IMPROVED PARALLEL WAVEGAN VOCODER WITH PERCEPTUALLY WEIGHTED SPECTROGRAM LOSS

Eunwoo Song[1]*, Ryuichi Yamamoto*[2]*, Min-Jae Hwang*[3]*, Jin-Seob Kim*[1]*, Ohsung Kwon*[1]*, Jae-Min Kim*[1]

[1]NAVER Corp., Seongnam, Korea
[2]LINE Corp., Tokyo, Japan
[3]Search Solutions Inc., Seongnam, Korea

ABSTRACT

This paper proposes a spectral-domain perceptual weighting technique for Parallel WaveGAN-based text-to-speech (TTS) systems. The recently proposed Parallel WaveGAN vocoder successfully generates waveform sequences using a fast non-autoregressive WaveNet model. By employing multi-resolution short-time Fourier transform (MR-STFT) criteria with a generative adversarial network, the light-weight convolutional networks can be effectively trained without any distillation process. To further improve the vocoding performance, we propose the application of frequency-dependent weighting to the MR-STFT loss function. The proposed method penalizes perceptually-sensitive errors in the frequency domain; thus, the model is optimized toward reducing auditory noise in the synthesized speech. Subjective listening test results demonstrate that our proposed method achieves 4.21 and 4.26 TTS mean opinion scores for female and male Korean speakers, respectively.

Index Terms— Text-to-speech, speech synthesis, neural vocoder, Parallel WaveGAN

1. INTRODUCTION

Generative models for raw speech waveforms have significantly improved the quality of neural text-to-speech (TTS) systems [1, 2]. Specifically, autoregressive generative models such as *WaveNet* have successfully replaced the role of traditional parametric vocoders [2–5]. Non-autoregressive versions, including *Parallel WaveNet*, provide a fast waveform generation method based on a teacher-student framework [6, 7]. In this method, the model is trained using a probability density distillation method in which the knowledge of an autoregressive teacher WaveNet is transferred to an inverse autoregressive flow student model [8].

In our previous work, we introduced generative adversarial network (GAN) training methods to the Parallel WaveNet framework [9], and proposed *Parallel WaveGAN* by combining the adversarial training with multi-resolution short-time Fourier transform (MR-STFT) criteria [10, 11]. Although it is possible to train GAN-based non-autoregressive models by only using adversarial loss function [12], employing the MR-STFT loss function has been proven to be advantageous for increasing the training efficiency [10, 13, 14]. Furthermore, because the Parallel WaveGAN only trains a WaveNet model without any density distillation, the entire training process becomes much easier than it is in the conventional methods, and the model can produce natural sounding speech waveforms with just a small number of parameters.

To further enhance the performance of the Parallel WaveGAN, this paper proposes a spectral-domain perceptual weighting method for optimizing the MR-STFT criteria. A frequency-dependent masking filter is designed to penalize errors near the spectral valleys, which are perceptually-sensitive to the human ear [15]. By applying this filter to the STFT loss function calculations in the training step, the network is guided to reduce the noise component in those regions. Consequently, the proposed model generates a more natural voice in comparison to the original Parallel WaveGAN. Our contributions can be summarized as follows:

- We propose a perceptually weighted MR-STFT loss function alongside a conventional adversarial training method. This approach improves the quality of the synthesized speech in the Parallel WaveGAN-based neural TTS system.
- Because the proposed method does not change the network architecture, it maintains the small number of parameters found in the original Parallel WaveGAN's and retain its fast inference speed. In particular, the system can generate a 24 kHz speech waveform 50.57 times faster than real-time in a single GPU environment with 1.83 M parameters.
- Consequently, our method achieved mean opinion score (MOS) results of 4.21 and 4.26 for female and male Korean speakers, respectively, in the neural TTS systems.

2. RELATED WORK

The idea of using STFT-based loss functions is not new. In their study of spectrogram inversion, Sercan et al. [16] first proposed *spectral convergence* and *log-scale STFT magnitude* losses, and our previous work proposed combing these in a multi-resolution form [9].

Moreover, perceptual noise-shaping filters have significantly improved the quality of synthesized speech in autoregressive WaveNet frameworks [17]. Based on the characteristics of the human auditory system, an external noise-shaping filter is designed to reduce perceptually-sensitive noise in the spectral valley regions. This filter acts as a pre-processor in the training step; thus, the WaveNet learns the distribution of *noise-shaped residual signal*. In the synthesis step, by applying its inverse filter to the WaveNet's output, the enhanced speech can be reconstructed.

However, it has been shown that the filter's effectiveness does not work for the non-autoregressive generation models, including WaveGlow [18] and the Parallel WaveGAN. One possible reason for this might be that the characteristics of the noise-shaped residual signal are difficult for the non-autoregressive model to capture without previous time-step information. To address this problem, the proposed system applies a frequency-dependent mask to the process of calculating the STFT loss functions. As this method does not change the target speech's distribution, the non-autoregressive WaveNet can be stably optimized, while significantly reducing the auditory noise components.

3. PARALLEL WAVEGAN

The Parallel WaveGAN jointly trains a non-causal WaveNet generator, G, and a convolutional neural network (CNN) discriminator, D, to generate a time-domain speech waveform from the corresponding input acoustic parameters. Specifically, the generator learns a distribution of realistic waveforms by trying to deceive the discriminator into recognizing the generated samples as real. The process is performed by minimizing the generator losses as follows:

$$L_{\mathrm{G}}(G,D) = L_{\mathrm{mr\_stft}}(G) + \lambda_{\mathrm{adv}} L_{\mathrm{adv}}(G,D), \qquad (1)$$

where $L_{\mathrm{mr\_stft}}(G)$ denotes an MR-STFT loss, which will be discussed in the next section; $L_{\mathrm{adv}}(G,D)$ denotes an adversarial loss; λ_{adv} denotes the hyperparameter balancing the two loss functions. The adversarial loss is designed based on least-squares GANs [19–22], as follows:

$$L_{\mathrm{adv}}(G,D) = \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}\left[(1 - D(G(\boldsymbol{z},\boldsymbol{h})))^2\right], \qquad (2)$$

where $\boldsymbol{z}$, $p_{\boldsymbol{z}}$, and $\boldsymbol{h}$ denote the input noise, a Gaussian distribution $N(\mathbf{0},\mathbf{I})$, and the conditional acoustic parameters, respectively.

The discriminator is trained to correctly classify the generated sample as *fake* while classifying the ground truth as *real* using the following optimization criterion:

$$L_{\mathrm{D}}(G,D) = \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}}[(1-D(\boldsymbol{x}))^2] + \mathbb{E}_{\boldsymbol{z}\sim p_{\mathrm{z}}}\left[D(G(\boldsymbol{z},\boldsymbol{h}))^2\right], \qquad (3)$$

where $\boldsymbol{x}$ and p_{data} denote the target speech waveform and its distribution, respectively.

3.1. Conventional MR-STFT loss

To guarantee the stability of the adversarial training method described above, it is crucial to incorporate an MR-STFT loss function into the generator's optimization process [10]. The MR-STFT loss function in equation (1) is defined in terms of the number of STFT losses, M, as follows:

$$L_{\mathrm{mr\_stft}}(G) = \frac{1}{M}\sum_{m=1}^{M} L_{\mathrm{stft}}^{(m)}(G), \qquad (4)$$

where $L_{\mathrm{stft}}^{(m)}(G)$ denotes the m^{th} STFT loss defined as follows:

$$L_{\mathrm{stft}}(G) = \mathbb{E}_{\boldsymbol{x}\sim p_{data},\hat{\boldsymbol{x}}\sim p_G}\left[L_{\mathrm{sc}}(\boldsymbol{x},\hat{\boldsymbol{x}}) + L_{\mathrm{mag}}(\boldsymbol{x},\hat{\boldsymbol{x}})\right], \qquad (5)$$

where $\hat{\boldsymbol{x}}$ denotes the generated sample drawn by probability distribution of generator, p_G; L_{sc} and L_{mag} denote *spectral convergence* and *log STFT magnitude* losses, respectively, which are defined as follows [16]:

$$L_{\mathrm{sc}}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \frac{\sqrt{\sum_{t,f}(|\mathbf{X}_{t,f}| - |\hat{\mathbf{X}}_{t,f}|)^2}}{\sqrt{\sum_{t,f}|\mathbf{X}_{t,f}|^2}}, \qquad (6)$$

$$L_{\mathrm{mag}}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \frac{\sum_{t,f}|\log|\mathbf{X}_{t,f}| - \log|\hat{\mathbf{X}}_{t,f}||}{T\cdot N}, \qquad (7)$$

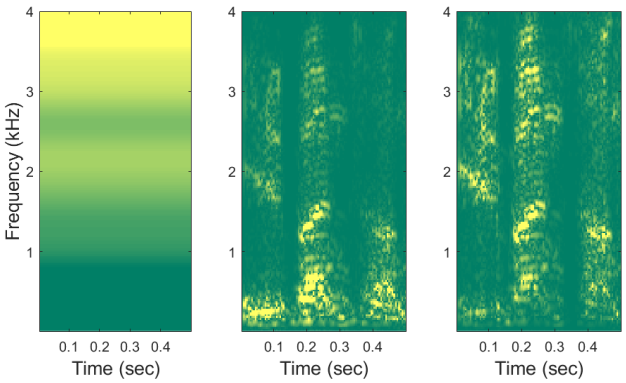
where $|\mathbf{X}_{t,f}|$ and $|\hat{\mathbf{X}}_{t,f}|$ denote the f^{th} STFT magnitude of $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ at the time frame t, respectively; T and N denote the number of frames and the number of frequency bins, respectively.

3.2. Proposed perceptual weighting for MR-STFT loss

To further enhance the performance of the Parallel WaveGAN, this paper proposes to apply a spectral-domain perceptual masking filter to the MR-STFT loss criteria as follows:

$$L_{\mathrm{sc}}^{w}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \frac{\sqrt{\sum_{t,f}(\mathbf{W}_{t,f}(|\mathbf{X}_{t,f}| - |\hat{\mathbf{X}}_{t,f}|))^2}}{\sqrt{\sum_{t,f}|\mathbf{X}_{t,f}|^2}}, \qquad (8)$$

$$L_{\mathrm{mag}}^{w}(\boldsymbol{x},\hat{\boldsymbol{x}}) = \frac{\sum_{t,f}|\log\mathbf{W}_{t,f}(\log|\mathbf{X}_{t,f}| - \log|\hat{\mathbf{X}}_{t,f}|)|}{T\cdot N}, \qquad (9)$$



(a) (b) (c)

Fig. 1: Magnitude distance (MD) obtained when calculating the spectral convergence: (a) The weight matrix of spectral mask, (b) the MD before applying the mask (conventional method), and (c) the MD after applying the mask (proposed method).

where $\mathbf{W}_{t,f}$ denotes a weight coefficient of the spectral mask. The weight matrix $\mathbf{W}$ is constructed by repeating a time-invariant frequency masking filter along the time axis, whose transfer function is defined as follows:

$$\boldsymbol{W}(z) = 1 - \sum_{k=1}^{p} \tilde{\alpha}_k z^{-k}, \tag{10}$$

where $\tilde{\alpha}_k$ denotes the k^{th} linear prediction (LP) coefficient with the order p, obtained by averaging all spectra extracted from the training data. As shown in Fig. 1a, the weight matrix of the spectral mask is designed to represent the global characteristics of the spectral formant structure. This enables an emphasis on losses at the frequency regions of the spectral valleys, which are more sensitive to the human ear. When calculating the STFT loss (Fig. 1b), this filter is used to penalize losses in those regions (Fig. 1c). As a result, the training process can guide the model to further reduce the perceptual noise in the synthesized speech[1].

The merits of the proposed method are presented in Fig. 2 which shows the log-spectral distance between the original and generated speech signals. The proposed perceptual weighting of MR-STFT losses enables an accurate estimation of speech spectra, and it is therefore expected that it will provide more accurate training and generation results, to be discussed further in the following section.

[1]Although the log-scale STFT-magnitude loss in equation (7) was designed to fit small amplitude components [16], our preliminary experiments verified that applying the masking filter to this loss was also beneficial to synthetic quality.

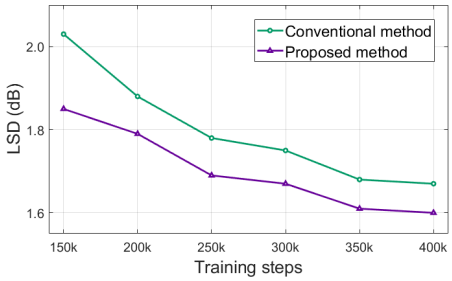


Fig. 2: Log-spectral distance (LSD; dB) between the original and generated speech signals

Table 1: Utterances in speech sets by Korean male (KRM) and Korean female (KRF) speakers (SPK).

SPK	Training	validation	Testing
KRF	5,085 (5.5 h)	360 (0.4 h)	180 (0.2 h)
KRM	5,382 (7.4 h)	290 (0.4 h)	140 (0.2 h)

4. EXPERIMENTS

4.1. Experimental setup

4.1.1. Database

The experiments used two phonetically and prosodically rich speech corpora recorded by Korean male and female professional speakers. The speech signals were sampled at 24 kHz, and each sample was quantized by 16 bits. Table 1 shows the number of utterances in each set. The acoustic features were extracted using an improved time-frequency trajectory excitation vocoder at the analysis intervals of 5 ms [23], and these features included 40-dimensional line spectral frequencies (LSFs), fundamental frequency, energy, voicing flag, a 32-dimensional slowly evolving waveform, and a 4-dimensional rapidly evolving waveform, all of which constituted a 79-dimensional feature vector.

4.1.2. Acoustic model

Although there are many state-of-the-art acoustic architectures available [24–26], we used a Tacotron model with phoneme alignment [27, 28] for its fast and stable generation and competitive synthesis quality. The left section of Fig. 3 presents the acoustic model which consists of three sub-modules, namely, context analysis, context embedding, and Tacotron decoding.

In the context analysis module, a grapheme-to-phoneme converter was applied to the input text by the Korean standard pronunciation grammar, and then phoneme-level feature vectors were extracted by the internal context information-labeling program. These were composed of 330 binary features for categorical linguistic contexts and 24 features for

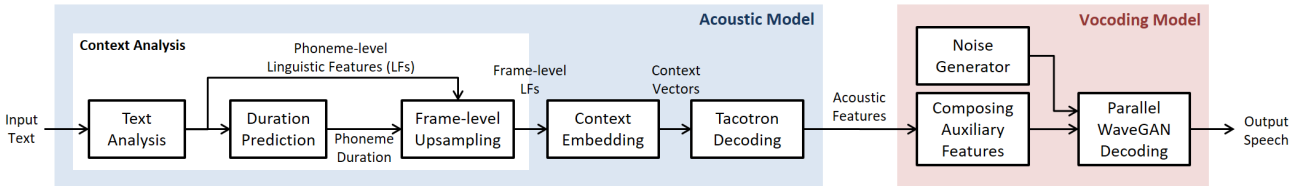


Fig. 3: Block diagram of the TTS framework.

Table 2: Vocoding model details, including size and inference speed: 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.

System	Model	MR-STFT loss	Perceptual weighting	Noise shaping	Number of layers	Model size	Inference speed
Baseline 1	WaveNet	-	-	-	24	3.71 M	0.34×10^{-2}
Baseline 2	WaveNet + NS	-	-	Yes	24	3.81 M	0.34×10^{-2}
Baseline 3	Parallel WaveGAN	Yes	-	-	30	1.83 M	50.57
Baseline 4	Parallel WaveGAN + NS	Yes	-	Yes	30	1.83 M	47.70
Proposal	Parallel WaveGAN + PW	Yes	Yes	-	30	1.83 M	50.57

numerical linguistic contexts. By inputting those linguistic features, the corresponding phoneme duration was estimated through three fully connected (FC) layers with 1,024, 512, 256 units followed by a unidirectional long short-term memory (LSTM) network with 128 memory blocks. Based on this estimated duration, the phoneme-level linguistic features were then up-sampled to frame-level by adding the two numerical vectors of phoneme duration and relative position.

In the context embedding module, the linguistic features are transformed into high-level context vectors. The module in this experiment consisted of three convolution layers with 10×1 kernels and 512 channels per layer, a bi-directional LSTM network with 512 memory blocks, and an FC layer with 512 units.

To generate the output acoustic features, we used a Tacotron 2 decoder network [25]. First, the previously generated acoustic features were fed into two FC layers with 256 units (i.e. the PreNet), and those features and the vectors from the context embedding module were then passed through two uni-directional LSTM layers with 1,024 memory blocks followed by two projection layers. Finally, to improve generation accuracy, five convolution layers with 5×1 kernels and 512 channels per layer were used as a post-processing network (i.e. the PostNet) to add the residual elements of the generated acoustic features.

Before training, the input and output features were normalized to have zero mean and unit variance. The weights were initialized using *Xavier* initialization [29] and *Adam* optimization was used [30]. The learning rate was scheduled to be decayed from 0.001 to 0.0001 via a decaying rate of 0.33 per 100 K steps.

4.1.3. Vocoding model

Table 2 presents details of the vocoding models including their size and inference speed. As baseline systems, we used two autoregressive WaveNet vocoders, namely, a plain WaveNet (Baseline 1) [3] and a WaveNet with noise-shaping (NS) method (Baseline 2) [17]. We adopted continuous Gaussian output distributions for both the baseline systems [7], instead of using the categorical distributions. These two approaches used the same network architecture but differed in the target output; the plain WaveNet system was designed to predict speech signals, whereas the latter method was designed to predict the noise-shaped residual signals. Note that a time-invariant noise-shaping filter was obtained by averaging all spectra extracted from the training data. 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. 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).

The experiment involved three Parallel WaveGAN systems, namely, the plain Parallel WaveGAN (Baseline 3) [10], a Parallel WaveGAN with the same noise-shaping method as before (Baseline 4), and the proposed method with the perceptually weighted (PW) criteria (Proposal). All had the same network architecture consisting of 30 dilated residual convolution block layers with three exponentially increasing dila-

Table 3: The details of the MR-STFT loss calculations. A Hanning window was applied before the FFT process.

STFT loss	FFT size	Window size	Frame shift
$L_{\mathrm{stft}}^{(1)}$	512	240 (10 ms)	50 ($\approx$ 2 ms)
$L_{\mathrm{stft}}^{(2)}$	1024	600 (25 ms)	120 (5 ms)
$L_{\mathrm{stft}}^{(3)}$	2048	1200 (50 ms)	240 (10 ms)

tion cycles. The number of residual and skip channels was set to 64, and the convolution filter size was three. The discriminator consisted of 10 layers of non-causal dilated 1-D convolutions with leaky ReLU activation function ($\alpha = 0.2$). The strides were set to 1, and linearly increasing dilations were applied to the 1-D convolutions, except the first and last layers, from 1 to 8. The number of channels and filter size were the same as the generator. We applied weight normalization to all convolutional layers for both the generator and the discriminator [31].

The MR-STFT loss was calculated by summing three STFT losses as shown in Table 3, which had been defined in its original version [10]. In the proposed method, to obtain the time-invariant masking filter in equation (10), all the LSFs ($p = 40$) collected from the training data were averaged, and converted to the corresponding LP coefficients [32]. For a stable convergence, the masking filter's magnitude response was normalized to have a range from 0.5 to 1.0 before applying it to the MR-STFT loss. The discriminator loss was computed by the average of per-time-step scalar predictions with the discriminator. The value of hyperparameter, λ_{adv}, in equation (1) was chosen to be 4.0. The models were trained for 400 K steps with RAdam optimization ($\epsilon = 1e^{-6}$) to stabilize training [33]. 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 8, 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.

Across all vocoding models, the input auxiliary features were up-sampled 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 [9, 34].

4.1.4. Text-to-speech generation

In the synthesis step, the acoustic feature vectors were predicted by the acoustic model with the given input text. To enhance spectral clarity, an LSF-sharpening filter was applied to the spectral parameters [23]. By using these features as the conditional inputs, vocoding models such as WaveNet and Parallel WaveGAN generated corresponding time-sequences of the waveforms.

Table 4: Naturalness MOS test results with 95% confidence intervals for the TTS systems with respect to the different vocoding models: The MOS results for the proposed system are in bold font. The KRF and KRM denote Korean female and male speakers, respectively.

Index	Model	KRF	KRM
Test 1	WaveNet	3.64±0.14	3.60±0.13
Test 2	WaveNet + NS	4.36±0.11	4.32±0.10
Test 3	Parallel WaveGAN	4.02±0.10	4.11±0.11
Test 4	Parallel WaveGAN + NS	2.34±0.10	1.72±0.09
Test 5	**Parallel WaveGAN + PW**	**4.26±0.10**	**4.21±0.10**
Test 6	Raw	4.64±0.07	4.59±0.09

4.2. Evaluations

Naturalness MOS tests were conducted to evaluate the perceptual quality of the proposed system[2]. 20 native Korean speakers were asked to make quality judgments about the synthesized speech samples using the five following possible responses: 1 = Bad; 2 = Poor; 3 = Fair; 4 = Good; and 5 = Excellent. In total, 30 utterances were randomly selected from the test set and synthesized using the different generation models.

Table 4 presents the MOS test results for the TTS systems with respect to the different vocoding models, and the analysis can be summarized as follows: First, in systems with autoregressive WaveNet vocoders, applying the noise-shaping filter performed significantly better than the plain systems (Tests 1 and 2). This confirms that reducing auditory noise in the spectral valley regions was beneficial to perceptual quality. However, the effectiveness of the noise-shaping filter was not evident for the Parallel WaveGAN systems (Tests 3 and 4). Since the training and generation processes are both non-autoregressive, it might be that the characteristics of a noise-shaped target signal were difficult for the model to capture without previous time-step information. Second, the systems with Parallel WaveGAN and the proposed perceptually weighted MR-STFT loss function demonstrated improved quality of synthesized speech (Tests 3 and 5). Because the weighting helped the model reduce generation errors in the spectral valleys, and because the adversarial training method helped capture the characteristics of realistic speech waveforms, the system was able to generate a natural voice within a non-autoregressive framework, providing the 14.87 K times faster inference than the best autoregressive model (Test2). Consequently, the TTS system with the proposed Parallel WaveGAN vocoder achieved 4.21 and 4.26 MOS results for female and male speakers, respectively.

[2]Generated audio samples are available at the following URL: `https://sewplay.github.io/demos/wavegan-pwsl`

5. CONCLUSIONS

This paper proposed a spectral-domain perceptual weighting technique for Parallel WaveGAN-based TTS systems. A frequency-dependent masking filter was applied to the MR-STFT loss function, enabling the system to penalize errors near the spectral valleys. As a result, the generation errors in those frequency regions were reduced, which improved the quality of the synthesized speech. The experimental results verified that a TTS system with the proposed Parallel WaveGAN vocoder performs better than systems with conventional methods. Future research includes further improving the Parallel WaveGAN's perceptual quality by replacing the time-invariant spectral masking filter with a signal-dependent adaptive predictor.

6. REFERENCES

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

[2] 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," *arXiv preprint arXiv:1609.03499*, 2016.

[3] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder," in *Proc. INTERSPEECH*, 2017, pp. 1118–1122.

[4] M.-J. Hwang, F. Soong, E. Song, X. Wang, H. Kang, and H.-G. Kang, "LP-WaveNet: Linear prediction-based WaveNet speech synthesis," *arXiv preprint arXiv:1811.11913*, 2018.

[5] E. Song, K. Byun, and H.-G. Kang, "ExcitNet vocoder: A neural excitation model for parametric speech synthesis systems," in *Proc. EUSIPCO*, 2019, pp. 1–5.

[6] 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 *et al.*, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. ICML*, 2018, pp. 3915–3923.

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

[8] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved variational inference with inverse autoregressive flow," in *Proc. NIPS*, 2016, pp. 4743–4751.

[9] 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.

[10] ——, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, 2020, pp. 6199–6203.

[11] R. Yamamoto, E. Song, M.-J. Hwang, and J.-M. Kim, "Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators," *arXiv preprint arXiv:2010.14151*, 2020.

[12] 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.

[13] 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," *arXiv preprint arXiv:2005.05106*, 2020.

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

[15] M. R. Schroeder, B. S. Atal, and J. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *Journal of Acoust. Soc. of America*, vol. 66, no. 6, pp. 1647–1652, 1979.

[16] S. Ö. Arık, H. Jun, and G. Diamos, "Fast spectrogram inversion using multi-head convolutional neural networks," *IEEE Signal Procees. Letters*, vol. 26, no. 1, pp. 94–98, 2019.

[17] 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.

[18] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Real-time neural text-to-speech with sequence-to-sequence acoustic model and WaveGlow or single Gaussian WaveRNN vocoders," in *Proc. INTERSPEECH*, 2019, pp. 1308–1312.

[19] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proc. ICCV*, 2017, pp. 2794–2802.

[20] Q. Tian, X. Wan, and S. Liu, "Generative adversarial network based speaker adaptation for high fidelity WaveNet vocoder," in *Proc. SSW*, 2019, pp. 19–23.

[21] 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.

[22] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proc. INTERSPEECH*, 2017, pp. 3642–3646.

[23] 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.

[24] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, 2017, pp. 4006–4010.

[25] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[26] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. T. Zhou, "Neural speech synthesis with Transformer network," in *Proc. AAAI*, 2019, pp. 6706–6713.

[27] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Tacotron-based acoustic model using phoneme alignment for practical neural text-to-speech systems," in *Proc. ASRU*, 2019, pp. 214–221.

[28] 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.

[29] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, 2010, pp. 249–256.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[31] 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.

[32] F. Soong and B. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. ICASSP*, 1984, pp. 37–40.

[33] 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.

[34] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, 2016. [Online]. Available: http://distill.pub/2016/deconv-checkerboard