INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Harmonic WaveGAN: GAN-Based Speech Waveform Generation Model with Harmonic Structure Discriminator

Kazuki Mizuta[1], *Tomoki Koriyama*[2], *Hiroshi Saruwatari*[2]

[1]Faculty of Engineering, The University of Tokyo, Japan
[2]Graduate School of Information Science and Technology, The University of Tokyo, Japan
hartmann76514@g.ecc.u-tokyo.ac.jp, t.koriyama@ieee.org

Abstract

This paper proposes Harmonic WaveGAN, a GAN-based waveform generation model that focuses on the harmonic structure of a speech waveform. Our proposed model uses two discriminators to capture characteristics of a speech waveform in a time domain and in a frequency domain, respectively. In one of them, a harmonic structure discriminator, a 2-D convolution layer called "harmonic convolution" is inserted to model a harmonic structure of a speech waveform. Although harmonic convolution has been shown to perform well in audio restoration tasks, this convolution layer has not yet been fully explored in the field of speech synthesis. Therefore, we seek to improve the perceptual quality of speech samples synthesized by the waveform generation model and investigate the usefulness of harmonic convolution in the field of speech synthesis. Mean opinion score tests showed that the Harmonic WaveGAN can synthesize more natural speech than conventional Parallel WaveGAN. We also showed that a spectrogram of a speech waveform showed a clearer harmonic structure when synthesized by our model than a speech waveform synthesized by the original Parallel WaveGAN.

Index Terms: text-to-speech, speech synthesis, generative adversarial network, neural vocoder, harmonic structure

1. Introduction

Generating natural speech waveforms is important for text-to-speech (TTS) synthesis. Waveform generation models that use deep neural networks (DNNs)—often called "neural vocoders"—have been widely studied and found to generate human-like speech. WaveNet [1] is an innovative waveform generation model composed of dilated convolution layers and autoregressive prediction. Although WaveNet can generate high-quality synthetic speech, the sample-by-sample autoregressive model causes very slow inference. To overcome the longer computation time of WaveNet, various waveform generation models have been proposed, such as distillation-based models [2], flow-based models [3], diffusion models [4, 5], and generative adversarial network (GAN)-based models [6, 7].

Parallel WaveGAN (PWG) [6] is a GAN-based model in which a waveform generator and a waveform discriminator are trained simultaneously with an adversarial strategy. The PWG achieves fast inference because the generator is a non-autoregressive model; thus, the inference can be computed in parallel. Furthermore, the PWG can reduce training time compared with other waveform generation models [6]. However, a gap in quality between generated and natural speech samples remains, and the generated speech signals often have artificial traces at a high frequency. In the GAN-based waveform generation models, discriminators play an important role in achieving naturalness. Hence, methods of improving discriminator architecture are widely studied to find ways to capture the characteristics of a speech signal and enhance the speech quality generated by GAN-based waveform generation models. MB-MelGAN [8] used multiple discriminators with different frequency bands. Hi-Fi GAN by Kong et al. [9] introduced reshaping speech signals into the discriminator to obtain periodic features. TFGAN [10] added the discriminator to the frequency domain.

In this study, we consider the function of the discriminator from a different perspective, so that the waveform generation model can more directly capture the characteristics of a speech waveform. Specifically, we focus on harmonic structure, which is important for human perception [11]. Regarding the models for capturing harmonic structure, harmonic convolution [12] is an operation in which convolution kernels are defined on the basis of a harmonic series instead of adjacent frequencies. Harmonic convolution has been proven effective in tasks such as audio restoration and sound separation.

In this paper, we propose Harmonic WaveGAN (HWG), a novel GAN-based waveform generation model that focuses on the harmonic structure of a speech waveform. To be specific, the HWG utilizes a time domain and a harmonic structure discriminator, where a harmonic convolution layer is inserted into the harmonic structure discriminator. The generator of our model is the same as that of the PWG, so the speed of speech synthesis remains fast. Subjective evaluation experiments showed that the perceptual quality of the speech synthesized by the HWG is higher than that of speech synthesized by the conventional PWG. We also determined that the spectrogram of speech synthesized by our model had a more distinct harmonic structure than that of speech synthesized by the original PWG.

2. Parallel WaveGAN

PWG [6] is trained to generate a natural speech waveform with an adversarial strategy by using a non-autoregressive WaveNet-based model conditioned on an auxiliary feature as a generator and using a non-causal dilated convolutional neural network as a discriminator. The generator is a non-autoregressive model, so speech samples can be generated in parallel. In addition to an adversarial loss of GAN, the generator is trained by an auxiliary loss that performs a short-time Fourier transform (STFT) [13] on the synthetic and target speech waveforms and penalizes the difference between them in the frequency domain, which is called "multi-resolution STFT loss." The generator is trained to minimize a generator loss L_G, a combination of the multi-resolution STFT loss L_{STFT} and adversarial loss L_{adv}, which is defined by the following equation:

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

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-583

where λ_{adv} is the hyperparameter balancing two loss terms. The adversarial loss L_{adv} is calculated by the following equation:

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

where $\boldsymbol{z}$ denotes the input noise, $p_{\boldsymbol{z}}$ denotes a Gaussian distribution $N(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{h}$ is the conditional acoustic features.

The discriminator is trained to minimize a discriminator loss L_D, which is defined by the following equation:

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

where $\boldsymbol{x}$ denotes a target speech waveform and p_{data} denotes its distribution. By training the generator with the sum of the adversarial loss and the multi-resolution STFT loss, the model is easily trained with relatively few parameters and improves the quality of the synthesized speech.

3. Harmonic convolution

The frequencies of a speech waveform are harmonically related, and the harmonic structure is confirmed to affect human perception [11]. Harmonic convolution [12] is a convolution layer that can capture the harmonic structure of an acoustic signal by defining the convolution kernel based on the harmonic series rather than the adjacent frequencies. Given the STFT spectrogram $X(\omega, \tau)$, in which ω is a frequency index and τ is a time index, harmonic convolution of $X(\omega, \tau)$ is defined by the following equations:

$$Y_n(\hat{\omega}, \hat{\tau}) = \sum_{k=1}^{K_f} \sum_{\tau=1}^{K_t} X(\frac{k\hat{\omega}}{n}, \hat{\tau} - \tau) K(k, \tau) \tag{4}$$

$$Y(\hat{\omega}, \hat{\tau}) = \sum_{n=1}^{N} \omega_n Y_n(\hat{\omega}, \hat{\tau}), \tag{5}$$

where $K(\cdot, \cdot)$ denotes the kernel of harmonic convolution, K_f and K_t denote the kernel size along the frequency and the time axes, respectively, and n is an anchor for computing the convolution with $\hat{\omega}/n$ as the base frequency. This anchor n makes it possible to capture the harmonic series of integer multiples of the base frequency $\hat{\omega}$ and the harmonic series of 1/n times the base frequency $\hat{\omega}/n$.

Harmonic convolution requires inefficient computation to access discrete data for convolution along a frequency axis, which increases the computation time. Therefore, Takeuchi et al. [14] focused on a method called *lowering* [15, 16], which avoids complex looping to speed up computation by converting a multidimensional array into a matrix. They succeeded in speeding up the computation of harmonic convolution, called "harmonic lowering." Furthermore, it was shown that harmonic lowering can be calculated faster by capturing the frequency domain in logarithm. This method that simplifies the computation of harmonic lowering is called logarithmic harmonic lowering. Logarithmic harmonic lowering of $X(\omega, \tau)$ is defined by the following equations:

$$X'_n(k, \hat{\omega}, \hat{\tau}) = X(\log \frac{k\hat{\omega}}{n}, \hat{\tau}) = X(\log \frac{k}{n} + \log \hat{\omega}, \hat{\tau}) \tag{6}$$

$$K'(k, 1, \tau) = K(k, \tau) \tag{7}$$

$$Y_n(\hat{\omega}, \hat{\tau}) = \sum_{k=1}^{K_f} \sum_{\omega=1}^{1} \sum_{\tau=1}^{K_t} X'(k, \hat{\omega}, \hat{\tau} - \tau) K'(k, \omega, \tau). \tag{8}$$

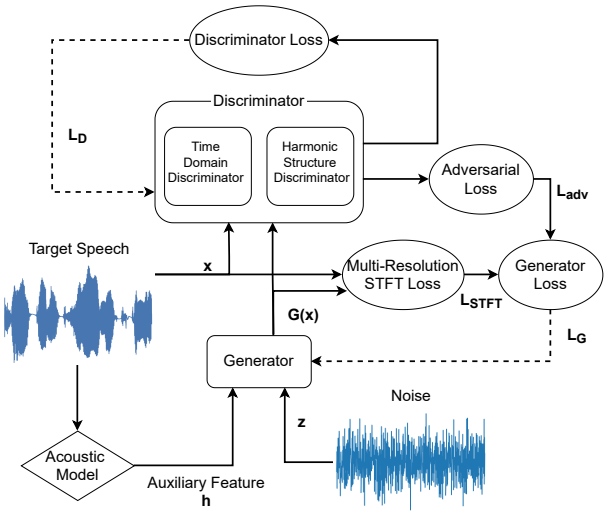

Figure 1: *Architecture of the proposed HWG.*

By transforming X into X'_n, the harmonic structure in a frequency domain can be accessed continuously; thus, the harmonic convolution can be computed efficiently in the same manner as a conventional 2-D convolution.

4. Harmonic WaveGAN

In the human auditory system, the organ called the cochlea is thought to perform a process equivalent to time-frequency analysis [17]. The conventional PWG tried to make the synthesized speech waveform resemble the target one in the time domain. Instead of the time domain resemblance, we assume that the waveform generation model can acquire a representation of speech waveform more efficiently by making the synthesized speech closer to the target speech in the frequency domain. In addition, the harmonic structure of speech affects human perception [11], so a waveform generation model should be able to synthesize more natural speech if it can focus on the harmonic structure. Therefore, we demonstrate that our HWG (based on the PWG) can effectively capture the harmonic structure of a speech waveform. The HWG utilizes not only a harmonic structure discriminator but also a time domain one to learn the harmonic and temporal characteristics of the speech waveform, which stabilizes the training of this model. The architecture of the HWG is shown in Fig. 1, and the details of its discriminators are shown in Fig. 2.

The time domain discriminator is the same as the conventional PWG discriminator, consisting of 1-D dilated convolution layers, and the input is a speech waveform. The harmonic structure discriminator takes the input from a STFT of the speech waveform. A logarithmic harmonic lowering is utilized for the lowest layer of this discriminator, and 2-D convolution layers are used as the subsequent layers. The input complex STFT coefficients are treated as two separate real-valued channels. In [12], harmonic convolution layers are inserted in succession; however, in this paper, only one harmonic convolution layer is used as the bottom layer, and 2-D convolution layers are used as the subsequent layers. It is expected that a single harmonic convolution layer is sufficient to capture the harmonic structure because the harmonic structure exists only in the bottom layer and harmonic convolution requires longer computation time for training than conventional convolution. The subsequent 2-D

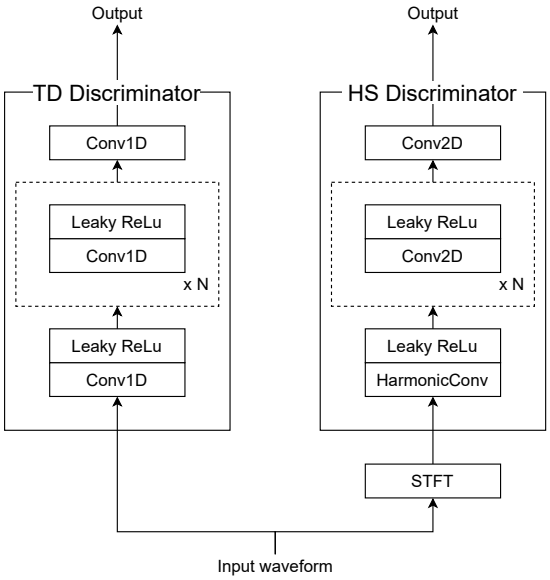

Figure 2: *Details of the HWG discriminators, TD Discriminator (time domain discriminator) and HS Discriminator (harmonic structure discriminator).*

convolution layers are dilated to expand the receptive field in both the time and the frequency directions.

The generator loss is the same as that of the PWG in Eq. (1), and the time domain discriminator loss is the same as the discriminator loss of the PWG in Eq. (3). The harmonic structure discriminator loss L_{HS} is calculated as follows:

$$L_{HS} = \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{data}}}[(1 - D_{HS}(\boldsymbol{x}))^2] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[D_{HS}(G(\boldsymbol{z}, \boldsymbol{h}))^2], \tag{9}$$

where D_{HS} denotes the harmonic structure discriminator and note that the output of this discriminator is in four dimensions: batch size, channel, frequency, and time. The discriminator loss L_D is obtained by adding the time domain and harmonic structure discriminator losses as follows:

$$L_D = (L_{TD} + \lambda_{\mathrm{har}} L_{HS})/2, \tag{10}$$

where L_{TD} denotes time domain discriminator loss and λ_{har} is the hyperparameter balancing two loss terms.

5. Experiments

5.1. Experimental setup

5.1.1. Database

In the experiments, we used the JSUT corpus [18], which contains the speech data of a single female Japanese speaker sampled at 24 kHz, and the JVS corpus [18], which contains the speech data of 100 Japanese speakers sampled at 24 kHz. With the JSUT corpus, 7,191 utterances (9.6 hours) were used for training, and 100 utterances (0.2 hours) were used for evaluation. With the JVS corpus, 10,046 utterances (20.3 hours) were used for training, and 100 utterances (0.2 hours) were used for evaluation.

The 80-band log-mel spectrograms with a band-limited frequency range (70 to 8000 Hz) were extracted and used as the input auxiliary features for the waveform generation models. The frame length was set to 50 ms and the shift length was set to 12.5 ms.

5.1.2. Model details

The baseline PWG [6] used in the experiments was an open-source implementation[1]. The generator of our HWG is the same as the generator of the PWG. It consisted of 30 layers of dilated residual convolution blocks with three exponentially increasing dilation cycles. The number of residual and skip channels were set to 64, and the convolution kernel size was set to three. The time domain discriminator had the same architecture as the PWG discriminator, which consisted of ten layers of non-causal dilated 1-D convolutions with a leaky ReLU activation function ($\alpha = 0.2$). The number of channels were set to 64, the convolution kernel size was set to three, the strides were set to one, and linearly increasing dilations were applied to the 1-D convolutions starting from one to eight except for the first and last layers. The harmonic structure discriminator consisted of a logarithmic harmonic lowering layer and nine layers of non-causal dilated 2-D convolutions with leaky ReLU activation functions ($\alpha = 0.2$). The logarithmic harmonic lowering had a frequency kernel size of seven and a time kernel size of seven. The anchor n in Eq. (4) was set to seven, and the stride was set to one. For the 2-D convolution layers, the number of channels were set to 64, the convolution kernel sizes were set to $(3, 3)$, the strides were set to $(1, 1)$, and linearly increasing dilations were applied to the 2-D convolutions starting from one to eight except for the last layer. All the convolution layers for the generator and the discriminators had weight normalization. The input to the harmonic structure discriminator was real and imaginary values of STFT coefficients with a window length of 1022 and a hop length of 64.

At the training stage, the multi-resolution STFT loss was computed by the sum of three different STFT losses based on the original PWG. The time domain discriminator loss was computed by the average of the per-time-step scalar predictions with the time domain discriminator. The harmonic structure discriminator loss was computed by the average of the per-time-step and per-frequency-bin scalar predictions with the harmonic structure discriminator. The hyperparameter λ_{adv} in Eq. (1) was chosen to be 4.0, and the hyperparameter λ_{har} in Eq. (10) was chosen to be 1.0. The waveform generation models were trained for 400 K steps with the RAdam optimizer ($\epsilon = 10^{-6}$) [19]. The discriminators were fixed for the first 100 K steps, and the generator and discriminators were jointly trained afterwards. The minibatch size was set to six, and the length of each audio clip was set to 24 K time samples. The initial learning rate was set to 0.0001 for the generator and 0.00005 for the discriminators. The learning rate was reduced by half for every 200 K steps.

5.2. Evaluation of analysis-synthesis

A mean opinion score (MOS)[2] test was performed to evaluate the perceptual quality of the generated speech samples. Forty 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. Five utterances were randomly selected from each evaluation set.

[1] https://github.com/kan-bayashi/ParallelWaveGAN

[2] Audio samples are available at the following URL: https://m-kazuki.github.io/HarmonicWaveGAN/

Table 1: *The MOS results with* 95% *confidence intervals: The acoustic features extracted from the recorded speech waveforms were used to compose the input auxiliary features*

Model	JSUT	JVS
Parallel WaveGAN	3.48 ± 0.11	3.11 ± 0.13
Harmonic WaveGAN	3.67 ± 0.11	3.29 ± 0.14
Recording	4.02 ± 0.12	4.30 ± 0.11

Model	JSUT
Harmonic WaveGAN w/o harmonic convolution	3.33 ± 0.12
Harmonic WaveGAN	3.69 ± 0.14
Recording	4.01 ± 0.12

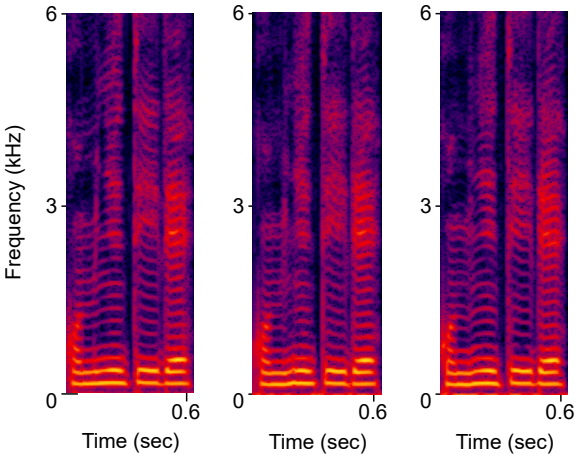

Figure 3: *Spectrograms of (a) target speech, (b) generated speech from the conventional PWG, and (c) generated speech from the proposed HWG.*

The results of the MOS tests for the different corpora are shown in the upper part of Table 1. These results show that the HWG achieved higher scores than the PWG with the JSUT and JVS corpora. The MOS of the JVS corpus was lower than that of the JSUT corpus. This is because the JVS corpus is a multi-speaker corpus, so the waveform generation model needs to acquire a generalized representation of speech samples. Therefore, the high performance of the HWG even in the JVS corpus suggests that our model has improved the waveform generation model by capturing the harmonic structure to obtain a more general representation of the speech waveform.

The lower part of Table 1 shows the results of the MOS test comparing HWG with a model in which the logarithmic harmonic lowering layer of HWG is replaced by a 2-D convolution layer. This result confirms that the high performance of HWG is not due to the use of two types of discriminators, but is largely due to the harmonic convolution.

Figure 3 shows the spectrograms of the target speech included in the JSUT corpus and the speech generated by the conventional PWG and our HWG. In comparing the spectrograms of the speech synthesized by the HWG and PWG, the harmonic structure is clearly visible in the spectrogram of the HWG even in the high frequency range. This is one reason for the high naturalness of the speech produced by our model because the spectrogram of the target speech also shows the clear harmonic structure in the high frequency range.

Table 2: *The MOS results with* 95% *confidence intervals: The acoustic features generated from the transformer TTS model were used to compose the input auxiliary features*

Model	JSUT
Parallel WaveGAN	3.48 ± 0.14
Harmonic WaveGAN	3.71 ± 0.14
Recording	4.16 ± 0.12

5.3. Evaluation of TTS

To verify the effectiveness of our model as a vocoder in the TTS framework, we performed a MOS test using the acoustic features extracted by the acoustic model as input to the waveform generation models. As the acoustic model, we used transformer TTS trained by the JSUT corpus [20, 21]. The test setup was the same as the one described in Section 5.2.

Table 2 shows the MOS test results. The HWG performed as a vocoder better than the PWG in the TTS framework. For the unvoice sounds, both PWG and HWG generated artificial noise, but our model suppressed the magnitude of the noise compared to the baseline model.

6. Conclusions

In this paper, we proposed the HWG (based on the PWG), which can focus on a harmonic structure specific to a speech waveform. In our model, the discriminator is divided into a time domain discriminator and a harmonic structure discriminator to ensure that the waveform generation model captures the speech characteristics in the time and frequency domains. In the harmonic structure discriminator, a harmonic convolution layer is inserted to capture the harmonic structure inherent in the speech waveform and reproduce it in the speech waveform synthesized by the generator. The experimental results showed that the HWG performed better than the original PWG. Therefore, introducing harmonic convolution into the waveform generation model improves the quality of synthesized speech.

The HWG has a larger model and higher training cost than the PWG; so, in future work, we will investigate whether the HWG can be made more compact while maintaining the quality of the synthesized speech. In this study, we used a time domain discriminator to stabilize the training of HWG, and controlled its effect using λ_{har} in Eq. (10). This λ_{har} has not been optimized sufficiently, and possibly the time domain discriminator is not necessary, and these should also be verified in the future. Furthermore, the effectiveness of our model should be tested with corpora of different languages and genders, especially a single male speaker corpus, which was not addressed in this paper. Since the effectiveness of the harmonic structure discriminator was demonstrated in this study, it should be studied whether it can be combined with the discriminator of other GAN-based waveform generation models for further performance improvement.

7. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP-19K20292. The source code for the logarithmic harmonic lowering used in this paper was provided by the author, Hirotoshi Takeuchi [14].

8. 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," *arXiv preprint arXiv:1609.03499*, 2016.

[2] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *Proc. ICLR*. PMLR, 2018, pp. 3918–3926.

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

[4] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A Versatile Diffusion Model for Audio Synthesis," in *Proc. ICLR*, 2021. [Online]. Available: https://openreview.net/forum?id=a-xFK8Ymz5J

[5] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating Gradients for Waveform Generation," in *Proc. ICLR*, 2021. [Online]. Available: https://openreview.net/forum?id=NsMLjcFaO8O

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

[7] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. Courville, "MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis," in *Proc. NIPS*, vol. 32, 2019.

[8] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 492–498.

[9] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Proc. NIPS*, vol. 33, 2020, pp. 17 022–17 033.

[10] Q. Tian, Y. Chen, Z. Zhang, H. Lu, L. Chen, L. Xie, and S. Liu, "TFGAN: Time and Frequency Domain Based Generative Adversarial Network for High-fidelity Speech Synthesis," *arXiv preprint arXiv:2011.12206*, 2020.

[11] S. Popham, D. Boebinger, D. P. Ellis, H. Kawahara, and J. H. McDermott, "Inharmonic speech reveals the role of harmonicity in the cocktail party problem," *Nature communications*, vol. 9, no. 1, pp. 1–13, 2018.

[12] Z. Zhang, Y. Wang, C. Gan, J. Wu, J. B. Tenenbaum, A. Torralba, and W. T. Freeman, "Deep Audio Priors Emerge From Harmonic Convolutional Networks," in *Proc. ICLR*, 2019.

[13] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[14] H. Takeuchi, K. Kashino, Y. Ohishi, and H. Saruwatari, "Harmonic Lowering for Accelerating Harmonic Convolution for Audio Signals," *Proc. Interspeech*, pp. 185–189, 2020.

[15] K. Chellapilla, S. Puri, and P. Simard, "High performance convolutional neural networks for document processing," in *Tenth International Workshop on Frontiers in Handwriting Recognition*. Suvisoft, 2006.

[16] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, "cudnn: Efficient primitives for deep learning," *arXiv preprint arXiv:1410.0759*, 2014.

[17] G. v. Békésy, "On the resonance curve and the decay period at various points on the cochlear partition," *The Journal of the Acoustical Society of America*, vol. 21, no. 3, pp. 245–254, 1949.

[18] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, 2020.

[19] 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.

[20] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[21] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *Proc. ICASSP*. IEEE, 2020, pp. 7654–7658.