



AF-Vocoder: Artifact-Free Neural Vocoder with Global Artifact Filter

Zhuangqi Chen¹, Xianjun Xia², Xiaohuai Le¹, Siyu Sun¹, Chuanzeng Huang²

¹Bytedance China, Shenzhen, China

²Bytedance Inc., Sydney, Australia

chenzhuangqi.audio@bytedance.com, xxjppjj@mail.ustc.edu.cn

Abstract

Recent studies have demonstrated the advantage of generative adversarial network (GAN)-based vocoders in high-fidelity speech synthesis and fast inference speed. However, they often suffer from audible artifacts such as aliasing and blurring. In this paper, we propose AF-Vocoder, a novel GAN-based vocoder that can synthesize high-fidelity speech with fewer artifacts. Specifically, we introduce a frequency-domain artifacts filter named GAFilter to achieve artifact removal. GAFilter incorporates a learnable frequency filter, which enforces a desired inductive bias of frequency control for artifact-free speech synthesis. Experimental results show that the proposed AF-Vocoder outperforms other GAN-based vocoders in speech reconstruction quality and artifact suppression on various datasets including out-of-domain speakers.

Index Terms: speech synthesis, vocoder, artifact-free, generative adversarial network, GAFilter

1. Introduction

The vocoder is designed to convert the acoustic feature (e.g., mel-spectrogram) into speech waveform, which is widely applied in a range of speech generation tasks such as text to speech and voice conversion [1, 2]. With the advancement of generative neural networks, numerous studies have attempted to apply these networks to vocoders and have attained remarkable outcomes. These efforts encompass methods rooted in Generative Adversarial Networks (GANs)-based models [3, 4, 5, 6], Diffusion-based models [7, 8, 9], and Flow-based models [10, 11, 12]. Among these methods, when conditioned on a given mel-spectrogram, the GAN-based approach can generate high-quality speech signals at a higher speed [5]. As a result, the GAN-based vocoder has emerged as a prevalent solution within the field. Despite significant advancements in GAN-based vocoders, generating high-quality artifact-free speech remains challenging. The artifacts [13, 14] can mainly be attributed to two aspects: 1) aliasing artifacts caused by imperfect upsampling and 2) blurring artifacts caused by loss of spectral details.

Numerous vocoders predominantly operate in the time domain [4, 5], which take mel-spectrogram as input and employ a series of upsampling layers to directly reconstruct waveform signal. However, due to the imperfections of upsampling, these methods are susceptible to spectral aliasing. Although several methods [13, 14] have been proposed to deal with spectral aliasing, these methods mainly focus on aliasing-related artifacts yet neglect the artifacts associated with spectral blurring. As a result, there is still room for these methods to be improved. To circumvent the influence of upsampling, some studies proposed to implement the vocoders in the frequency

domain [15, 16], where they first convert mel-spectrogram into amplitude and phase information, and then use inverse short-time Fourier transform to recover the waveform signal. Those methods preserve the same time-domain resolution during processing and utilize ISTFT for waveform conversion, effectively avoiding the problem of aliasing artifacts. Nevertheless, compared to the time-domain vocoder that introduces periodic inductive bias, the harmonic details generated by those methods are insufficient, particularly in the high frequency band.

In order to generate higher-quality speech signals, many researches have enhanced the modeling ability of generators by incorporating some well-designed discriminators and auxiliary loss functions [17, 13, 18]. While those approaches substantially improve the model's overall performance, they don't specifically deal with the problem of spectral blurring, resulting in the continued presence of blurring artifacts in the results. In our preliminary experiments, we also found that while enhancing the discriminator and loss function can contribute to the improvement of overall metrics, it still fails to effectively solve the issue of blurring artifacts. Thus, there is still considerable room for the optimization of the model's performance.

In this paper, we investigate whether the generative networks can automatically filter out the unexpected artifacts by introducing some inductive biases of frequency control. Since artifacts are primarily the unreasonable components generated during the model processing, for instance, aliasing artifacts stem from the mirror images of the spectrogram itself which is caused by the imperfect upsampling, and blurring artifacts mainly originate from the unreasonable components among harmonics. Based on this, we propose to incorporate additional modules to empower the model to remove these undesired components. To this end, we propose a global artifacts filter, GAFilter, which can handle artifacts through learned frequency control. Based on GAFilter, we introduce an artifact-free GAN-based vocoder, AF-Vocoder, that allows synthesizing speech with fewer artifacts. The results show that AF-Vocoder can effectively filter out abnormal artifacts, thereby reaching a higher upper bound of performance. Our contributions are three-fold:

- We propose an artifact-free neural vocoder, AF-Vocoder, that can generate high quality speech with reduced artifacts.
- We introduce a novel inductive bias of frequency control and propose a simple yet effective artifacts filter, GAFilter, which helps to generate artifact-free results.
- We demonstrate that AF-Vocoder can effectively eliminate abnormal artifacts and synthesize higher quality speech. AF-Vocoder outperforms other SOTA GAN-based vocoders with comparable model size for both in-domain and out-of-domain scenarios. In addition, GAFilter not only enhances performance but also ensures stable model training.

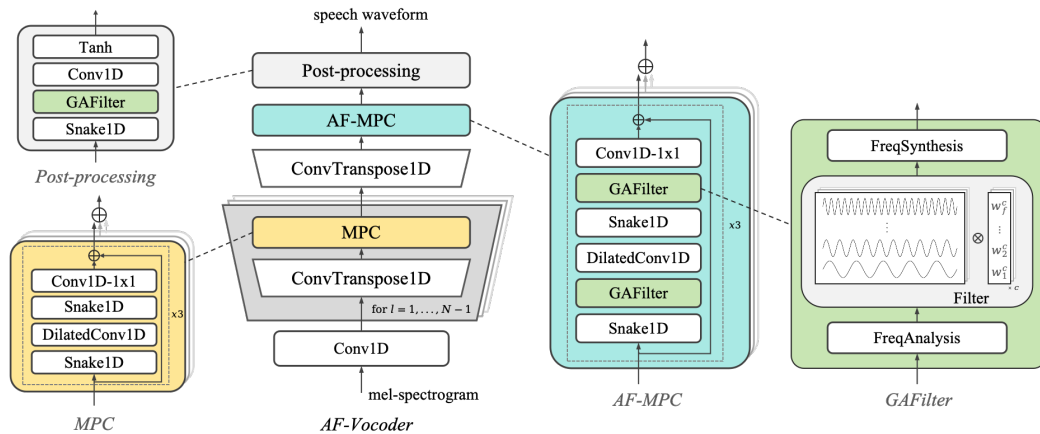


Figure 1: Overall architecture of AF-Vocoder. AF-Vocoder consists of stacked upsampling layers for recovering speech waveform from mel-spectrogram, where the Snake function is utilized for providing periodic inductive bias and a global artifacts filter (GAFilter) is designed to remove unexpected artifacts. The GAFilter module is a frequency filter providing desired inductive bias of frequency control, which facilitates the synthesis of artifact-free speech.

2. Method

2.1. Overall framework

In this paper, we focus on the GAN-based vocoder. Specifically, we aim to construct a generative model $\Psi_\theta : M \rightarrow S$ that can convert a given mel-spectrogram $m \in M$ into artifact-free and high-fidelity speech waveform $s \in S$. In the GAN-based training paradigm, there is one generator and multiple discriminators that are trained adversarially, combined with additional GAN losses for performance improvement.

As depicted in Figure 1, we present AF-Vocoder as the generator, which is an artifact-free neural vocoder that effectively eliminates synthesis distortions while maintaining high waveform fidelity. Similar to BigVGAN [5], the AF-Vocoder employs stacked ConvTranspose1D layers to upsample input features and progressively restore temporal resolution. Following each upsampling layer, we introduce a multi-periodicity composition (MPC) module that incorporates the Snake activation function [19] to provide periodic inductive bias, while employing varied kernel sizes and dilation rates to capture multi-scale receptive fields. The MPC module consists of parallel residual blocks, and each block contains dilated Conv1D layers, Conv1D-1x1 layers and Snake activations. To generate artifact-free results, we introduce a frequency control inductive bias and propose a frequency-domain artifact filter named GAFilter. This filter is integrated with the last MPC module to construct an Artifact-Free MPC (AF-MPC) architecture. Specifically, the processing pipeline of AF-Vocoder operates as follows: Given an input mel-spectrogram $m \in \mathbb{R}^{F \times T}$, a Conv1D layer first projects the frequency dimension F into a higher-dimensional latent space, facilitating enhanced extraction of speech’s intrinsic representations. Subsequently, a series of ConvTranspose1D layers following the MPC modules progressively restore temporal resolution while reducing frequency dimensionality, where the last MPC module is substituted with AF-MPC to eliminate residual artifacts. Finally, a post-processing stage is performed to reconstruct the final waveform, which comprises sequential operations: a Snake activation layer, GAFilter module, Conv1D layer, and Tanh activation.

Notably, while AF-Vocoder shares architectural similarities with BigVGAN, two key distinctions emerge: (1) Architectural

innovation: a dedicated GAFilter module is incorporated in AF-Vocoder to ensure artifact-free waveform generation. (2) Module simplification: the upsampling-and-downsampling operations present in BigVGAN - strategically positioned after each Snake activation function to suppress aliasing artifacts and high-frequency distortions - are intentionally omitted in our framework. This design philosophy stems from GAFilter’s comprehensive artifact suppression capability, which alleviates the network’s burden of intermediate feature purification during generation, thereby achieving structural simplification without compromising output quality.

Prioritizing artifact-free generator development, we strategically leverage established discriminators from BigVGAN [5] and BigVSAN [6] to validate AF-Vocoder’s efficacy. BigVGAN is a large-scale neural vocoder that employs Least Squares GAN [20] training with hybrid discriminators: a Multi-Period Discriminator (MPD) [4] and a Multi-Resolution Discriminator (MRD) [21, 22]. In contrast, BigVSAN enhances BigVGAN through the Slicing Adversarial Network (SAN) training framework [23], which optimizes discriminators with discriminative projections to achieve state-of-the-art performance.

2.2. GAFilter: Global Artifact Filter

The design rationale of GAFilter stems from the empirical observation that diverse artifacts consistently manifest themselves as anomalous residual components within the frequency domain. We consider introducing a frequency-domain filter that enables the model to autonomously suppress artifacts through training. To this end, we introduce an inductive bias of frequency control and propose a learnable global artifacts filter (GAFilter), where the term *global* means the filter operates on full frequency bands with temporally invariant characteristics. In addition, we apply GAFilter only after the last upsampling layer, based on two considerations: (1) The temporal resolution has been fully restored after last upsampling layer, while the intermediate features in other MPC modules exhibiting limited interpretability. (2) This design relaxes constraints on intermediate feature learning to help enhance performance potential while reducing model complexity. Our preliminary experiments also revealed that inserting GAFilter in intermediate layers ad-

Table 1: Comparison with other SOTA methods on LibriTTS-dev set.

Models	Params (M)	M-STFT (\downarrow)	PESQ (\uparrow)	Periodicity (\downarrow)	V/UV F1 (\uparrow)	UTMOS (\uparrow)
FreeV [15]	18.2	0.9492	3.260	0.1368	0.9424	3.1610
Vocos [16]	13.5	0.8603	3.613	0.1142	0.9532	3.5488
BigVGAN (<i>Lee et al.</i> [5])	112.44	0.7997	4.027	0.1018	0.9598	-
BigVGAN (<i>reprod.</i>)	112.44	0.8056	3.974	0.1035	0.9606	3.5139
BigVGAN+GAFilter	112.56	0.7933	4.021	0.0966	0.9621	3.6199
AF-Vocoder V1	112.56	0.7844	4.051	0.0946	0.9641	3.6499
AF-Vocoder V2	98.17	0.7865	4.058	0.0928	0.9640	3.6534
AF-Vocoder V3	13.44	0.8641	3.711	0.1201	0.9494	3.5628
BigVSAN [6]	112.44	0.7800	4.123	0.0905	0.9657	3.4649
AF-Vocoder-SAN	112.56	0.7713	4.129	0.0886	0.9668	3.5170

versely affects model convergence, thereby validating our architectural decision.

As shown in Figure 1, the proposed GAFilter consists of an STFT module Γ , a learnable filter, and an ISTFT module Γ^{-1} . Given a hidden feature $z \in \mathbb{R}^{B \times C \times T}$, the STFT module first transforms it into the frequency domain feature $Z \in \mathbb{C}^{B \times C \times F \times T}$. Then, a frequency filter with learnable parameters $W \in \mathbb{R}^{1 \times C \times F \times 1}$ is applied to Z to obtain an artifact-free feature Z' . The ISTFT is finally employed to recover the time resolution and output an artifact-free feature z' . This process can be formulated as:

$$z' = \Gamma^{-1}(\Gamma(z) * W) \quad (1)$$

where $*$ denotes broadcasting element-wise multiplication.

2.3. Training objective

For GAN training, we follow the official implementation of BigVGAN and BigVSAN, where the overall objective \mathcal{L}_G for generator and \mathcal{L}_D for discriminator can be formulated as:

$$\mathcal{L}_{adv}(G) = \sum_k [\mathcal{L}_{adv}(G; D_k) + \lambda_{fm} \mathcal{L}_{fm}(G; D_k)] \quad (2)$$

$$\mathcal{L}_G = \mathcal{L}_{adv}(G) + \lambda_{mel} \mathcal{L}_{mel}(G) \quad (3)$$

$$\mathcal{L}_D = \sum_k [\mathcal{L}_{adv}(D_k; G)] \quad (4)$$

where \mathcal{L}_{adv} denotes the least-square GAN in BigVGAN or least-squares SAN in BigVSAN, \mathcal{L}_{fm} and \mathcal{L}_{mel} represent the feature matching and mel-spectrogram loss, respectively, and the λ_{fm} and λ_{mel} are the control scalars.

3. Experiments

3.1. Datasets

We utilize the full *train* set in the LibriTTS dataset [24] as training data, and use a subset derived from the *dev-other* and *dev-clean* set for performance evaluation and comparison. The data selection configuration is consistent with that of the BigVGAN to ensure the uniformity of experimental conditions. In addition, to further analyze the robustness of the model on unknown speakers, we also conduct objective evaluation on the LJSpeech [25], VCTK [26], and MUSDB18-HD [27] datasets, which contain various out-of-distribution scenarios. 100 samples are randomly selected from the LJSpeech and VCTK datasets, respectively. From each of the 50 samples within the vocal tracks of

the MUSDB18-HD test set, 10-second audio clips are randomly segmented. All data are resampled to a sampling rate of 24kHz.

3.2. Experimental setups

To ensure a fair comparison, the configuration of model training follows the official settings of BigVGAN, including optimizer, learning rate, batch size, segment length. The 100-band log-mel spectrogram is used with the frequency range of [0, 12] KHz, 1024 FFT length, 1024 hanning window and 256 hop size. We train all models for 1M steps.

To verify the effectiveness of the proposed method, we first trained our proposed AF-Vocoder with the same discriminators and loss as BigVGAN, where three models with varying complexities are constructed: 1) **AF-Vocoder V1**: the number of upsampling channels is [768,384,192,96,48,24], which follows the setting in BigVGAN; 2) **AF-Vocoder V2**: the number of upsampling channels reduced to [720,360,180,96,48,24] and the number of channel in first Conv1D is 1280, which has smaller model size than BigVGAN but still maintaining a high complexity; 3) **AF-Vocoder V3**: the number of upsampling channels reduced to [256,128,64,32,24] and the number of channel in first Conv1D is 512, in which we significantly reduces the model size. In addition, we also trained AF-Vocoder with the SAN framework (**AF-Vocoder-SAN**) used in BigVGAN. Furthermore, we applied our GAFilter to BigVGAN (**BigVGAN+GAFilter**) to show the its generality. The frame length and frame shift used in GAFilter are set to 20ms and 10ms respectively, and the kernel sizes and dilation rates follow the setting in BigVGAN. Several latest GAN-based vocoders are employed for comparison, where the official implementations of BigVGAN, BigVSAN and Vocos are used. We re-trained FreeV on larger LibriTTS dataset with the same batch size and mel features and also retrained BigVGAN for fair ablation studies.

To evaluate the model performance, five objective metrics are used: (1) M-STFT [28] which measures the spectral distance with different resolutions; (2) PESQ [29], a widely used assessment of speech quality; (3) Periodicity error; (4) V/UV F1 [30], the F1 score for voiced/unvoiced classification; and (5) UTMOS [31], an objective mean opinion score (MOS) prediction for evaluating synthesized speech. A subjective Comparison MOS (CMOS) test is also conducted using Comparison Category Rating (CCR) method [32] with eight raters.

3.3. Results and analysis

As shown in Table 1, we first compare the proposed AF-Vocoder with other SOTA GAN-based vocoders on LibriTTS

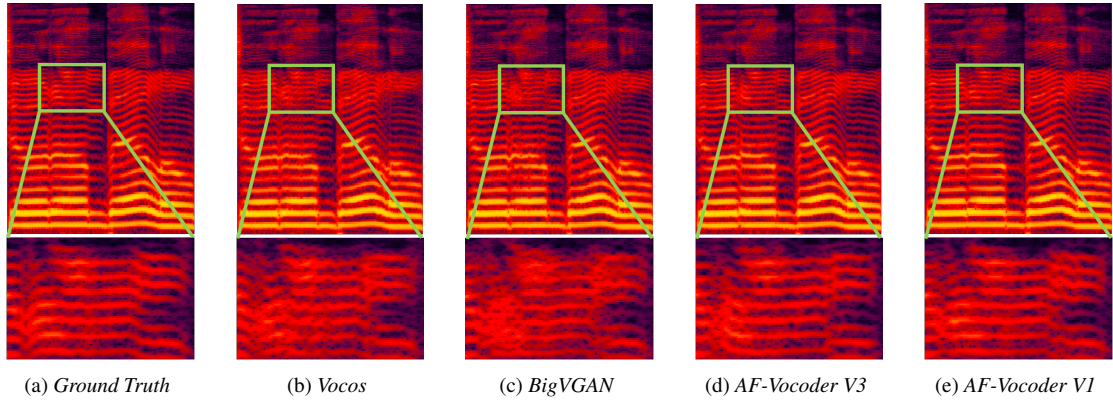


Figure 2: Visualization of the model output.

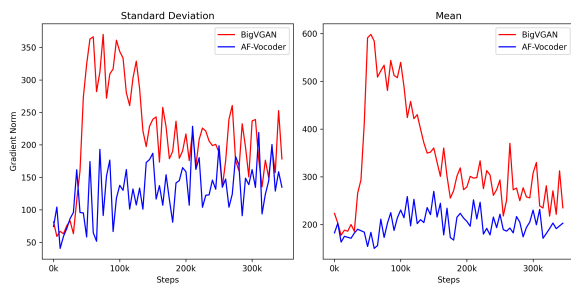


Figure 3: Gradient curves of BigVGAN (red) and Ours (blue).

dev set, and the results show that AF-Vocoder significantly outperform other SOTA methods on all objective metrics. And the improvement is not due to the increase in model size, since the model AF-Vocoder V2 with less model complexity can achieve higher performance than BigVGAN. In addition, by simply applying the proposed GAFilter to BigVGAN, all objective scores show improvements. Furthermore, AF-Vocoder can achieve higher objective metrics than the improved BigVGAN with GAFilter, where the improved BigVGAN has larger model complexity caused by the additional upsampling-and-downsampling operations. The result demonstrates that upon the introduction of GAFilter, the upsampling-and-downsampling operations can be removed despite the fact that it has been proven to help eliminate aliasing. The subjective results shown in Table 2 also confirm the perceptual improvement. It’s worth mentioning that we found the results contain more audible artifacts under the SAN training framework. Therefore, we take the models with least-square GAN training for further illustration.

In order to verify whether AF-Vocoder can effectively filter out unexpected artifacts, we visualize the model outputs, and the results are shown in Figure 2. There are many unexpected components in the results of BigVGAN, and the harmonics are unclear. While from the results of AF-Vocoder, the blurring artifacts almost disappear, the harmonic structure is more regular, and the generated speech signal is closer to groundtruth. Even after reducing the model complexity, the result of model output still shows fewer artifacts. This verifies that the proposed AF-Vocoder can effectively filter out unexpected artifacts, which helps the model obtain a higher upper bound of performance.

Furthermore, we found that the training procedure of AF-

Table 2: Subjective test results on MUSDB test set.

Models	CMOS	Models	CMOS
BigVGAN	0	BigVSAN	0
AF-Vocoder V1	+0.328	AF-Vocoder-SAN	+0.559

Table 3: Results on unseen scenarios.

Datasets	Models	PESQ	Periodicity	UTMOS
VCTK	BigVGAN	3.990	0.0815	3.8577
	Ours	4.084	0.0740	3.9541
LJSpeech	BigVGAN	4.104	0.0906	4.0749
	Ours	4.154	0.0890	4.1463
MUSDB	BigVGAN	3.836	0.1104	1.5080
	Ours	3.921	0.1012	1.5448

Vocoder is more stable. As shown in Figure 3, in the early stages of training, the gradients of BigVGAN are more divergent and unstable, which has higher deviation. However, by replacing the AF-Vocoder as generator, the gradients become more stable, which helps the model converge better. This improvement may be attributed to the learnable frequency control of GAFilter, which effectively filters out abnormal components and enables precise signal reconstruction in the frequency domain.

In addition, we also conduct performance evaluation on the out-of-distribution datasets, and the results are shown in Table 3. With GAFilter, our AF-Vocoder can achieve superior performance compared to BigVGAN in unseen speaker scenarios, which verifies the robustness of the proposed method.

4. Conclusion

In this paper, we explore the GAN-based artifact-free vocoder and present a high-fidelity neural vocoder AF-Vocoder. AF-Vocoder employs an artifacts filter, GAFilter, for eliminating the unexpected artifacts. The GAFilter is a learnable frequency filter, which controls the generated frequencies and helps to reduce artifacts. We have performed extensive experiments, and the results show that the AF-Vocoder yields significantly superior scores across various metrics of speech quality assessment compared to other state-of-the-art vocoders for both seen and unseen speaker scenarios.

5. References

- [1] 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 *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.
- [2] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, “Wavenet vocoder with limited training data for voice conversion,” in *Interspeech 2018*, 2018, pp. 1983–1987.
- [3] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. De Brebisson, Y. Bengio, and A. C. Courville, “Melgan: Generative adversarial networks for conditional waveform synthesis,” *Advances in neural information processing systems*, vol. 32, 2019.
- [4] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.
- [5] S. gil Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, “BigVGAN: A universal neural vocoder with large-scale training,” in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=iTtGCMDEzS_
- [6] T. Shibuya, Y. Takida, and Y. Mitsufuji, “Bigvsan: Enhancing gan-based neural vocoders with slicing adversarial network,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 121–10 125.
- [7] Z. Chen, X. Tan, K. Wang, S. Pan, D. Mandic, L. He, and S. Zhao, “Infergrad: Improving diffusion models for vocoder by considering inference in training,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8432–8436.
- [8] T. D. Nguyen, J.-H. Kim, Y. Jang, J. Kim, and J. S. Chung, “Fregard: Lightweight and fast frequency-aware diffusion vocoder,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 736–10 740.
- [9] N. Takahashi, M. Kumar, Y. Mitsufuji *et al.*, “Hierarchical diffusion models for singing voice neural vocoder,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [10] H. Lim, S. Oh, K. Byun, and H.-G. Kang, “A study on conditional features for a flow-based neural vocoder,” in *2020 54th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2020, pp. 662–666.
- [11] 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 *Interspeech 2020*, 2020, pp. 3545–3549.
- [12] M. Luong and V. A. Tran, “Flowvocoder: A small footprint neural vocoder based normalizing flow for speech synthesis,” in *Interspeech 2022*, 2022, pp. 1576–1580.
- [13] R. Shen, Y. Ren, and Z. Sun, “Fa-gan: Artifacts-free and phase-aware high-fidelity gan-based vocoder,” in *Interspeech 2024*, 2024, pp. 3884–3888.
- [14] H. Cho, J. Lee, and W. Jung, “Jengan: Stacked shifted filters in gan-based speech synthesis,” in *Interspeech 2024*, 2024, pp. 3879–3883.
- [15] Y. Lv, H. Li, Y. Yan, J. Liu, D. Xie, and L. Xie, “Freev: Free lunch for vocoders through pseudo inversed mel filter,” in *Interspeech 2024*, 2024, pp. 3869–3873.
- [16] H. Siuzdak, “Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis,” in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=vY9nzQmQBw>
- [17] T. Bak, J. Lee, H. Bae, J. Yang, J.-S. Bae, and Y.-S. Joo, “Avocodo: Generative adversarial network for artifact-free vocoder,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 12 562–12 570.
- [18] Y. Gu, X. Zhang, L. Xue, and Z. Wu, “Multi-scale sub-band constant-q transform discriminator for high-fidelity vocoder,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 616–10 620.
- [19] L. Ziyin, T. Hartwig, and M. Ueda, “Neural networks fail to learn periodic functions and how to fix it,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 1583–1594, 2020.
- [20] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, “Least squares generative adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.
- [21] 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 *Interspeech 2021*, 2021, pp. 2207–2211.
- [22] M. Liu, Z. Chen, X. Yan, Y. Lv, X. Xia, C. Huang, Y. Xiao, and L. Xie, “Rad-net 2: A causal two-stage repairing and denoising speech enhancement network with knowledge distillation and complex axial self-attention,” in *Interspeech 2024*, 2024, pp. 1700–1704.
- [23] Y. Takida, M. Imaizumi, T. Shibuya, C.-H. Lai, T. Uesaka, N. Murata, and Y. Mitsufuji, “SAN: Inducing metrizable of GAN with discriminative normalized linear layer,” in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=eiF7TU1E8E>
- [24] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “Libritts: A corpus derived from librispeech for text-to-speech,” in *Interspeech 2019*, 2019, pp. 1526–1530.
- [25] K. Ito and L. Johnson, “The lj speech dataset,” <https://keithito.com/LJ-Speech-Dataset/>, 2017.
- [26] C. Veaux, J. Yamagishi, and K. MacDonald, “Str vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit,” 2017. [Online]. Available: <https://datashare.ed.ac.uk/handle/10283/2651>
- [27] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, “Musdb18-hq - an uncompressed version of musdb18,” Aug. 2019. [Online]. Available: <https://doi.org/10.5281/zenodo.3338373>
- [28] 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 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.
- [29] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.
- [30] M. Morrison, R. Kumar, K. Kumar, P. Seetharaman, A. Courville, and Y. Bengio, “Chunked autoregressive GAN for conditional waveform synthesis,” in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=v3aelsY_vVX
- [31] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, “Utmos: Utokyo-sarulab system for voicemos challenge 2022,” in *Interspeech 2022*, 2022, pp. 4521–4525.
- [32] ITU-T, “P.800 methods for subjective determination of transmission quality,” 1996. [Online]. Available: <https://www.itu.int/rec/T-REC-P.800-199608-1>