Avocodo: Generative Adversarial Network for Artifact-free Vocoder

Taejun Bak, Junmo Lee, Hanbin Bae, Jinhyeok Yang, Jae-Sung Bae, Young-Sun Joo

Aug. 05. 2022. (FRI)

Youngwon Choi

Speech and Audio Processing Lab.

Contents

1. Introduction

Introduction

- Speech synthesis also known as text-to-speech(TTS) generates speech waveforms that correspond to the input text.
  - At first, a TTS model generates acoustic features such as a mel-spectrogram corresponding to the input text.
  - A vocoder then converts the acoustic features into a speech waveform.
- Recently, GAN-based vocoders with non-autoregressive convolutional architectures have been propos ed.
  - Comparing the previous neural vocoders, these models are faster, lighter, and can generate high-quality wavef orms.
  - Ex) MelGAN (multi-scale discriminator), HiFi-GAN (multi-period discriminator)
- Because the speech spectrum in the low-frequency bands has a much more important impact on perc eptual quality, GAN-based vocoders perform multi-scale analysis that evaluates the downsampled wav eforms along with the full-band waveform.
  - The multi-scale analysis allows the generator to focus on the speech spectrum in the low-frequency bands.

Introduction

- However, GAN-based vocoders suffer from two major problems.
  - The first is that of the degraded reproducibility of the harmonic components.
  - The second problem is that of a lack of reproducibility at high-frequency bands.

- To address these issues, author propose a neural vocoder called Avocodo, which specializes in learning various frequency features.
  - Author propose two discriminators; a collaborative multi-band discriminator (CoMBD) and a sub-band discrimi nator (SBD).
  - Additionally, author utilize a pseudo quadrature mirror filter bank (PQMF) equipped with high stopband atten uation suppressing aliasing to obtain downsampled and decomposed waveforms in the training process.

- Owing to the proposed discriminators and the utilization of the PQMF, the generator learns exactly the speech spectrum not only in the low-frequency bands but also in the high-frequency bands.

2. Artifacts in GAN-based Vocoders

Artifacts in GAN-based Vocoders

2.1 Aliasing in downsampling

- □ GAN-based vocoders use discriminators to evaluate downsampled waveforms to learn the spectral inf ormation in low-frequency bands.
  - – Typical downsampling methods include the average pooling or the equally spaced sampling.

- □ However, aliasing was observed in the downsampled waveforms using the above methods.

Experiments

*2.1 **Aliasing** in downsampling*

- When downsampling using equally spaced sampling (Figure 1c), high-frequency components that are supposed to be removed, fold back and distort the harmonic frequency components at a low-frequency band.
- In the case of the average pooling (Figure 1d), which is a composition of a simple low-pass filtering and a decimation, aliasing is not that noticeable at a low-frequency band but harmonic components over 800Hz are distorted.
- To avoid aliasing, downsampling using a band-pass filter equipped with high stopband attenuation is required.
  - The PQMF is a digital filter that satisfies this requirement.

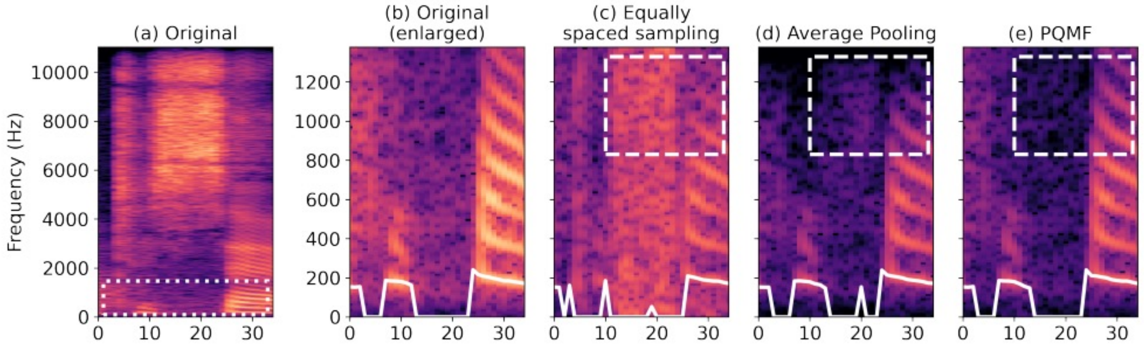

Figure 1: The spectrograms of original and downsampled audio samples. White solid lines are contours of F_0. We perform downsampling of (a) the original waveform with (c) the equally spaced sampling, (d) the average pooling, and (e) PQMF.

Experiments

2.2 Imaging artifact in upsampling

- GAN-based vocoders include upsampling layers in their structure to increase the rate of input f eatures, such as a mel-spectrogram, up to the s ampling rate of the waveform.
- During the upsampling process, low-frequency components are mirrored to the high-frequenc y bands after an expansion by zero-insertion, as shown in Figure 2b.
  - Then, in the filtering stage, these frequency com ponents should be removed.
- In GAN-based vocoders, upsampling layers such as a transposed convolution take these process es.

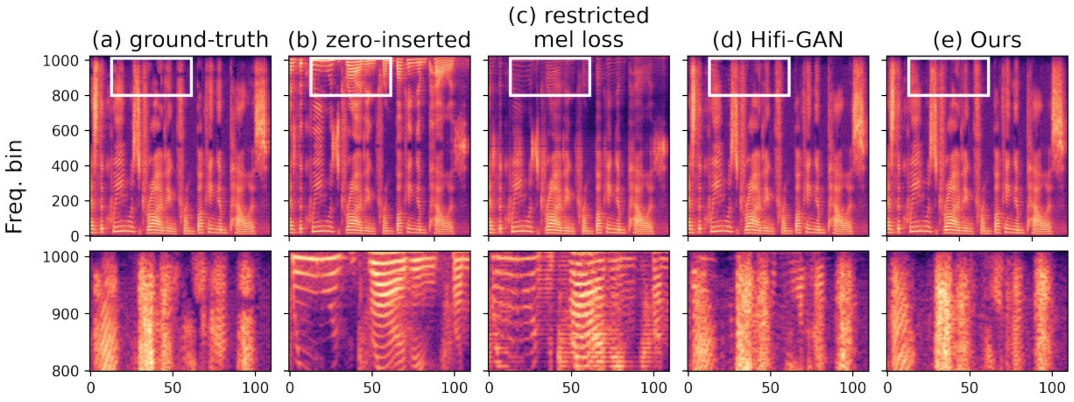

Figure 2: Sub-figures in the first row show spectrograms of (a) a ground truth, generated waveforms from (b) a zero-stuffing, (c) model trained with restricted mel-reconstruction loss, (d) HiFi-GAN, and (e) proposed methods. The enlarged version of the white rectangular box is depicted in the second row, mirrored low-frequencies in (b) still exist in results from (c) and (d), but not from (e).

Experiments

2.2 Imaging artifact in upsampling

- However, because the upsampling layers are in sufficient to remove them, unintended frequen cy components remain.
  - In this paper, we call these remained frequency c omponents at the high-frequency bands as **imagi ng artifacts**.
  - The imaging artifacts also degrade the speech qu ality, causing distortions in the high-frequency ba nd.

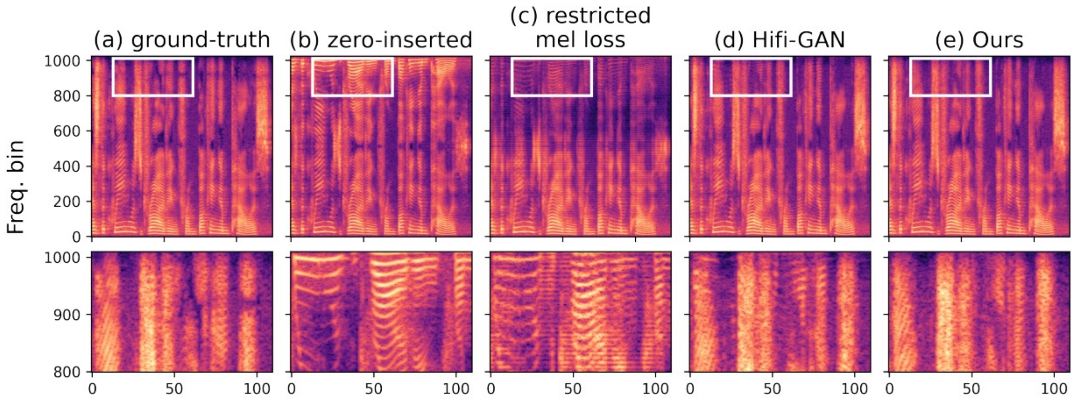

Figure 2: Sub-figures in the first row show spectrograms of (a) a ground truth, generated waveforms from (b) a zero-stuffing, (c) model trained with restricted mel-reconstruction loss, (d) HiFi-GAN, and (e) proposed methods. The enlarged version of the white rectangular box is depicted in the second row, mirrored low-frequencies in (b) still exist in results from (c) and (d), but not from (e).

3. Proposed Method

Proposed Method

- □ Avocodo has a single generator and the proposed two discriminators(CoMBD, SBD).
- □ Taking a mel-spectrogram as input, the **generator** outputs not only a full-resolution waveform but also intermediate outputs.
- □ Then 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.

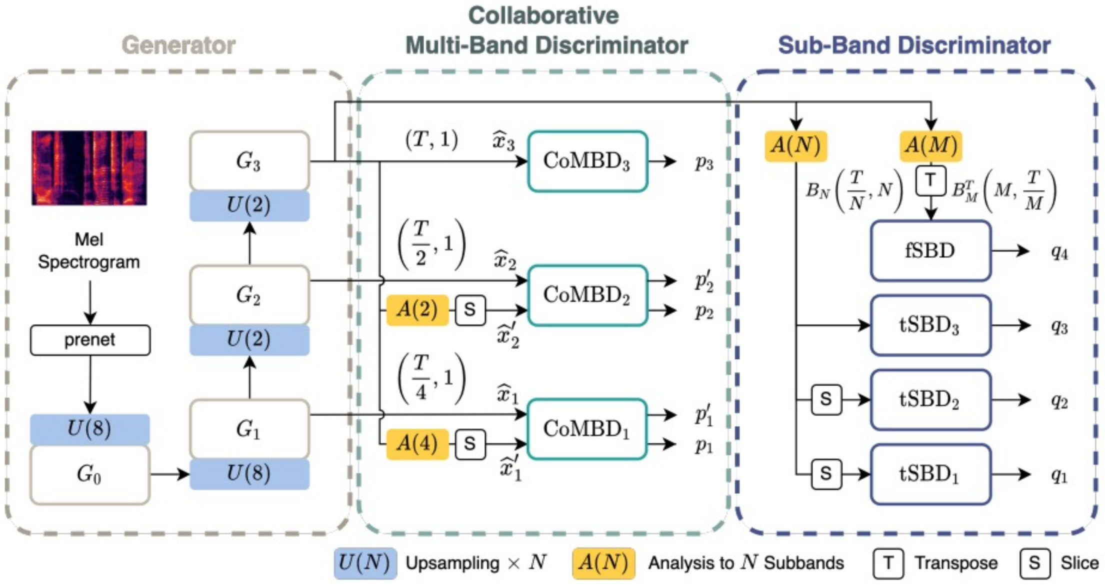

Figure 3: Overall Architecture of Avocodo.

Proposed Method

3.1 Generator

- □ The proposed generator mainly follows the stru cture of the HiFi-GAN generator.
- □ The generator has four subblocks, three of whic h $G_k(1 \leq k \leq 3)$ generate waveforms $\widehat{x^k}$ with the corresponding resolution of $\frac{1}{2^{3-k}}$ of the full-reso lution.
- □ Each sub-block is composed of multi-receptive f ield fusion (MRF) blocks and transposed convol ution layers.
  - – The MRF blocks consist of multiple residual block s of diverse kernel sizes and dilation rates to capt ure the spatial features of the input.

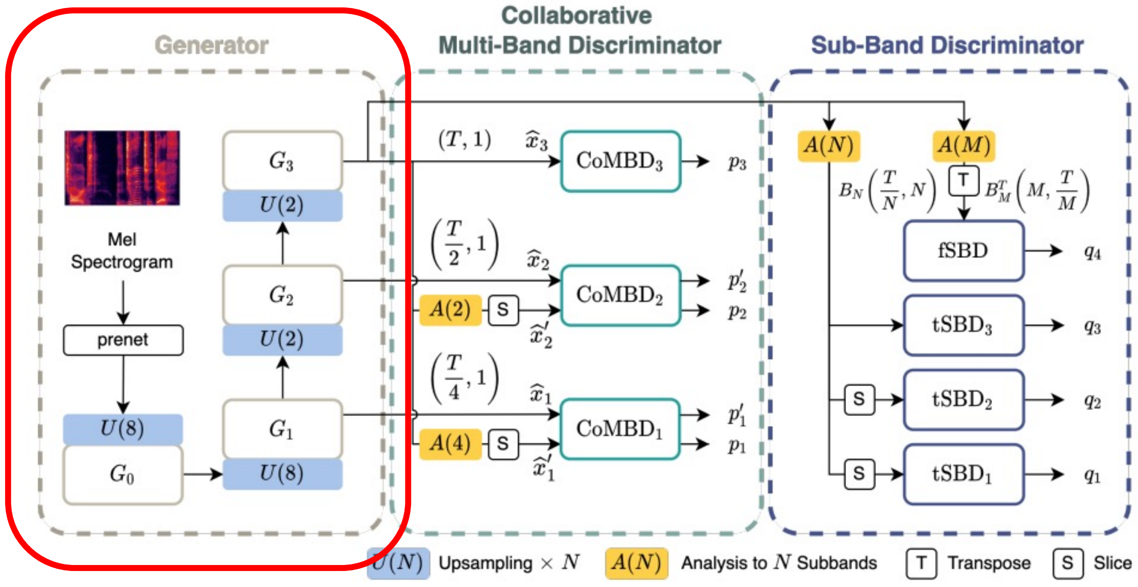

Figure 3: Overall Architecture of Avocodo.

Proposed Method

3.2 Collaborative Multi-Band Discriminator

□ In Avocodo, authors combine a multi-scale stru cture or a hierarchical structure which are com monly used in conventional GAN based neural v ocoders, respectively.

- The multi-scale structure helps the generator foc us on the spectral features in low-frequency band.
- The hierarchical structure helps the generator lea rn the various levels of acoustic properties in a ba lanced manner.
  - Suppress the imaging artifacts mentioned in 2.2

□ This collaborative structure of multi-scale and h ierarchical arrangements helps the generator sy nthesize high-quality waveforms with reduced artifacts.

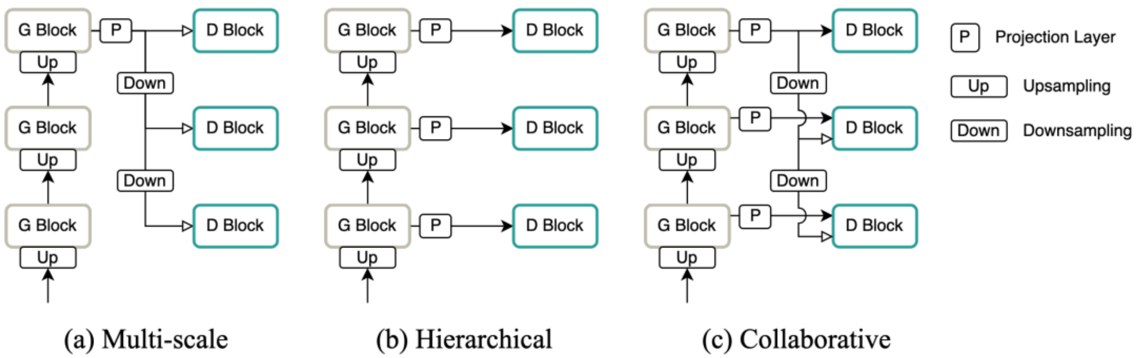

Figure 4: Comparison on various structure of discriminators.

Proposed Method

3.2 Collaborative Multi-Band Discriminator

- For the collaborative structure, the sub-modules at low resolution 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.
  - This structure intends that the intermediate output waveforms and downsampled waveforms become the same as each other.

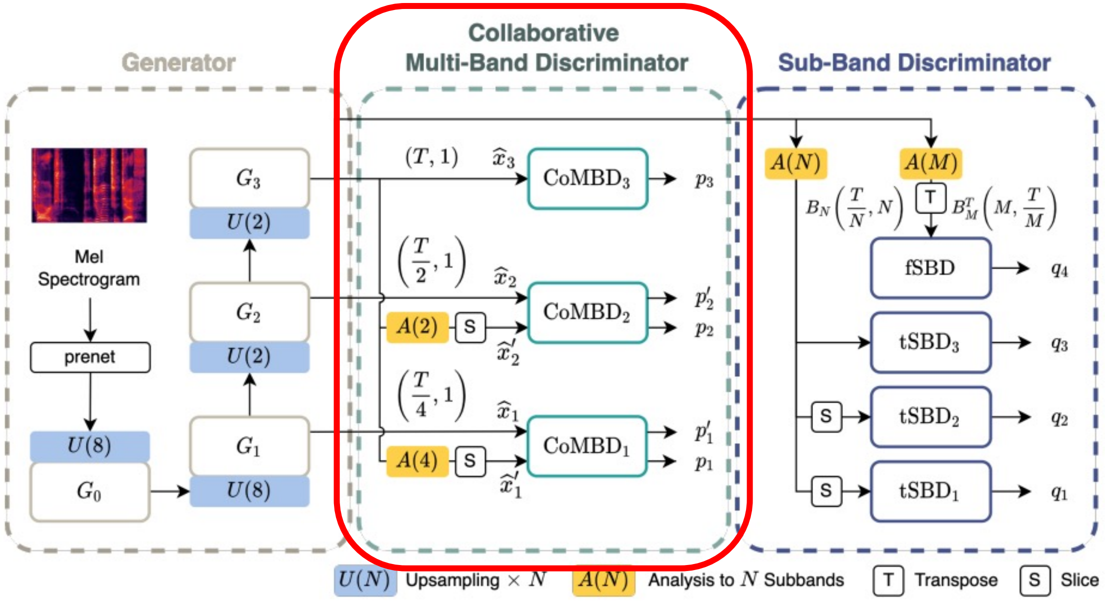

Figure 3: Overall Architecture of Avocodo.

Proposed Method

3.2 Collaborative Multi-Band Discriminator

- To further improve speech quality by reducing artifacts, authors adopt a differentiable PQMF to obtain downsampled waveform with restricted aliasing.
  - First, author decompose a full-resolution speech waveform into N sub-band signals $B_N\ (b_1, \ldots, b_4)$ by using the PQMF analysis.
  - Then, author select the first sub-band signal b_1 corresponding to the lowest frequency band.

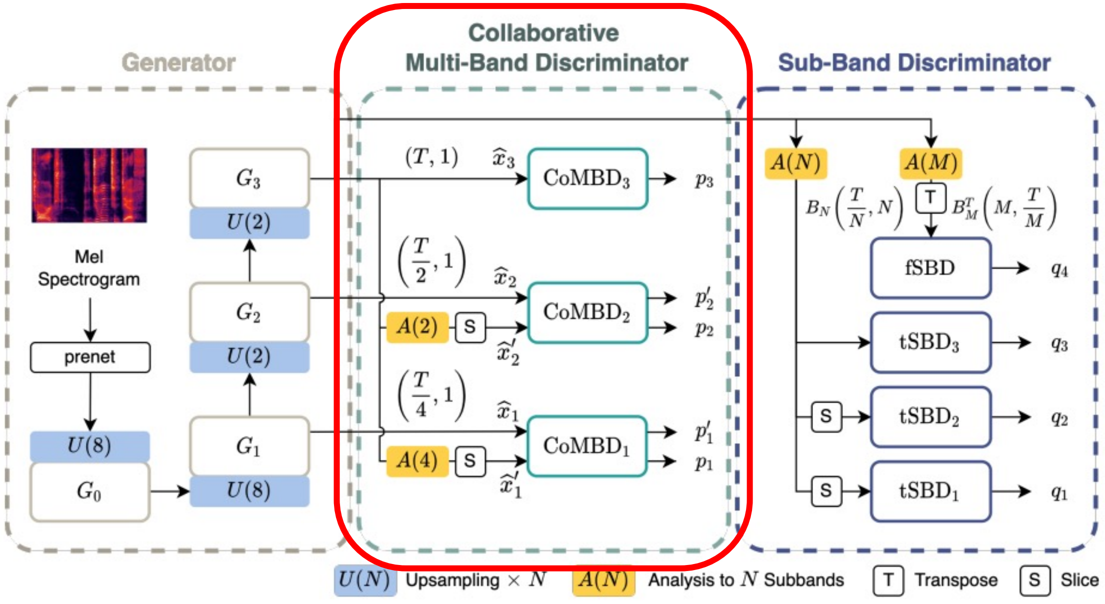

Figure 3: Overall Architecture of Avocodo.

Proposed Method

3.3 Sub-Band Discriminator

□ The PQMF enables the n th sub-band signal bn to contain frequency information correspondin g to the range from $\frac{(n-1)f_s}{2N}$ to $\frac{nf_s}{2N}$, where f_s is the sampling frequency and *N* is the number of subbands.

□ Sub-modules of the SBD learn various discrimin ative features by using different ranges of the s ub-band signals.

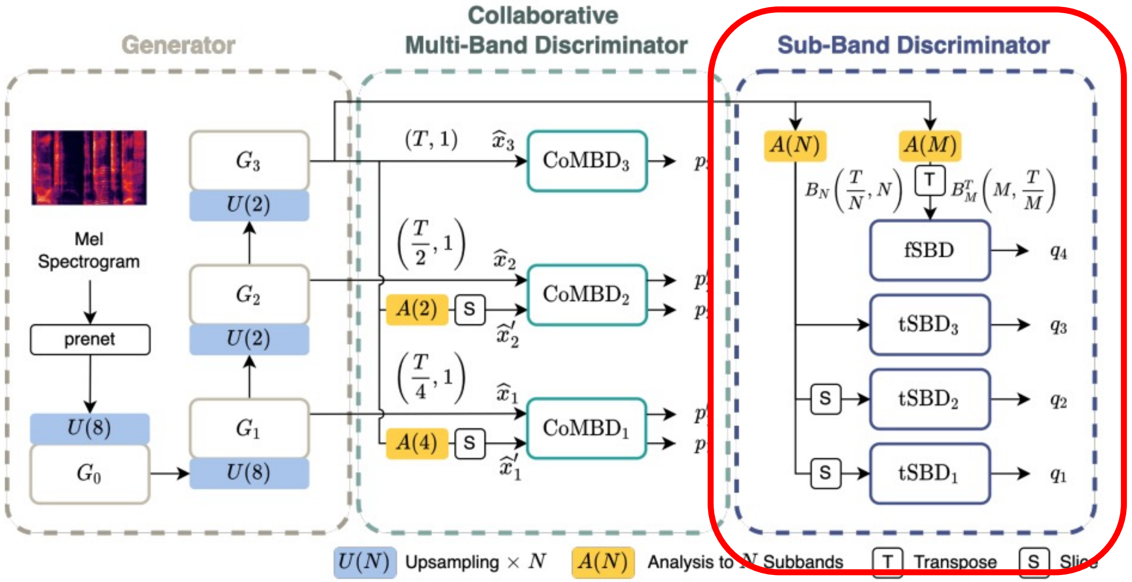

Figure 3: Overall Architecture of Avocodo.

Proposed Method

3.3 Sub-Band Discriminator

□ tSBD

- Takes B_N as its input and performs time-domain convolution with it.
- Each submodule can learn the characteristics of the specific frequency range by diversifying the sub-band ranges.

□ fSBD

- takes the transposed version of M channel sub-bands B_M^T.
- The composition of fSBD is inspired by the spectral features of the speech waveform, such as harmonics and formants.

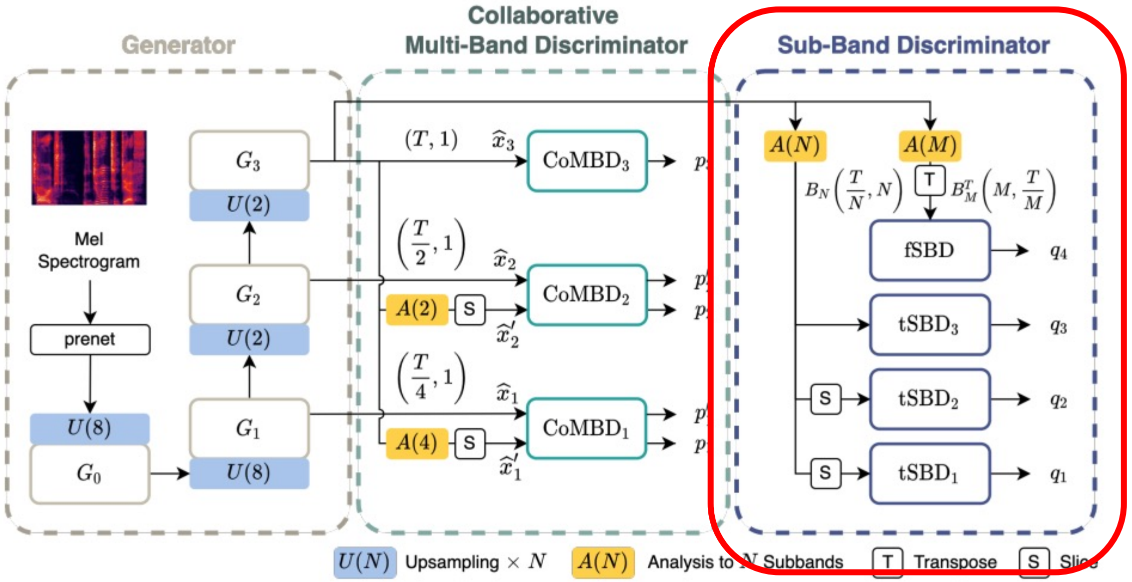

Figure 3: Overall Architecture of Avocodo.

Proposed Method

3.4 Training Objectives

Final Loss = GAN Loss + Feature Matching Loss + Reconstruction Loss

□ GAN Loss

– Author used LSGAN that replaces a sigmoid cross-entropy term of the GAN training objective with the least square for stable GAN training.

– The GAN losses V for multi-scale outputs and W for downsampled waveform are defined as follows:

$$V(D_k;G) = \mathrm{E}_{(x_k,s)}\left[(D_k(x_k)-1)^2 + \left(D_k(\hat{x}_k)\right)^2\right], W(G;D_k) = \mathrm{E}_s[(D_k(\hat{x}_k)-1)^2]$$

$$V(D_k;G) = \mathrm{E}_{(x_k,s)}\left[(D_k(x_k)-1)^2 + \left(D_k(\hat{x}'_k)\right)^2\right], W(G;D_k) = \mathrm{E}_s[(D_k(\hat{x}'_k)-1)^2]$$

where x_k represents the k th downsampled ground-truth waveform, and s denotes the speech representation.

Proposed Method

3.4 Training Objectives

□ Feature Matching Loss

- Feature matching loss, a perceptual loss for GAN training, has been used in GAN-based vocoder systems.
- Feature matching loss can be defined as follows:

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

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

Proposed Method

3.4 Training Objectives

□ Reconstruction Loss

- Reconstruction loss based on a mel-spectrogram increases the stability and efficiency in the training of wavefo rm generation.
- Reconstruction loss can be defined as follows:

$$L_{spec}(G) = \mathrm{E}_{x,s}[\|\phi(x) - \phi(\hat{x})\|_1],$$

where ϕ represents a function of the transform to the mel-spectrogram.

Proposed Method

3.4 Training Objectives

□ Final Loss

Final Loss = GAN Loss + Feature Matching Loss + Reconstruction Loss

– Final loss for the overall system training can be established from the aforementioned loss terms and defined a s follows:

$$L_D^{total} = \sum_{p=1}^{P} V\left(D_p^C; G\right) + \sum_{p=1}^{P-1} W\left(D_p^C; G\right) + \sum_{q=1}^{Q} W\left(D_q^S; G\right)$$

$$L_G^{total} = \sum_{p=1}^{P} \left[V\left(G; D_p^C\right) + \lambda_{fm} L_{fm}\left(G; D_p^C\right)\right] + \sum_{p=1}^{P-1} \left[W\left(G; D_p^C\right) + \lambda_{fm} L_{fm}\left(G; D_p^C\right)\right] + \sum_{q=1}^{Q} \left[V\left(G; D_Q^S\right) + \lambda_{fm} L_{fm}\left(G; D_q^S\right)\right] + \lambda_{spec} L_{sepc}(G),$$

where D_p^C and D_p^C denote pth sub-module of CoMBD and qth sub-module of SBD, respectively.

– In this paper, $\lambda_{fm} = 2$ and $\lambda_{spec} = 45$.

4. Experiments

Experiments

4.1 Datasets

- □ Single speaker speech synthesis: LJSpeech dataset
  - – Recorded by native English-speaking female speaker with total amount of 24 hours.
  - – Contains 13,100 audio samples, 150 samples taken for the test dataset.
- □ Singing voice synthesis: internal dataset
  - – Contains of about 8500 samples recorded by 16 speakers.
- □ Unseen speaker speech synthesis: internal multi-speaker Korean dataset
  - – Contains 156 gender-balanced speakers with amount of about 244 hours long.
  - – 16 (unseen) speakers were excluded from training.
  - – Datasets sampled at 22,050Hz, 16bit PCM. ㅇ

Experiments

4.2 Training Setup

□ Baseline model

– HiFi-GAN, VocGAN

□ Data processing

– Calculated 80 bands of mel-spectrograms.

  ▪ STFT parameters: 1024(FFT), 1024(window), 245(hop size)

– Segmentation

  ▪ 8192 samples (0.4 seconds long)
  ▪ 65,536 (3 seconds long) for singing dataset due to a long vowel duration.

□ Model

– Trained for 3M steps.

– Used AdamW optimizer ($\beta_1 = 0.8, \beta_2 = 0.99$)

  ▪ Used exponential learning rate decay(0.999) with initial learning rate of 0.002

– HiFi-GAN and Avocodo have two version;

  ▪ V1 is larger than V2.

– The number of sub-band N is 16 for tSBD and is M = 64 for fSBD.

– The parameters of the PQMF were empirically sel ected.

5. Results

Results

5.1 Audio Quality & Comparison

□ Subjective Measure: MOS, CMOS

- 19 native English speakers participated via Amazon Mechanical Turk for English dataset
- 19 native Korean speakers participated for Korean datasets.

□ Objective Measure

- Measured *F0* RMSE(root mean square error), false positive and negative rate of the voice/unvoice classificatio n (VUV_{fpr}, VUV_{fnr}), to validate the reproducibility of the fundamental frequency.
- Calculate the mel-cepstral distortion (MCD), structural similarity index (SSIM), perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) to measure the perceived quality of the synthesize d speech.

Results

□ Single speaker speech synthesis & Unseen speaker synthesis.

Table 1: The results of subjective evaluations with 95% CI, the number of parameters and inference speed on CPU and GPU.

Model	MOS (CI)		# G Param (M)	# D Param (M)	Inference Speed (CPU)	Inference Speed (GPU)
	LJ	Unseen				
Ground Truth	4.373±0.06	4.562±0.05	-	-	-	-
VocGAN	4.162±0.06	4.049±0.07	7.06	12.03	3.26x	235.0x
HiFi-GAN $V1$	4.270±0.06	3.709±0.07	13.94	70.72	2.98x	157.6x
HiFi-GAN $V2$	4.010±0.06	3.516±0.07	0.93	-	10.56x	541.2x
Avocodo $V1$	**4.285±0.06**	**4.051±0.06**	13.94	27.07	2.93x	156.3x
Avocodo $V2$	4.087±0.06	3.558±0.07	0.93	-	10.09x	539.6x

Table 2: Results of objective evaluations

Single speaker speech synthesis							
Model	F_0 RMSE(↓)	MCD(↓)	VUV_{fpr}(↓)	VUV_{fnr}(↓)	SSIM(↑)	PESQ(↑)	STOI(↑)
VocGAN	37.51	2.63	20.154	12.445	0.882	3.25	0.9614
HiFi-GAN $V1$	35.96	2.25	18.670	11.133	0.939	3.64	0.9819
HiFi-GAN $V2$	37.26	2.86	20.618	12.174	0.878	2.98	0.9648
Avocodo $V1$	**33.98**	**2.06**	**17.741**	**10.115**	**0.953**	**3.81**	**0.9866**
Avocodo $V2$	37.63	2.59	20.691	11.478	0.899	3.11	0.9709

Results

□ Singing voice synthesis

Singing voice synthesis

Model	F_0 RMSE(↓)	MCD(↓)	$\mathrm{VUV}_{\mathrm{fpr}}$(↓)	$\mathrm{VUV}_{\mathrm{fnr}}$(↓)	SSIM(↑)	PESQ(↑)	STOI(↑)
HiFi-GAN $V1$	27.86	2.67	12.075	2.044	0.9155	3.48	**0.8125**
Avocodo $V1$	**26.88**	**2.42**	**10.57**	**1.74**	**0.931**	**3.55**	0.8052

Table 3: CMOS results of singing voice synthesis with 95% CI.

(−)	CMOS (CI)	(+)
HiFi-GAN $V1$	0.403 (±0.06)	Avocodo $V1$

Results

5.2 Ablation Study

Table 5: Results of objective evaluations for ablation study. Every models were trained with the generator of $V2$.

Model		F_0 RMSE(↓)	MCD(↓)	VUV_{fpr}(↓)	VUV_{fnr}(↓)	SSIM(↑)	PESQ(↑)	STOI(↑)
MSD[11]		36.45	3.91	21.29	12.33	0.830	2.58	0.951
MPD[12]		37.02	3.91	22.15	12.07	0.840	2.62	0.953
AP	Multi-scale	39.26	2.93	22.65	12.06	0.867	2.72	0.961
AP	Hierarchical	39.65	3.12	22.29	12.10	0.842	2.60	0.956
PQMF	Multi-scale	38.30	3.02	21.78	11.70	0.855	2.70	0.959
PQMF	Hierarchical	36.92	3.10	21.30	11.91	0.847	2.62	0.957
CoMBD		37.20	2.85	21.74	11.58	0.870	2.88	0.965
tSBD		36.08	2.78	20.65	11.57	0.887	2.95	0.964
tSBD+fSBD		36.05	2.84	21.49	11.20	0.882	2.97	0.964

6. Conclusions

Conclusions

- In this paper, we proposed an artifact-free GAN-based vocoder, Avocodo.

- Two artifacts which degrade the synthesized speech quality were defined as aliasing and imaging artifa ct. Authors designed two novel discriminators, CoMBD and SBD, to solve these problems.

- In both subjective and objective evaluations, Avocodo outperformed the baseline vocoders both in sin gle and unseen speaker synthesis tasks.
  - Although Avocodo showed improved rendition compared to the baseline vocoders, the proposed methods we re limited to increasing the performance with a smaller generator.

- In singing speaker synthesis, discontinuities on F0 were observed.
  - Author assume that it is a limitation of a generator structure (i.e., hidden dimension and receptive field size).

Thank you for listening
Q & A