model G outputs a collection of synthetic data points $\hat{\boldsymbol{x}} = G(\boldsymbol{z}, \boldsymbol{c})$ at all timescales, given a Gaussian noise input $\boldsymbol{z}$ and time-varying conditioning $\boldsymbol{c}$ (derived from the acoustic feature input).

The goal of the generator is to produce $\hat{\boldsymbol{x}}$ that appears correct at each timescale, while the discriminator provides useful learning signals to match the real and generated data distributions (also at each timescale). We use the Wasserstein GAN [14] loss

$$\mathcal{L}_D^{\mathrm{W}} = -\mathbb{E}_{x\sim p_{\mathcal{D}}}\left[D(\boldsymbol{x}, \boldsymbol{c})\right] + \mathbb{E}_{\hat{x}\sim p_g}\left[D(\hat{\boldsymbol{x}}, \boldsymbol{c})\right] \tag{1}$$

for the discriminator and $\mathcal{L}_G^{\mathrm{W}} = -\mathcal{L}_D^{\mathrm{W}}$ for the generator. To keep the discriminator Lipschitz continuous, we use a one-sided gradient penalty (also proposed in [14], see their Appendix C). We prevent the penalty from activating for magnitudes below one to avoid cyclic, non-convergent training dynamics described in [11]. The masked gradient penalty is given by

$$\mathcal{L}_D^{\mathrm{GP}} = \mathbb{E}_{x\sim p_{\mathcal{D}}, \hat{x}\sim p_g}\left[\left(\max\{0, \|\nabla_{\tilde{\boldsymbol{x}}} D(\tilde{\boldsymbol{x}}, \boldsymbol{c})\| - 1\}\right)^2\right], \tag{2}$$

where $\tilde{\boldsymbol{x}} = \varepsilon\boldsymbol{x} + (1-\varepsilon)\hat{\boldsymbol{x}}$ is sampled randomly along the line segment between $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$. Additionally, to encourage convergence, we use the "R1" penalty [11] that penalizes large discriminator gradient magnitudes for the real data samples

$$\mathcal{L}_D^{\mathrm{R1}} = \mathbb{E}_{x\sim p_{\mathcal{D}}}\left[\|\nabla_{\boldsymbol{x}} D(\boldsymbol{x}, \boldsymbol{c})\|^2\right] \tag{3}$$

Finally, we add an optional loss term based on the mean squared error of FFT magnitudes. Similar auxiliary FFT-based losses have been found helpful for waveform generation in e.g. [8, 25, 9]. We only apply the FFT loss on the full-resolution signal:

$$\mathcal{L}_G^{\mathrm{FFT}} = \mathbb{E}\left[\left(|\mathrm{FFT}(\boldsymbol{x}_n)| - |\mathrm{FFT}(\hat{\boldsymbol{x}}_n)|\right)^2\right] \tag{4}$$

Notably, the FFT-based loss only uses the spectral magnitude, so all learning of phase information is due to the GAN loss. We found that the spectral loss stabilizes the training both for speech waveforms and glottal excitations, but is not strictly necessary for the latter.

The total training objective is to minimize $\mathcal{L}_{G,C} = \mathcal{L}_G^{\mathrm{W}} + \lambda_1\mathcal{L}_G^{\mathrm{FFT}}$ by updating the generator and the conditioning model, while minimizing $\mathcal{L}_D = \mathcal{L}_D^{\mathrm{W}} + \lambda_2\mathcal{L}_D^{\mathrm{GP}} + \lambda_3\mathcal{L}_D^{\mathrm{R1}}$ by updating the discriminator. We use alternating gradient descent with Adam

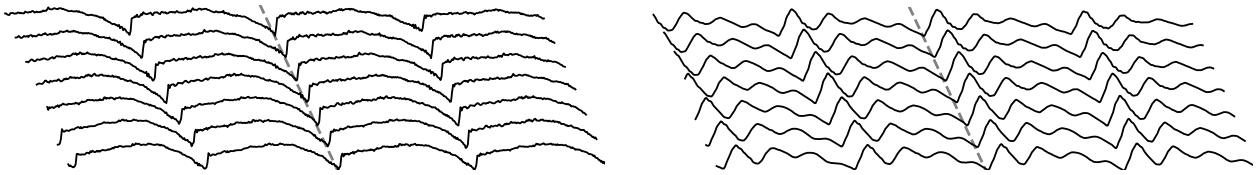

Fig. 1: A waterfall plot showing consecutive frames of GCI-centered glottal excitation (left) and speech (right) waveforms.

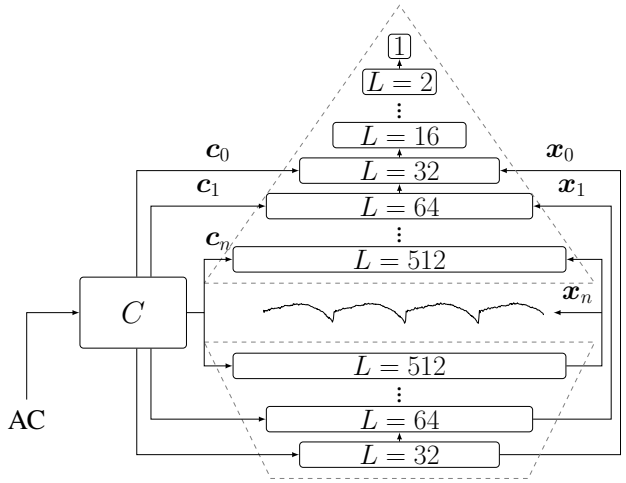

Fig. 2: GAN architecture overview. Generator (bottom) performs progressive upsampling and outputs waveforms at multiple timescales, while the discriminator (top) evaluates the waveforms at all timescales. Both models have access to the same acoustic conditioning, provided by a conditioning model C which collaborates with the generator.

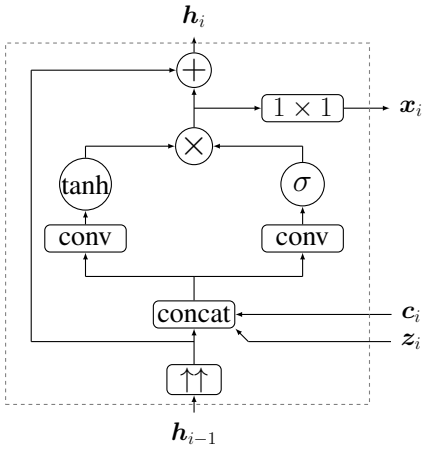

Fig. 3: Generator upsampling block. Input hidden features $\boldsymbol{h}_{i-1}$ are upsampled linearly before concatenating them with a latent noise input z_u and conditioning c_i. Each block applies a gated convolution to the concatenated features and outputs a signal $\boldsymbol{x}_i$ at its respective timescale

(LR=1e-4, β_1=0.9, β_2=0.999) and loss weights λ_1=1, λ_2=10, λ_3=1. Similarly to [10], we only apply one discriminator update per generator update and use an additional batch standard deviation feature in the penultimate discriminator layer. The models were trained for 100k iterations, where a single iteration contains 150 consecutive frames of speech, each associated with a 512 point full resolution waveform.

4. EVALUATION

We conducted subjective listening tests to evaluate the synthetic speech quality and voice similarity to a natural reference. Our main interest is the comparison between the proposed method, applied to glottal excitation signals (named "GlotGAN"), and two established neural vocoder methods, GlottDNN and WaveNet. For the similarity DMOS test, we included a direct waveform variant of the proposed method (called here "WaveGAN") and a classical SPSS vocoder, STRAIGHT [37] (which uses its distinct feature set and acoustic model). The latter two were excluded from the pair-wise quality CCR test, as the number of system pairings would have grown to be unpractical. Audio samples and code are available.[1]

4.1. Reference methods

The GlottDNN and Wavenet vocoders are both conditioned on the same acoustic feature set as the proposed model. GlottDNN uses a DNN to predict glottal excitation pulse waveshapes in a two period pitch synchronous format [30]. The excitation model outputs conditional average pulse shapes, where the lack of high frequency stochastic content is compensated by adding shaped noise as indicated by the HNRs, and by applying a spectral envelope matching filter [22]. We use the configuration from [29], where the excitation model consists of a single BLSTM layer (size 128), followed by three fully connected layers (size 512), finally outputting a 400 point pulse. Our WaveNet vocoder uses a model configuration from [21] (30 residual blocks in three dilation groups, 64 residual channels, 128 post-net channels). However, the model is trained with 8-bit softmax cross-entropy on μ-law companded quantized speech, and is trained on the same speaker dependent data as the proposed GlotGAN model.

4.2. Listening test

Listening tests were conducted on the Figure Eight[2] crowd-sourcing platform. Each test case was evaluated by 50 listeners, and listener quality was maintained by artificial low-quality anchor cases and post-screening of zero-variance listener responses. 15 test set utterances were randomly selected for the listening experiments.

[1] https://github.com/ljuvela/multiscale-GAN

[2] https://www.figure-eight.com/

Voice similarity between synthetic speech and a natural reference was evaluated in a DMOS-like test. The listeners were asked to rate the voice similarity of a test sample to a natural reference (with the same linguistic contents) using a 5-level absolute category rating scale, ranging from "Bad"(1) to "Excellent"(5). Figure 4 shows system mean ratings with 95% confidence intervals and normalized histograms for the rating categories. The Mann-Whitney U-tests indicate that differences between systems are statistically significant (at a Bonferroni corrected 95% confidence level), except GlotGAN–GlottDNN for "Jenny" and all pairings of WaveNet–GlotGAN–GlottDNN for "Nick".

A category comparison rating (CCR) [38] test was performed to evaluate the synthetic speech quality. The listeners were presented with a pair of test samples and asked to rate the comparative quality from -3 ("Much worse") to 3 ("Much better"). The CCR scores are obtained by re-ordering and pooling together all ratings the system received. Figure 5 shows the mean scores with 95% confidence intervals. U-tests (following a similar procedure to above) indicate that all differences are statistically significant.

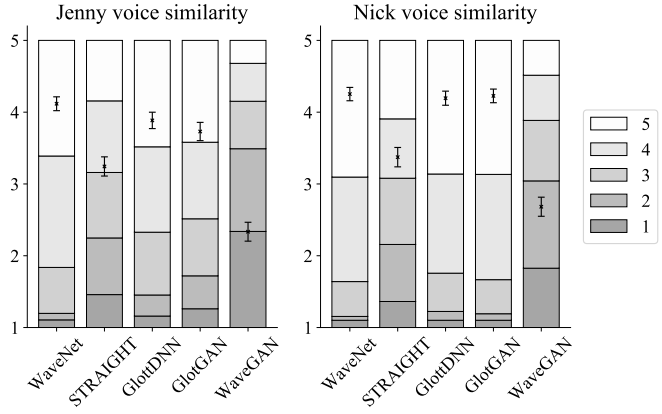

Fig. 4: Voice similarity ratings for "Jenny" (left) and "Nick" (right). The plot shows mean ratings with 95% confidence intervals, along with stacked rating distribution histograms.

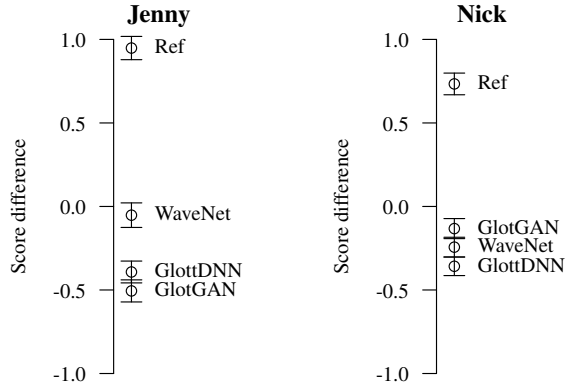

Fig. 5: Combined score differences obtained from the quality comparison CCR test for "Jenny" (left) and "Nick" (right). Error bars are t-statistic based 95% confidence intervals for the mean.

5. DISCUSSION

A characteristic issue we have identified with frame-based GAN audio generation is related to local GAN mode collapse (i.e. lack of variety). In this case, the Generator network fails to reproduce the correct stochastic properties of the data, and instead outputs a "mode" waveshape that appears appropriate when examined in isolation, but causes audible artefacts when combined with adjacent frames. Perceptually, this repeated "frozen noise" can result in buzzy robotic artefacts and ringing, reminiscent of pure impulse train excited vocoders. This issue is especially prominent in unvoiced sounds, and for the listening experiments we resorted to synthesizing the unvoiced parts using pure white noise excitation filtered with the predicted vocal tract envelope. This effect is also somewhat audible in the voiced parts of GlotGAN for "Jenny", which contributes to GlotGAN under-performing the reference methods. As for direct waveform synthesis, the quality is further degraded, although the synthetic speech remains intelligible.

Another potential cause to these artefacts is overfitting and the mismatch between natural and synthetic acoustic features at training and test time, respectively. Furthermore, the acoustic model performance varies between different speakers [22], as reflected in the larger gap between natural and synthetic voices for "Jenny" compared with "Nick". In contrast, the difference in neural vocoder performance in less prominent with the same speech material, when using natural acoustic features in a copy-synthesis setup [25].

6. CONCLUSION

This paper presented a multi-scale GAN architecture for generating glottal excitation (GlotGAN) and speech waveforms (WaveGAN) pitch-synchronously. The proposed model is evaluated as a neural vocoder for a statistical parametric speech synthesis system and listening test results show that GlotGAN can achieve similar performance to a WaveNet vocoder. Future work includes upgrading the TTS system acoustic mapping to use sequence-to-sequence models [1] and applying joint generative acoustic model and neural vocoder training [18]. Progressive upsampling should also naturally extend to higher sample rates, while still allowing fast parallel inference.

7. ACKNOWLEDGMENT

This study was supported by the Academy of Finland (project 312490). JY was partially supported by JST CREST Grant Number JPMJCR18A6, Japan and by MEXT KAKENHI Grant Numbers (16H06302, 17H04687, 18H04120, 18H04112, 18KT0051), Japan. We acknowledge the computational resources provided by the Aalto Science-IT project.

8. REFERENCES

[1] Jonathan Shen, Ruoming Pang, Ron Weiss, et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[2] Jose Sotelo, Soroush Mehri, Kundan Kumar, et al., "Char2Waw: End-to-end speech synthesis," in *ICLR workshop*, 2017.

[3] Aäron van den Oord, Sander Dieleman, Heiga Zen, et al., "WaveNet: A generative model for raw audio," *arXiv pre-print*, 2016.

[4] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, "Speaker-dependent WaveNet vocoder," in *Proc. Interspeech*, 2017, pp. 1118–1122.

[5] Xin Wang, Jaime Lorenzo-Trueba, Shinji Takaki, Lauri Juvela, and Junichi Yamagishi, "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," in *Proc. ICASSP*, 2018, pp. 4804–4808.

[6] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, et al., "Efficient neural audio synthesis," in *Proc. ICML*, 2018, vol. 80 of *Proceedings of Machine Learning Research*, pp. 2410–2419.

[7] Sercan O. Arik, Mike Chrzanowski, Adam Coates, et al., "Deep Voice: Real-time neural text-to-speech," in *Proc. ICML*, 2017.

[8] Aäron van den Oord, Yazhe Li, Igor Babuschkin, et al., "Parallel WaveNet: Fast high-fidelity speech synthesis," *arXiv pre-print*, 2017.

[9] Wei Ping, Kainan Peng, and Jitong Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," *arXiv pre-print*, 2018.

[10] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. ICLR*, 2018.

[11] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin, "Which training methods for GANs do actually converge?," in *Proc. ICML*, 2018, vol. 80, pp. 3481–3490.

[12] Yifan Wang, Federico Perazzi, Brian McWilliams, et al., "A fully progressive approach to single-image super-resolution," in *CVPR Workshop*, June 2018.

[13] Emily L. Denton, Soumith Chintala, Arthur Szlam, and Robert Fergus, "Deep generative image models using a Laplacian pyramid of adversarial networks," *arXiv pre-print*, 2015.

[14] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville, "Improved training of Wasserstein GANs," *arXiv pre-print*, 2017.

[15] Santiago Pascual, Antonio Bonafonte, and Joan Serr, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[16] Kou Tanaka, Takuhiro Kaneko, Nobukatsu Hojo, and Hirokazu Kameoka, "WaveCycleGAN: Synthetic-to-natural speech waveform conversion using cycle-consistent adversarial networks," *arXiv pre-print*, 2018.

[17] Takuhiro Kaneko, Hirokazu Kameoka, Nobukatsu Hojo, et al., "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *Proc. ICASSP*, 2017, pp. 4910–4914.

[18] Yi Zhao, Shinji Takaki, Hieu-Thi Luong, et al., "Wasserstein GAN and waveform loss-based acoustic model training for multi-speaker text-to-speech synthesis systems using a WaveNet vocoder," *IEEE Access*, pp. 1–1, 2018.

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

[20] Tuomo Raitio, Heng Lu, John Kane, et al., "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *Proc. EUSIPCO*, 2014.

[21] Lauri Juvela, Vassilis Tsiaras, Bajibabu Bollepalli, et al., "Speaker-independent raw waveform model for glottal excitation," in *Proc. Interspeech*, 2018, pp. 2012–2016.

[22] Manu Airaksinen, Lauri Juvela, Bajibabu Bollepalli, Junichi Yamagishi, and Paavo Alku, "A comparison between STRAIGHT, glottal, and sinusoidal vocoding in statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1658–1670, Sept 2018.

[23] Yang Cui, Xi Wang, Lei He, and Frank K. Soong, "A new glottal neural vocoder for speech synthesis," in *Proc. Interspeech*, 2018, pp. 2017–2021.

[24] Bajibabu Bollepalli, Lauri Juvela, and Paavo Alku, "Generative adversarial network-based glottal waveform model for statistical parametric speech synthesis," in *Proc. Interspeech*, 2017, pp. 3394–3398.

[25] Lauri Juvela, Bajibabu Bollepalli, Xin Wang, et al., "Speech waveform synthesis from MFCC sequences with generative adversarial networks," in *Proc. ICASSP*, 2018, pp. 5679–5683.

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

[27] Korin Richmond, Robert AJ Clark, and Susan Fitt, "Robust LTS rules with the Combilex speech technology lexicon," in *Proc. Interspeech*, Brighton, 2009, pp. 1295–1298.

[28] Heiga Zen, Takashi Nose, Junichi Yamagishi, et al., "The HMM-based speech synthesis system version 2.0," in *Proc. ISCA SSW6*, 2007, pp. 294–299.

[29] Lauri Juvela, Bajibabu Bollepalli, Junichi Yamagishi, and Paavo Alku, "Reducing mismatch in training of DNN-based glottal excitation models in a statistical parametric text-to-speech system," in *Proc. Interspeech*, 2017, pp. 1368–1372.

[30] Lauri Juvela, Bajibabu Bollepalli, Manu Airaksinen, and Paavo Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," in *Proc. ICASSP*, 2016, pp. 5120–5124.

[31] David Talkin, "REAPER: Robust Epoch And Pitch EstimatoR," https://github.com/google/REAPER, 2015.

[32] Min-Jae Hwang, Eunwoo Song, Jin-Seob Kim, and Hong-Goo Kang, "A unified framework for the generation of glottal signals in deep learning-based parametric speech synthesis systems," in *Proc. Interspeech*, 2018, pp. 912–916.

[33] Eric Moulines and Jean Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, no. 2, pp. 175–205, 1995.

[34] Werner Verhelst and Marc Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. ICASSP*, 1993, vol. 2, pp. 554–557.

[35] Augustus Odena, Vincent Dumoulin, and Chris Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016.

[36] Dario Rethage, Jordi Pons, and Xavier Serra, "A WaveNet for speech denoising," in *Proc. ICASSP*, 2018, pp. 5069–5073.

[37] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain 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.

[38] "Methods for Subjective Determination of Transmission Quality," Recommendation P.800, ITU-T SG12, Geneva, Switzerland, 1996.