HiFi-GAN:
Generative Adversarial Networks for Effi cient and High Fidelity Speech Synthesis

2022. 08. 16. (TUE)

Youngwon Choi

모두의연구소

풀잎스쿨 Hands-on TTS

Contents

1. Introduction

Introduction

□ Previous work

– Several recent work on speech synthesis have employed generative adversarial networks (GANs) to produce raw waveforms.

– Although such methods improve the sampling efficiency and memory usage, their sample quality has not yet reached that of autoregressive and flow-based generative models.

– Despite of the sophisticated GANs (MelGAN, Parallel WaveGAN, etc.,), there is still a gap in sample quality between the GAN models and AR or flow-based models.

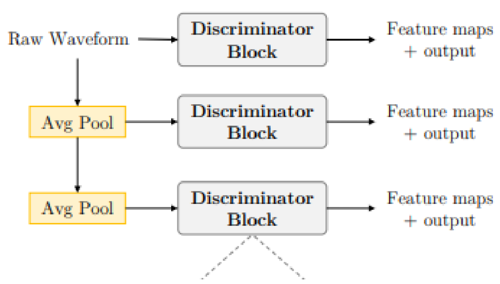
□ Authors propose HiFi-GAN, which achieves both higher computational efficiency and sample quality than AR or flow-based models.

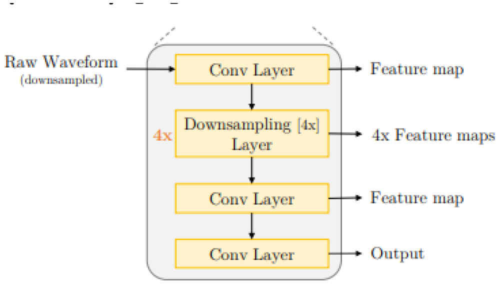
– As speech audio consists of sinusoidal signals with various periods, modeling the periodic patterns matters to generate realistic speech audio. Therefore, we propose a discriminator which consists of small sub-discriminators, each of which obtains only a specific periodic parts of raw waveforms

2. HifiGAN

HiFi-GAN

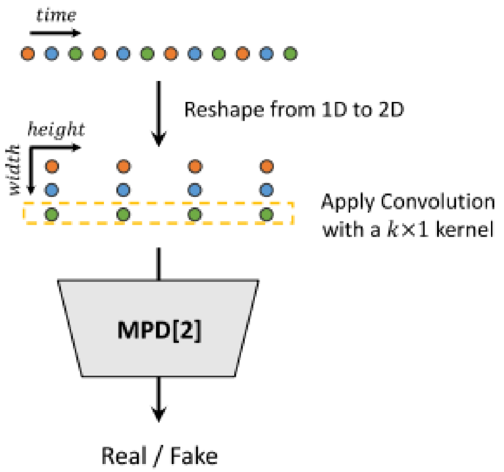
- HiFi-GAN consist of one generator and two discriminators.
  – Discriminators: multi-scale and multi-period discriminator
  – Multi-scale discriminator

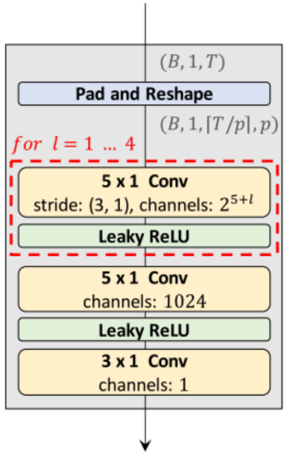




(b) Discriminator

  – Multi-period discriminator





(b)

HiFi-GAN

□ Generator

– The generator is a fully convolutional neural network.

– It uses a mel-spectrogram as input and upsamples it through transposed convolutions until the length of the output sequence matches the temporal resolution of raw waveforms

– Multi-Receptive Field Fusion (MRF)

- Different kernel sizes and dilation rates are selected for each residual block to form diverse receptive field patterns.
- MRF module returns the sum of outputs from multiple residual blocks.

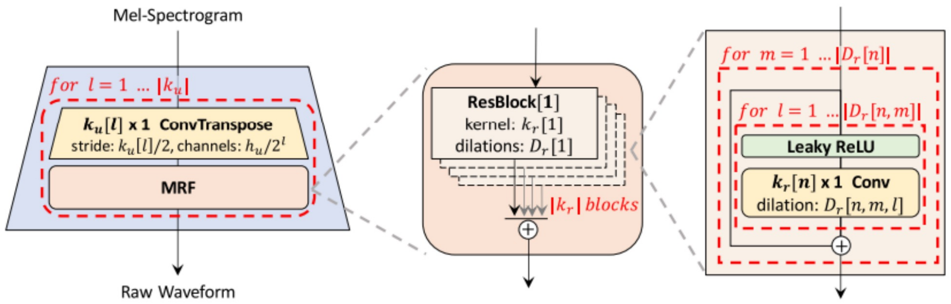
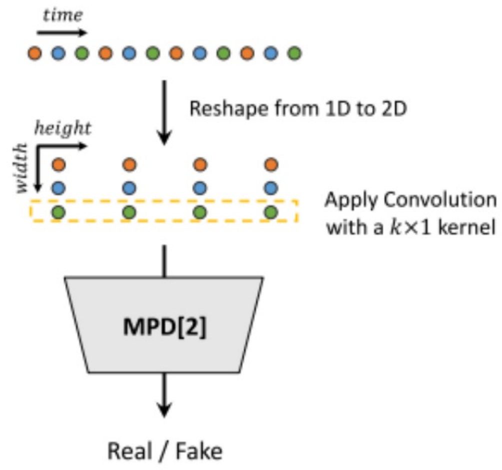


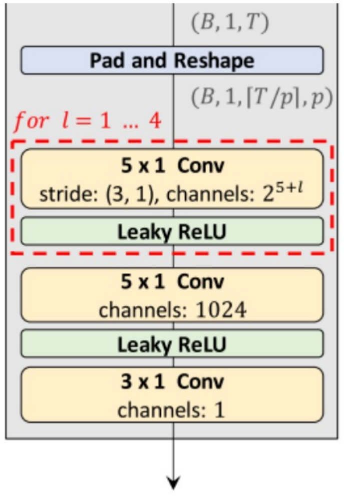
Figure 1: The generator upsamples mel-spectrograms up to $|k_u|$ times to match the temporal resolution of raw waveforms. A MRF module adds features from $|k_r|$ residual blocks of different kernel sizes and dilation rates. Lastly, the n-th residual block with kernel size $k_r[n]$ and dilation rates $D_r[n]$ in a MRF module is depicted.

HiFi-GAN

- Multi-period discriminator
  - MPD is a mixture of sub-discriminators, each of which only accepts equally spaced samples of an input audio; the space is given as period p.
    - period p: [2, 3, 5, 7, 11]
    - The sub-discriminators are designed to capture different implicit structures from each other by looking at different parts of an input audio
  - We first reshape 1D raw audio of length T into 2D data of height T/p and width p and then apply 2D convolutions to the reshaped data.
  - In every convolutional layer of MPD, we restrict the kernel size in the width axis to be 1 to process the periodic samples independently.
  - By reshaping the input audio into 2D data instead of sampling periodic signals of audio, gradients from MPD can be delivered to all time steps of the input audio.



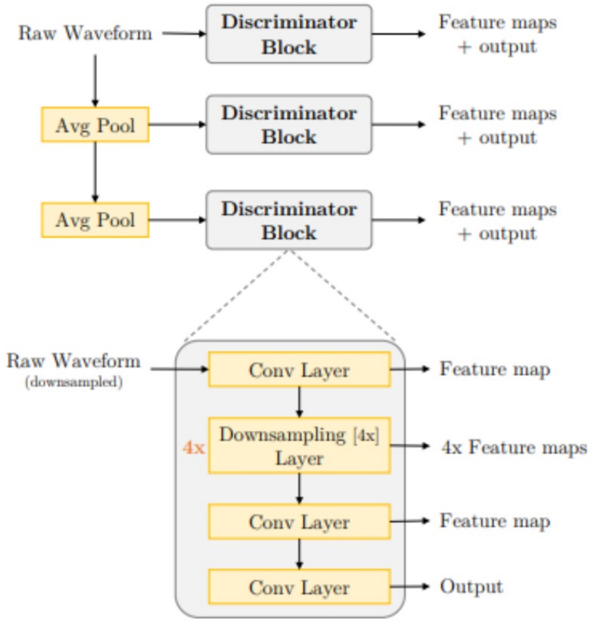
(b)



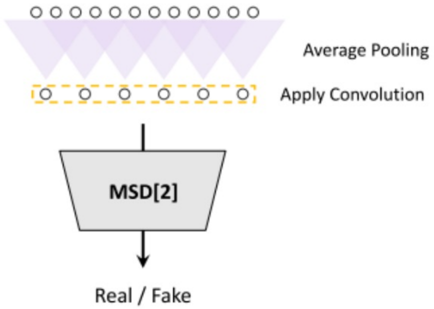
(b)

HiFi-GAN

- Multi-scale discriminator
  - The architecture of MSD is drawn from that of MelGAN.
  - Because each sub-discriminator in MPD only accepts disjoint samples, we add MSD to consecutively evaluate the audio sequence.
  - MSD is a mixture of three sub-discriminators operating on different input scales:
    - raw audio, ×2 average-pooled audio, and ×4 average-pooled audio.
    - Note that MPD operates on disjoint samples of raw waveforms, whereas MSD operates on smoothed waveforms.



(b) Discriminator



HiFi-GAN

□ Training loss terms

– G: Generator, D: Discriminator

– x: ground truth audio, s: mel-spectrogram of the ground truth audio

– Final Loss = GAN Loss + Mel-Spectrogram loss + Feature matching loss

□ GAN Loss

– For training stability, the objectives follow the least-squares GAN (LSGAN). (1 for real, 0 for fake)

– Discriminator

$$\mathcal{L}_{Adv}(D;G) = \mathbb{E}_{(x,s)}\left[(D(x)-1)^2 + \left(D\left(G(s)\right)\right)^2\right]$$

– Generator

$$\mathcal{L}_{Adv}(G;D) = \mathbb{E}_{(x,s)}\left[\left(D\left(G(s)\right)-1\right)^2\right]$$

HiFi-GAN

□ Mel Spectrogram loss

- Reconstruction loss
- The mel-spectrogram loss is the L1 distance between the mel-spectrogram of a waveform synthesized by the generator and that of a ground truth waveform.

$$\mathcal{L}_{Mel}(G) = \mathbb{E}_{(x,s)}[\|\phi(x) - \phi(G(s))\|_1]$$

- The mel-spectrogram loss helps the generator to synthesize a realistic waveform corresponding to an input condition, and also stabilizes the adversarial training process from the early stages.

- The mel-spectrogram loss can be expected to have the effect of focusing more on improving the perceptual quality due to the characteristics of the human auditory system.

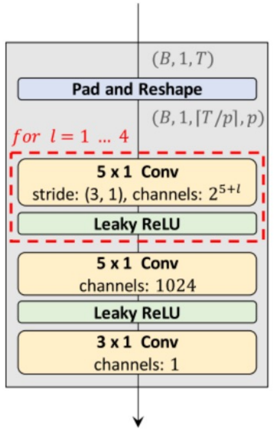
HiFi-GAN

□ Feature matching loss

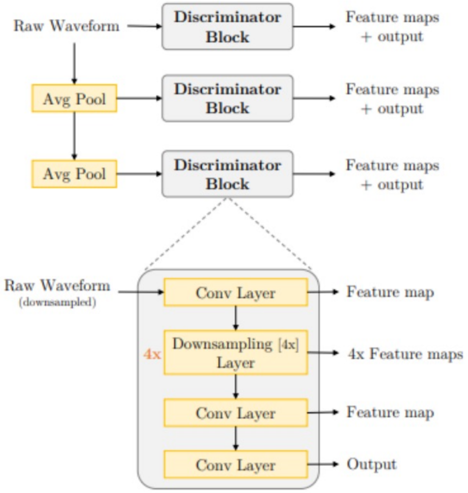
- The feature matching loss is a learned similarity metric measured by the difference in features of the discriminator between a ground truth sample and a generated sample.

- This objective minimizes the L1 distance between discriminator feature maps of reals and synthetic speech.

$$\mathcal{L}_{FM}(G;D) = \mathbb{E}_{(x,s)}\left[\sum_{i=1}^{T}\frac{1}{N_i}\left\|D^i(x) - D^i(G(s))\right\|_1\right]$$



(b)



(b) Discriminator

HiFi-GAN

□ Final Loss

– Final Loss = GAN Loss + Mel-Spectrogram loss + Feature matching loss

$$\mathcal{L}_G = \mathcal{L}_{Adv}(G; D) + \lambda_{fm} \cdot \mathcal{L}_{FM}(G; D) + \lambda_{mel} \cdot \mathcal{L}_{Mel}(G)$$

$$\mathcal{L}_D = \mathcal{L}_{Adv}(D; G)$$

Where $\lambda_{fm} = 2, \lambda_{mel} = 45$

3. Experiment & Result

Experiment & Result

□ The three variations of the generator $V1$, $V2$ and $V3$:

- V1 : h_u = 512, k_r = [3, 7, 11], k_u = [16, 16, 4, 4], D_r =[[1,1], [3,1], [5,1]] x 3
- V2 : The small version of $V1$, h_u = 128.
- V3: h_u = 256, k_r = [3, 5, 7], k_u = [16, 16, 8], D_r =[[1], [2]], [[2], [6]], [[3], [12]]

□ 기타 Experiment configuration 은 생략하겠습니다.

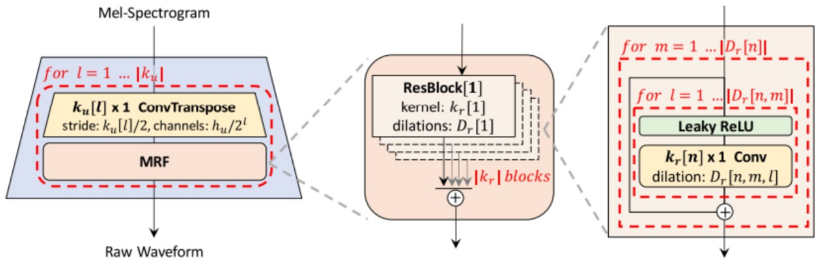


Figure 1: The generator upsamples mel-spectrograms up to $|k_u|$ times to match the temporal resolution of raw waveforms. A MRF module adds features from $|k_r|$ residual blocks of different kernel sizes and dilation rates. Lastly, the n-th residual block with kernel size $k_r[n]$ and dilation rates $D_r[n]$ in a MRF module is depicted.

h_u: hidden dimension
k_r: kernel size of standard convolution
k_u: kernel size of transposed convolution
D_r: dilation rates

Experiment & Result

□ MOS Test

Table 1: Comparison of the MOS and the synthesis speed. Speed of n kHz means that the model can generate $n \times 1000$ raw audio samples per second. The numbers in () mean the speed compared to real-time.

Model	MOS (CI)	Speed on CPU (kHz)	Speed on GPU (kHz)	# Param (M)
Ground Truth	4.45 (±0.06)	–	–	–
WaveNet (MoL)	4.02 (±0.08)	–	0.07 (×0.003)	24.73
WaveGlow	3.81 (±0.08)	4.72 (×0.21)	501 (×22.75)	87.73
MelGAN	3.79 (±0.09)	145.52 (×6.59)	14,238 (×645.73)	4.26
HiFi-GAN $V1$	**4.36** (±0.07)	31.74 (×1.43)	3,701 (×167.86)	13.92
HiFi-GAN $V2$	4.23 (±0.07)	214.97 (×9.74)	16,863 (×764.80)	**0.92**
HiFi-GAN $V3$	4.05 (±0.08)	**296.38** (×**13.44**)	**26,169** (×**1,186.80**)	1.46

Experiment & Result

□ Ablation Study

Model	MOS (CI)
Ground Truth	4.57 (±0.04)
Baseline (HiFi-GAN $V3$)	4.10 (±0.05)
w/o MPD	2.28 (±0.09)
w/o MSD	3.74 (±0.05)
w/o MRF	3.92 (±0.05)
w/o Mel-Spectrogram Loss	3.25 (±0.05)
MPD p=[2,4,8,16,32]	3.90 (±0.05)
MelGAN	2.88 (±0.08)
MelGAN with MPD	3.35 (±0.07)

Experiment & Result

□ Generalization to Unseen Speakers

Table 3: Quality comparison of synthesized utterances for unseen speakers.

Model	MOS (CI)
Ground Truth	3.79 (±0.07)
WaveNet (MoL)	3.52 (±0.08)
WaveGlow	3.52 (±0.08)
MelGAN	3.50 (±0.08)
HiFi-GAN $V1$	**3.77** (±0.07)
HiFi-GAN $V2$	3.69 (±0.07)
HiFi-GAN $V3$	3.61 (±0.07)

Experiment & Result

□ End to End speech synthesisß

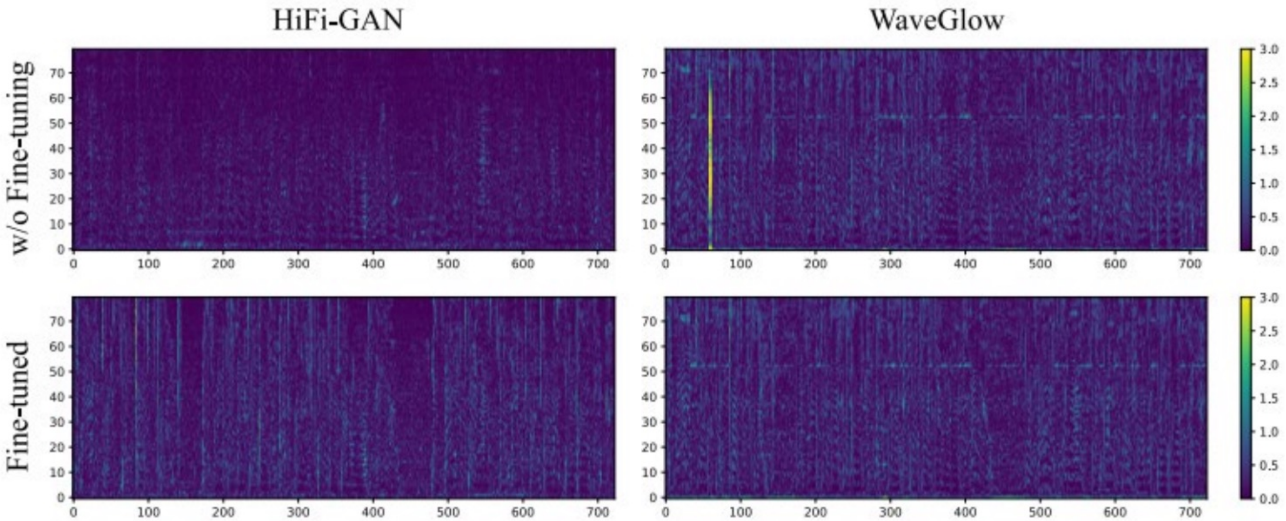


Figure 3: Pixel-wise difference in the mel-spectrogram domain between generated waveforms and a mel-spectrogram from Tacotron2. Before fine-tuning, HiFi-GAN generates waveforms corresponding to input conditions accurately. After fine-tuning, the error of the mel-spectrogram level increased, but the perceptual quality increased.

Table 4: Quality comparison for end-to-end speech synthesis.

Model	MOS (CI)
Ground Truth	4.23 (±0.07)
WaveGlow (w/o fine-tuning)	3.69 (±0.08)
HiFi-GAN $V1$ (w/o fine-tuning)	3.91 (±0.08)
HiFi-GAN $V2$ (w/o fine-tuning)	3.88 (±0.08)
HiFi-GAN $V3$ (w/o fine-tuning)	3.89 (±0.08)
WaveGlow (find-tuned)	3.66 (±0.08)
HiFi-GAN $V1$ (find-tuned)	**4.18** (±0.08)
HiFi-GAN $V2$ (find-tuned)	4.12 (±0.07)
HiFi-GAN $V3$ (find-tuned)	4.02 (±0.08)

Thank you for listening!
Q&A