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

Fre-GAN: Adversarial Frequency-consistent Audio Synthesis

Ji-Hoon Kim[1], *Sang-Hoon Lee*[2], *Ji-Hyun Lee*[1], *Seong-Whan Lee*[1,2]

[1]Department of Artificial Intelligence, Korea University, Seoul, Korea
[2]Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea
{jihoon_kim, sh_lee, jihyun-lee, sw.lee}@korea.ac.kr

Abstract

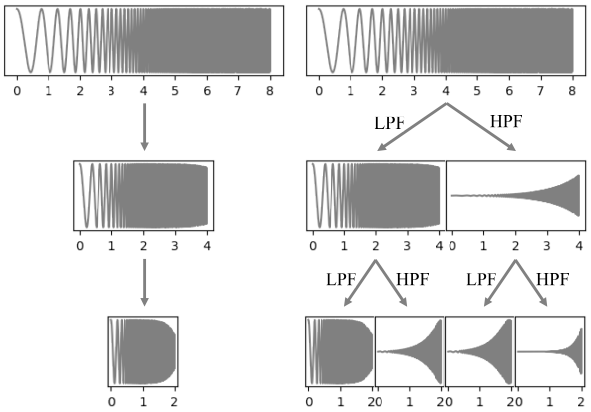
Although recent works on neural vocoder have improved the quality of synthesized audio, there still exists a gap between generated and ground-truth audio in frequency space. This difference leads to spectral artifacts such as hissing noise or reverberation, and thus degrades the sample quality. In this paper, we propose Fre-GAN which achieves frequency-consistent audio synthesis with highly improved generation quality. Specifically, we first present resolution-connected generator and resolution-wise discriminators, which help learn various scales of spectral distributions over multiple frequency bands. Additionally, to reproduce high-frequency components accurately, we leverage discrete wavelet transform in the discriminators. From our experiments, Fre-GAN achieves high-fidelity waveform generation with a gap of only 0.03 MOS compared to ground-truth audio while outperforming standard models in quality.

Index Terms: audio synthesis, neural vocoder, generative adversarial networks, discrete wavelet transform

1. Introduction

Deep generative models have revolutionized neural vocoder that aims to transform acoustic features into intelligible human speech. Especially, autoregressive models [1, 2] have shown exceptional performance in terms of quality and replaced the role of conventional approaches [3, 4]. Nevertheless, they suffer from slow inference speed owing to their autoregressive nature. To address the structural limitation of them, flow-based vocoders are proposed [5–7]. While they can produce natural waveform in real-time because of their ability to convert noise sequence into raw waveform in parallel, they require a heavy computation due to the complex architecture [8].

The other approach is based on Generative Adversarial Networks (GANs) [9–13]. MelGAN [9] adopts Multi-Scale Discriminator (MSD) [14] operating on multiple scales of waveforms modulated by Average Pooling (AP). The MSD has been proven to be advantageous for capturing consecutive patterns of audio [11–13]. Parallel WaveGAN [10] proposes multi-resolution spectrogram loss which helps to stabilize adversarial training. Recently, HiFi-GAN [12] identifies the periodic patterns of audio through Multi-Period Discriminator (MPD) and synthesizes high-fidelity audio. It further improves the audio quality by applying a Multi-Receptive field Fusion (MRF) module in the generator which observes various receptive field patterns in parallel. The model outperforms the autoregressive [1] and flow-based vocoder [6] in terms of both quality and inference speed. Despite the recent advances, there still exists a gap between synthesized and ground-truth audio in frequency space. This gap leads to spectral artifacts such as hissing noise or robotic sound, because audio is made of a complicated mixture of various frequencies.



(a) Average Pooling (b) Discrete Wavelet Transform

Figure 1: *Comparison of Average Pooling (AP) and Discrete Wavelet Transform (DWT). Here, LPF and HPF refer to Low-Pass and High-Pass Filter, respectively. In this example, an up-chirp signal whose frequency increases from 0 Hz at time $t = 0$ to 150 Hz at time $t = 8$ is downsampled by AP and DWT.*

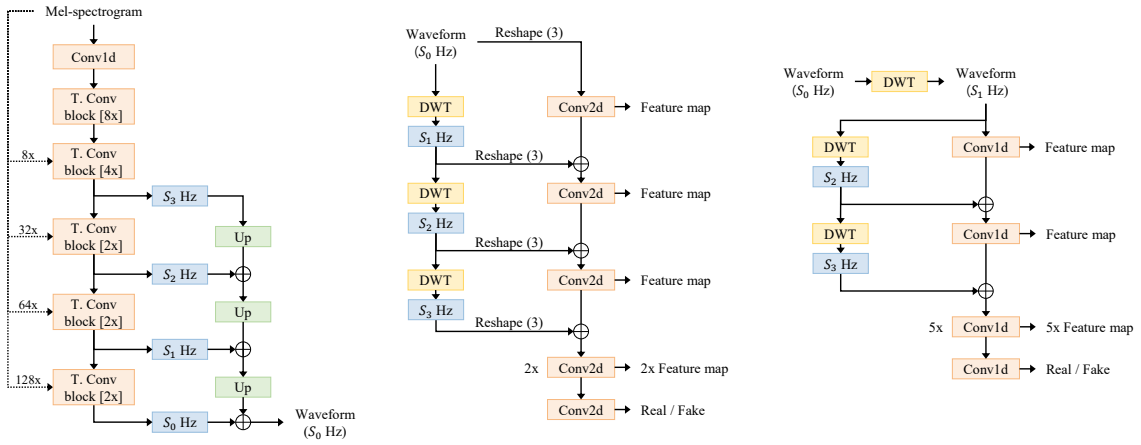
In this paper, we propose Fre-GAN which synthesizes frequency-consistent audio on par with ground-truth audio. Fre-GAN employs resolution-connected generator and resolution-wise discriminators to learn various levels of spectral distributions over multiple frequency bands. The generator upsamples and sums multiple waveforms at different resolutions. Each waveform is adversarially evaluated in the corresponding resolution layers of the discriminators. To further facilitate the training of discriminators, we provide the downsampled audio to each resolution layer of the discriminators [15–17]. Based on this architecture, we use Discrete Wavelet Transform (DWT) as the downsampling method. While the conventional downsampling method, i.e. AP, washes the high-frequency components away, DWT guarantees that all the information can be kept due to its biorthogonal property. In Fig. 1, we provide evidence for the above statement. Unlike AP, DWT can safely deconstruct the signal into low-frequency and high-frequency sub-bands without losing high-frequency contents. In the experiments, the effectiveness of Fre-GAN is demonstrated on various metrics.

2. Fre-GAN

2.1. Resolution-connected Generator

Fre-GAN generator takes a mel-spectrogram as input and upsamples it through transposed convolution blocks until the temporal resolution of the output sequence matches that of

http://dx.doi.org/10.21437/Interspeech.2021-845



(a) RCG (b) RPD (c) RSD

Figure 2: *Fre-GAN network architecture. (a) The RCG. (b) The second sub-discriminator of RPD. (c) The second sub-discriminator of RSD. T. Conv block denotes transposed convolution block. In this research,* $256\times$ *upsampling is conducted in* 5 *stages of* $8\times$*,* $4\times$*,* $2\times$*,* $2\times$*, and* $2\times$*. UP denotes nearest neighbor upsampler which consists of nearest neighbor interpolation and* 1×1 *convolution. Reshape* (p) *refers to reshaping the raw audio from* $1d$ *into* $2d$ *data with period* p*.*

the raw waveform. The transposed convolution blocks consist of transposed convolutional layer followed by MRF module proposed in HiFi-GAN [12] and leaky-relu activation.

Inspired by StyleGAN2 [17], we adopt skip-connections to the generator and call it Resolution-Connected Generator (RCG). The RCG upsamples and sums top-K waveform outputs corresponding to different resolutions, as illustrated in Fig. 2(a). To upsample lower scale waveforms, we use Nearest Neighbor (NN) upsampler [18] which has been proven to alleviate tonal artifacts caused by transposed convolutions [19]. In addition, we directly condition input mel-spectrogram to each top-K transposed convolution block. This allows the multiple waveforms to be consistent with the input mel-spectrogram. In this research, we set K to four.

We investigate that the RCG structure has several benefits. Firstly, it captures various levels of spectral distributions by explicitly summing multiple waveforms at different resolutions. This instigates the model to effectively learn acoustic properties across multiple frequency bands. Secondly, the RCG is trained progressively [20]. The training of RCG starts by focusing on low resolution and then progressively shifts attention to higher resolutions, as we will validate it in Sec. 3.2. This allows the model to first discover the easier coarse structure and then shift its focus to learn increasingly finer details, instead of learning all scales at once. By gradually increasing the resolution, we can speed up and greatly stabilize adversarial training.

2.2. Resolution-wise Discriminators

Fre-GAN employs two discriminators: Resolution-wise MPD (RPD) and Resolution-wise MSD (RSD), whose architectures are drawn from HiFi-GAN [12]. RPD comprises five sub-discriminators each of which accepts specific periodic parts of input audio; the period is given by $p \in \{2, 3, 5, 7, 11\}$ to avoid overlaps [12]. To be specific, the input audio of length T is first reshaped to $2d$ data of height T/p and width p, and then applied to $2d$ convolutions. Whereas RSD consists of three sub-discriminators operating on different input scales: raw audio, $2\times$ downsampled audio, and $4\times$ downsampled audio. Note that RPD captures periodic patterns of audio, and RSD observes consecutive patterns and long-term dependencies of audio.

By setting the resolutions of convolution layers in each sub-discriminator to match that of top-K waveforms in the RCG, we encourage a specific layer in each sub-discriminator to evaluate the waveform of the corresponding resolution. This resolution-wise adversarial evaluation provokes the RCG to learn the mapping from input mel-spectrogram to audio at various scales. Moreover, we provide downsampled audio to the corresponding resolution layer of each sub-discriminator, as shown in Fig. 2(b) and 2(c). As proven in recent works [15–17], this residual connection facilitates the discriminators training. Based on this structure, we use Discrete Wavelet Transform (DWT) to downsample audio without losing any information.

2.3. Discrete Wavelet Transform

In previous works using MSD, they have used AP to downsample raw audio [9, 11–13]. However, the AP ignores the sampling theorem [21], and high-frequency contents are aliased and become invalid [22]. This makes the generator lack the incentives from the MSD to learn high-frequency components, resulting in a spectral distortion in high-frequency bands.

To alleviate high-frequency loss, we replace problematic AP with Discrete Wavelet Transform (DWT) [23] as our down-sampling method. The DWT is an efficient but effective way of downsampling non-stationary signals into several frequency sub-bands. In 1d DWT, the signal is convolved by two filters: low-pass filter (g) and high-pass filter (h). According to Nyquist's rule [24], half the samples of the convolution results can be discarded since half the frequencies of the signal are removed. This gives two $2\times$ downsampled signals representing low-frequency and high-frequency components, respectively. They can be further decomposed by DWT as follows:

$$y_{low}[n] = \sum_k x[k]g[2n-k] \tag{1}$$

$$y_{high}[n] = \sum_k x[k]h[2n-k] \tag{2}$$

Table 1: *Evaluation results. The MOS are presented with* 95% *confidence intervals. Higher is better for MOS and speed, and lower is better for the other metrics. Speed of* n *kHz means that the model can synthesize* $n \times 1000$ *audio samples per second. The numbers in () denote real-time factor. Values in bold represent the best results for each metric.*

Model	**MOS** (↑)	**MCD**$_{13}$ (↓)	**RMSE**$_{f0}$ (↓)	**FDSD** (↓)	**Speed on CPU** (↑)	**Speed on GPU** (↑)
Ground Truth	4.40 ± 0.04	–	–	–	–	–
WaveNet	4.20 ± 0.06	2.293	43.347	0.725	–	0.12 ($\times$0.005)
WaveGlow	3.94 ± 0.07	2.048	40.463	0.542	15.83 ($\times$0.72)	319 ($\times$14.47)
HiFi-GAN V1	4.30 ± 0.05	1.231	39.947	0.190	53.20 ($\times$2.41)	2,351 ($\times$106.62)
HiFi-GAN V2	4.19 ± 0.05	1.606	40.258	0.282	**215.43** ($\times$**9.77**)	**10,730** ($\times$**486.62**)
Fre-GAN V1	$\mathbf{4.37 \pm 0.04}$	**1.060**	**38.339**	**0.150**	60.72 ($\times$2.75)	2,284 ($\times$103.58)
Fre-GAN V2	4.28 ± 0.05	1.308	38.843	0.205	192.07 ($\times$8.71)	10,458 ($\times$474.29)

where $y_{low}[n]$ and $y_{high}[n]$ are subsequent outputs of g and h, respectively. n and k denote levels of DWT and index of signal x, respectively. Due to the biorthogonal property of DWT, the signal can be deconstructed safely without information loss.

After each level of DWT, all the frequency sub-bands are channel-wise concatenated and passed to convolutional layers [25, 26]. When all the sub-bands are taken into account, Fre-GAN can avoid information loss, especially in high frequencies. In our implementation, Daubechies1 wavelet [23] is adopted.

2.4. Training Objectives

To train Fre-GAN, we use the least-squares GAN objective because of its training stability [27]. The training objectives for the discriminators and generator are defined as:

$$\begin{aligned}\mathcal{L}_D = &\sum_{n=0}^{4} \mathbb{E}[\|D_n^P(x) - 1\|_2 + \|D_n^P(\hat{x})\|_2] \\ &+ \sum_{m=0}^{2} \mathbb{E}[\|D_m^S(\phi^m(x) - 1)\|_2 + \|D_m^S(\phi^m(\hat{x}))\|_2]\end{aligned} \tag{3}$$

$$\begin{aligned}\mathcal{L}_G = &\sum_{n=0}^{4} \mathbb{E}[\|D_n^P(\hat{x}) - 1\|_2 + \lambda_{fm}\mathcal{L}_{fm}(G; D_n^P)] \\ &+ \sum_{m=0}^{2} \mathbb{E}[\|D_m^S(\hat{x}) - 1\|_2 + \lambda_{fm}\mathcal{L}_{fm}(G; D_m^S)] \\ &+ \lambda_{mel}\mathcal{L}_{mel}(G)\end{aligned} \tag{4}$$

where x and $\hat{x}$ denote ground-truth and generated audio, respectively. D^P and D^S are RPD and RSD, respectively. ϕ^m represents m-level DWT. In the experiments, we set $\lambda_{fm} = 2$ and $\lambda_{mel} = 45$ which balance the adversarial losses, the feature matching loss ($\mathcal{L}_{fm}$), and the mel-spectrogram loss ($\mathcal{L}_{mel}$) defined as follows:

$$\mathcal{L}_{fm}(G; D_k) = \mathbb{E}\Big[\sum_{i=0}^{T-1} \frac{1}{N_i}\|D_k^{(i)}(x) - D_k^{(i)}(\hat{x})\|_1\Big] \tag{5}$$

$$\mathcal{L}_{mel}(G) = \mathbb{E}\Big[\|\psi(x) - \psi(\hat{x})\|_1\Big] \tag{6}$$

Here, T denotes the number of layers in the discriminator. $D_k^{(i)}$ is the i^{th} layer feature map of the k^{th} sub-discriminator, N_i is the number of units in each layer, and ψ is the STFT function to convert raw audio into the corresponding mel-spectrogram.

The feature matching loss minimizes $L1$ distance between the discriminator feature maps of real and generated audio [28]. As it was successfully adopted to neural vocoder [9, 12], we use it as an auxiliary loss to improve the training efficiency. In addition, we add mel-spectrogram loss to further improve the sample quality and training stability. It is the reconstruction loss that minimizes $L1$ distance between the mel-spectrogram of synthesized audio and that of ground-truth audio. Referring to previous works [10, 29], applying reconstruction loss to GANs training helps to generate realistic results and stabilize the adversarial training process from the early stages.

3. Experimental Results

We conducted experiments on LJSpeech dataset at a sampling rate of $22,050$ Hz. The dataset contains $13,100$ audio samples of a single English speaker, and we randomly split the dataset into train (80%), validation (10%), and test (10%) sets. Fre-GAN was compared against several neural vocoders trained on the same dataset: the popular open-source implementation of mixture of logistics WaveNet [30], and the official implementation of WaveGlow [31] and HiFi-GAN [32]. All the models were trained for 3100 epochs.

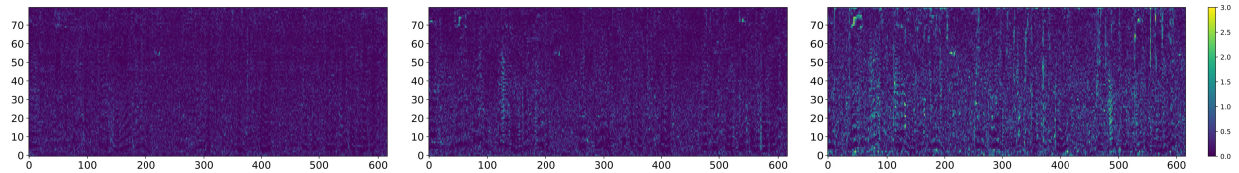
Similar to HiFi-GAN [12], we conducted experiments based on two variations of the generator: V1, V2 with the same discriminator configuration. Each variation resembles that of HiFi-GAN but we set the kernel sizes of transposed convolutions to $[16, 8, 4, 4, 4]$ and dilation rates of MRF to $[[1, 1], [3, 1], [5, 1], [7, 1] \times 3]$. 80 bands mel-spectrogram was transformed with 1024 of window size, 256 of hop size, and 1024 points of Fourier transform. We used AdamW optimizer [33] with $\beta_1 = 0.8$, $\beta_2 = 0.999$, 16 of batch size, and followed the same learning rate schedule in [12]. The synthesized audio samples are available at `https://prml-lab-speech-team.github.io/demo/FreGAN`

3.1. Audio Quality and Inference Speed

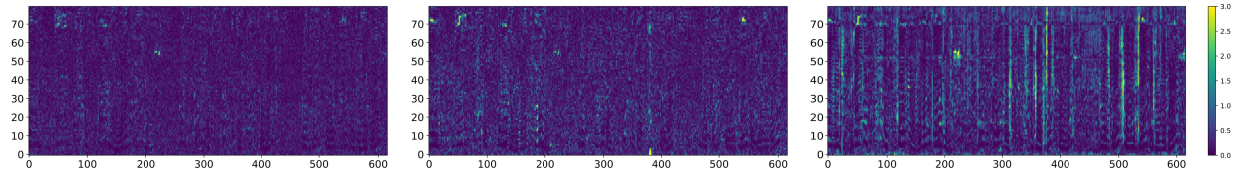
We assessed Fre-GAN on various metrics. As a subjective test, we performed 5-scale MOS tests via Amazon MTurk where at least 20 raters were asked to rate the naturalness of audio. As an objective quality evaluation, we used MCD$_{13}$ [34], RMSE$_{f0}$ [35], and FDSD [36]. 50 and 200 synthesized utterances were used for subjective and objective evaluation, respectively. We also measured the synthesis speed on Intel Xeon Gold 6148 2.40 GHz CPU and a single NVIDIA Titan Xp GPU.

The results are presented in Table 1. Above all, Fre-GAN outperforms other models in terms of quality. Specifically, Fre-GAN V1 demonstrates similarity to ground-truth audio with a gap of only 0.03 MOS; this implies that the generated audio is highly similar to the real audio. In terms of synthesis speed, Fre-GAN was nearly as fast as HiFi-GAN, and all the variations of Fre-GAN were faster than WaveNet and WaveGlow.

Moreover, we investigated the pixel-wise difference in the

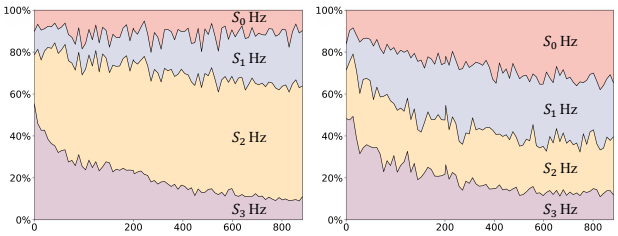


(a) Fre-GAN V1 (b) HiFi-GAN V1 (c) WaveNet



(d) Fre-GAN V2 (e) HiFi-GAN V2 (f) WaveGlow

Figure 3: *Pixel-wise difference in the mel-spectrogram space between generated and ground-truth audio. Fre-GAN reproduces the desired spectral distributions.*



(a) w/o DWT (b) w/ DWT

Figure 4: *Contribution of each resolution to the output of the RCG as the training proceeds. The x- and y-axis depicts the training epochs and standard deviations, respectively.*

mel-spectrogram domain between generated and ground-truth audio, as illustrated in Fig. 3. We observed that the error of the mel-spectrogram was highly reduced in Fre-GAN, which indicates that Fre-GAN generates frequency-consistent audio corresponding to the input mel-spectrogram.

3.2. Ablation Study

We performed an ablation study to observe the effect of each Fre-GAN component. Fre-GAN V2 was used as a generator, and each model was trained for $500k$ steps. In Table 2, all the Fre-GAN components contribute to the sample quality. Especially, MOS was largely dropped after replacing DWT with AP, whereas using MPD and MSD (instead of RPD and RSD) shows a relatively small but perceptible degradation. Removing RCG architecture in the generator shows the worst performance in MCD_{13}. The absence of mel-spectrogram condition and replacing NN upsampler with transposed convolution also lead to metallic noise and quality degradation.

Fig. 4 further verifies the advantages of RCG and DWT. As argued in Sec. 2.1, the RCG benefits from progressive learning. To validate this, we quantified the relative importance of multiple waveforms at different resolutions (S_0, S_1, S_2, and S_3 Hz) by measuring their contribution to the final audio. We calculated the standard deviations of the audio samples as a function of training epochs and normalized the values so that they summed to 100%. As shown, the RCG initially focuses on learning low-resolution and then slowly shifts its attention to higher resolutions. One might expect that the highest resolution will be dominant towards the end of training. However, the RCG fails to fully utilize the target resolution when we replace DWT with AP. This implies that replacing DWT with AP washes high-frequency components away, and thus the RCG lacks the incentives to learn high-frequency details.

Table 2: *MOS, MCD_{13}, and FDSD results of ablation study.*

Model	**MOS**	**MCD_{13}**	**FDSD**
Ground Truth	4.39 ± 0.03	–	–
Fre-GAN V2	$\mathbf{4.25 \pm 0.04}$	**1.383**	**0.209**
w/o RCG	4.14 ± 0.05	1.602	0.214
w/o NN upsampler	4.15 ± 0.05	1.562	0.238
w/o mel condition	4.18 ± 0.05	1.561	0.212
w/o RPD & RSD	4.20 ± 0.04	1.581	0.256
w/o DWT	4.12 ± 0.05	1.592	0.236
HiFi-GAN V2	4.08 ± 0.05	1.793	0.288

4. Conclusions

In this paper, we presented Fre-GAN which can synthesize high-fidelity audio with realistic spectra. We observed the inconsistency in frequency space between generated and ground-truth audio, and addressed the problem through the proposed network architecture with a lossless downsampling method. From the experiments, we verified each Fre-GAN component contributes to the sample quality. Additionally, Fre-GAN outperforms standard neural vocoders with only a gap of 0.03 MOS compared to real audio. For future work, we will apply our proposed method to end-to-end TTS systems to improve the quality of synthesized speech.

5. Acknowledgements

This work was supported by Institute for Information & communications Technology Planning & evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00079, Department of Artificial Intelligence, Korea University) and the Netmarble AI Center.

6. References

[1] A. v. d. 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:1609.03499*, 2016.

[2] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient Neural Audio Synthesis," in *Proc. International Conference on Machine Learning (ICML)*, 2018, pp. 2410–2419.

[3] D. Griffin and J. Lim, "Signal Estimation from Modified Short-time Fourier Transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, 1984.

[4] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-based High-quality Speech Synthesis System for Real-time Applications," *IEICE Trans. on Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[5] S. Kim, S.-G. Lee, J. Song, J. Kim, and S. Yoon, "FloWaveNet: A Generative Flow for Raw Audio," in *Proc. International Conference on Machine Learning (ICML)*, 2019, pp. 3370–3378.

[6] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A Flow-based Generative Network for Speech Synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3617–3621.

[7] H.-W. Yoon, S.-H. Lee, H.-R. Noh, and S.-W. Lee, "Audio Dequantization for High Fidelity Audio Generation in Flow-based Neural Vocoder," in *Proc. Interspeech*, 2020, pp. 3545–3549.

[8] S.-H. Lee, H.-W. Yoon, H.-R. Noh, J.-H. Kim, and S.-W. Lee, "Multi-SpectroGAN: High-Diversity and High-Fidelity Spectrogram Generation with Adversarial Style Combination for Speech Synthesis," in *Proc. AAAI Conference on Artificial Intelligence*, 2021.

[9] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[10] 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. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

[11] J. Yang, J. Lee, Y. Kim, H. Cho, and I. Kim, "VocGAN: A High-Fidelity Real-time Vocoder with a Hierarchically-nested Adversarial Network," in *Proc. Interspeech*, 2020, pp. 200–204.

[12] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[13] W. Jang, D. Lim, and J. Yoon, "Universal MelGAN: A Robust Neural Vocoder for High-Fidelity Waveform Generation in Multiple Domains," *arXiv:2011.09631*, 2020.

[14] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8798–8807.

[15] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved Training of Wasserstein GANs," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[16] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," in *Proc. International Conference on Learning Representation (ICLR)*, 2018.

[17] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8110–8119.

[18] J. Pons, S. Pascual, G. Cengarle, and J. Serrà, "Upsampling Artifacts in Neural Audio Synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[19] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and Checkerboard Artifacts," *Distill*, vol. 1, no. 10, 2016.

[20] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," in *Proc. International Conference on Learning Representation (ICLR)*, 2017.

[21] R. Zhang, "Making Convolutional Networks Shift-invariant Again," in *Proc. International Conference on Machine Learning (ICML)*, 2019, pp. 7324–7334.

[22] Y. Chen, G. Li, C. Jin, S. Liu, and T. Li, "SSD-GAN: Measuring the Realness in the Spatial and Spectral Domains," in *Proc. AAAI Conference on Artificial Intelligence*, 2021.

[23] I. Daubechies, "Orthonormal Bases of Compactly Supported Wavelets," *Commun. Pure. Appl. Math.*, vol. 41, no. 7, pp. 909–996, 1988.

[24] H. Nyquist, "Certain Topics in Telegraph Transmission Theory," *Trans. AIEE*, vol. 47, no. 2, pp. 617–644, 1928.

[25] H.-I. Suk and S.-W. Lee, "Subject and Class Specific Frequency Bands Selection for Multi-Class Motor Imagery Classification," *International Journal of Imaging Systems and Technology*, vol. 21, no. 2, pp. 123–130, 2011.

[26] M.-H. Lee, J. Williamson, D.-O. Won, S. Fazli, and S.-W. Lee, "A High Performance Spelling System based on EEG-EOG Signals with Visual Feedback," *IEEE Trans. on Neural Syst. Rehabil. Eng.*, vol. 26, no. 7, pp. 1443–1459, 2018.

[27] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least Squares Generative Adversarial Networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2794–2802.

[28] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond Pixels using a Learned Similarity Metric," in *Proc. International Conference on Machine Learning (ICML)*, 2016, pp. 1558–1566.

[29] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image Translation with Conditional Adversarial Networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1125–1134.

[30] R. Yamamoto, "wavenet vocoder," https://github.com/r9y9/wavenet_vocoder, 2018.

[31] R. Valle, "waveglow," https://github.com/NVIDIA/waveglow, 2018.

[32] J. Kong, "hifi-gan," https://github.com/jik876/hifi-gan, 2020.

[33] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proc. International Conference on Learning Representation (ICLR)*, 2019.

[34] R. Kubichek, "Mel-cepstral Distance Measure for Objective Speech Quality Assessment," in *Proc. IEEE Pacific Rim Conference on Communications Computers and Signal Processing (PACRIM)*, 1993, pp. 125–128.

[35] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An Investigation of Multi-speaker Training for WaveNet Vocoder," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 712–718.

[36] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High Fidelity Speech Synthesis with Adversarial Networks," in *Proc. International Conference on Learning Representation (ICLR)*, 2020.