Unified Source-Filter GAN: Unified Source-filter Network Based On Factorization of Quasi-Periodic Parallel WaveGAN

Reo Yoneyama[1], *Yi-Chiao Wu* [1], *Tomoki Toda*[2]

[1]Nagoya University, Japan
[2]Information Technology Center, Nagoya University, Japan

[1]{yoneyama.reo, yichiao.wu}@g.sp.m.is.nagoya-u.ac.jp, [2]tomoki@ics.nagoya-u.ac.jp

Abstract

We propose a unified approach to data-driven source-filter modeling using a single neural network for developing a neural vocoder capable of generating high-quality synthetic speech waveforms while retaining flexibility of the source-filter model to control their voice characteristics. Our proposed network called unified source-filter generative adversarial networks (uSFGAN) is developed by factorizing quasi-periodic parallel WaveGAN (QPPWG), one of the neural vocoders based on a single neural network, into a source excitation generation network and a vocal tract resonance filtering network by additionally implementing a regularization loss. Moreover, inspired by neural source filter (NSF), only a sinusoidal waveform is additionally used as the simplest clue to generate a periodic source excitation waveform while minimizing the effect of approximations in the source filter model. The experimental results demonstrate that uSFGAN outperforms conventional neural vocoders, such as QPPWG and NSF in both speech quality and pitch controllability.

Index Terms: Speech synthesis, neural vocoder, source-filter model, generative adversarial networks, Parallel WaveGAN

1. Introduction

Currently, neural vocoders [1–26] usually achieve very high-fidelity speech generation by directly modeling raw speech waveforms using advanced neural networks without ad hoc designs. On the other hand, because of the data-driven nature and unified network architecture [27, 28], the speech components controllability of the neural vocoders are usually inferior to the conventional source-filter vocoders [29, 30]. Therefore, it is desired to develop a neural vocoder capable of high-fidelity and controllable speech generation.

To improve the controllability, there have been proposed many generation models integrating conventional parametric-based source-filter models with deep neural network architectures [21–26]. For example, neural source-filter (NSF) [21, 22] realizes speech generation based on non-autoregressive modeling by non-linear filtering of parametrically generated source excitation signals with multiple dilated convolutional layers. LPCNet [23] adopts a WaveRNN [19]-like architecture to generate residual signals while a linear filtering process is applied to generate speech waveforms as in the conventional linear predictive coding (LPC) vocoder [31, 32]. Generative adversarial network (GAN) based neural homomorphic vocoder (NHV) [24] first develops neural-based linear-time-variant (LTV) filters with the input pulse trains and white noise to generate mixed source excitations, and then a trainable causal finite impulse response (FIR) filter is applied to the excitations for generating the output waveforms. Although these hybrid neural vocoders

have successfully improved the controllability by integrating parametric-based approaches, the synthetic speech quality of these vocoders tends to be inferior to that of data-driven unified neural vocoders. Moreover, there is still room for improvements in the controllability.

To achieve high-fidelity and high-controllability speech generation, we propose a GAN-based framework to introduce the source-filter model with fewer ad hoc designs into a single neural network. The generator is designed by factorizing quasi-periodic Parallel WaveGAN (QPPWG) [28] into two cascaded networks corresponding to the source excitation generation and resonance filtering, and these two networks are jointly optimized in the training stage. Only a sinusoidal waveform is additionally used as the simplest clue to generate a periodic source excitation waveform while minimizing the effect of approximations in the source filter model. Moreover, to generate reasonable source excitation signals, an additional auxiliary loss is applied to the source excitation network. The main contributions of this paper are summarized as follows:

- We propose a unified framework for neural vocoders attaining an interpretable and tractable source-filter-like architecture, making it possible to well models excitation generation and resonance filtering while keeping the simplicity of training.

- The proposed method achieves better fundamental frequency (F_0) controllability than the conventional neural vocoders, such as QPPWG and NSF, while attaining high-fidelity speech generation even in F_0 transformation scenarios.

2. Related work

This chapter describes non-AR neural vocoders: Parallel Wave-GAN (PWG) [3], QPPWG [28], and NSF [21]. PWG and QP-PWG are the basis of our method, and we use QPPWG as one of the baseline methods. NSF is a semi-parametric neural vocoder based on the source-filter model, which is used as the other baseline method.

2.1. Parallel WaveGAN (PWG)

PWG is a GAN-based method for generating raw waveforms. It is a compact model without an autoregressive structure or a causal mechanism, and can achieve fast speech generation with high fidelity. The model consists of two networks, generator (G) and discriminator (D). The WaveNet-based generator, which is conditioned by auxiliary features, learns to make the discriminator recognize the generated sample as *real*. This process can be written as follows:

$$\mathcal{L}_{adv}(G, D) = \mathbb{E}_{z \sim \mathcal{N}(0, I)} \left[(1 - D(G(z)))^2 \right], \quad (1)$$

where z is random noise distributed from Gaussian distribution. Note that all auxiliary features of the generator are omitted in this paper for simplicity.

The discriminator learns to identify the generated sample as $fake$ and the natural sample as $real$. This process can be written as follows:

$$\mathcal{L}_D(G, D) = \mathbb{E}_{\boldsymbol{x} \sim p_{data}} \left[(1 - D(\boldsymbol{x}))^2 \right] + \mathbb{E}_{\boldsymbol{z} \sim \mathcal{N}(0,I)} \left[D(G(\boldsymbol{z}))^2 \right], \quad (2)$$

where $\boldsymbol{x}$ denotes the natural samples and p_{data} denotes the data distribution of the natural samples. PWG also adopts multi-STFT loss [3] as an auxiliary loss $\mathcal{L}_{aux}(G)$ to improve the training stability. In conclusion, the final loss function of the generator can be written as a weighted sum of $\mathcal{L}_{aux}$ and $\mathcal{L}_{adv}$ as follows:

$$\mathcal{L}_G(G, D) = \mathcal{L}_{aux}(G) + \lambda_{adv} \mathcal{L}_{adv}(G, D) \quad (3)$$

where λ_{adv} is hyperparameter for weight and empirically set to 4.0 in this paper.

2.2. Quasi-periodic parallel WaveGAN (QPPWG)

Although PWG achieves high-fidelity speech generation, the fully data-driven nature makes PWG lack the explicit controllability of each speech component especially when unseen auxiliary features are given such as F_0 outside the F_0 range of training data. To alleviate this issue, in QPPWG pitch-dependent dilated convolution neural networks (PDCNNs), which dynamically changes dilation sizes adapting to pitch, are introduced to PWG. PDCNN facilitates QPPWG to capture the very long-term dependencies of periodic components and makes the pitch of the QPPWG-generated speech more consistent with the auxiliary F_0.

For the ordinary dilated convolution neural networks, the dilation size d, is predefined and time-invariant. The dilation sizes d_t of PDCNNs are dynamically defined at each time step t as follows:

$$d_t = d \times f_s / (f_{0,t} \times a) \quad (4)$$

where f_s is the sampling rate, $f_{0,t}$ is an F_0 value at time t, and a is a hyperparameter called dense factor, which determines the sparsity of the PDCNNs, and empirically set to 4.0 in this paper.

2.3. Neural source-filter (NSF)

NSF is a neural vocoder based on source-filter model, and divided into three modules: a condition module, a source module, and a filter module. An F_0 sequence $f_{0,1}, \cdots, f_{0,T}$ and a spectral feature sequence (e.g. mel-spectrogram) are used as an input of NSF. The condition module upsamples the input features and extracts feature embeddings for the resonance filtering. In the source module, by treating $f_{0,t}$ as the instantaneous frequency, the fixed number of sinusoidal wave basis signals is generated, where the h-th basis signal $e_t^{(h)}$ is given by

$$e_t^{(h)} = \begin{cases} \sin\left(\sum_{k=1}^{t} 2\pi \frac{h f_{0,k}}{f_s} + \phi\right) + n_t & \text{if } f_{0,t} > 0 \\ \frac{1}{3\sigma} n_t & \text{if } f_{0,t} = 0, \end{cases} \quad (5)$$

where f_s is sampling frequency, $\phi \in [-\pi, \pi]$ is a random initial phase, $n_t \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian noise, and h is the scale factor of the fundamental frequency. These signals are merged using a feed-forward network to output the source excitations.
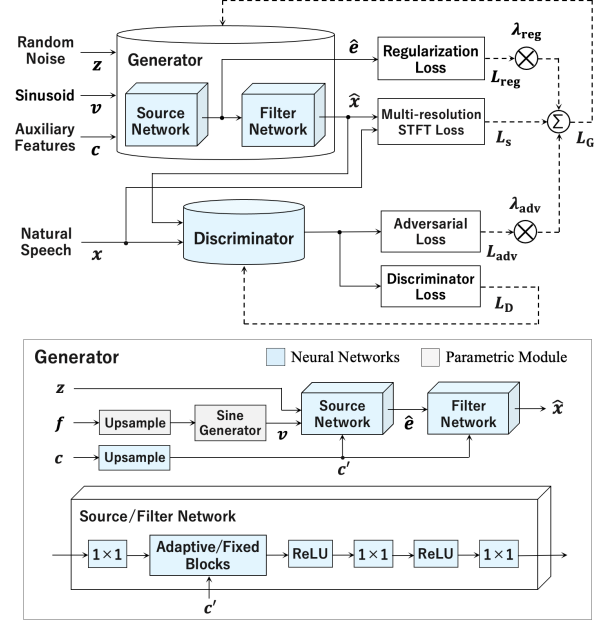

Figure 1: *Architecture of the proposed method, uSFGAN.*

The filter module modulates the source signal using multiple stages of dilated convolution and affine transformations similar to those in ClariNet [20]. NSF adopts multi-resolution STFT loss to learn the difference between the output and the target waveform in the spectral domain. Unlike that of PWG, it calculates the mean square error (MSE) for the log power spectrum.

3. Proposed method: unified source-filter GAN (uSFGAN)

Our proposed method, uSFGAN is based on QPPWG, but differs on the generator in several ways: (1) the generator is explicitly split into a source-network and a filter-network; (2) a sinusoidal signal is used as an additional input of the source-network, and (3) a regularization term for the output of the source-network is added to the auxiliary loss. On the other hand, the discriminator is the same as that of QPPWG.

3.1. Network architecture

As the proposed architecture shown in Fig.1, the generator of uSFGAN receives random noise $\boldsymbol{z}$ sampled from Gaussian distribution, an F_0 sequence $\boldsymbol{f}$, and an auxiliary feature sequence $\boldsymbol{c}$ as the input, where $\boldsymbol{f}$ and $\boldsymbol{c}$ are assumed to be extracted per frame. A sinusoidal signal $\boldsymbol{v} = v_1, \cdots, v_T$ is first generated on the basis of upsampled $\boldsymbol{f}$ as follows:

$$v_t = \begin{cases} \sin\left(\sum_{k=1}^{t} 2\pi \frac{f_{0,k}}{f_s}\right) & \text{if } f_{0,t} > 0 \\ 0 & \text{if } f_{0,t} = 0, \end{cases} \quad (6)$$

where $f_{0,t}$ is the instantaneous frequency at time t, and f_s is the sampling frequency. Unlike QPPWG adopting only noise inputs, the sinusoidal signal input is used to make the estimation of the harmonic components easier and improve the learning efficiency of the proposed source-network. Then, $\boldsymbol{v}$ is combined with $\boldsymbol{z}$ as a two-channel input of the source-network. The source-network performs the pitch-dependent dilated convolution conditioned on the upsampled auxiliary features $\boldsymbol{c}$ to output

the source excitation signal $\hat{e}$. The generated source excitation signal is used as the input to the filter-network and is also used to calculate the spectral envelope regularization loss, which is used for the auxiliary loss, as mentioned in Section 3.2. In the filter-network, non-causal dilated convolution with fixed dilation sizes is performed. The output waveform is input to the discriminator and also used to calculate the multi-resolution STFT auxiliary loss.

3.2. Spectral envelope regularization loss

To encourage the source-network to output a reasonable source excitation signal, one constraint is imposed on the output of the source-network. As in the traditional source-filter vocoders, such as STRAIGHT [29] and WORLD [30], we assume that the spectral structure of the source excitation signal consists of harmonic components and stochastic components, and its spectral envelope is flat, *i.e.*, the power of spectral envelope is constant over all frequency. We adopt regularization to satisfy this assumption on the spectral envelope of the source excitation signal in the output of the source-network.

We use a simplified algorithm of cheaptrick [33] to extract the spectral envelope from the output signal of source-network. The original algorithm of cheaptrick is composed of three steps: (1) F_0 adaptive windowing and calculation of log power spectrum, (2) F_0 adaptive smoothing in the spectral domain, and (3) F_0 adaptive liftering in the cepstrum domain. In order to speed up the spectral envelope estimation process, we apply several modifications to the cheaptrick algorithm. First, we directly use the F_0 values given as the auxiliary feature $\boldsymbol{f}$ rather than extracting F_0 values from the output signal. Those F_0 values are further rounded to integers, and then, the corresponding windows and liftering functions are obtained in advance. Moreover, the step (2) is omitted because this process requires a relatively large processing time and the F_0 adaptive spectral envelope extraction is still performed by the F_0 adaptive liftering in the step (3). Although these modification causes slight degradation of the spectral envelope estimation accuracy, it doesn't cause any significant issues as the precise spectral envelope estimation is not necessary for the regularization.

The spectral envelope regularization loss is given by

$$\mathcal{L}_{reg}(G) = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \hat{E}_k^{(n)2}, \qquad (7)$$

where $\hat{E}_k^{(n)}$ is the k-th frequency component of log power spectral envelope extracted from $\hat{e}$ at n-th time frame by the simplified cheaptrick algorithm. Note that when this loss reaches to 0, linear power values of the spectral envelope are 1 over all frequency and time frames.

3.3. Training criteria

The same adversarial losses as QPPWG but different auxiliary losses are adopted in uSFGAN training. Our method uses two types of auxiliary losses: multi-resolution STFT loss and the spectral envelope regularization loss. The STFT loss is defined as follows:

$$\mathcal{L}_s(G) = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \left[\log \frac{\mathrm{Re}(Y_k^{(n)})^2 + \mathrm{Im}(Y_k^{(n)})^2}{\mathrm{Re}(\hat{Y}_k^{(n)})^2 + \mathrm{Im}(\hat{Y}_k^{(n)})^2} \right]^2, \quad (8)$$

where Y_k^n and $\hat{Y}_k^n$ are the k-th STFT component at the n-th time frame of a natural waveform and the output waveform. Re

Table 1: *Parameter settings for multi-resolution STFT loss. We apply a hanning window before the FFT process.*

STFT loss	Frame shift	Frame size	DFT bins
$L_s^{(1)}$	80 (5 ms)	320 (20 ms)	512
$L_s^{(2)}$	40 (2.5 ms)	80 (5 ms)	128
$L_s^{(3)}$	640 (40 ms)	1920 (120 ms)	2048

and Im denote real part and imaginary part, respectively. This STFT loss is different from that of PWG [3], but the same as that of NSF [21, 22]. Finally, our auxiliary loss is represented as follows:

$$\mathcal{L}_{aux}(G) = \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}_s^{(m)}(G) + \lambda_{reg} \mathcal{L}_{reg}(G), \qquad (9)$$

where M is the number of STFT losses using various STFT parameters, and λ_{reg} is a hyperparameter balancing the two auxiliary losses and empirically set to 1.0 in this paper.

4. Experimental evaluation

4.1. Experimental conditions

To investigate the effectiveness of our proposed method, we compared four different models: publicly available pretrained NSF (hn-sinc-nsf-9 [34]) model referred to as PT-NSF, NSF with WORLD features referred to as WORLD-NSF, QPPWG, uSFGAN, and uSFGAN without the spectral envelope regularization loss. We adopted F_0 conversion to evaluate the controllability.

For the training data, we used 4000 utterances from CMU-ARCTIC database [35] consisting of more than 1000 utterances each from four speakers: slt, bdl, clb, and rms. We used a set of 264 utterances consisting of 66 utterances from each speaker as validation data and another set of 264 utterances as test data. The sampling frequency was set to 16000 Hz by downsampling.

WORLD-NSF used the same architecture as PT-NSF. QPPWG used 10 adaptive blocks and 10 fixed blocks as the optimized setting. uSFGAN used 30 adaptive blocks for the source-network, and 30 fixed blocks for the filter-network. uSFGAN was trained with the RAdam optimizer [36] ($\epsilon = 10^{-6}$) with 400 k iterations as in QPPWG. The generator of uSFGAN is trained with only auxiliary loss in the first 100 k iterations, and then trained with the adversarial loss as well as the auxiliary loss in the remaining 300 k iterations. The parameter settings of the multi-resolution STFT loss are shown in Table 1.

In PT-NSF, F_0 extracted by YAAPT [37] and mel-spectrum extracted by STFT-based method were used for the input. In the other models, F_0, spectral envelope, and aperiodicity extracted by WORLD [30] were used. The window length was set to 64 ms and the shift length was set to 5 ms in all models. The spectral envelope was parameterized into 25-dimensional mel-cepstral coefficients, and aperiodicity was coded into 1-dimension. The unvoiced/voiced intervals were represented as a binary feature. For PT-NSF, the target speech used for training was normalized after feature extraction. On the other hand, for the other models, no normalization process was applied to the target speech.

4.2. Objective evaluations

As objective evaluation indexes, we used root mean square error of log F_0: RMSE [Hz], unvoiced/voiced decision error: U/V

Table 2: *Results of objective evaluation.*

model	RMSE	U/V	MCD	LSD
	$1.0 \times F_0$			
PT-NSF	0.09	10	3.01	1.85
WORLD-NSF	**0.06**	**9**	**2.70**	**1.69**
QPPWG	0.08	11	3.07	1.83
uSFGAN (ours)	0.08	10	2.79	1.73
uSFGAN w/o L_{reg}	0.08	11	2.83	1.71
	$2.0 \times F_0$			
PT-NSF	0.15	38	4.53	–
WORLD-NSF	0.09	**14**	3.70	–
QPPWG	0.33	36	4.26	–
uSFGAN (ours)	**0.06**	**14**	3.81	–
uSFGAN w/o L_{reg}	0.16	22	**3.67**	–
	$0.5 \times F_0$			
PT-NSF	0.69	24	3.57	–
WORLD-NSF	0.68	55	3.60	–
QPPWG	0.17	37	3.45	–
uSFGAN (ours)	**0.14**	40	3.08	–
uSFGAN w/o L_{reg}	0.38	**34**	**3.06**	–

Table 3: *Speech quality MOS evaluations with 95% CI.*

	$1.0 \times F_0$	$2.0 \times F_0$	$0.5 \times F_0$
Natural	4.58 ± 0.18	–	–
WORLD	3.93 ± 0.25	$\mathbf{2.71 \pm 0.25}$	2.66 ± 0.27
PT-NSF	3.75 ± 0.29	2.21 ± 0.27	2.09 ± 0.25
QPPWG	3.66 ± 0.27	1.48 ± 0.21	2.41 ± 0.32
uSFGAN	$\mathbf{4.07 \pm 0.26}$	2.15 ± 0.24	$\mathbf{2.94 \pm 0.31}$

[%], mel-cepstral distortion: MCD [dB], and log spectral distortion: LSD [dB]. We conducted all calculations after normalizing the power. For the calculation of RMSE when F_0 was transformed, we considered F_0 values extracted from natural speech multiplied by the scale factor as reference F_0 values.

The objective evaluation results are shown in Table 2. It shows that uSFGAN tends to generate speech well conveying the information of the given auxiliary features with higher accuracy than PT-NSF, WORLD-NSF, and QPPWG. In addition, the spectral envelope regularization loss significantly improves the F_0 transformation accuracy in uSFGAN. An example of spectrograms and waveforms of the output source excitation signals from uSFGAN with and without the spectral envelope regularization loss are shown in Figs. 2 and 3. We can find that the use of the spectral envelope regularization loss is effective for generating a reasonable source excitation signal of which spectral envelopes are flat and both harmonic components and periodic waveform shape well correspond to F_0 values.

4.3. Subjective evaluations

We conducted an opinion test on speech quality. Natural speech and synthetic speech from WORLD and three models: PT-NSF, QPPWG, and uSFGAN were evaluated by 10 subjects. We evaluated 160 utterances per each method per F_0 scaling factor. The synthetic speech was generated by scaling F_0 values by 1.0, 2.0 and 0.5 times.

As the experimental result, mean opinion scores (MOS) are shown in Table 3. It shows that uSFGAN significantly outperforms WORLD, PT-NSF, and QPPWG in both $1.0 \times F_0$ and $0.5 \times F_0$. In $2.0 \times F_0$, although WORLD still achieves the best speech quality, uSFGAN achieves comparable speech quality to PT-NSF and significantly better speech quality than QPPWG.

To investigate the controllability, we conducted an ABX test to evaluate the perceptual accuracy of F_0 conversion by 10 subjects. We evaluated 100 utterances per each method per F_0 scaling factor. The synthetic speech by WORLD was used as a
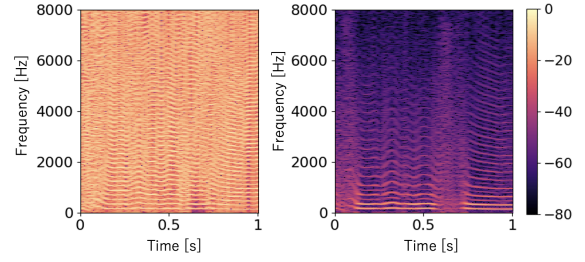
Figure 2: *Spectrograms of source signals output from the source-networks of uSFGAN w/ (left) and w/o (right) $\mathcal{L}_{reg}$.*
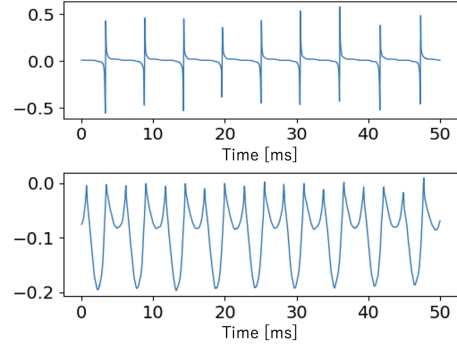
Figure 3: *Source waveform output from the source-networks of uSFGAN w/ (top) and w/o (bottom) $\mathcal{L}_{reg}$. Given F_0 values over this segment are around 200 Hz.*

Table 4: *Perceptive F_0 accuracy ABX evaluations with 95% CI.*

	$2.0 \times F_0$	$0.5 \times F_0$
PT-NSF / QPPWG	**68** / 32 ± 1.8 [%]	14 / **86** ± 1.4 [%]
QPPWG / uSFGAN	7 / **93** ± 1.0 [%]	16 / **84** ± 1.4 [%]
uSFGAN / PT-NSF	**65** / 35 ± 1.9 [%]	**95** / 5 ± 0.9 [%]

reference speech for three models: PT-NSF, QPPWG, and uSFGAN. The subjects selected a speech sample of which pitch was closer to that of the reference speech.

The experimental results are shown in Table 4. Each entry in the table shows the number of times selected as closer to the reference pitch in each pair comparison from the three models. It shows that uSFGAN significantly outperforms the other models in both $2.0 \times F_0$ and $0.5 \times F_0$. Our samples can be found on our demo page [38].

5. Conclusions

In this paper, we have proposed a unified neural vocoder framework based on the source-filter model, which is called unified source-filter GAN. In the proposed neural vocoder, a sinusoidal signal as the additional input, the pitch-dependent dilated convolution, and spectral envelope regularization loss have been implemented for factorizing the overall network into the source-network and the filter-network. The experimental evaluation results have demonstrated that the proposed neural vocoder can significantly improve F_0 transformation accuracy while achieving high speech quality. Thus, uSFGAN can be applied to entertaining TTS by combining it with an acoustic model that outputs vocoder features.

6. Acknowledgements

7. References

[1] 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," in *Proc. SSW9*, Sept. 2016, p. 125.

[2] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. ICML*, July 2018, pp. 3915–3923.

[3] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, May 2020, pp. 6199–6203.

[4] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," in *Proc. ICLR*, Apr. 2017.

[5] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "FFTNet: A real-time speaker-dependent neural vocoder," in *Proc. ICASSP*, Apr. 2018, pp. 2251–2255.

[6] S. Kim, S.-G. Lee, J. Song, J. Kim, and S. Yoon, "FloWaveNet : A generative flow for raw audio," in *Proc. ICML*, June 2019, pp. 3370–3378.

[7] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. NeurIPS*, Dec. 2019, pp. 14 910–14 921.

[8] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech," in *Proc. SLT*, Jan. 2021.

[9] N.-Q. Wu and Z.-H. Ling, "WaveFFJORD: FFJORD-based vocoder for statistical parametric speech synthesis," in *Proc. ICASSP*, May 2020, pp. 7214–7218.

[10] Y. Ai and Z.-H. Ling, "A neural vocoder with hierarchical generation of amplitude and phase spectra for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 839–851, 2020.

[11] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, May 2019, pp. 3617–3621.

[12] O. McCarthy and Z. Ahmed, "HooliGAN: Robust, high quality neural vocoding," *arXiv preprint arXiv:2008.02493*, 2020.

[13] J. Yang, J. Lee, Y. Kim, H. Cho, and I. Kim, "Vocgan: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network," in *Proc. INTERSPEECH*, 2020, pp. 200–204.

[14] R. Yamamoto, E. Song, and J.-M. Kim, "Probability density distillation with generative adversarial networks for high-quality parallel waveform generation," in *Proc. INTERSPEECH*, Sept. 2019, pp. 699–703.

[15] Q. Tian, X. Wan, and S. Liu, "Generative Adversarial Network based Speaker Adaptation for High Fidelity WaveNet Vocoder," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 19–23.

[16] K. Oura, K. Nakamura, K. Hashimoto, Y. Nankaku, and K. Tokuda, "Deep neural network based real-time speech vocoder with periodic and aperiodic inputs," in *Proc. SSW10*, Sept. 2019, pp. 13–18.

[17] B. Bollepalli, L. Juvela, and P. Alku, "Generative adversarial network-based glottal waveform model for statistical parametric speech synthesis," in *Proc. INTERSPEECH*, 2017, pp. 3394–3398.

[18] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "Waveform generation for text-to-speech synthesis using pitch-synchronous multi-scale generative adversarial networks," in *Proc. ICASSP*, May 2019, pp. 6915–6919.

[19] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *Proc. ICML*, July 2018, pp. 2415–2424.

[20] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *Proc. ICLR*, May 2019.

[21] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," in *Proc. ICASSP*, May 2019, pp. 5916–5920.

[22] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter waveform models for statistical parametric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 402–415, 2020.

[23] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, May 2019, pp. 5891–5895.

[24] Z. Liu, K. Chen, and K. Yu, "Neural homomorphic vocoder," in *Proc. INTERSPEECH*, Oct. 2020, pp. 240–244.

[25] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, "GELP: GAN-excited linear prediction for speech synthesis from mel-spectrogram," in *Proc. INTERSPEECH*, Sept. 2019, pp. 694–698.

[26] L. Juvela, B. Bollepalli, V. Tsiaras, and P. Alku, "Glotnet—a raw waveform model for the glottal excitation in statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 27, no. 6, p. 1019–1030, Jun. 2019.

[27] Y. C. Wu, T. Hayashi, P. L. Tobing, K. Kobayashi, and T. Toda, "Quasi-periodic wavenet: An autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1134–1148, 2021.

[28] Y.-C. Wu, T. Hayashi, T. Okamoto, H. Kawai, and T. Toda, "Quasi-periodic parallel wavegan: A non-autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 792–806, 2021.

[29] 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. 3-4, pp. 187–207, 1999.

[30] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[31] D. Wong, B.-H. Juang, and A. Gray, "An 800 bit/s vector quantization LPC vocoder," *IEEE Transactions onAcoustics, Speech, and Signal Processing*, vol. 30, no. 5, pp. 770–780, 1982.

[32] A. V. McCree and T. P. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, 1995.

[33] M. Morise, "Cheaptrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Communication*, vol. 67, pp. 1–7, 2015.

[34] "X. Wang, nii-yamagishilab/project-NN-Pytorch-scripts, Accessed: 2021. [Online]." Available: https://github.com/nii-yamagishilab/project-NN-Pytorch-scripts/.

[35] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases for speech synthesis research," in *Tech. Rep. CMU-LTI-03-177*, 2003.

[36] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proc. ICLR*, Apr. 2020.

[37] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4559–4571, 2008.

[38] "R. Yoneyama, uSFGAN demo, Accessed: 2021. [Online]." Available: https://chomeyama.github.io/UnifiedSourceFilterGAN_demo/.