Avocodo: Generative Adversarial Network for Artifact-Free Vocoder

Taejun Bak[1*], **Junmo Lee**[2*†], **Hanbin Bae**[3†], **Jinhyeok Yang**[4†], **Jae-Sung Bae**[3†], **Young-Sun Joo**[1]

[1]AI Center, NCSOFT, Seongnam, Korea
[2]SK Telecom, Seoul, Korea
[3]Samsung Research, Seoul, Korea
[4]Supertone Inc., Seoul, Korea
happyjun@ncsoft.com, ljun4121@sk.com

Abstract

Neural vocoders based on the generative adversarial neural network (GAN) have been widely used due to their fast inference speed and lightweight networks while generating high-quality speech waveforms. Since the perceptually important speech components are primarily concentrated in the low-frequency bands, most GAN-based vocoders perform multi-scale analysis that evaluates downsampled speech waveforms. This multi-scale analysis helps the generator improve speech intelligibility. However, in preliminary experiments, we discovered that the multi-scale analysis which focuses on the low-frequency bands causes unintended artifacts, e.g., aliasing and imaging artifacts, which degrade the synthesized speech waveform quality. Therefore, in this paper, we investigate the relationship between these artifacts and GAN-based vocoders and propose a GAN-based vocoder, called Avocodo, that allows the synthesis of high-fidelity speech with reduced artifacts. We introduce two kinds of discriminators to evaluate speech waveforms in various perspectives: a collaborative multi-band discriminator and a sub-band discriminator. We also utilize a pseudo quadrature mirror filter bank to obtain downsampled multi-band speech waveforms while avoiding aliasing. According to experimental results, Avocodo outperforms baseline GAN-based vocoders, both objectively and subjectively, while reproducing speech with fewer artifacts.

Introduction

Speech synthesis generates speech waveforms that correspond to the input text. An acoustic model initially generates acoustic features corresponding to the input text (Wang et al. 2017; Li et al. 2019; Ren et al. 2019, 2021). A vocoder then converts the acoustic features into a speech waveform (Masanori, Yokomori, and Ozawa 2016; van den Oord et al. 2016). With the emergence of deep learning, neural vocoders can generate high-fidelity speech waveforms that are indistinguishable from human recordings (Prenger, Valle, and Catanzaro 2019; Kim et al. 2019; van den Oord et al. 2018). Recently, vocoders based on generative adversarial network (GAN) (Goodfellow et al. 2014) with non-autoregressive convolutional architectures have been proposed (Yamamoto, Song, and Kim 2020; Kumar et al. 2019; Yang et al. 2020; Kong, Kim, and Bae 2020; Yang et al. 2021; Mustafa, Pia, and Fuchs 2021; Kim et al. 2021). Compared to other neural vocoders (van den Oord et al. 2016; Prenger, Valle, and Catanzaro 2019; Kim et al. 2019; van den Oord et al. 2018), GAN-based vocoders are faster and lighter while still maintaining a high level of synthesized speech quality. Specifically, a generator converts input features such as random noise or a mel-spectrogram into speech waveforms. A discriminator then evaluates the generated speech waveforms.

Because the speech spectrum in the low-frequency bands is much more crucial to perceptual quality, most GAN-based vocoders perform multi-scale analysis that evaluates the downsampled speech waveforms. Multi-scale analysis allows a generator to focus on the speech spectrum in low-frequency bands; downsampling limits the frequency range of speech by decreasing the sampling rate (Shannon 1949). In MelGAN (Kumar et al. 2019), a multi-scale discriminator (MSD) evaluates the downsampled waveforms that used an average pooling technique. In HiFi-GAN (Kong, Kim, and Bae 2020), a multi-period discriminator (MPD) specializing in periodic components was proposed. It discriminates downsampled waveforms obtained by using an equally spaced sampling technique with various periods. Consequently, these GAN-based vocoders have successfully increased the quality of synthesized speech (Yang et al. 2020; Kim et al. 2021; Jang et al. 2021).

In preliminary experiments, however, we discovered that, GAN-based vocoders suffer from two major issues. The first issue is the artifacts caused by the upsampling layer (Pons et al. 2021). For example, artifacts in high-frequency bands degrade the quality of speech by introducing noise. The second issue is the degraded reproducibility of the harmonic components. The fundamental frequency (F_0) of synthesized speech is often inaccurate (Morrison et al. 2022) with the aliasing during simple downsampling, such as an average pooling or an equally spaced sampling, being one of the reasons behind this problem. These artifacts significantly reduce the perceptual quality, when synthesizing speech with large pitch variation (Lorenzo-Trueba et al. 2019; Zaïdi et al. 2022). The preceding issues are analyzed further in the following section.

To address these issues, we propose Avocodo, a GAN-based vocoder that synthesizes high-quality speech wave-

*These authors contributed equally.
†Work performed at NCSOFT.

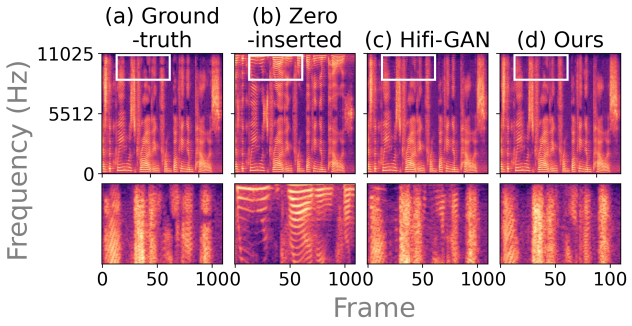


Figure 1: Sub-figures in the first row show spectrograms of (a) a ground truth, generated waveforms from (b) a zero-stuffing, (c) HiFi-GAN, and (d) proposed method. Enlarged versions of the white rectangular boxes are depicted in the second row; note that mirrored low frequencies in (b) still exist in results from (c), but not from (d).

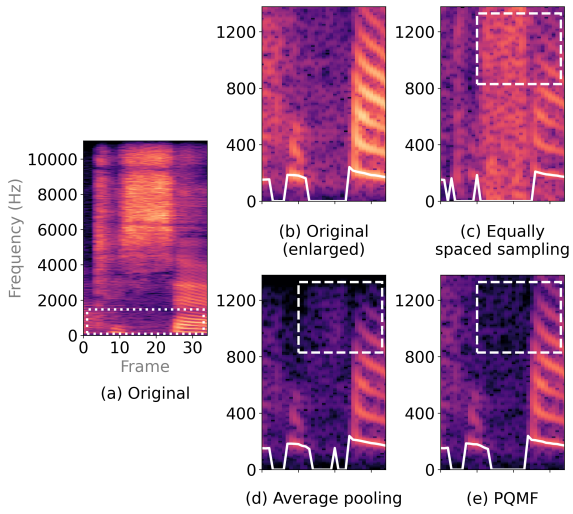


Figure 2: Spectrograms of original and downsampled audio samples. Downsampling is performed for (a,b) the original waveform with (c) the equally spaced sampling, (d) the average pooling, and (e) PQMF. White solid lines are the F_0 contours.

form by minimizing artifacts. Avocodo is designed to factor in and suppress artifacts that should be considered in the digital signal processing. The Avocodo contains two discriminators; a collaborative multi-band discriminator (CoMBD) and a sub-band discriminator (SBD). (1) The CoMBD comprises a novel structure for multi-scale analysis and suppressing upsampling artifacts. Since the CoMBD discriminates full-resolution waveform and intermediate outputs altogether, it takes two advantages. First, it helps the generator to focus on the spectral features in low-frequency bands by the multi-scale analysis. Second, the generator is trained to learn to suppress artifacts caused by the upsampling layer. (2) The SBD improves the sound quality by discriminating frequency-wise decomposed waveforms. It allows the generator to learn the speech spectrum, not only in low-frequency bands, but in high-frequency bands as well. In addition, to further improve the sound quality, we utilize a pseudo quadrature mirror filter bank (PQMF) (Nguyen 1994) equipped with high stopband attenuation as a downsampling method, as opposed to the simple downsampling methods that cause aliasing, which are commonly used in conventional GAN-based vocoders.

We evaluated Avocodo's performance with both objective and subjective evaluations. The subjective evaluation shows that Avocodo can synthesize high-quality speech and be robust in unseen speaker synthesis. In addition, in the objective evaluation, accuracy in F_0 reconstruction and the quality of high-frequency bands are improved. Finally, an analysis on the effect of alias-free methods on the artifacts is described.

Artifacts in GAN-Based Vocoders

Upsampling Artifacts

GAN-based vocoders incorporate upsampling layers to increase the rate of input features, such as a mel-spectrogram, up to the sampling rate of waveform (Kumar et al. 2019; Kong, Kim, and Bae 2020; Yang et al. 2020). However, upsampling layer, such as transposed convolution, causes several artifacts, tonal artifacts (Pons et al. 2021) being one example. Tonal artifacts appears as a horizontal line on spectrogram. Additionally, mirrored low frequencies are observed in high-frequency bands, which are called *imaging artifacts* in this paper. In digital signal processing, the signal is upsampled by inserting zeros between neighboring samples, and then applying low-pass filtering (Schafer and Rabiner 1973). Without the filtering, low-frequency components appear in high-frequency bands, as shown in Figure 1b, because the spectrum repeats over a cycle of sampling rate. The upsampling layer should also remove enough of the unintended frequency components, but it is unable to meet that criteria. As shown in Figure 1c, the unintended frequency components, which are imaging artifacts, eventually degrade the speech quality by distortion in high-frequency bands. Such artifacts are similar to texture sticking of image generative models, reported in (Karras et al. 2021).

To address these artifacts in GAN-based synthesis, several studies have proposed modifying the structure of the upsampling layer (Pons et al. 2021; Karras et al. 2021; Donahue, McAuley, and Puckette 2019; Stoller, Ewert, and Dixon 2018). However, these methods either increase the model complexity or are insufficient in suppressing artifacts. Therefore in this study, a novel discriminator and loss functions that do not modify the upsampling layer are designed to suppress artifacts.

Aliasing in Downsampling

GAN-based vocoders use discriminators to evaluate downsampled waveforms to learn the spectral information in low-frequency bands. Typical downsampling methods, such as the average pooling used in (Kumar et al. 2019; Kong, Kim, and Bae 2020; Yang et al. 2020) or the equally spaced sampling used in (Kong, Kim, and Bae 2020; Kim et al. 2021;

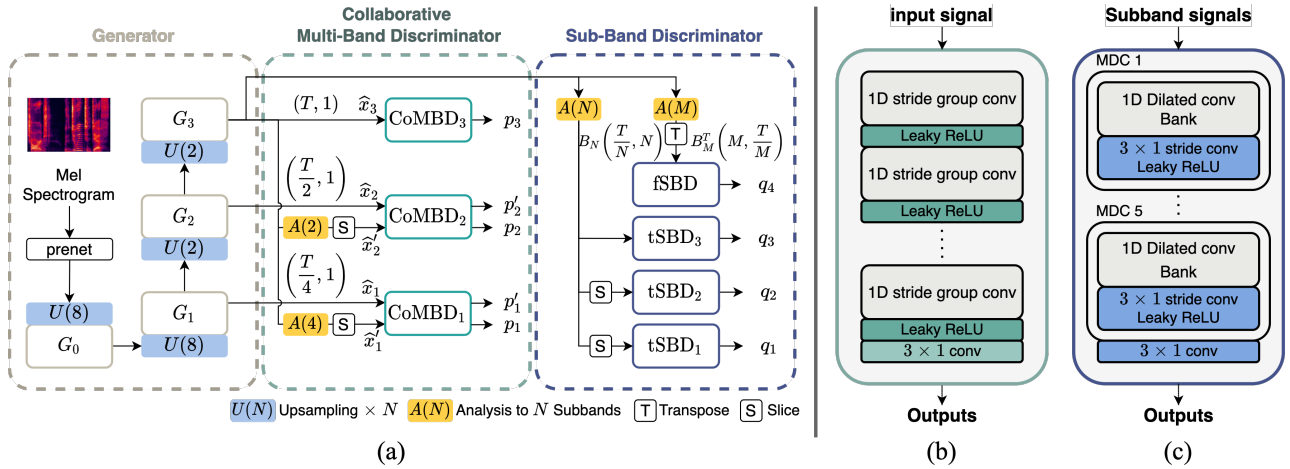


Figure 3: Overall architecture of Avocodo (a). Avocodo comprises the generator and two discriminators: CoMBD and SBD. Detailed architectures of each sub-module of CoMBD and SBD are depicted in (b) and (c), respectively.

Jang et al. 2021), are easy to implement and efficient for obtaining band-limited speech waveforms. In preliminary experiments, however, aliasing was observed in the downsampled waveforms using the aforementioned methods. Figure 2 illustrates examples of the downsampled waveforms using several approaches; the downsampling factor is set to 8. Taking into consideration downsampling uses equally spaced sampling (Figure 2c), high-frequency components, which are supposed to be removed, fold back and distort the harmonic frequency components at low-frequency bands. In the case of the average pooling (Figure 2d), which is a composition of a simple low-pass filtering and a decimation, aliasing is not as apparent in low-frequency bands but harmonic components over 800Hz are distorted. As the downsampling factor increases, the artifacts increase too. Using these distorted downsampled waveforms during the training makes it difficult for the model to generate accurate waveforms.

To avoid this aliasing, downsampling using a band-pass filter equipped with a high stopband attenuation is required. PQMF, a digital filter, satisfies this requirement (Yu et al. 2020; Yang et al. 2021). As shown in Figure 2d, downsampling using the PQMF preserves the harmonics well.

Proposed Method

Figure 3a describes the overall architecture of Avocodo. It consists of a single generator and the two proposed discriminators. Taking a mel-spectrogram as input, the generator outputs not only full-resolution waveforms but also intermediate outputs. Subsequently, the CoMBD discriminates the full-resolution waveform and its downsampled waveforms along with the intermediate outputs; the PQMF is used as a low-pass filter to downsample the full-resolution waveform. Additionally, the SBD discriminates sub-band signals obtained by the PQMF analysis.

Generator

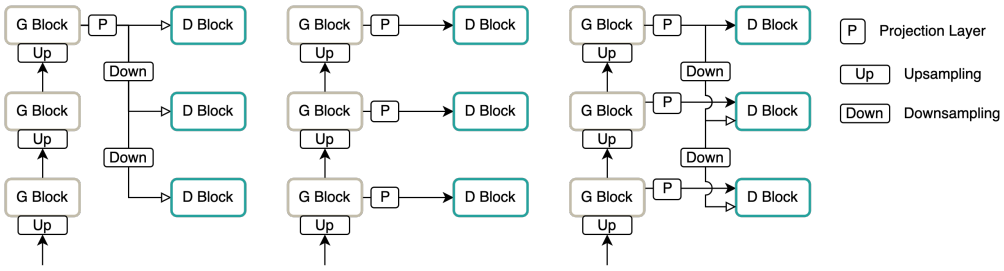
The generator has the same structure as the HiFi-GAN generator, but it produces multi-scale outputs that is composed of both high-resolution and intermediate waveforms. The generator has four sub-blocks, three of which $G_k (1 \leq k \leq 3)$ generate waveforms $\hat{x}_k$ with the corresponding resolution of $\frac{1}{2^{3-k}}$ of the full resolution. To elaborate, $\hat{x}_3$ is a full-resolution waveform; moreover, $\hat{x}_1$ and $\hat{x}_2$ denote intermediate outputs. Each sub-block comprises multi-receptive field fusion (MRF) blocks (Kong, Kim, and Bae 2020) and transposed convolution layers. The MRF blocks contain multiple residual blocks of diverse kernel sizes and dilation rates to capture the spatial features of input. Additional projection layers are added, unlike HiFi-GAN, after each sub-block to return the intermediate outputs. Please note that our approach can be applied to any GAN-based vocoder using upsampling layers. In this paper, HiFi-GAN's generator is selected due to its acceptable performance.

Collaborative Multi-Band Discriminator

The proposed CoMBD discriminates multi-scale outputs from the generator. It comprises identical sub-modules, which evaluate waveforms at different resolutions. Additionally, each sub-module is based on the discriminator module of MSD. The module comprises fully convolutional layers and a leaky ReLU activation function.

Either a multi-scale structure (Figure 4a) or a hierarchical structure (Figure 4b) is commonly used in conventional GAN-based neural vocoders; however, in this paper, the two structures are combined to take advantage of each structure, as shown in Figure 4c. This collaborative structure helps the generator to synthesize high-quality waveforms with reduced artifacts.

The multi-scale structure increases speech quality by discriminating not only the full-resolution waveform but also the downsampled waveform (Kumar et al. 2019; Kong, Kim, and Bae 2020; Yang et al. 2021; Jang et al. 2021). In particular, the discrimination of waveforms downsampled into multiple scales helps the generator to focus on the spectral features in low-frequency bands (Kumar et al. 2019). Meanwhile, the hierarchical structure uses intermediate out-



(a) Multi-scale (b) Hierarchical (c) Collaborative

Figure 4: Comparison of various structures for discriminators.

put waveforms of each generator sub-block, helping the generator to learn the various levels of acoustic properties in a balanced manner (Yang et al. 2020; Zhang, Xie, and Yang 2018). In particular, the generator sub-blocks are trained to learn expansion and filtering in a balanced way by inducing the sub-blocks of the generator to generate a band-limited waveform. Therefore, upsampling artifacts are expected to be suppressed by adopting the hierarchical structure.

For the proposed collaborative structure, the sub-modules at low resolution, i.e., CoMBD_1 and CoMBD_2, take both the intermediate outputs $\hat{x}$ and the downsampled waveforms $\hat{x}'$ as their inputs. For each resolution, both inputs share the sub-module. For example, as shown in Figure 3, the intermediate output $\hat{x}_2$ and the downsampled waveform $\hat{x}'_2$ share the weights of CoMBD_2 for output p_2 and p'_2, respectively. The intermediate output and downsampled waveform are intended to match each other after collaboration. Note that no additional parameters are necessary for collaborating the two structures because of the weight-sharing process (Liu and Tuzel 2016).

To further improve speech quality by reducing artifacts, a differentiable PQMF is adopted to obtain downsampled waveform with restricted aliasing. First, a full-resolution speech waveform is decomposed into K sub-band signals B_K by using the PQMF analysis. The B_K comprise single-band signals $b_1, ..., b_K$ with a length of $\frac{T}{K}$, where T is the length of the full-resolution waveform (Nguyen 1994). Next, the first sub-band signal b_1 is selected corresponding to the lowest frequency band.

Sub-Band Discriminator

An SBD is introduced that discriminates multiple sub-band signals by PQMF analysis. The PQMF enables the n^{th} sub-band signal b_n to contain frequency information corresponding to the range from $(n-1)f_s/2N$ to $nf_s/2N$, where f_s is the sampling frequency and N is the number of sub-bands. Inspired by this characteristic of sub-band signals, the SBD sub-modules learn various discriminative features by using different ranges of the sub-band signals.

Two types of sub-modules are designed: one captures the changes in spectral feature over the time axis and the other captures the relationship between each sub-band signal. These two sub-modules are referred to as tSBD and fSBD, respectively, in Figure 3a. The tSBD takes B_N as its input and performs time-domain convolution with B_N. By diversifying sub-band ranges, each sub-module can be designed to learn the characteristics of a specific frequency range. In other words, tSBD_k takes a certain number of sub-band signals $b_{i_k:j_k}$ as its input. In contrast, fSBD takes the transposed version of M channel sub-bands ${B_M}^T$. The composition of fSBD is inspired by the spectral features of speech waveform, such as harmonics and formants.

Each sub-module of the SBD comprises stacked multi-scale dilated convolution banks (Brock et al. 2017) to evaluate sub-band signals. The dilated convolution bank contains convolution layers with different dilation rates that cover diverse receptive fields. Moreover, the SBD architecture follows an inductive bias as an accurate analysis on speech waveforms requires various receptive fields for each frequency range. Consequently, different dilation factors are prepared for each sub-module.

Several neural vocoders utilize filter-banks to decompose speech waveforms and utilize discriminators to inspect the sub-band signals (Yang et al. 2021; Mustafa, Pia, and Fuchs 2021; Kim et al. 2021). In particular, the SBD is similar to the filter-bank random window discriminators (FB-RWDs) of the StyleMelGAN (Mustafa, Pia, and Fuchs 2021) as both obtain sub-band signals using the PQMF. However, SBD and FB-RWDs are considerably different. Each sub-module of SBD evaluates a different range of sub-band signals, whereas the FB-RWDs vary the number of sub-band signals for each discriminator. In addition, the SBD has many types of blocks: blocks used to observe a lower frequency band, a whole range of frequency bands, and a relationship between frequency bands. Consequently, the SBD can evaluate signals more effectively than FB-RWDs.

Training Objectives

GAN Loss For training GAN networks, the least square adversarial objective (Mao et al. 2017) is used, which replaces a sigmoid cross-entropy term of the GAN training objective proposed in (Goodfellow et al. 2014) with the least square for stable GAN training. The GAN losses, V for multi-scale outputs and W for downsampled waveforms, are defined as follows:

$$V(D_k; G) = \mathbb{E}_{(x_k,s)}\Big[(D_k(x_k)-1)^2 + (D_k(\hat{x}_k))^2\Big] \quad (1)$$

$$V(G;D_k) = \mathbb{E}_s\Big[(D_k(\hat{x}_k)-1)^2\Big] \quad (2)$$

$$W(D_k;G) = \mathbb{E}_{(x_k,s)}\Big[(D_k(x_k)-1)^2 + (D_k(\hat{x}'_k))^2\Big] \quad (3)$$

$$W(G;D_k) = \mathbb{E}_s\Big[(D_k(\hat{x}'_k)-1)^2\Big], \quad (4)$$

where x_k represents the k^{th} downsampled ground-truth waveform, and s denotes the speech representation. In this paper, mel-spectrogram is utilized.

Feature Matching Loss Feature matching loss is a perceptual loss for GAN training (Salimans et al. 2016), which has been used in GAN-based vocoder systems (Kumar et al. 2019; Kong, Kim, and Bae 2020; Yang et al. 2020). Moreover, the feature matching loss of a sub-module in the discriminator can be established with L1 differences between the intermediate feature maps of the ground-truth and predicted waveforms. The loss can be defined as follows:

$$L_{fm}(G;D_t) = \mathbb{E}_{(x,s)}\left[\sum_{t=1}^{T}\frac{1}{N_t}||D_t(x) - D_t(\hat{x})||\right], \quad (5)$$

where T denotes the number of layers in a sub-module. D_t and N_t represent the t^{th} feature map and the number of elements in feature map, respectively.

Reconstruction Loss Reconstruction loss based on a mel-spectrogram increases the stability and efficiency in the training of waveform generation (Yamamoto, Song, and Kim 2020). For that, L1 differences are calculated between the mel-spectrograms of the ground-truth x and predicted $\hat{x}$ speech waveforms. The reconstruction loss can be expressed as follows:

$$L_{spec}(G) = \mathbb{E}_{(x,s)}\Big[||\phi(x) - \phi(\hat{x})||_1\Big]. \quad (6)$$

where $\phi(\cdot)$ denotes the transform function to mel-spectrogram.

Final Loss Final loss for the overall system training can be established from the aforementioned loss terms and defined as follows:

$$\begin{aligned} L_D^{total} = &\sum_{p=1}^{P} V(D_p^C;G) \\ &+ \sum_{p=1}^{P-1} W(D_p^C;G) + \sum_{q=1}^{Q} V(D_q^S;G) \end{aligned} \quad (7)$$

$$\begin{aligned} L_G^{total} = &\sum_{p=1}^{P}\Big[V(G;D_p^C) + \lambda_{fm}L_{fm}(G;D_p^C)\Big] \\ &+ \sum_{p=1}^{P-1}\Big[W(G;D_p^C) + \lambda_{fm}L_{fm}(G;D_p^C)\Big] \\ &+ \sum_{q=1}^{Q}\Big[V(G;D_q^S) + \lambda_{fm}L_{fm}(G;D_q^S)\Big] \\ &+ \lambda_{spec}L_{spec}(G), \end{aligned} \quad (8)$$

where D_p^C and D_q^S denote the p^{th} sub-module of CoMBD and the q^{th} sub-module of SBD, respectively. λ_{fm} and λ_{spec} denote the loss scales for feature matching and reconstruction losses, respectively. λ_{fm} and λ_{spec} are set as 2 and 45, respectively.

Experimental Setup

Datasets

Single Speaker Speech Synthesis The LJSpeech (Ito and Johnson 2017) dataset was used for a single speaker experiment. The dataset contains 13,100 audio samples recorded by a native English-speaking female speaker, which amounted to a total recording time of 24h and the audio samples are sampled at 22,050Hz with 16bit. For the testset, 150 samples are randomly selected.

Unseen Speaker Speech Synthesis Public English dataset, i.e., VCTK (Yamagishi, Veaux, and MacDonald 2019), (Unseen(EN)) and internal Korean dataset (Unseen(KR)) were used to evaluate the generalization of the proposed model. VCTK consists of audio samples recorded by 109 speakers, and the total amount of samples is 44h long. 9 speakers were selected for the testset. The internal Korean dataset contains 156 speakers, amounting to an approximately 244h long recording. Among them, 16 unseen speakers were excluded from the training. The voice style of dataset comprises a variety of reading, daily conversations, and acting. The audio samples of VCTK and internal datasets were resampled at 24,000Hz and 22,050Hz, respectively.

Training Setup

As the baseline models, we selected HiFi-GAN[1], VocGAN, and StyleMelGAN[2]. HiFi-GAN utilizes discriminators based on multi-scale structure downsampling with average pooling and equally spaced sampling. VocGAN uses a discriminator based on hierarchical structure downsampling with average pooling. StyleMelGAN utilizes discriminators that discriminate the sub-band signals of random window selected signal obtained by PQMF analysis. For the single speaker speech synthesis, Avocodo[3] and HiFi-GAN were both trained up to 3M steps. VocGAN and StyleMelGAN were trained up to 2.5M and 1.5M steps, respectively. Next, for the unseen speaker synthesis, all models were trained up to 1M steps.

The hyper-parameters of Avocodo's generator are the same as that of the HiFi-GAN. The HiFi-GAN generator has two versions with an identical architecture but different number of parameters: $V1$ is larger than $V2$, and Avocodo also follows this rule. The number of sub-bands N is 16 for tSBD and M is 64 for fSBD. Moreover, the parameters of PQMF were selected empirically. An AdamW optimizer (Loshchilov and Hutter 2019) was used with an initial learning rate of 2×10^{-4}. The optimizer parameters

[1] https://github.com/jik876/hifi-gan

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

[3] Source code is available at https://github.com/ncsoft/avocodo.

Model	MOS (CI)			# G Param (M)	# D Param (M)	Inference Speed (CPU)	Inference Speed (GPU)
	LJ	Unseen(EN)	Unseen(KR)				
Ground Truth	4.362($\pm$0.07)	4.173($\pm$0.08)	4.690($\pm$0.05)	-	-	-	-
VocGAN	4.135($\pm$0.08)	3.638($\pm$0.09)	3.770($\pm$0.07)	7.06	12.03	18.20x	235.00x
StyleMelGAN	3.663($\pm$0.09)	3.597($\pm$0.09)	1.990($\pm$0.08)	3.55	5.90	14.33x	180.09x
HiFi-GAN $V1$	4.150($\pm$0.08)	3.940($\pm$0.09)	3.810($\pm$0.06)	13.94	70.72	15.95x	157.54x
Avocodo $V1$	**4.258($\pm$0.08)**	**4.080($\pm$0.08)**	**3.972($\pm$0.07)**	13.94	27.07	15.45x	156.21x

Table 1: MOS results with 95% CI, the number of parameters and inference speed of CPU and GPU.

Single speaker speech synthesis								
Model	F_0 RMSE(↓)	F_0 AE-STD(↓)	VUV_{fpr}(↓)	VUV_{fnr}(↓)	MCD(↓)	PESQ(↑)	LSD-LF(↓)	LSD-HF(↓)
VocGAN	37.51	38.19	20.15	12.45	2.63	3.25	7.61	9.50
StyleMelGAN	36.60	36.42	19.78	14.12	3.82	2.30	8.61	10.14
HiFi-GAN $V1$	35.96	37.11	18.67	11.13	2.25	3.64	7.05	9.72
Avocodo $V1$	**33.98**	**34.97**	**17.74**	**10.12**	**2.06**	**3.81**	**6.90**	**9.13**

Table 2: Objective evaluations results for the single speaker speech synthesis task.

(β_1, β_2) were set as $(0.8, 0.99)$, and an exponential learning rate decay of 0.999 was applied (Kong, Kim, and Bae 2020).

Input features are ground-truth mel-sepctrogram extracted from recorded speech waveform. 80 bands of mel-spectrograms were calculated from audio samples using the short-time Fourier transform (STFT). The STFT parameters for $22{,}050$Hz were set as $1{,}024$, $1{,}024$, 256 for the number of STFT bin, window sizes, and hop sizes, respectively. For 24kHz, the parameters were set as $2{,}048$, $1{,}200$, 300, respectively. Each audio sample was sliced with the random window selection method. The segment size was $8{,}192$, which is about 0.4s long.

Experimental Results

Audio Quality & Comparison

The performance of the proposed model for each dataset was assessed using various subjective and objective measurements[4].

Subjective Evaluation 5-scale mean opinion score (MOS) tests were conducted for single and unseen speaker syntheses. For the English dataset, 15 native English speakers and 19 native Korean speakers participated for the Korean dataset. All participants were requested to assess the sound quality of 20 audio samples randomly selected from each testset.

Table 1 lists subjective evaluation results. We can see Avocodo $V1$ performs the best performance in both single and unseen speaker synthesis tasks. For synthesizing high-quality speech waveform of unseen speakers, learning generalized characteristic of speech signals is crucial. Because artifacts inhibit the generator from learning the generalized characteristics of speech signals, Avocodo's approaches for suppressing artifacts are much more robust than baseline models. In particular, Avocodo outperforms baseline models even in Unseen(KR). Since the dataset includes various speech styles, the overall differences from the ground truth are larger than other datasets. Note that StyleMelGAN even failed to train for Unseen(KR).

Objective Evaluation Objective evaluations were conducted to quantitatively compare vocoders. To validate the reproducibility of F_0, we measured the F_0 root mean square error (F_0 RMSE) in the voiced frame. Because the artifacts exist in very short regions (only a few frames), the average value of F_0 RMSE is insufficient to represent the artifacts. Therefore, we further calculated the standard deviation of F_0 absolute error (F_0 AE-STD); a low F_0 AE-STD value means less distortion exists in harmonics. For evaluating the accuracy in voiced/unvoiced (VUV) frame, false positive and negative rates of the VUV classification (VUV_{fpr}, VUV_{fnr}) were measured. To measure the perceived quality of the synthesized speech, the mel-cepstral distortion (MCD)(Kubichek 1993) and perceptual evaluation of speech quality (PESQ)(Rix et al. 2001) were calculated. Additionally, we measured the log-spectral distance (Rabiner and Juang 1993; Han and Lee 2022) in low-frequency bands from 0Hz to 5.5kHz (LSD-LF) and in high-frequency bands from 5.5kHz to 11.02kHz Hz (LSD-HF); the low value of LSD-HF means the imaging artifacts are less.

Table 2 shows that Avocodo $V1$ also outperformed baseline models in overall results. In particular, due to the methods for suppressing upsampling artifacts, LSD-HF results show that Avocodo improves reproducibility in high-frequency bands. Avocodo also takes advantage of reduced aliasing in training. Training with aliased waveform makes it easy to distort harmonic components. Because of anti-aliasing methods of Avocodo, F_0 AE-STD and VUV errors are improved. Despite the smaller number of parameters in discriminators, Avocodo also performs better than HiFi-GAN. Inference times for these two models are almost

[4]Audio samples are available at https://nc-ai.github.io/speech/publications/Avocodo.

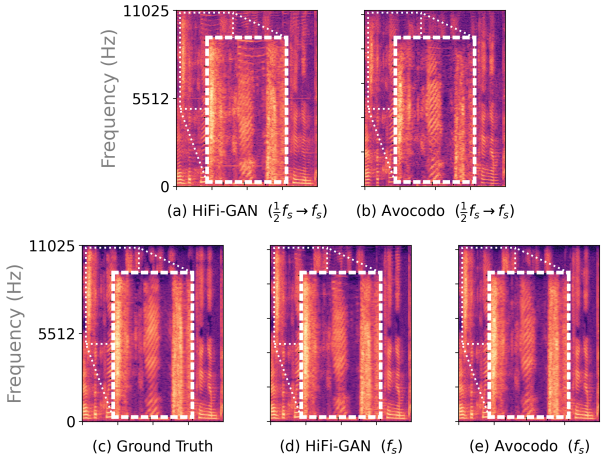


Figure 5: Linear-scale spectrograms of synthesized audio samples. In case of HiFi-GAN, artifacts from upsampling layer in (a) remain and distort final outputs as shown in (d). Meanwhile, artifacts are suppressed in (b) and final outputs (e).

Model	MOS (CI)
MSD(Kumar et al. 2019)	3.743 ($\pm$0.07)
MPD(Kong, Kim, and Bae 2020)	3.675 ($\pm$0.08)
CoMBD	4.156 ($\pm$0.05)
SBD	4.130 ($\pm$0.06)

Table 3: MOS results of the discriminator-wise comparison with 95% CI.

the same in CPU (Intel i7 CPU 3.00GHz) and single-GPU (NVIDIA V100) environments.

Discriminator-Wise Comparison

MOS test was conducted for single speaker synthesis task to compare the performances of the proposed discriminators, MSD of MelGAN and MPD of HiFi-GAN. All discriminators were trained with Avocodo's generator $V2$. λ_{fm} and λ_{spec} are observed to empirically affect the training, therefore they were adjusted to 2 and 10, respectively. However, λ_{spec} was adjusted to 20 for the CoMBD which has a larger loss value owed to weight-sharing. Table 3 shows that each Avocodo discriminator contributes to the generator synthesizing higher-quality speech with fewer artifacts.

Analysis on Artifacts

Upsampling Artifacts The ability of Avocodo to suppress artifacts is explained by observing the upsampling artifacts occurring in intermediate upsampling layers of the generator. Audio samples are generated with HiFi-GAN and Avocodo, and their linear-scale spectrograms are depicted in Figure 5. Audio samples of the first row of Figure 5 are the projected output of the last transposed convolution layer for upsampling from $\frac{1}{2}f_s$ to f_s while skipping the last MRF block; Figure 5a and Figure 5b correspond to samples from

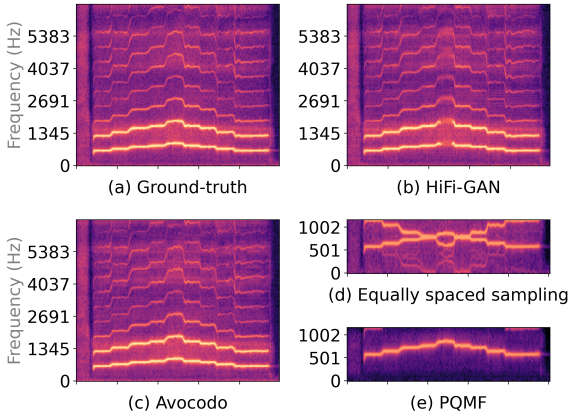


Figure 6: Examples of failed F_0 reconstruction. Due to aliasing of downsampled waveform in (d), HiFi-GAN fails to synthesize high F_0 over 750Hz (b).

HiFi-GAN and Avocodo, respectively. Meanwhile, samples of the second row of Figure 5 are obtained by the complete generator. Tonal and imaging artifacts caused by transpose convolution exist in audio samples from HiFi-GAN as shown in Figure 5a. Imaging artifacts still remain in the final output as shown in Figure 5d. However, Avocodo's generator learns to remove artifacts occurring in intermediate upsampling layers from CoMBD. Therefore, no artifacts are present in neither Figure 5b nor Figure 5e.

Aliasing To observe the distortion in F_0 caused by aliasing, we trained GAN-based vocoders with singing voice datasets with a large range of F_0. Large-scale downsampling for adequately modeling low-frequency components causes incomplete F_0 reconstruction; for example, HiFi-GAN and VocGAN downsample by a factor of up to 11 and 16, respectively. In Figure 6d, the harmonic components of the downsampled waveforms are distorted due to the aliasing caused by downsampling, while downsampled waveforms with anti-aliasing PQMF preserve F_0 as shown in Figure 6e. Therefore, HiFi-GAN (Figure 6b), trained using distorted downsampled waveforms, fails to reconstruct F_0 higher than 750Hz, while Avocodo (Figure 6c) preserves F_0 contour.

Conclusions

In this paper, an artifact-free GAN-based vocoder, Avocodo, is proposed. Artifacts, such as upsampling artifacts and aliasing, are observed to originate from the limitation of the upsampling layer and the objective function biased towards the low-frequency bands obtained by naive downsampling methods. To solve these problems, two novel discriminators, namely CoMBD and SBD, are designed. The CoMBD performs multi-scale analysis with a collaborative structure of multi-scale and hierarchical structures. The SBD discriminates the sub-band signals decomposed by PQMF analysis in both time and frequency aspects. Furthermore, PQMF is utilized for downsampling and PQMF analysis. Various experimental results proved that these discriminators and the PQMF effectively reduce the artifacts in synthesized speech.

References

Brock, A.; Lim, T.; Ritchie, J. M.; and Weston, N. 2017. Neural Photo Editing with Introspective Adversarial Networks. In *International Conference on Learning Representations*.

Donahue, C.; McAuley, J.; and Puckette, M. 2019. Adversarial Audio Synthesis. In *International Conference on Learning Representations*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. *Advances in neural information processing systems*, 27.

Han, S.; and Lee, J. 2022. NU-Wave 2: A General Neural Audio Upsampling Model for Various Sampling Rates. In *Proc. INTERSPEECH 2022*, 4401–4405.

Ito, K.; and Johnson, L. 2017. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/. Accessed: 2017.

Jang, W.; Lim, D.; Yoon, J.; Kim, B.; and Kim, J. 2021. UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation. In *Proc. INTERSPEECH 2021*, 2207–2211.

Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-Free Generative Adversarial Networks. *Advances in neural information processing systems*, 34.

Kim, J.-H.; Lee, S.-H.; Lee, J.-H.; and Lee, S.-W. 2021. FreGAN: Adversarial Frequency-Consistent Audio Synthesis. In *Proc. INTERSPEECH 2021*, 2197–2201.

Kim, S.; Lee, S.-G.; Song, J.; Kim, J.; and Yoon, S. 2019. FloWaveNet : A Generative Flow for Raw Audio. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 3370–3378. PMLR.

Kong, J.; Kim, J.; and Bae, J. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Advances in neural information processing systems*, 33: 17022–17033.

Kubichek, R. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, 125–128. IEEE.

Kumar, K.; Kumar, R.; de Boissiere, T.; Gestin, L.; Teoh, W. Z.; Sotelo, J.; de Brébisson, A.; Bengio, Y.; and Courville, A. C. 2019. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. *Advances in neural information processing systems*, 32.

Li, N.; Liu, S.; Liu, Y.; Zhao, S.; and Liu, M. 2019. Neural Speech Synthesis with Transformer Network. In *Proc. AAAI Conference on Artificial Intelligence*, volume 33.

Liu, M.-Y.; and Tuzel, O. 2016. Coupled Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, volume 29.

Lorenzo-Trueba, J.; Drugman, T.; Latorre, J.; Merritt, T.; Putrycz, B.; Barra-Chicote, R.; Moinet, A.; and Aggarwal, V. 2019. Towards Achieving Robust Universal Neural Vocoding. In *Proc. Interspeech 2019*, 181–185.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *International Conference on Computer Vision*, 2794–2802.

Masanori, M.; Yokomori, F.; and Ozawa, K. 2016. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7): 1877–1884.

Morrison, M.; Kumar, R.; Kumar, K.; Seetharaman, P.; Courville, A.; and Bengio, Y. 2022. Chunked Autoregressive GAN for Conditional Waveform Synthesis. In *International Conference on Learning Representations*.

Mustafa, A.; Pia, N.; and Fuchs, G. 2021. StyleMelGAN: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization. In *International Conference on Acoustics, Speech and Signal Processing*, 6034–6038. IEEE.

Nguyen, T. 1994. Near-perfect-reconstruction pseudo-QMF banks. *IEEE Transactions on Signal Processing*, 42(1): 65–76.

Pons, J.; Pascual, S.; Cengarle, G.; and Serrà, J. 2021. Upsampling artifacts in neural audio synthesis. In *International Conference on Acoustics, Speech and Signal Processing*, 3005–3009. IEEE.

Prenger, R.; Valle, R.; and Catanzaro, B. 2019. WaveGlow: A flow-based generative network for speech synthesis. In *International Conference on Acoustics, Speech and Signal Processing*, 3617–3621. IEEE.

Rabiner, L.; and Juang, B.-H. 1993. *Fundamentals of speech recognition*. Prentice-Hall, Inc.

Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T. 2021. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *International Conference on Learning Representations*.

Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T. 2019. FastSpeech: Fast, Robust and Controllable Text to Speech. In *Advances in Neural Information Processing Systems*, volume 32.

Rix, A. W.; Beerends, J. G.; Hollier, M. P.; and Hekstra, A. P. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *International Conference on Acoustics, Speech and Signal Processing*, volume 2, 749–752. IEEE.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. *Advances in neural information processing systems*, 29: 2234–2242.

Schafer, R. W.; and Rabiner, L. R. 1973. A digital signal processing approach to interpolation. *Proceedings of the IEEE*, 61(6): 692–702.

Shannon, C. E. 1949. Communication in the presence of noise. *Proceedings of the IRE*, 37(1): 10–21.

Stoller, D.; Ewert, S.; and Dixon, S. 2018. Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR*, 334–340.

van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A. W.; and Kavukcuoglu, K. 2016. WaveNet: A Generative Model for Raw Audio. In *ISCA Speech Synthesis Workshop*, 125. ISCA.

van den Oord, A.; Li, Y.; Babuschkin, I.; Simonyan, K.; Vinyals, O.; Kavukcuoglu, K.; van den Driessche, G.; Lockhart, E.; Cobo, L.; Stimberg, F.; Casagrande, N.; Grewe, D.; Noury, S.; Dieleman, S.; Elsen, E.; Kalchbrenner, N.; Zen, H.; Graves, A.; King, H.; Walters, T.; Belov, D.; and Hassabis, D. 2018. Parallel WaveNet: Fast High-Fidelity Speech Synthesis. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 3918–3926. PMLR.

Wang, Y.; Skerry-Ryan, R. J.; Stanton, D.; Wu, Y.; Weiss, R. J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; Le, Q. V.; Agiomyrgiannakis, Y.; Clark, R.; and Saurous, R. A. 2017. Tacotron: Towards End-to-End Speech Synthesis. In *Proc. INTERSPEECH*, 4006–4010.

Yamagishi, J.; Veaux, C.; and MacDonald, K. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. https://doi.org/10.7488/ds/2645. Accessed: 2019-11-13.

Yamamoto, R.; Song, E.; and Kim, J.-M. 2020. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *International Conference on Acoustics, Speech and Signal Processing*, 6199–6203. IEEE.

Yang, G.; Yang, S.; Liu, K.; Fang, P.; Chen, W.; and Xie, L. 2021. Multi-band MelGAN: Faster Waveform Generation for High-Quality Text-to-Speech. In *Spoken Language Technology Workshop*, 492–498. IEEE.

Yang, J.; Lee, J.; Kim, Y.; Cho, H.-Y.; and Kim, I. 2020. VocGAN: A High-Fidelity Real-Time Vocoder with a Hierarchically-Nested Adversarial Network. In *Proc. INTERSPEECH*, 200–204.

Yu, C.; Lu, H.; Hu, N.; Yu, M.; Weng, C.; Xu, K.; Liu, P.; Tuo, D.; Kang, S.; Lei, G.; Su, D.; and Yu, D. 2020. DurIAN: Duration Informed Attention Network for Speech Synthesis. In *Proc. INTERSPEECH 2020*, 2027–2031.

Zaïdi, J.; Seuté, H.; van Niekerk, B.; and Carbonneau, M.-A. 2022. Daft-Exprt: Cross-Speaker Prosody Transfer on Any Text for Expressive Speech Synthesis. In *Proc. Interspeech 2022*, 4591–4595.

Zhang, Z.; Xie, Y.; and Yang, L. 2018. Photographic Text-to-Image Synthesis with a Hierarchically-Nested Adversarial Network. In *Conference on Computer Vision and Pattern Recognition*, 6199–6208.