

PARALLEL WAVEGAN: A FAST WAVEFORM GENERATION MODEL BASED ON GENERATIVE ADVERSARIAL NETWORKS WITH MULTI-RESOLUTION SPECTROGRAM

Ryuichi Yamamoto¹, Eunwoo Song² and Jae-Min Kim²

¹LINE Corp., Tokyo, Japan.

²NAVER Corp., Seongnam, Korea

ABSTRACT

We propose Parallel WaveGAN, a distillation-free, fast, and small-footprint waveform generation method using a generative adversarial network. In the proposed method, a non-autoregressive WaveNet is trained by jointly optimizing multi-resolution spectrogram and adversarial loss functions, which can effectively capture the time-frequency distribution of the realistic speech waveform. As our method does not require density distillation used in the conventional teacher-student framework, the entire model can be easily trained. Furthermore, our model is able to generate high-fidelity speech even with its compact architecture. In particular, the proposed Parallel WaveGAN has only 1.44 M parameters and can generate 24 kHz speech waveform 28.68 times faster than real-time on a single GPU environment. Perceptual listening test results verify that our proposed method achieves 4.16 mean opinion score within a Transformer-based text-to-speech framework, which is comparative to the best distillation-based Parallel WaveNet system.

Index Terms— Neural vocoder, text-to-speech, generative adversarial networks, Parallel WaveNet, Transformer

1. INTRODUCTION

Deep generative models in text-to-speech (TTS) frameworks have significantly improved the quality of synthetic speech signals [1–3]. Remarkably, autoregressive generative models such as WaveNet have shown much superior performance over traditional parametric vocoders [4–8]. However, they suffer from slow inference speed due to their autoregressive nature and thus are limited in their applications to real-time scenarios.

One approach to address the limitation is to utilize fast waveform generation methods based on a teacher-student framework [9–11]. In this framework, a bridge defined as probability density distillation transfers the knowledge of an autoregressive teacher WaveNet to an inverse autoregressive flow (IAF)-based student model [12]. Although the IAF student can achieve real-time generation of speech with reasonable perceptual quality, there remain problems in the training process: it requires not only a well trained teacher model, but also a trial and error methodology to optimize the complicated density distillation process.

To overcome the aforementioned problems, we propose a *Parallel WaveGAN*¹, a simple and effective parallel waveform generation method based on a generative adversarial network (GAN) [14]. Unlike the conventional distillation-based methods, the Parallel WaveGAN does not require the two-stage, sequential teacher-

student training process. In the proposed method, only a non-autoregressive WaveNet model is trained by optimizing the combination of multi-resolution short-time Fourier transform (STFT) and adversarial loss functions that enable the model to effectively capture the time-frequency distribution of the realistic speech waveform. As a result, the entire training process becomes much easier than the conventional methods, as well as the model can produce natural sounding speech waveforms with a small number of model parameters. Our contributions are summarized as follows:

- We propose a joint training method of the multi-resolution STFT loss and the waveform-domain adversarial loss. This approach effectively works for the conventional distillation-based Parallel WaveNet (e.g., ClariNet), as well as for the proposed distillation-free Parallel WaveGAN.
- As the proposed Parallel WaveGAN can be simply trained without any teacher-student framework, our approach significantly reduces both the training and inference time. In particular, the training process becomes 4.82 times faster (from 13.5 days to 2.8 days with two NVIDIA Tesla V100 GPUs) and the inference process becomes 1.96 times faster (from 14.62 to 28.68 real-time² to generate 24 kHz speech waveforms with a single NVIDIA Tesla V100 GPU) compared with the conventional ClariNet model.
- We combined the proposed Parallel WaveGAN with a TTS acoustic model based on a Transformer [15–17]. The perceptual listening tests verify that the proposed Parallel WaveGAN achieves 4.16 MOS, which is competitive to the best distillation-based ClariNet model.

2. RELATED WORK

The idea of using GAN in the Parallel WaveNet framework is not new. In our previous work, the IAF student model was incorporated as a generator and jointly optimized by minimizing the adversarial loss along with the Kullback-Leibler divergence (KLD) and auxiliary losses [11]. As the GAN learns the distribution of realistic speech signals, the method significantly improves the perceptual quality of synthetic signal. However, the complicated training stage based on density distillation limits its utilization.

Our aim is to minimize the effort to train the two-stage pipeline of the conventional teacher-student framework. In other words, we propose a novel method to train the Parallel WaveNet without any

¹ Note that our work is not closely related to an unsupervised waveform synthesis model, WaveGAN [13].

² The inference speed defined as k means that the system can generate waveforms k times faster than real-time.

distillation process. Juvela et al [18] has proposed a similar approach (e.g., GAN-excited linear prediction; GELP) that generates glottal excitations by using the adversarial training method. However, since GELP requires linear prediction (LP) parameters to convert glottal excitations to speech waveform, quality degradation may occur when the LP parameters contain inevitable errors caused by the TTS acoustic model. To avoid this problem, our method is designed to directly estimate the speech waveform. As it is very difficult to capture the dynamic nature of speech signal including both the vocal cord movement and the vocal tract resonance (represented by glottal excitations and LP parameters in GELP, respectively), we propose a joint optimization method between the adversarial loss and multi-resolution STFT loss in order to capture the time-frequency distributions of the realistic speech signal. As a result, the entire model can be easily trained even with a small number of parameters while effectively reducing the inference time and improving perceptual quality of synthesized speech.

3. METHOD

3.1. Parallel waveform generation based on GAN

GANs are generative models that are composed of two separate neural networks: a generator (G) and a discriminator (D) [14]. In our method, a WaveNet-based model conditioned on an auxiliary feature (e.g., mel-spectrogram) is used as the generator, which transforms the input noise to the output waveform in parallel. The generator differs from the original WaveNet in that: (1) we use non-causal convolutions instead of causal convolutions; (2) the input is random noise drawn from a Gaussian distribution; (3) the model is non-autoregressive at both training and inference steps.

The generator learns a distribution of realistic waveforms by trying to deceive the discriminator to recognize the generator samples as *real*. The process is performed by minimizing the adversarial loss³ (L_{adv}) as follows:

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

where z denotes the input white noise. Note that the auxiliary feature for G is omitted for brevity.

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

$$L_D(G, D) = \mathbb{E}_{\mathbf{x} \sim p_{data}} [(1 - D(\mathbf{x}))^2] + \mathbb{E}_{z \sim N(0, I)} [D(G(z))^2], \quad (2)$$

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

3.2. Multi-resolution STFT auxiliary loss

To improve the stability and efficiency of the adversarial training process, we propose a multi-resolution STFT auxiliary loss. Fig. 1 shows our framework combining the multi-resolution STFT loss with the adversarial training method as described in section 3.1.

Similar to the previous work [11], we define a single STFT loss as follows:

$$L_s(G) = \mathbb{E}_{z \sim p(z), \mathbf{x} \sim p_{data}} [L_{sc}(\mathbf{x}, \hat{\mathbf{x}}) + L_{mag}(\mathbf{x}, \hat{\mathbf{x}})], \quad (3)$$

³ Our method adopts least-squares GANs thanks to its training stability [19–22].

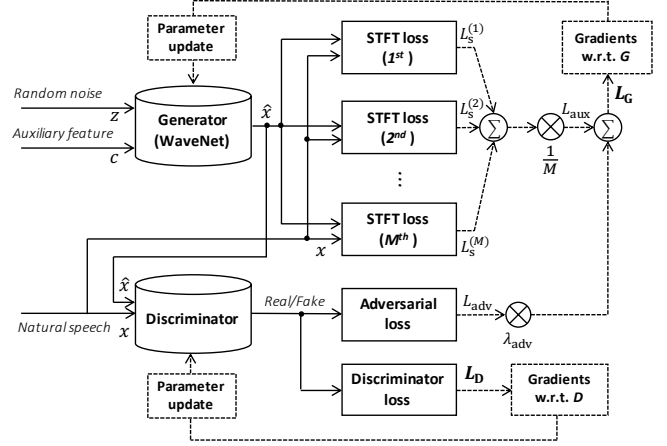


Fig. 1: An illustration of our proposed adversarial training framework with the multi-resolution STFT loss.

where $\hat{\mathbf{x}}$ denotes the generated sample (i.e., $G(z)$), and L_{sc} and L_{mag} denote *spectral convergence* and *log STFT magnitude* loss, respectively, which are defined as follows [23]:

$$L_{sc}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{\| |\text{STFT}(\mathbf{x})| - |\text{STFT}(\hat{\mathbf{x}})| \|_F}{\| |\text{STFT}(\mathbf{x})| \|_F}, \quad (4)$$

$$L_{mag}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{N} \| \log |\text{STFT}(\mathbf{x})| - \log |\text{STFT}(\hat{\mathbf{x}})| \|_1, \quad (5)$$

where $\| \cdot \|_F$ and $\| \cdot \|_1$ denote the Frobenius and L_1 norms, respectively; $|\text{STFT}(\cdot)|$ and N denote the STFT magnitudes and number of elements in the magnitude, respectively.

Our multi-resolution STFT loss is the sum of the STFT losses with different analysis parameters (i.e., FFT size, window size, and frame shift). Let M be the number of STFT losses, the multi-resolution STFT auxiliary loss (L_{aux}) is represented as follows:

$$L_{aux}(G) = \frac{1}{M} \sum_{m=1}^M L_s^{(m)}(G). \quad (6)$$

In the STFT-based time-frequency representation of signals, there is a trade-off between time and frequency resolution; e.g., increasing window size gives higher frequency resolution while reducing temporal resolution [24]. By combining multiple STFT losses with different analysis parameters, it greatly helps the generator to learn the time-frequency characteristics of speech [25]. Moreover, it also prevents the generator from being overfit to a fixed STFT representation, which may result in suboptimal performance in the waveform-domain.

Our final loss function for the generator is defined as a linear combination of the multi-resolution STFT loss and the adversarial loss as follows:

$$L_G(G, D) = L_{aux}(G) + \lambda_{adv} L_{adv}(G, D), \quad (7)$$

where λ_{adv} denotes the hyperparameter balancing the two loss terms. By jointly optimizing the waveform-domain adversarial loss and the multi-resolution STFT loss, the generator can learn the distribution of the realistic speech waveform effectively.

4. EXPERIMENTS

4.1. Experimental setup

4.1.1. Database

In the experiments, we used a phonetically and prosaically balanced speech corpus recorded by a female professional Japanese speaker. The speech signals were sampled at 24 kHz, each sample was quantized by 16 bits. In total, 11,449 utterances (23.09 hours) were used for training, 250 utterances (0.35 hours) were used for validation, and another 250 utterances (0.34 hours) were used for evaluation. The 80-band log-mel spectrograms with band-limited frequency range⁴ (70 to 8000 Hz) were extracted and used as the input auxiliary features for waveform generation models (i.e., local-conditioning [4]). The frame and shift lengths were set to 50 ms and 12.5 ms, respectively. The mel-spectrogram features were normalized to have zero mean and unit variance before training.

4.1.2. Model details

The proposed Parallel WaveGAN consisted of 30 layers of dilated residual convolution blocks with exponentially increasing three dilation cycles [4]. The number of residual and skip channels were set to 64 and the convolution filter size was set to three. The discriminator consisted of ten layers of non-causal dilated 1-D convolutions with leaky ReLU activation function ($\alpha = 0.2$). The strides were set to one and linearly increasing dilations were applied for the 1-D convolutions starting from one to eight except for the first and last layers. 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 [26].

At the training stage, the multi-resolution STFT loss was computed by the sum of three different STFT losses as described in Table 1. The discriminator loss was computed by the average of per-time step scalar predictions with the discriminator. The hyperparameter λ_{adv} in equation (7) was chosen to be 4.0 based on our preliminary experiments. Models were trained for 400 K steps with RAdam optimizer ($\epsilon = 1e^{-6}$) to stabilize training [27]. Note that the discriminator was fixed for the first 100K steps, and two models were jointly trained afterwards. The minibatch size was set to eight 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 for every 200 K steps.

As baseline systems, we used both the autoregressive Gaussian WaveNet and the parallel one (i.e., ClariNet) [10,11]. The WaveNet consisted of 24 layers of dilated residual convolution blocks with four dilation cycles. The number of residual and skip channels were set to 128 and the filter size was set to three. The model was trained for 1.5 M steps with RAdam optimizer. The learning rate was set to 0.001, and it was reduced by half for every 200 K steps. The minibatch size was set to eight and the length of each audio clip was set to 12 K time samples (0.5 second).

To train the baseline ClariNet, the autoregressive WaveNet described above was used as the teacher model. The ClariNet was based on Gaussian IAFs [10], which consisted of six flows. Each flow was parameterized by ten layers of dilated residual convolution blocks with an exponentially increasing dilation cycle. The number of residual and skip channels were set to 64 and the filter

⁴We empirically found that using the band-limited features alleviates the over-smoothing problem caused by acoustic models in TTS.

Table 1: The details of the multi-resolution STFT loss. A Hanning window was applied before the FFT process.

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

size was set to three. The weight coefficients to balance the KLD and STFT auxiliary losses were set to 0.5 and 1.0, respectively. The model was trained for 400 K steps with the same optimizer settings of the Parallel WaveGAN. We also investigated ClariNet with adversarial loss as a hybrid approach of GAN and density distillation [11]. The model structure was the same as the baseline ClariNet, but it was trained with the mixture of KLD, STFT and adversarial losses, where the weight coefficients to balance them were set to 0.05, 1.0 and 4.0, respectively. The model was trained for 200K steps with the fixed discriminator, and the generator and discriminator were jointly optimized for the rest 200 K steps.

Throughout the waveform generation models, the input auxiliary features were upsampled by nearest neighbor upsampling followed by 2-D convolutions so that the time-resolution of auxiliary features matches the sampling rate of the speech waveform [11,28]. Note that the auxiliary features were not used for discriminators. All the models were trained using two NVIDIA Tesla V100 GPUs. Experiments were conducted on the NAVER smart machine learning (NSML) platform [29].

4.2. Evaluation

To evaluate the perceptual quality, we performed mean opinion score (MOS)⁵ tests. Eighteen native Japanese speakers were asked to make quality judgments about the synthesized speech samples using the following five possible responses: 1 = Bad; 2 = Poor; 3 = Fair; 4 = Good; and 5 = Excellent. In total, 20 utterances were randomly selected from the evaluation set and were then synthesized using the different models.

Table 2 shows the inference speed and the MOS test results with respect to different generation models. The findings can be summarized as follows: (1) the systems trained with the STFT loss performed better than ones trained without the STFT loss (i.e., the autoregressive WaveNet). Note that most listeners were unsatisfied with the high-frequency noise caused by the autoregressive WaveNet system. This could be explained by the fact that only the band-limited (70 - 8000 Hz) mel-spectrogram was used for local-conditioning in the WaveNet, while the other systems were able to directly learn full-band frequency information via STFT loss. (2) The proposed multi-resolution STFT loss-based models showed higher perceptual quality than the conventional single STFT loss-based ones (comparing System 3 and 6 with System 2 and 5, respectively). This confirms that the multi-resolution STFT loss effectively captured the time-frequency characteristics of the speech signal, enabling to achieve better performance. (3) The proposed adversarial loss did not work well with the ClariNet. However, its advantage could be found when it was combined with a TTS framework, which will be discussed in the next section. (4) Finally, the proposed Parallel WaveGAN achieved 4.06 MOS. Although its perceptual quality was relatively worse than the ClariNet's, the

⁵Audio samples are available at the following URL:

<https://r9y9.github.io/demos/projects/icassp2020/>

Table 2: The inference speed and the MOS results with 95% confidence intervals: Acoustic features extracted from the recorded speech signal were used to compose the input auxiliary features. The evaluation was conducted on a server with a single NVIDIA Tesla V100 GPU. Note that the inference speed k means that the system was able to generate waveforms k times faster than real-time.

System index	Model	KLD-based distillation	STFT loss	Adversarial loss	Number of layers	Model size	Inference speed	MOS
System 1	WaveNet	-	-	-	24	3.81 M	0.32×10^{-2}	3.61 ± 0.12
System 2	ClariNet	Yes	$L_s^{(1)}$	-	60	2.78 M	14.62	3.88 ± 0.11
System 3	ClariNet	Yes	$L_s^{(1)} + L_s^{(2)} + L_s^{(3)}$	-	60	2.78 M	14.62	4.21 ± 0.09
System 4	ClariNet	Yes	$L_s^{(1)} + L_s^{(2)} + L_s^{(3)}$	Yes	60	2.78 M	14.62	4.21 ± 0.09
System 5	Parallel WaveGAN	-	$L_s^{(1)}$	Yes	30	1.44 M	28.68	1.36 ± 0.07
System 6	Parallel WaveGAN	-	$L_s^{(1)} + L_s^{(2)} + L_s^{(3)}$	Yes	30	1.44 M	28.68	4.06 ± 0.10
System 7	Recording	-	-	-	-	-	-	4.46 ± 0.08

Table 3: Training time comparison: All the experiments were conducted on a server with two NVIDIA Tesla V100 GPUs. Each vocoder model corresponds to System 1, 3, 4, and 6 described in Table 2, respectively. Note that the times for ClariNets include the training time for the teacher WaveNet.

Model	Training time (days)
WaveNet	7.4
ClariNet	12.7
ClariNet-GAN	13.5
Parallel WaveGAN (ours)	2.8

Parallel WaveGAN was able to generate speech signal 1.96 times faster than the ClariNet. Furthermore, the benefit of the proposed method could be found in its simple training procedure. We measured the total training time for obtaining the optimal models, as described in Table 3. Because the Parallel WaveGAN did not require any complicated density distillation, it only took 2.8 training days to be optimized, which was 2.64 and 4.82 times faster than the autoregressive WaveNet and the ClariNet, respectively.

4.3. Text-to-speech

To verify the effectiveness of the proposed method as the vocoder of the TTS framework, we combined the Parallel WaveGAN with the Transformer-based parameter estimator [15–17].

To train the Transformer, we used the phoneme sequences as input and mel-spectrograms extracted from the recorded speech as output. The model consisted of a six-layer encoder and a six-layer decoder, each was based on multi-head attention (with eight heads). The configuration followed the prior work [17], but the model was modified to accept accent as an external input for pitch accent language (e.g., Japanese) [30]. The model was trained for 1000 epochs using RAdam optimizer with warmup learning rate scheduling [15]. Initial learning rate was set to 1.0 and dynamic batch size (average 64) strategy was used to stabilize training.

In the synthesis step, the input phoneme and accent sequences were converted to the corresponding mel-spectrograms by the Transformer TTS model. By inputting resulting acoustic parameters, vocoder models generated the time-domain speech signal.

To evaluate the quality of the generated speech samples, we performed MOS tests. The test setups were the same as those described in section 4.2, but we used the autoregressive WaveNet and the parallel generation models trained with the multi-resolution STFT loss in the test (System 1, 3, 4, and 6 described in Table 2, respectively). The results of the MOS tests are shown in Table 4, of which findings can be summarized as follows: (1) the ClariNet

Table 4: MOS results with 95% confidence intervals: Acoustic features generated from the Transformer TTS model were used to compose the input auxiliary features.

Model	MOS
Transformer + WaveNet	3.33 ± 0.11
Transformer + ClariNet	4.00 ± 0.10
Transformer + ClariNet-GAN	4.14 ± 0.10
Transformer + Parallel WaveGAN (ours)	4.16 ± 0.09
Recording	4.46 ± 0.08

trained with the adversarial loss performed better than the system trained without the adversarial loss, although their perceptual qualities were almost same in the analysis/synthesis case (System 3 and 4 shown in Table 2). This implies that the use of adversarial loss was advantageous for improving the model’s robustness to the prediction errors caused by the acoustic model. (2) The merits of the adversarial training were also beneficial to the proposed Parallel WaveGAN system. Consequently, the Parallel WaveGAN with the Transformer TTS model achieved 4.16 MOS, which was comparable to the best distillation-based Parallel WaveNet system (i.e., ClariNet-GAN).

5. CONCLUSION

We proposed Parallel WaveGAN, a distillation-free, fast, and small-footprint waveform generation method based on GAN. By jointly optimizing waveform-domain adversarial loss and multi-resolution STFT loss, our model was able to learn how to generate realistic waveforms without any complicated probability density distillation. Experimental results demonstrated that our proposed method achieved 4.16 MOS within the Transformer-based TTS framework competitive to the conventional distillation-based approaches, generating 24 kHz speech waveform 28.68 times faster than real-time with only 1.44 M model parameters. Future research includes improving the multi-resolution STFT auxiliary loss to better capture the characteristics of speech (e.g., introducing phase-related loss), and verifying its performance to a variety of corpora, including expressive ones.

6. ACKNOWLEDGEMENTS

The work was supported by Clova Voice, NAVER Corp., Seongnam, Korea. The authors would like to thank Adrian Kim, Jung-Woo Ha, Muhammad Ferjad Naeem, and Xiaodong Gu at NAVER Corp., Seongnam, Korea, for their support.

7. 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] 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.
- [3] 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.
- [4] 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.
- [5] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. 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] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, “Speaker-dependent WaveNet vocoder,” in *Proc. INTERSPEECH*, 2017, pp. 1118–1122.
- [7] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, “An investigation of multi-speaker training for WaveNet vocoder,” in *Proc. ASRU*, 2017, pp. 712–718.
- [8] 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.
- [9] 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.
- [10] W. Ping, K. Peng, and J. Chen, “ClariNet: Parallel wave generation in end-to-end text-to-speech,” in *Proc. ICLR*, 2019.
- [11] 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.
- [12] 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.
- [13] C. Donahue, J. McAuley, and M. Puckette, “Adversarial audio synthesis,” in *Proc. ICLR*, 2019.
- [14] 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.
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. NIPS*, 2017, pp. 5998–6008.
- [16] 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.
- [17] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, “FastSpeech: Fast, robust and controllable text to speech,” *arXiv preprint arXiv:1905.09263*, 2019.
- [18] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku, “GELP: GAN-excited linear prediction for speech synthesis from mel-spectrogram,” in *Proc. INTERSPEECH*, 2019, pp. 694–698.
- [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. Serr, “SEGAN: Speech enhancement generative adversarial network,” in *Proc. INTERSPEECH*, 2017, pp. 3642–3646.
- [23] S. Ö. Arık, H. Jun, and G. Diamos, “Fast spectrogram inversion using multi-head convolutional neural networks,” *IEEE Signal Process. Letters*, vol. 26, no. 1, pp. 94–98, 2019.
- [24] I. Daubechies, “The wavelet transform, time-frequency localization and signal analysis,” *IEEE trans. on information theory*, vol. 36, no. 5, pp. 961–1005, 1990.
- [25] X. Wang, S. Takaki, and J. Yamagishi, “Neural source-filter-based waveform model for statistical parametric speech synthesis,” in *Proc. ICASSP*, 2019, pp. 5916–5920.
- [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] 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.
- [28] A. Odena, V. Dumoulin, and C. Olah, “Deconvolution and checkerboard artifacts,” *Distill*, 2016. [Online]. Available: <http://distill.pub/2016/deconv-checkerboard>
- [29] H. Kim, M. Kim, D. Seo, J. Kim, H. Park, S. Park, H. Jo, K. Kim, Y. Yang, Y. Kim *et al.*, “NSML: Meet the mlaas platform with a real-world case study,” *arXiv preprint arXiv:1810.09957*, 2018.
- [30] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, “Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language,” in *Proc. ICASSP*, 2019, pp. 6905–6909.