MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis

Kundan Kumar, Rithesh Kummer, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo,
Alexandre de Brebisson, Yoshua Bengio, Aaron Courville

NeurIPS 2019

정영석

CONTENTS

Research Background

- **일반적 음성신호 처리 연구에서 raw audio를 사용하는 일은 거의 없음**
  - ✓ High temporal resolution (16,000 samples per second)
  - ✓ 시간을 기준으로 한 장단기 의존성문제 발생 ↑
  - → 대부분의 연구에서 lower-resolution representation feature를 modeling
- **Lower-resolution representation**
  - ✓ 시간 관점의 해상도에서 효율적으로 연산 되어야 함
  - ✓ 원래 raw audio로 회귀하는 것이 가능해야 함
  - → 음성신호 처리에서 Aligned linguistic features, Mel-spectrogram등의 feature가 자주 사용됨

Research Background

- Audio modeling에서는 앞서 설명한 이유로 두가지 step으로 나누어 모델링 하는 경우가 많음
  1. Text → intermediate representation (Mel-Spectrogram, Spectrogram)
  2. Intermediate representation → Audio

- Mel-Spectrogram Inversion
  - ✓ Pure Signal processing approaches
  - ✓ Autogressive neural-networks-based models
  - ✓ Non autoregressive models

Research Background

- Pure signal processing approaches
  - ✓ Griffin-Lim (STFT → temporal signal)
  - ✓ WORLD vector (temporal signa을 생성하기 위한 중간 표현; Char2Wav에서 사용됨)
- Autoregressive neural-networks-based models
  - ✓ WaveNet, WaveRNN
  - ✓ SampleRNN (music)
    - → 학습, 추론 속도가 매우 느림
- Autoregressive neural-networks-based models
  - ✓ Parallel Wavenet, Clarinet
  - ✓ WaveGlow

Research Background

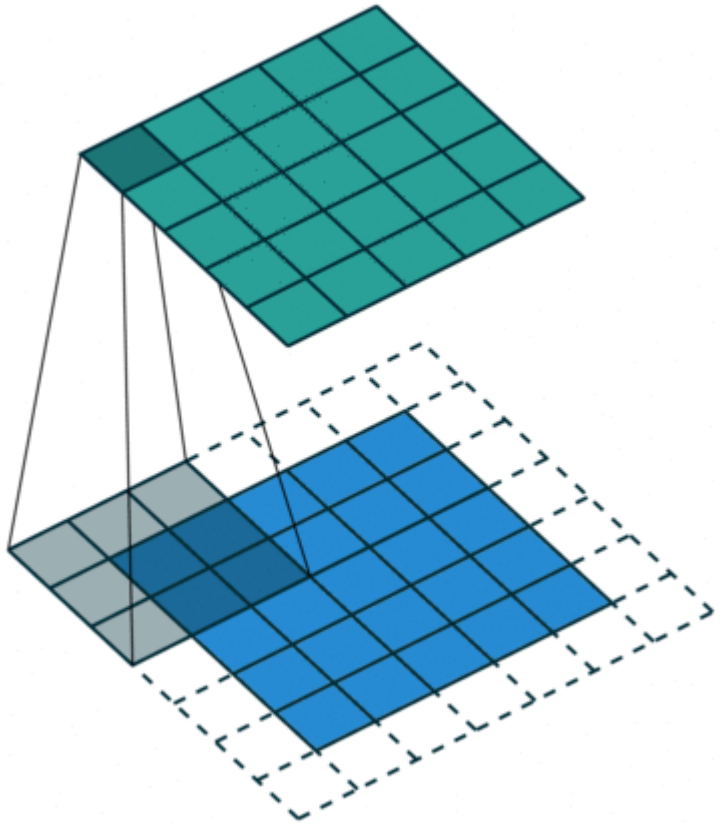
- GANs for audio
  - ✓ Music timbre generator (STFT, Phase angle 생성) [Engel et al. 2019]
  - ✓ Simple Magnitude Generator [Neekhara et al. 2019]
    - → Not sufficient for high quality waveform generation
- Main contribution
  - ✓ 최초의 GAN 기반 waveform generation model
  - ✓ 다양한 환경(music generation, text-to-speech…)에서 사용될 수 있는 autoregressive 모델
  - ✓ 다른 모델들에 비해 추론 속도가 빠름 (기존의 가장 빠른 모델보다 10배 빠름)

Methodology (Generator)

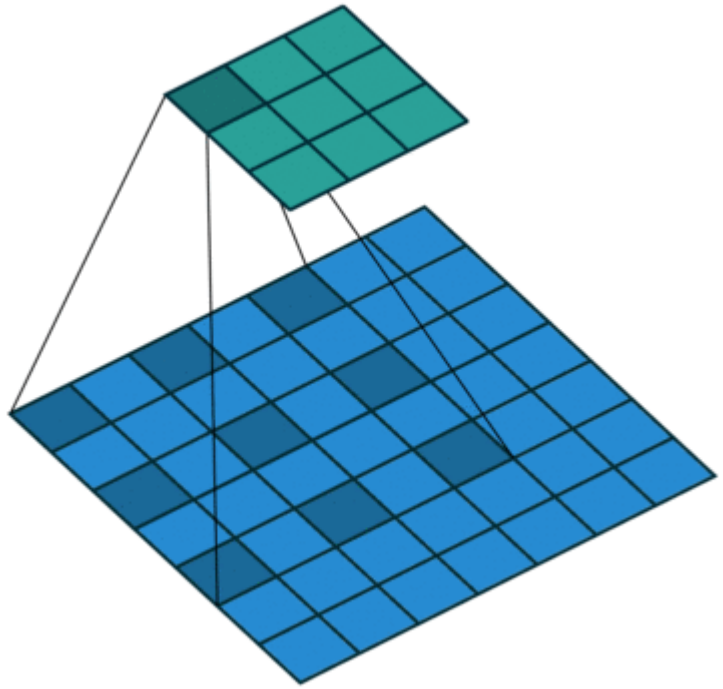
- Architecture
  - ✓ s(spectrogram)을 입력으로 하고, x(wave form)을 출력으로 하는 CNN 기반 모델
  - ✓ Time dimension을 증폭하기 위해 transposed convolution을 사용함
  - ✓ Residual block 과 dilated convolution을 사용
  - ✓ 일반적인 GAN 모델과 다르게 noise vector를 사용하지 않음
    - Mel-spectrogram은 손실압축의 결과이기 때문에, 많은 audio를 생성하는 것이 가능
    - 하지만, 최근 연구[Mathieu et al. 2015; Isola et al 2017]에 따르면 condition information이 강한 경우 noise vector는 불필요
    - Noise vector 삽입 시 변형된 audio가 생성됨

Methodology (Generator)

- Induced receptive field



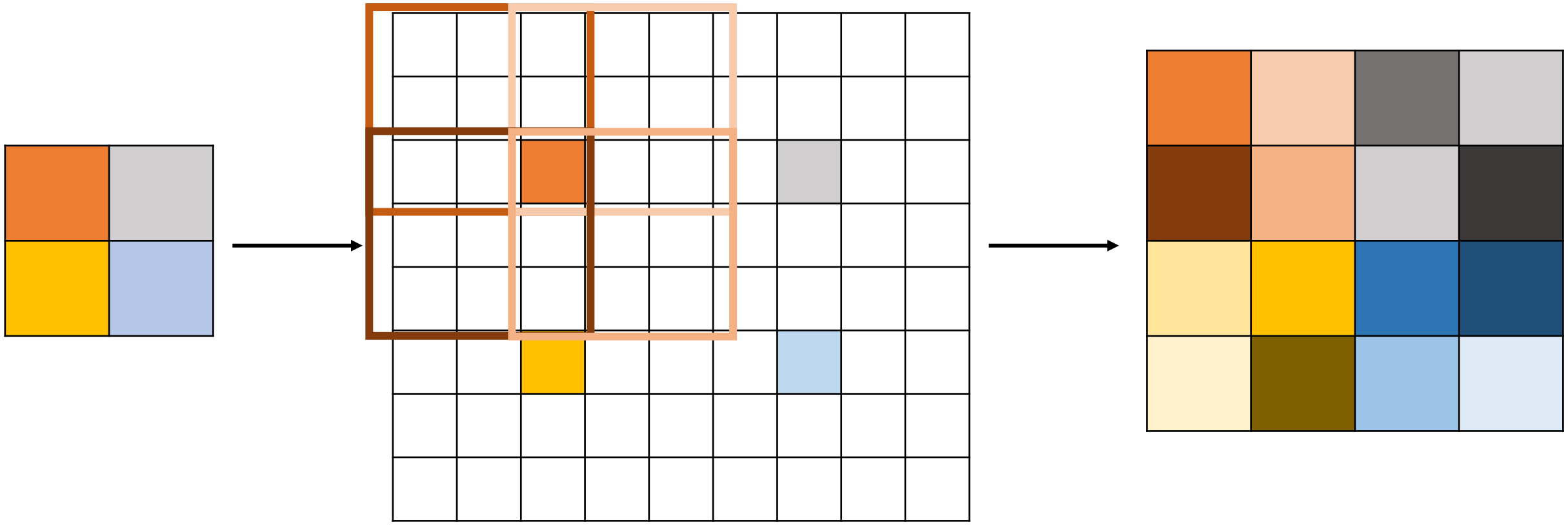
original convolution filter



dilated convolution filter

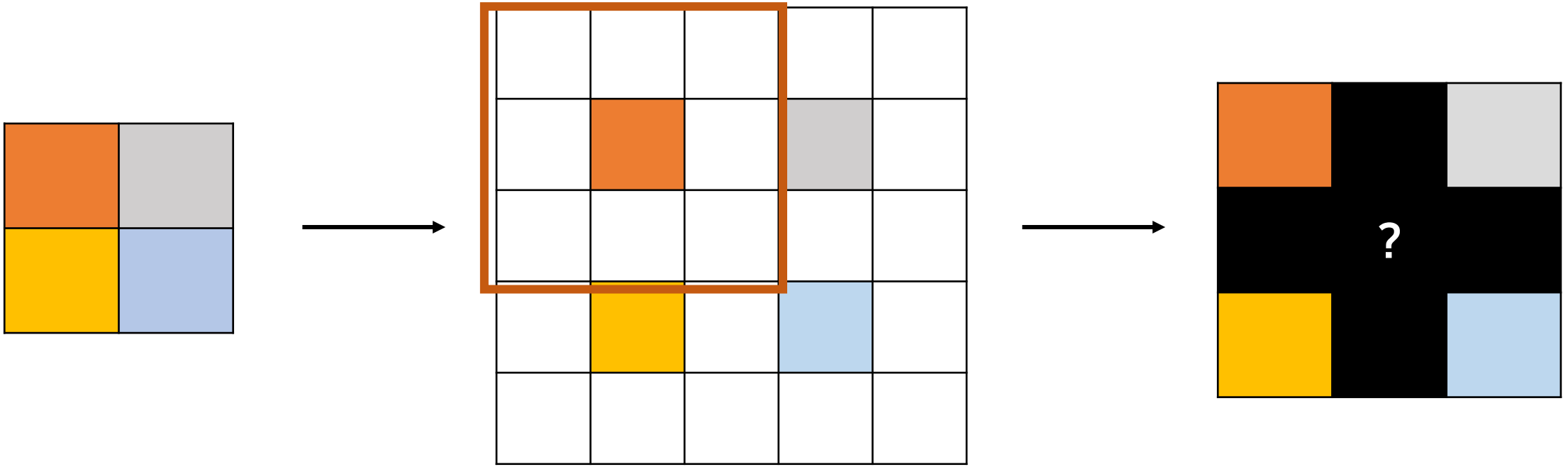
Methodology (Generator)

- Checkerboard artifacts (Good example)



Methodology (Generator)

- Checkerboard artifacts (Bad example)



Methodology (Generator)

- Normalizing Technique

  ✓ Instance normalizing : spectrogram feature의 pitch정보를 잃어 소리가 metallic하게 변환됨

  ✓ Spectral normalizing : Lipshitz 상수가 Discriminator의 feature matching objective에 영향을 끼침

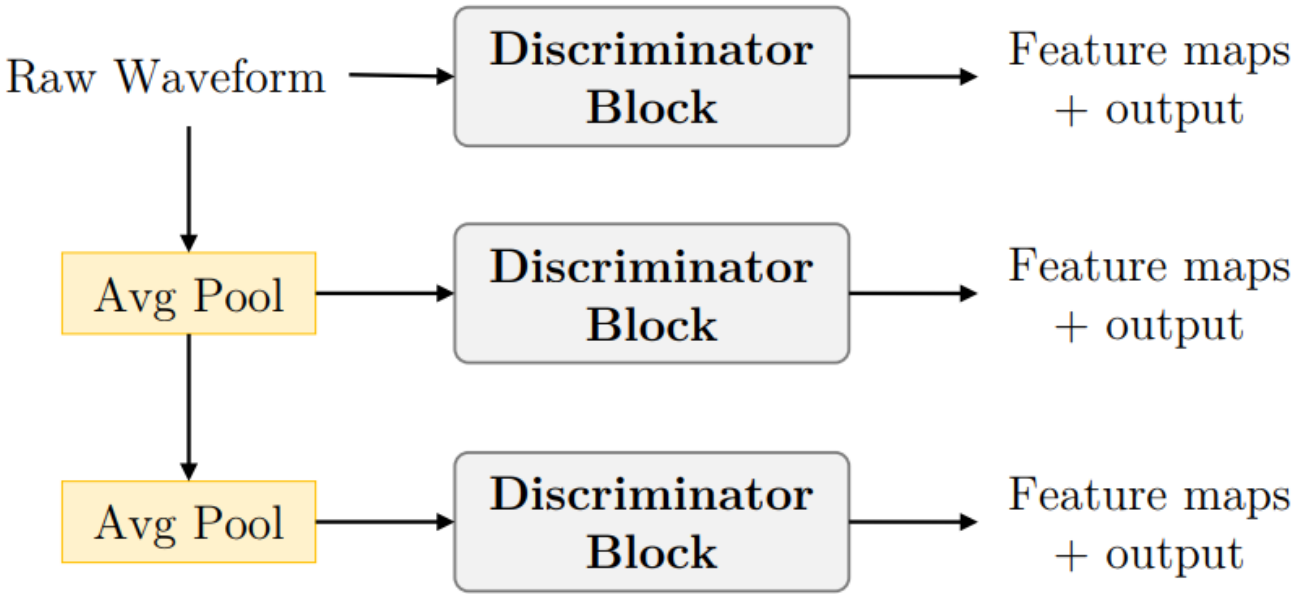
  $$|f(x) - f(y)| \leq M|x - y|, (M : \ Lipshitz\ constant, f(x): Lipshiz\ function)$$

  ✓ Weight normalizing : activation value, discriminator에 영향을 주지 않아 모델에 적합한 normalizing 기법

Methodology (Discriminator)

- Multi-Scale Architecture

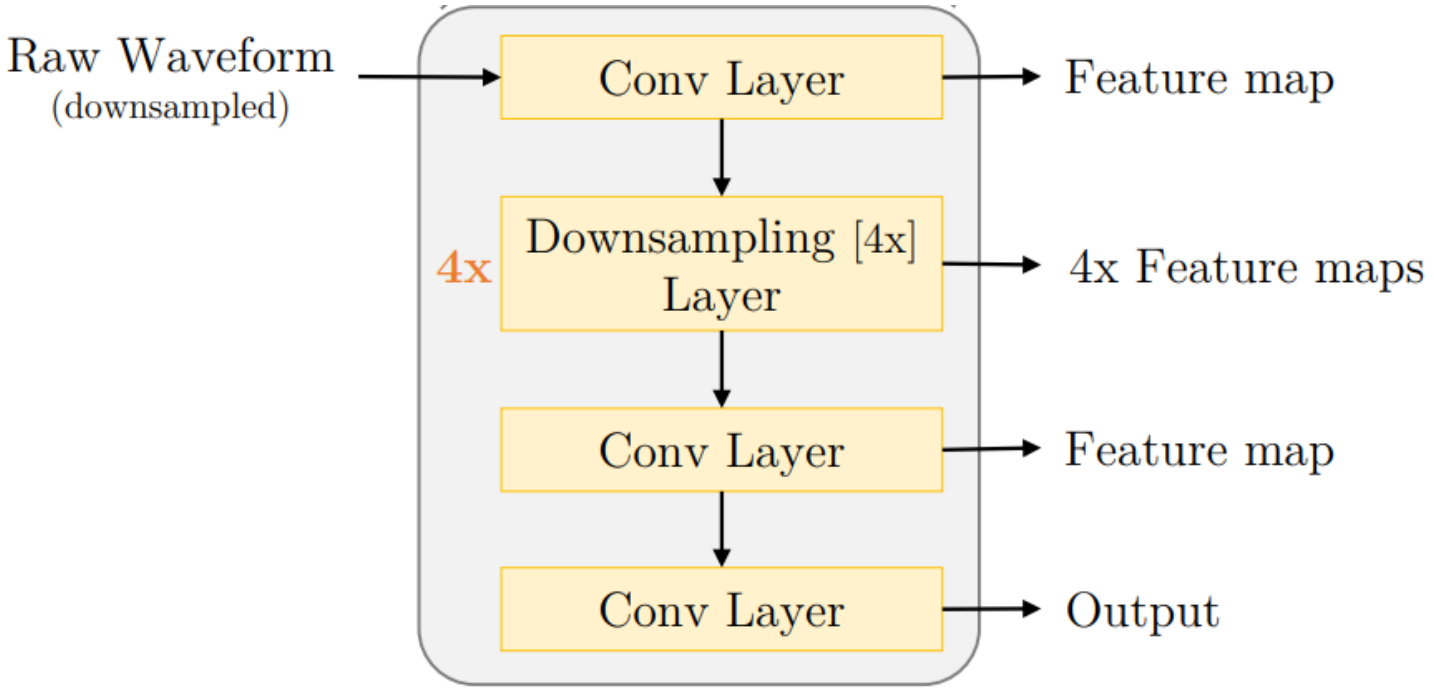
  ✓ 특정 주파수에 편향되는 것을 방지 하기 위해 multi-scale architecture구조를 사용



Methodology (Discriminator)

- Window-based objectives

  ✓ Learns to maintain coherence across patches



Methodology (Training Objective)

- Training objective

  ✓ Adversarial objective function (LSGAN hinge version)

$$\min_{D_k} E_x\left[\max\left(0, 1 - D_k(x)\right)\right] + E_{s,z}[\max(0, 1 + D_k(G(s,z))], \quad \forall k = 1,2,3$$

$$\min_{D_k} E_{s,z}[\sum_{k=1,2,3} -D_k(G(s,z))]$$

  ✓ Feature Matching

$$\mathcal{L}_{FM} = E_{x,s\sim p_{data}}[\sum_i^T \frac{1}{N_i} \left|\left| D_k^{(i)}(x) - D_k^{(i)}\left(G(s)\right)\right|\right|_1]$$

  ✓ Generator

$$\min_{D_k} E_{s,z}[\sum_{k=1,2,3} -D_k(G(s,z))] + \lambda \sum_{k=1}^3 \mathcal{L}_{FM}(G, D_k)$$

Experiments

- Number of parameters and inference speed

Table 1: Comparison of the number of parameters and the inference speed. Speed of n kHz means that the model can generate $n \times 1000$ raw audio samples per second. All models are benchmarked using the same hardware [3].

Model	Number of parameters (in millions)	Speed on CPU (in kHz)	Speed on GPU (in kHz)
Wavenet (Shen et al., 2018)	24.7	0.0627	0.0787
Clarinet (Ping et al., 2018)	10.0	1.96	221
WaveGlow (Prenger et al., 2019)	87.9	1.58	223
MelGAN (ours)	**4.26**	**51.9**	**2500**

Experiments

- Ablation study

Table 2: Mean Opinion Score of ablation studies. To evaluate the biases induced by each component, we remove them one at a time, and train the model for 500 epochs each. Evaluation protocol/details can be found in appendix B.

Model	MOS	95% CI	
w/ Spectral Normalization	1.33	±0.07	metallic
w/ L1 loss (audio space)	2.59	±0.11	metallic + 고주파 잡음
w/o Window-based Discriminator	2.29	±0.10	고주파 잡음
w/o Dilated Convolutions	2.60	±0.10	고주파 잡음
w/o Multi-scale Discriminator	2.93	±0.11	metallic + missing word
w/o Weight Normalization	3.03	±0.10	metallic
Baseline (MelGAN)	**3.09**	± **0.11**	

Experiments

- Benchmarking competing models

Table 3: Mean Opinion Scores

Model	MOS	95% CI
Griffin Lim	1.57	±0.04
WaveGlow	4.11	±0.05
WaveNet	4.05	±0.05
MelGAN	3.61	±0.06
Original	**4.52**	± **0.04**

- Generalization to unseen speakers

Table 4: Mean Opinion Scores on the VCTK dataset (Veaux et al., 2017).

Model	MOS	95% CI
Griffin Lim	1.72	±0.07
MelGAN	3.49	±0.09
Original	**4.19**	± **0.08**

Experiments

- Non-autoregressive decoder for VQ-VAE (piano music generatioin)