A Multi-Scale Time-Frequency Spectrogram Discriminator for GAN-based Non-Autoregressive TTS

Haohan Guo, Hui Lu, Xixin Wu, Helen Meng

The Chinese University of Hong Kong, Hong Kong SAR, China

{hguo,luhui,wuxx,hmmeng}@se.cuhk.edu.hk

Abstract

The generative adversarial network (GAN) has shown its outstanding capability in improving Non-Autoregressive TTS (NAR-TTS) by adversarially training it with an extra model that discriminates between the real and the generated speech. To maximize the benefits of GAN, it is crucial to find a powerful discriminator that can capture rich distinguishable information. In this paper, we propose a multi-scale time-frequency spectrogram discriminator to help NAR-TTS generate high-fidelity Mel-spectrograms. It treats the spectrogram as a 2D image to exploit the correlation among different components in the time-frequency domain. And a U-Net-based model structure is employed to discriminate at different scales to capture both coarse-grained and fine-grained information. We conduct subjective tests to evaluate the proposed approach. Both multi-scale and time-frequency discriminating bring significant improvement in the naturalness and fidelity. When combining the neural vocoder, it is shown more effective and concise than fine-tuning the vocoder. Finally, we visualize the discriminating maps to compare their difference to verify the effectiveness of multi-scale discriminating.

Index Terms: Non-Autoregressive TTS, Speech Synthesis, Mel-Spectrogram, GAN, End-to-End Model

1. Introduction

Neural Text-to-speech (TTS) technology has achieved significant improvement with the introduction of the auto-regressive model [1, 2, 3]. As a sequential generative model, it effectively enhances TTS in naturalness and fidelity. However, due to recursive generation and "exposure bias" [4], inference speed and stability are also affected seriously. To solve this problem, the non-autoregressive TTS (NAR-TTS) has been attracted increasing attention for better stability and parallelizability, such as [5, 6, 7, 8]. It directly up-samples the encoded text features to the frame-level sequence with the explicit duration information, then decodes them to the acoustic features in parallel. But the removal of the auto-regressive mechanism also degrades generative capability, and hence affects the naturalness and fidelity. To address this problem, other generative models have been introduced, such as glow [9], VAE [10], diffusion model [11], etc... Unfortunately, due to their special model design, these approaches have the reduced flexibility, and cannot be easily adapted to arbitrary NAR-TTS models.

To address the aforementioned problem, the generative adversarial network (GAN) shows great potential, which has been widely applied in speech synthesis, including statistical speech synthesis [12, 13], spectrogram post-filter [14], spectrogram super-resolution [15], and neural vocoder [16, 17, 18]. In this framework, NAR-TTS can be enhanced by only using a discriminator. As shown in Fig.1, in training, the discriminator is introduced to distinguish between the generated speech and authentic speech, and applied to the adversarial training to narrow the gap between these two domains. So, to maximize the benefits derived from GAN, a well-designed discriminator capturing rich distinguishable information in training is critical.

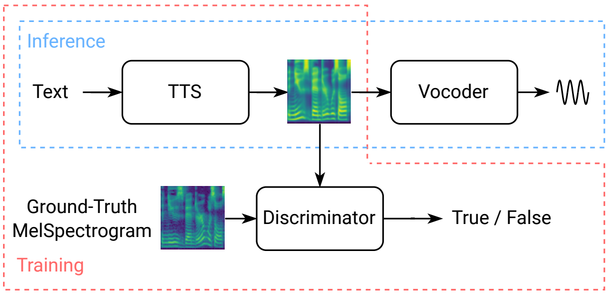


Figure 1: *The framework of GAN-based TTS. The waveform denotes the generated audio. The dotted boxes indicate the modules and operations used in training or inference.*

GAN has been widely used in neural vocoders, which adopts various structures to process the input waveform [16, 17]. However, for NAR-TTS, the current designs of the discriminator are still preliminary, which cannot effectively incorporate the characteristics of the spectrogram. The spectrogram is a 2-D image in the time-frequency space. There is strong relationship among different elements in this image, which affects the spectral shape, harmonics, pronunciation, prosody, and timbre, locally and globally in the time-frequency space. Hence, to generate a realistic spectrogram, it is necessary to ensure that the discriminator can effectively capture this multi-scale relationship in the time-frequency domain. However, different from the waveform that is down-sampled or converted to spectrograms with various STFT parameters for multi-scale discriminating, this 2-D spectrogram is hard to be processed in the same way. To achieve this goal, in this paper, we propose a multi-scale time-frequency spectrogram discriminator. A U-Net-based structure [19, 20] is adopted to discriminate at both fine-grained and coarse-grained levels. Meanwhile, we treat the spectrogram as a 2-D image along the time and frequency axis to better exploit its information in this space.

This paper is organized as follows: We will first introduce our proposed discriminator, including the model structure and training algorithm for NAR-TTS. Then we will present the experiments based on ParallelTacotron, including its model structure and the training setup. The results of the preference tests validate that both multi-scale and time-frequency discriminating improve the training quality to NAR-TTS. When applied to the TTS system using the neural vocoder, GAN is shown as a more effective and concise approach than fine-tuning the vocoder. Finally, we visualize the output maps of the discrimi-

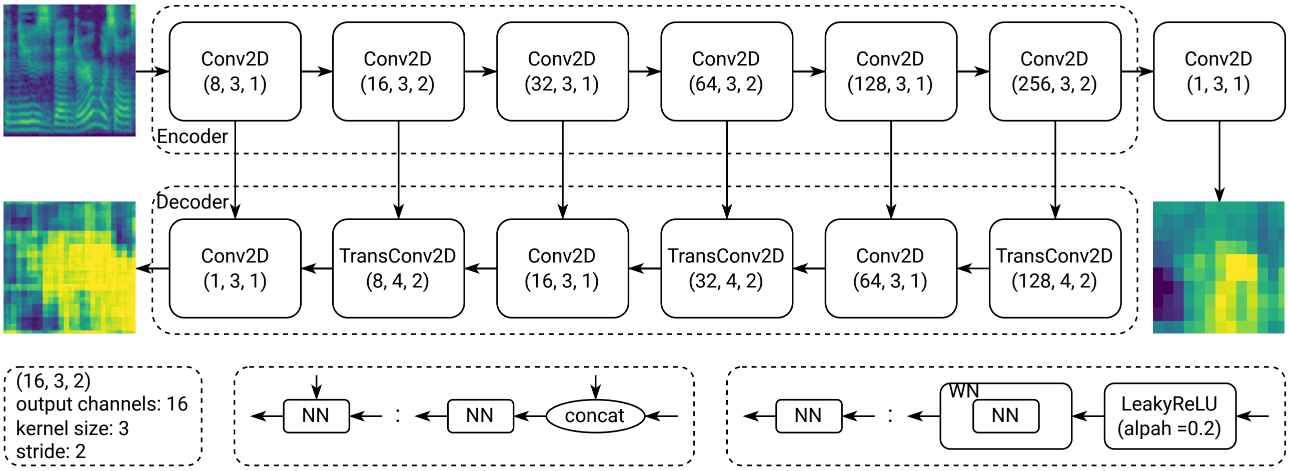


Figure 2: *The architecture of the multi-scale time-frequency discriminator. The upper-left spectrogram denotes the input ground-truth or predicted Mel spectrogram. The two below heatmaps denote the fine-grained (left) and coarse-grained (right) discriminator outputs. Dotted boxes provide a detailed explanation of parameters or operations for the encoder and decoder. "concat" denote the concatenation operation. "WN" and "NN" denote the weight normalization layer and an arbitrary neural network layer.*

nator to analyze the difference between the coarse-grained and fine-grained discriminating outputs.

2. Multi-Scale Time-Frequency Spectrogram Discriminator

In this section, We will first illustrate the U-Net based model structure of the proposed discriminator, and then introduce the corresponding training algorithm to NAR-TTS.

2.1. Model Architecture

The model is illustrated in Fig.2, which has a U-Net based encoder-decoder structure. Firstly, an encoder with a stack of convolutional layers is employed to down-sample the input spectrogram with the shape $(1, T, N)$, i.e. 1 channel, T frames, N frequency bins, into the feature map with shape the $(256, T/8, N/8)$. Then we use a convolutional output layer to compute the coarse-grained discriminator output map. The decoder has a structure symmetrical to the encoder, where the strided convolution is replaced with the transposed convolution. For each layer, it concatenates the output of the previous layer and the resolution-matched hidden feature map in the encoder as the input. In this way, the local high-resolution feature can be better extracted with the guide of the low-resolution information given from the encoder. After up-sampling in the decoder, a fine-grained discriminating output map with the same resolution as the input spectrogram can be achieved. Here, each convolutional layer is wrapped by a weight normalization layer [21], which helps stabilize adversarial training. The LeakyReLU with $\alpha = 0.2$ is set as the activation function for all layers except for the input layer.

Most GAN-based vocoders are also trained with multi-scale discriminators by pre-processing the waveform into waveforms with different sample rates [16] or spectrograms with different STFT parameters [18]. However, it is difficult to do so in the training of the acoustic model, since the Mel spectrogram is hard to be down-sampled well or converted to other features with different scales. The introduction of the U-Net based spectrogram discriminator makes it possible to discriminate one sequence at multiple scales directly.

2.2. Training Algorithm

In training, we first input the text and the ground-truth durations to the TTS model to generate a fake spectrogram S_f for the discriminator. The target spectrogram of the text is set as the real input S_r. They are fed to the discriminator respectively to get the discriminating results and all hidden vectors as follows:

$$S_f = TTS(text) \tag{1}$$

$$C_r, F_r, H_r = Discriminator(S_r) \tag{2}$$

$$C_f, F_f, H_f = Discriminator(S_f) \tag{3}$$

where C_r, F_r, H_r and C_f, F_f, H_f denote the coarse-grained output, fine-grained output, and hidden vectors of the real and fake spectrogram, respectively.

Then we update the discriminator with the LS-GAN loss function L_D:

$$\begin{aligned} L_d = MSE(1, C_r) + MSE(1, F_r) \\ + MSE(0, C_f) + MSE(0, F_f) \end{aligned} \tag{4}$$

Before updating the TTS model, we use the updated discriminator to extract those features again, and then calculate losses as follows:

$$L_a = MSE(1, C_f) + MSE(1, F_f) \tag{5}$$

$$L_f = MAE(H_f, H_r) \tag{6}$$

$$L_g = L_{tts} + \lambda_a L_a + \lambda_f L_f \tag{7}$$

Adversarial loss L_a is used to fool the discriminator by making C_f and F_f close to 1. Feature matching loss L_f is an effective loss function to improve stablity and quality of adversrial training [16, 22]. It calculates MAE loss for N pairs of the hidden vectors of H_f and H_r, then averages them. L_{tts} is the loss function for NAR-TTS, e.g. a MSE loss between the predicted and the target spectrogram. Finally, we get L_g by combining these losses with two weight parameters λ_a and λ_f.

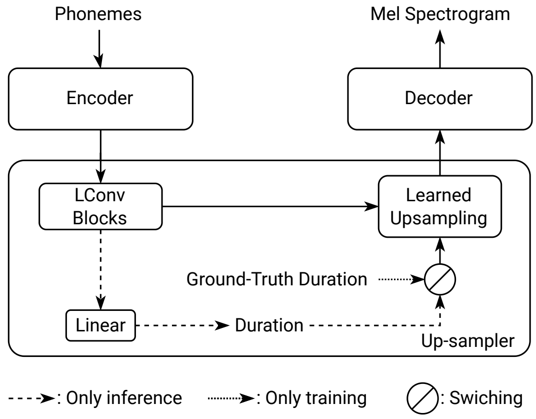


Figure 3: *The structure of the simplified Parallel-Tacotron2. The phoneme sequence is processed by the encoder firstly, then further processed by LConv Blocks in the up-sampler, and up-sampled with the ground-truth or predicted durations according to the running mode (training or inference), finally decoded to the Mel spectrogram.*

3. Experimental Protocol

Our experiments are all conducted on a standard single-speaker English speech dataset, LJSpeech, with over 10 hours of recordings. After screening and pre-processing, we collect 11000 pairs of (text, 16kHz audio) as the training set.

3.1. Non-Autoregressive TTS

As shown in Fig.3, the model is implemented based on ParallelTacotron2 [8], but removes speaker embedding and residual encoder for simplification. In this model, the learned upsampling module can up-sample the input, a phoneme sequence with punctuations, to the frame-level features according to explicitly predicted phoneme-level durations. Then we use the decoder to generate the 80-dim log-scale Mel-spectrogram with 12.5ms frameshift and 50ms frame length.

Notably, we avoid using Soft-DTW loss in our experiments due to its huge cost on computing resources. Instead, in the training stage, we use ground-truth duration[1] as input and add an extra loss function between the predicted and target duration. It is also an effective approach with less computing cost for duration learning. To reconstruct waveform from Mel-spectrogram, Griffin-Lim [23] and Hifi-GAN [17] trained on the same dataset are used in the tests[2].

3.2. Training Setup

We use MSE and MAE loss functions for the iterative loss of Mel-spectrogram and the duration loss as follows:

$$L_{spec} = MSE(S, \hat{S}) + MAE(S, \hat{S}) \tag{8}$$

$$L_{dur} = MSE(D, \hat{D}) + MAE(D, \hat{D}) \tag{9}$$

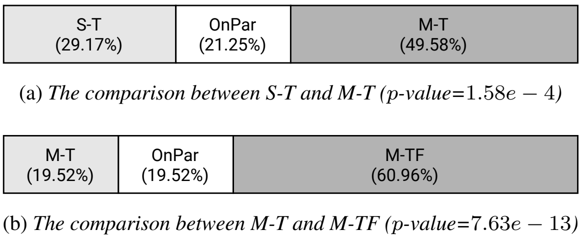
$$L_{tts} = L_{spec} + \lambda_{dur} L_{dur} \tag{10}$$

where L_{spec} denotes the loss function between the target spectrogram S and the predicted spectrograms $\hat{S}$, L_{dur} denotes the loss function between the target duration D and the predicted duration $\hat{D}$. λ_{dur} is set to balance these two losses, which is 0.02 in our experiments. The output Mel-spectrogram of the last layer in the decoder is involved in the adversarial training. λ_a and λ_f are set to 0.2 and 2 for all experiments with U-Net based discriminators, otherwise are 1 and 10.

RAdam [24] with ($\beta_1 = 0.9, \beta_2 = 0.999$) and Lookahead [25] with ($k = 5, \alpha = 0.5$) are combined as the optimizer to provide more stable training process. The learning rate is exponentially decayed from $1e^{-3}$ to $1e^{-5}$ after 20,000 iterations. All models are trained for 200,000 iterations with a batch size of 64.

4. Results

We conduct subjective tests using Amazon Mechanical Turk. 80 utterances in the dataset which are disjoint from training set are used as the test set. For each test, each listener can only rate 30 sets of utterances to ensure good test quality. [3]



(a) *The comparison between S-T and M-T (p-value=*$1.58e-4$*)*

(b) *The comparison between M-T and M-TF (p-value=*$7.63e-13$*)*

Figure 4: *The preference test for different discriminators*

4.1. Discriminators

The preference tests are conducted to validate that multi-scale and time-frequency discrimination are both effective for the adversarial training. Three discriminators are involved in the comparison:

1. S-T: Single-Scale Time Discriminator, which only uses the encoder part and 1-D convolutions along the time axis. It has been validated effective in [22].
2. M-T: Multi-Scale Time Discriminator, which is based on S-T, and uses both encoder and decoder.
3. M-TF: Multi-Scale Time-Frequency Discriminator.

We first compare S-T and M-T, then compare M-T and M-TF. Here, all samples are generated using Griffin-Lim [23] to avoid the bias brought from data-driven neural vocoder.

In the comparison between S-T and M-T shown in Fig.4a, M-T achieves significant preference with the voting rate of 49.58%. We find that they generate similar timbre and rhythm overall, but multi-scale discrimination makes the formant smoother and clearer, and produces better prosody with higher naturalness and diversity, hence receives higher listener preference. The comparison between M-T and M-TF in Fig.4b validates that the time-frequency operation is an effective approach by the higher voting rate of 61.11%. It obviously improves the fidelity, including the spectral clarity, smoothness, and continuity, which can be easily noticed by listeners. It shows that operating the spectrogram as a 2-D image along both time and frequency axes can better exploit spectral information.

[1] We get the duration using MFA at `https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner`

[2] The code of HifiGAN is available at `https://github.com/jik876/hifi-gan`

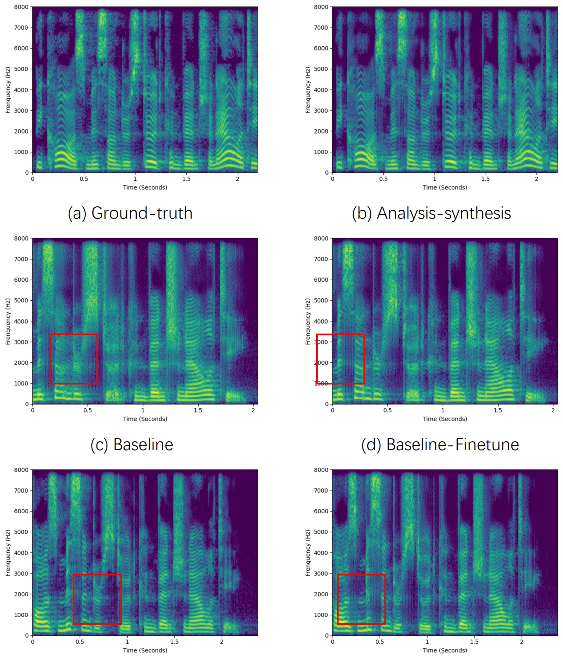
[3] Samples are available at `https://hhguo.github.io/DemoUGANTTS`

Table 1: *The MOS test results (✓ and × denote the corresponding approach is used or not. ± indicates 95% CI)*

GAN	Finetune	MOS
×	×	2.91 ± 0.14
×	✓	3.43 ± 0.15
✓	×	$\mathbf{3.81 \pm 0.15}$
✓	✓	3.71 ± 0.14
Analysis-Synthesis		3.97 ± 0.17

4.2. GAN or Fine-tuning Vocoder? Or Both?

In mainstream TTS systems, the neural vocoder is usually adopted for waveform generation due to its high-quality generation. It is trained with ground-truth spectrograms but is used based on the predicted ones. The gap between them causes errors in vocoding, such as noise and distortion. To narrow the gap, fine-tuning the vocoder with TTS predictions is often used, but also leads to more costs on computing and storage resources, and more complicated TTS training and update. Instead, it is more concise to directly improve the fidelity of the generated spectrogram, e.g. GAN training. In this paper, both our approach and fine-tuning vocoder are evaluated in this aspect via an MOS test.



(a) Ground-truth

(b) Analysis-synthesis

(c) Baseline

(d) Baseline-Finetune

(e) GAN

(f) GAN-Finetune

Figure 5: *The magnitude spectrograms of the ground-truth audio (a), the analysis-synthesis audio (b) and audios generated by baseline models w/ (c) or w/o (d) fine-tuned vocoder (d), GAN-based models w/ (e) or w/o (f) fine-tuned vocoder.*

he MOS test result and spectrograms generated by these models are shown in Table.1 and Fig.5, respectively. Firstly, the high-fidelity reconstruction of analysis-synthesis in Fig.5(b) shows that the vocoder is trained well. The baseline system, without GAN training and fine-tuned vocoder, receives the worst score of 2.91. The fuzzy spectrogram with unsmoothed low-frequency harmonics leads to serious degradation in naturalness and fidelity, which is shown in Fig.5(c). After fine-tuning the vocoder, the harmonics in the middle and low-frequency parts are enhanced significantly, hence improving its output quality with a much higher MOS of 3.43.

For the GAN-based system, it already achieves a much higher score of 3.81 without fine-tuning. The spectrogram shown in Fig.5(e) presents both clearer, smoother harmonics and richer details in high frequency. Meanwhile, the system using both GAN and fine-tuning obtains a worse score of 3.71. The harmonics in middle frequency are slightly fuzzier, hence degrading the fidelity. Since the vocoder is fine-tuned to map the predicted spectrogram to the target waveform, the larger mismatch between them makes the fine-tuning more challenging. GAN training increases the variety of the generated spectrogram, but also enlarges this mismatch, and degrades the fine-tuning quality. In conclusion, GAN can bring benefits to synthesis, while finetuning may not bring consistent.

4.3. Discriminating Visualization

As shown in Fig.6, we present a Mel-spectrogram (a) and its coarse-grained (b) and fine-grained (c) discriminating outputs. The high-lightness area indicates that it has a higher probability classified as the real one. The up-sampled coarse-grained output map in (b) shows a smooth and averaged heatmap, which provides coarse-grained, global discriminating information. In comparison, (c) shows a sharper heatmap that has higher resolution and more attention on local parts. This fine-grained discriminating enhances fidelity-related information like formants.

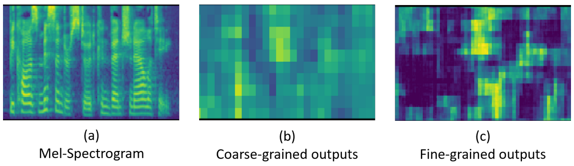


Figure 6: *The visualization of the multi-scale time-frequency discriminator outputs.*

5. Conclusion

This paper proposes a multi-scale time-frequency spectrogram discriminator to provide better GAN training for Non-Autoregressive TTS. It operates the Mel-spectrogram in time-frequency domain at different scales to exploit richer information for better discrimination. Preference tests validate the effectiveness of multi-scale and time-frequency discriminating. An MOS test is conducted to investigate the impact of GAN and vocoder fine-tuning on NAR-TTS. The results show that GAN training significantly improves TTS with the higher MOS of 3.81, but fine-tuning may cause negative effects. In addition, coarse-grained and fine-grained discriminator output maps are visualized to investigate their differences, and verify that they can provide richer discriminative information at different scales.

6. Acknowledgments

This work is supported by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Fund.

7. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017.

[2] 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*. IEEE, 2018, pp. 4779–4783.

[3] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. AAAI*, vol. 33, no. 01, 2019, pp. 6706–6713.

[4] H. Guo, F. K. Soong, L. He, and L. Xie, "A new gan-based end-to-end tts training algorithm," in *Proc. Interspeech*, 2019.

[5] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Proc. NeurIPS*, 2019, pp. 3165–3174.

[6] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2021.

[7] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. J. Weiss, and Y. Wu, "Parallel tacotron: Non-autoregressive and controllable tts," in *Proc. ICASSP*. IEEE, 2021, pp. 5709–5713.

[8] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. Skerry-Ryan, and Y. Wu, "Parallel Tacotron 2: A Non-Autoregressive Neural TTS Model with Differentiable Duration Modeling," in *Proc. Interspeech*, 2021, pp. 141–145.

[9] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," in *Proc. NeurIPS*, 2020.

[10] H. Lu, Z. Wu, X. Wu, X. Li, S. Kang, X. Liu, and H. Meng, "VAENAR-TTS: Variational Auto-Encoder Based Non-AutoRegressive Text-to-Speech Synthesis," in *Proc. Interspeech*, 2021, pp. 3775–3779.

[11] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-TTS: A Denoising Diffusion Model for Text-to-Speech," in *Proc. Interspeech*, 2021, pp. 3605–3609.

[12] S. Yang, L. Xie, X. Chen, X. Lou, X. Zhu, D. Huang, and H. Li, "Statistical parametric speech synthesis using generative adversarial networks under a multi-task learning framework," in *Proc. ASRU*, 2017, pp. 685–691.

[13] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, 2017.

[14] T. Kaneko, S. Takaki, H. Kameoka, and J. Yamagishi, "Generative adversarial network-based postfilter for stft spectrograms," in *Proc. Interspeech*, August 2017.

[15] L. Sheng, D. Huang, and E. N. Pavlovskiy, "High-quality speech synthesis using super-resolution mel-spectrogram," *CoRR*, vol. abs/1912.01167, 2019.

[16] 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. 14 910–14 921.

[17] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020.

[18] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation," in *Proc. Interspeech*, 2021, pp. 2207–2211.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[20] E. Schonfeld, B. Schiele, and A. Khoreva, "A u-net based discriminator for generative adversarial networks," in *Proc. CVPR*, 2020, pp. 8207–8216.

[21] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Proc. NeurIPS*, 2016, pp. 901–909.

[22] J. Yang, J.-S. Bae, T. Bak, Y.-I. Kim, and H.-Y. Cho, "GANSpeech: Adversarial Training for High-Fidelity Multi-Speaker Speech Synthesis," in *Proc. Interspeech*, 2021, pp. 2202–2206.

[23] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 236–243, 1984.

[24] 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*, 2019.

[25] M. R. Zhang, J. Lucas, J. Ba, and G. E. Hinton, "Lookahead optimizer: k steps forward, 1 step back," in *Proc. NeurIPS*, 2019, pp. 9593–9604.