



High-fidelity Parallel WaveGAN with Multi-band Harmonic-plus-Noise Model

Min-Jae Hwang^{1*}, Ryuichi Yamamoto^{2*}, Eunwoo Song³ and Jae-Min Kim³

¹Search Solutions Inc., Seongnam, Korea

²LINE Corp., Tokyo, Japan

³NAVER Corp., Seongnam, Korea

Abstract

This paper proposes a multi-band harmonic-plus-noise (HN) Parallel WaveGAN (PWG) vocoder. To generate a high-fidelity speech signal, it is important to well-reflect the harmonic-noise characteristics of the speech waveform in the time-frequency domain. However, it is difficult for the conventional PWG model to accurately match this condition, as its single generator inefficiently represents the complicated nature of harmonic-noise structures. In the proposed method, the HN WaveNet models are employed to overcome this limitation, which enable the separate generation of the harmonic and noise components of speech signals from the pitch-dependent sine wave and Gaussian noise sources, respectively. Then, the energy ratios between harmonic and noise components in multiple frequency bands (i.e., subband harmonicities) are predicted by an additional harmonicity estimator. Weighted by the estimated harmonicities, the gain of harmonic and noise components in each subband is adjusted, and finally mixed together to compose the full-band speech signal. Subjective evaluation results showed that the proposed method significantly improved the perceptual quality of the synthesized speech.

Index Terms: Speech synthesis, neural vocoder, Parallel WaveGAN, multi-band harmonic-plus-noise model

1. Introduction

The neural vocoder, which generates speech waveform from conditional acoustic features, has significantly improved the quality of text-to-speech (TTS) systems [1–7]. Neural vocoders mainly consist of two classes: an auto-regressive (AR) model that recursively generates a single speech sample conditioned by previously generated samples [1–4] and a non-AR model that generates a speech waveform in parallel [5–7]. Recently, non-AR neural vocoders have attracted interest thanks to their fast generation speed and reasonable quality of synthesis.

In our previous work, we proposed a Parallel WaveGAN (PWG) vocoder that combines a non-causal WaveNet model with the generative adversarial networks (GANs) [7–9]. In this model, the WaveNet generator efficiently learns the time-frequency characteristics of realistic speech waveform by involving a multi-resolution short-time Fourier transform (MR-STFT) loss to the adversarial training process. As the model is trained without any complicated distillation process, the PWG can provide an easily trainable and fast waveform generation method compared to conventional methods.

However, a single generator is insufficient to learn the complicated nature of speech signal such as harmonic and noise characteristics. As a result, the generated speech often suffers from unnatural artifacts. For instance, harmonic structure can be appeared to unvoiced regions, since the model mainly learns the behavior of periodic voice that are domi-

nant in speech. Thus, it is important to design the system to effectively represent harmonic and noise characteristics.

In this paper, we propose a harmonic-plus-noise PWG (HN-PWG) vocoder where two WaveNet generators jointly learn the harmonic-noise characteristics of target speech based on the *harmonic-plus-noise model (HNM)* within a GAN framework [10–12]. In the proposed method, one WaveNet receives the pitch-dependent sine wave as a source signal for generating a harmonic component; whereas the other receives the Gaussian noise for generating a noise component. Then, each waveform is mixed together to compose output speech. As the harmonic and noise components of the speech signal are separately modeled by the individual generators, the quality of the output speech becomes more stable than with conventional PWG.

To further enhance vocoding performance, we also propose a multi-band HN-PWG model, which combines the idea of HN-PWG with the multi-band approach. In this method, each harmonic and noise component is decomposed into its subband signals through a set of band-pass filters (BPFs). Then, the *subband harmonicities*, which are defined as the energy ratios between the harmonic and noise components in each subband, are predicted by an additional harmonicity estimator. Weighted by the estimated subband harmonicity, the gain of harmonic and noise components in each subband is adjusted, and then mixed together to compose the full-band speech signal. As the complicated frequency-dependent harmonic-noise structure of speech signal is captured by the external harmonicity estimator, the performance of the entire model can be effectively improved.

We verified the outperforming performance of the proposed multi-band HN-PWG in comparison to the conventional methods through subjective evaluations. Specifically, it provided a 4.03 mean opinion score (MOS) result in the TTS scenario, which is 13% higher than that of the conventional system.

2. Related work

There have been several studies to apply an HNM to neural vocoding systems. For instance, the harmonic-plus-noise neural source-filter (hn-NSF) model first generates harmonic and noise components separately. Then, it merges them by using digital low- and high-pass filters [13, 14]. On the other hand, the neural homomorphic model adopts a similar HNM structure within a GAN framework [15], and shows better quality than hn-NSF models. In the PeriodNet model [16], the HNM is applied to the PWG vocoder, for which results also show improvements in the perceptual quality of synthesized speech.

Even though our model is similar to those vocoders in terms of adopting the HNM, the clear difference is that our method proposes a *multi-band* HNM for efficiently capturing the frequency band-wise harmonic-noise characteristics of target speech signal. Note that our multi-band approach is similar to that of traditional parametric vocoders such as

* equal contribution

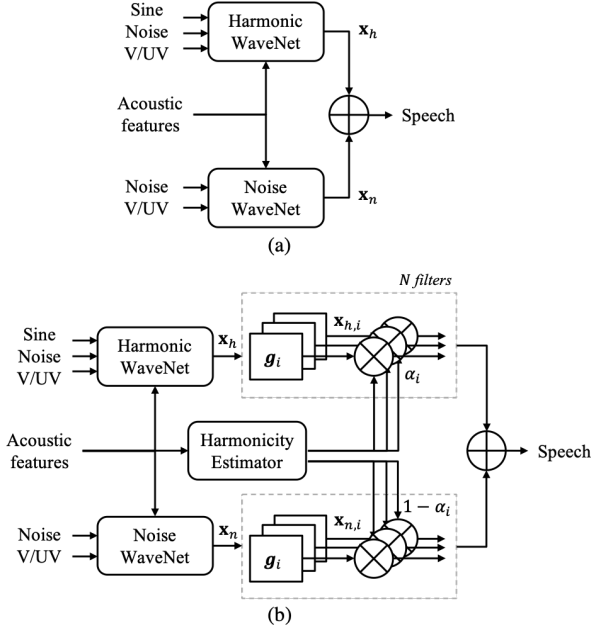


Figure 1: The waveform generators of (a) HN-PWG and (b) multi-band HN-PWG.

multi-band mixed excitation models [17, 18]. The difference is that the proposed model directly mixes speech signals rather than excitation signals.

3. Parallel WaveGAN with harmonic-plus-noise models

PWG is a non-AR WaveNet model that generates a time-domain speech waveform from the corresponding conditional acoustic parameters [7–9]. It consists of a non-causal WaveNet generator and a convolutional neural network (CNN)-based discriminator. By combining adversarial training and the MR-STFT loss function within a GAN framework [19], PWG efficiently learns the time-frequency characteristics of realistic speech.

3.1. Harmonic-plus-noise Parallel WaveGAN (HN-PWG)

To improve the performance of PWG, we propose an HN-PWG model, which involves the HNM to the PWG framework. As illustrated in Fig. 1, the proposed HN-PWG model divides a single WaveNet generator into harmonic and noise WaveNets for modeling the harmonic and noise components of a speech waveform, respectively. In particular, the harmonic WaveNet receives the sine wave, Gaussian noise¹, and voicing flags² as source signals, whereas the noise WaveNet receives the Gaussian noise and the voicing flags as source signals. To generate the sine wave, we adopt the method of neural source filter model [13], which designs the instantaneous frequency of sine wave to follow the fundamental fre-

¹Adding the Gaussian noise to the harmonic WaveNet is beneficial to improve the synthetic quality, especially when the proposed method is used for TTS applications. This will be further discussed in Section 4.3.

²The voicing flags are upsampled from frame-level to sample-level by nearest neighbor upsampling to match the time-resolution with sine wave and noise signals. Note that the usage of voicing flags enables each WaveNet to be effectively aware of the voicing states.

quency of target speech signal.

After harmonic and noise components are generated, they are mixed to compose a final speech waveform. As illustrated in Fig. 1-(a), the easiest way to compose harmonic and noise components is by simply adding them as follows:

$$\mathbf{x} = \mathbf{x}_h + \mathbf{x}_n, \quad (1)$$

where \mathbf{x} , \mathbf{x}_h , and \mathbf{x}_n denote output speech, harmonic, and noise waveforms, respectively. We can refer this type (i.e., Eq. (1)) to a *full-band model*, since the harmonic and noise components in the entire frequency range are equivalently mixed without considering their band-wise harmonic-noise property.

The performance of HN-PWG can be improved by considering this property through the multi-band harmonic-noise analysis during its training and generation processes. This will be discussed in the following section.

3.2. Multi-band HN-PWG

To further improve the performance of HN-PWG, we propose a multi-band HN-PWG model, which structurally represents the band-wise harmonic-noise characteristics of a speech signal. As illustrated in Fig. 1-(b), the multi-band HN-PWG decomposes the generated harmonic and noise components into the N number of subbands through a set of BPFs as follows:

$$\begin{aligned} \mathbf{x}_{h,i} &= \mathbf{x}_h \otimes \mathbf{g}_i, \\ \mathbf{x}_{n,i} &= \mathbf{x}_n \otimes \mathbf{g}_i, \end{aligned} \quad (2)$$

where \otimes denotes the convolution operation; $\mathbf{x}_{h,i}$, $\mathbf{x}_{n,i}$, and \mathbf{g}_i denote the harmonic and the noise waveforms and the BPF coefficients at the i^{th} subband, respectively. Then, the output speech signal is obtained through a weighted summation between the subband signals as follows:

$$\mathbf{x} = \sum_{i=0}^{N-1} [\alpha_i \cdot \mathbf{x}_{h,i} + (1 - \alpha_i) \cdot \mathbf{x}_{n,i}], \quad (3)$$

where α_i indicates a subband harmonicity that balances the energy between harmonic and noise components in the i^{th} subband.

Note that the subband harmonicity α_i can be treated as a heuristic parameter, which can be estimated by rule-based analysis methods [17, 18]. Alternatively, we design a *harmonicity estimator* consisting of small CNN blocks to predict the optimal value of α_i from input acoustic features. Since the subband harmonicity is now conditioned by acoustic features, it can efficiently learn the harmonic-noise characteristic of target speech, which is aligned with the characteristics of acoustic features.

To decouple the full-band signal into the subband components, we adopt the SincNet approach [20] that parameterizes each BPF by using a sinc function as follows:

$$g_i[k] = 2f_{i+1} \text{sinc}(2\pi f_{i+1}k) - 2f_i \text{sinc}(2\pi f_i k), \quad (4)$$

where $[f_i, f_{i+1}]$ denote the cutoff frequencies of the i^{th} subband and the sinc function is defined as $\text{sinc}(x) = \sin(x)/x$. Note that as the sinc function has a rectangular passband in the magnitude response, it can effectively minimize the aliasing effect between adjacent BPFs. For practical implementation, the filter coefficients are truncated by using a hamming window as follows:

$$\hat{g}_i[k] = g_i[k] * w[k], \quad (5)$$

where $\hat{g}_i[k]$ denotes truncated filter coefficients and $w[k] = 0.54 - 0.46 \cos(2\pi k/L)$ denotes a hamming window with a length of L . In the original SincNet, the cut-off frequencies of each BPF are initialized with Mel-scale and optimized during the training process. However, in the proposed system, we simply use the fixed cutoff frequencies that are uniformly divided by the N number of passbands in the frequency domain³.

Note that all of the operations proposed in the multi-band HN-PWG model, such as the set of BPFs and the harmonic-ity estimator, are fully differentiable. Therefore, two HNM-based generators, the harmonicity estimator, and the discriminators can be jointly optimized during the training process.

4. Experiments

4.1. Experimental setups

4.1.1. Speech database

In the experiments, a phonetically and prosodically balanced TTS corpus recorded by a female Korean professional speaker was used. The speech signals were sampled at 24 kHz with 16-bit quantization. In total, 5,087 utterances (5.5 hours), 550 utterances (36 minutes), and 130 utterances (6 minutes) were used for the training, validation, and test sets, respectively.

The acoustic features were extracted using an improved time-frequency trajectory excitation vocoder [22] at analysis intervals of 5 ms, including 40-dimensional line spectral frequencies, the fundamental frequency, the energy, the binary voicing flag, a 32-dimensional slowly evolving waveform, and a 4-dimensional rapidly evolving waveform, all of which constituted a 79-dimensional feature vector. The acoustic features were then normalized to have zero mean and unit variance using the statistics of the training data.

4.1.2. Neural vocoders

Table 1 presents the vocoding models including their model size and inference speed. As a baseline system, two WaveNet-based neural vocoders, an AR Gaussian WaveNet vocoder with a noise-shaping method [21] (S1) and a plain PWG vocoder [7] (S2) were tested.

For the Gaussian WaveNet, a time-invariant noise-shaping filter was obtained by averaging all spectra extracted from the training data to apply the noise-shaping method. This external filter was used to extract the residual signal before the training process, and its inverse filter was applied to reconstruct the speech signal in the synthesis step. The WaveNet systems consisted of 24 layers of dilated residual convolution blocks with four dilation cycles. There were 128 residual and skip channels, and the filter size was set to three. The model was trained for 1 M steps with a RAdam optimizer [23]. The learning rate was set to 0.001, and this was reduced by half every 200 K steps. The minibatch size was set to eight, and each audio clip was set to 12 K time samples (0.5 seconds).

For the plain PWG, the WaveNet generator consisting of 30 dilated residual blocks with three exponentially increasing dilation cycles was used. The number of residual and skip channels was set to 64, and the convolution filter size was five.

For the proposed HN-PWG, we tested the two cases of HN-PWG when the harmonic generator’s noise source is used or not (S3 and S4, respectively) to examine the importance of

additional noise as source signal. Note that the model without additional noise (S3) provides the same generator configuration with PeriodNet [16], where the sine wave is only used for the periodic (i.e., harmonic) generator. In detail, the harmonic WaveNet consisted of 20 dilated residual blocks with two exponentially increasing dilation cycles; whereas the noise WaveNet consisted of 10 residual blocks with one exponentially increasing dilation cycle. Similar to the plain PWG, the number of residual and skip channels was set to 64, and the convolution filter size was five. Note that the network size was also set to be the same as the plain PWG for a fair comparison. To provide continuously varying voicing information to the harmonic and noise WaveNets, the moving average filter with a 5 ms filter tap was applied to the upsampled voicing flag.

For the proposed multi-band HN-PWG (S5), the structures of harmonic and noise WaveNets were set to be the same as with HN-PWG. A total of 16 BPFs were parameterized by windowed sinc functions with 255 filter taps. The harmonicity estimator consisted of a 1-D CNN block with three convolution layers followed by output sigmoid layer. Each convolution layer consisted of 64 channels and five convolution filters interleaved with ReLU activation. To stabilize training, the last convolution layer was initialized with zeros, so as to equivalently mix the harmonic and noise components at early training stage (i.e., $\alpha_i = 0.5$). Across all vocoding models, the input auxiliary features were upsampled by nearest neighbor up-sampling followed by 2-D convolutions so that the time-resolution of the auxiliary features matched the sampling rate of the speech waveforms [24, 25]. Weight normalization was applied to all convolutional layers for all neural vocoders [26].

During the training of GAN-based vocoders, we used voicing-aware conditional discriminators, which efficiently guide the generator to learn voiced and unvoiced characteristics of speech signal. The detailed setup was the same as in original paper [9]. In addition, the MR-STFT loss was computed by summing three different STFT losses, as described for the original PWG [7]. The weight of the generator’s adversarial loss term was chosen to be 4.0. The models were trained for 400 K steps with a RAdam optimizer [23]. The discriminator was fixed for the first 100 K steps, and both the generator and discriminator were jointly trained afterwards. The minibatch size was set to 4, and the length of each audio clip was set to 24 K time samples (1.0 second). The initial learning rate was set to 0.0001 and 0.00005 for the generator and discriminator, respectively. The learning rate was reduced by half every 200 K steps.

4.2. Evaluation

To evaluate the perceptual quality of the vocoder itself, MOS listening tests in the analysis-synthesis scenario⁴. In particular, the speech samples were first generated by vocoders using ground-truth acoustic features. Then, a total of 20 native Korean listeners were asked to score the randomly selected 15 synthesized utterances from the test set one of the following five possible MOS responses: 1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent. Table 1 summarizes the MOS test results, the trends of which are analyzed as follows: (1) Both HN-PWG and multi-band HN-PWG provided significantly better perceptual quality of synthesized speech than the conventional PWG model while maintaining models complexity (S2 vs. S3, S4, and S5). In particular, the proposed multi-band HN-PWG achieved 4.29 MOS, which was 23% higher than the plain PWG (S2 vs. S5). (2) The quality of HN-PWG

³In our preliminary experiments, we also tried BPFs defined by Mel-scale cutoff frequencies, but found that there was no clear difference in their perceptual quality.

⁴Generated audio samples are available at the following URL: <https://min-jae.github.io/interspeech2021/>

Table 1. The model size, inference speed, and MOS results with 95% confidence intervals: Acoustic features extracted from the recorded speech signal were used to compose the input acoustic features. The MOS results for highest score is in bold font.

Label	Model	Use of HN model	Input signals for H-WaveNet	Type of HN model	Model size (M)	Inference speed	MOS
S1	WaveNet [21]	–	–	–	3.81	0.34×10^{-2}	4.22 ± 0.12
S2	PWG [7]	–	–	–	0.94	50.38	3.46 ± 0.37
S3	HN-PWG w/o noise [16]	Yes	Sine + V/UV	Full-band	0.94	47.91	4.02 ± 0.14
S4	HN-PWG	Yes	Sine + noise + V/UV	Full-band	0.94	47.93	4.18 ± 0.15
S5	Multi-band HN-PWG	Yes	Sine + noise + V/UV	Multi-band	0.99	47.87	4.29 ± 0.12
S6	Recordings	–	–	–	–	–	4.41 ± 0.12

S*i*: *i*th system; HN: harmonic-plus-noise; PWG: Parallel WaveGAN; H-WaveNet: harmonic WaveNet; V/UV: voicing flags upsampled from frame-level to sample-level. Note that inference speed, *k*, indicates that a system was able to generate waveforms *k* times faster than real-time. This evaluation was conducted on a server with a single NVIDIA Tesla V100 GPU.

Table 2. Subjective MOS test results with 95% confidence intervals for the TTS systems with respect to the different vocoding models. The MOS results for highest score is in bold font.

Label	Model	MOS
S-T1	WaveNet [21]	4.03 ± 0.19
S-T2	PWG [7]	3.56 ± 0.28
S-T3	HN-PWG w/o noise	2.60 ± 0.22
S-T4	HN-PWG	4.01 ± 0.17
S-T5	Multi-band HN-PWG	4.03 ± 0.16
S6	Recordings	4.41 ± 0.12

S-T*i*: *i*th system that generates speech waveform from the acoustic features predicted by TTS model.

could be improved by using an additional noise source for the harmonic WaveNet (S3 vs. S4). (3) The quality of multi-band HN-PWG was better than HN-PWG (S4 vs. S5). This indicates that the proposed multi-band approach was beneficial for improving the quality of HN-PWG. (4) The quality of multi-band HN-PWG was even better than the baseline AR WaveNet (S1 vs. S5), which was not for plain PWG (S1 vs. S2).

As shown in Fig. 2, the harmonic and noise components generated by the multi-band HN-PWG model were clearly decorrelated compared to those generated by the HN-PWG model. We conjecture that this is because our multi-band method efficiently guided each harmonic and noise WaveNet to learn the desired components during the training process.

4.3. Text-to-speech

To evaluate vocoding performance in the TTS scenario, we used an acoustic model based on Tacotron 2 with an external duration predictor [27, 28] for fast and stable generation as well as competitive synthesis quality.

To generate acoustic features, linguistic features were first extracted from the input text sequence. Then, the durations of each phoneme were predicted by a long short-term memory (LSTM)-based duration predictor. Based on the estimated durations, the phoneme-level linguistic features were upsampled to that of the frame level. Finally, the Tacotron2-style acoustic model predicted the acoustic features from the upsampled linguistic features. To improve the spectral clarity of the synthesized speech, the spectral domain sharpening filter [22] was applied as a post-processing technique. By inputting the resulting acoustic parameters, the vocoder models generated the time-domain speech signal. More setup details for the acoustic model are given in our previous work [27].

To evaluate the quality of the generated speech samples, the MOS tests were performed. The test setups were the same as those described in Section 4.2. Table 2 shows the results

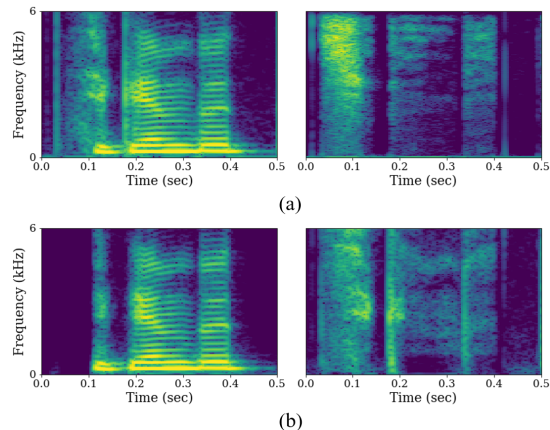


Figure 2: The spectrograms of harmonic (left side) and noise (right side) waveforms generated by the (a) HN-PWG (S4) and (b) multi-band HN-PWG (S5) models.

of the MOS tests, the findings of which are summarized as follows: (1) In the TTS scenario, the HN-PWG (w/o noise) provided significantly degraded quality compared to the system using noise source (S-T3 vs. S-T4). We found that the additional noise source was crucial for improving the robustness of HN-PWG when the acoustic features contain distortion through TTS prediction, which was not discovered in the study of PeriodNet [16]. (2) Even though the input acoustic features contained prediction errors, the HN-PWGs still presented better quality than the plain PWG (S-T2 vs. S-T4 and S-T5). (3) Finally, the multi-band HN-PWG within a TTS framework achieved 4.03 MOS, which was 13% higher than the plain PWG (S-T2 vs. S-T5).

5. Conclusion

In this paper, we proposed a multi-band HN-PWG vocoder to improve the PWG-based non-AR neural vocoding system. By guiding the neural vocoder to learn the complicated multiple frequency band-dependent harmonic and noise characteristics of speech signal, we successfully improved the quality of the PWG vocoder. The experimental results verified that the proposed multi-band HN-PWG model provided better synthesis quality within both analysis-synthesis and TTS scenarios. Future work includes improving the synthesis speed of HN-PWG by utilizing the knowledge of speech signal processing.

6. Acknowledgment

This work was supported by Clova Voice, NAVER Corp., Seongnam, Korea.

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 *Arxiv*, 2016. [Online]. Available: <https://arxiv.org/pdf/1609.03499.pdf>
- [2] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, “Speaker-dependent WaveNet vocoder,” in *Proc. Interspeech*, 2017, pp. 1118–1122.
- [3] E. Song, K. Byun, and H.-G. Kang, “ExcitNet vocoder: A neural excitation model for parametric speech synthesis systems,” in *Proc. EUSIPCO*, 2019, pp. 1179–1183.
- [4] M.-J. Hwang, F. Soong, E. Song, X. Wang, H. Kang, and H.-G. Kang, “LP-WaveNet: Linear prediction-based WaveNet speech synthesis,” in *Proc. APSIPA*, 2020, pp. 810–814.
- [5] 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*, 2018, pp. 3915–3923.
- [6] 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*, 2019, pp. 14 881–14 892.
- [7] R. Yamamoto, E. Song, and J. Kim, “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in *Proc. ICASSP*, 2020, pp. 6199–6203.
- [8] E. Song, R. Yamamoto, M.-J. Hwang, J.-S. Kim, O. Kwon, and J.-M. Kim, “Improved Parallel WaveGAN vocoder with perceptually weighted spectrogram loss,” in *Proc. SLT*, 2021, pp. 470–476.
- [9] R. Yamamoto, E. Song, M.-J. Hwang, and J.-M. Kim, “Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators,” in *Proc. ICASSP*, 2021, pp. 6039–6043.
- [10] Y. Stylianou, “Modeling speech based on harmonic plus noise models,” in *Nonlinear Speech Modeling and Applications*. Springer Berlin Heidelberg, 2005, pp. 244–260.
- [11] J. Laroche, Y. Stylianou, and E. Moulines, “HNS: Speech modification based on a harmonic+noise model,” in *Proc. ICASSP*, 1993, pp. 550–553.
- [12] A. Abrantes, J. Marques, and I. Trancoso, “Hybrid sinusoidal modeling of speech without voicing decision,” in *Proc. EUROSPEECH*, 1991, pp. 231–234.
- [13] X. Wang, S. Takaki, and J. Yamagishi, “Neural source-filter waveform models for statistical parametric speech synthesis,” *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 28, pp. 402–415, 2020.
- [14] X. Wang and J. Yamagishi, “Neural harmonic-plus-noise waveform model with trainable maximum voice frequency for text-to-speech synthesis,” in *Proc. SSW*, 2019, pp. 1–6.
- [15] Z. Liu, K. Chen, and K. Yu, “Neural homomorphic vocoder,” in *Proc. Interspeech*, 2020, pp. 240–244.
- [16] Y. Hono, S. Takaki, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “PeriodNet: A non-autoregressive waveform generation model with a structure separating periodic and aperiodic components,” in *Proc. ICASSP*, 2021, pp. 6049–6053.
- [17] A. V. McCree and T. P. Barnwell III, “A mixed excitation LPC vocoder model for low bit rate speech coding,” *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 4, pp. 242–250, 1995.
- [18] H. Kawahara, “Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited,” in *Proc. ICASSP*, 1997, pp. 1303–1306.
- [19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proc. NIPS*, 2014, pp. 2672–2680.
- [20] M. Ravanelli and Y. Bengio, “Speaker recognition from raw waveform with SincNet,” in *Proc. SLT 2018*, 2018, pp. 1021–1028.
- [21] K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, “An investigation of noise shaping with perceptual weighting for WaveNet-based speech generation,” in *Proc. ICASSP*, 2018, pp. 5664–5668.
- [22] E. Song, F. K. Soong, and H.-G. Kang, “Effective spectral and excitation modeling techniques for LSTM-RNN-based speech synthesis systems,” *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 11, pp. 2152–2161, 2017.
- [23] 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,” *arXiv preprint arXiv:1908.03265*, 2019.
- [24] A. Odena, V. Dumoulin, and C. Olah, “Deconvolution and checkerboard artifacts,” *Distill*, 2016. [Online]. Available: <http://distill.pub/2016/deconv-checkerboard>
- [25] R. Yamamoto, E. Song, and J.-M. Kim, “Probability density distillation with generative adversarial networks for high-quality parallel waveform generation,” in *Proc. Interspeech*, 2019, pp. 699–703.
- [26] T. Salimans and D. P. Kingma, “Weight normalization: A simple reparameterization to accelerate training of deep neural networks,” in *Proc. NIPS*, 2016, pp. 901–909.
- [27] E. Song, M.-J. Hwang, R. Yamamoto, J.-S. Kim, O. Kwon, and J.-M. Kim, “Neural text-to-speech with a modeling-by-generation excitation vocoder,” in *Proc. Interspeech*, 2020, pp. 3570–3574.
- [28] M. Hwang, R. Yamamoto, E. Song, and J. Kim, “TTS-by-TTS: TTS-driven data augmentation for fast and high-quality speech synthesis,” in *Proc. ICASSP*, 2021, pp. 6598–6602.