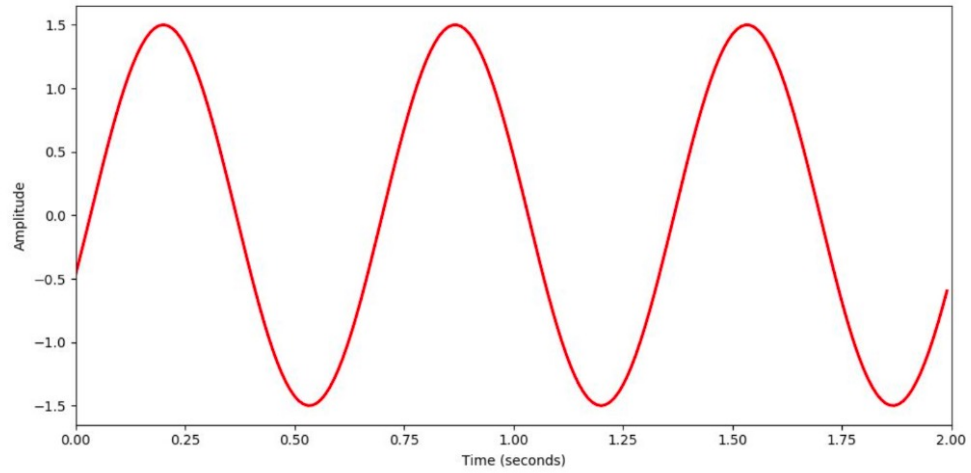
오디오 데이터 처리와 MelGAN

Voice Conversion

뉴보이스 이용준 김가은 이승빈

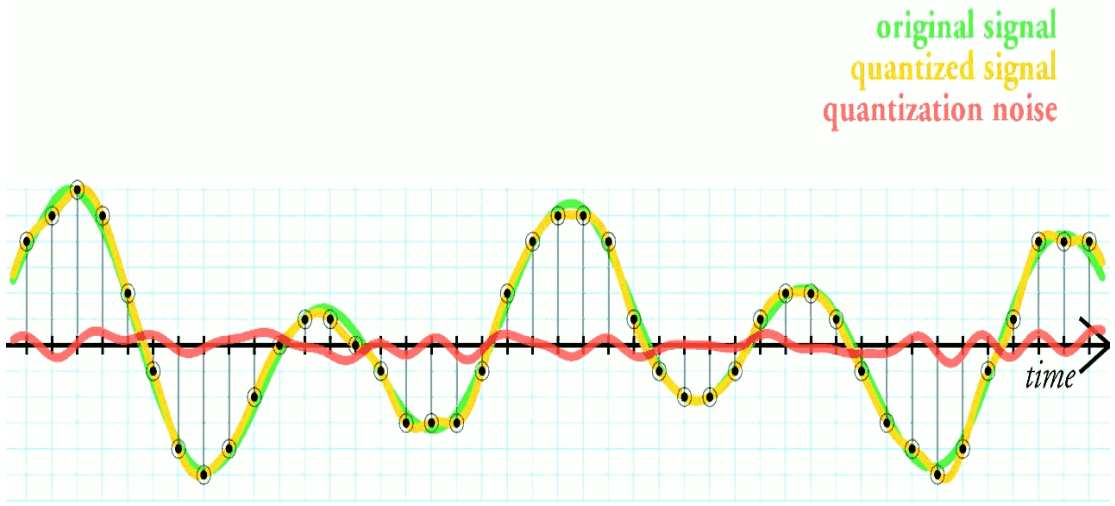
오디오 데이터 처리

raw waveform



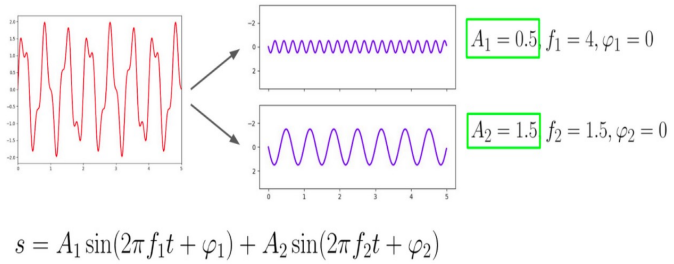
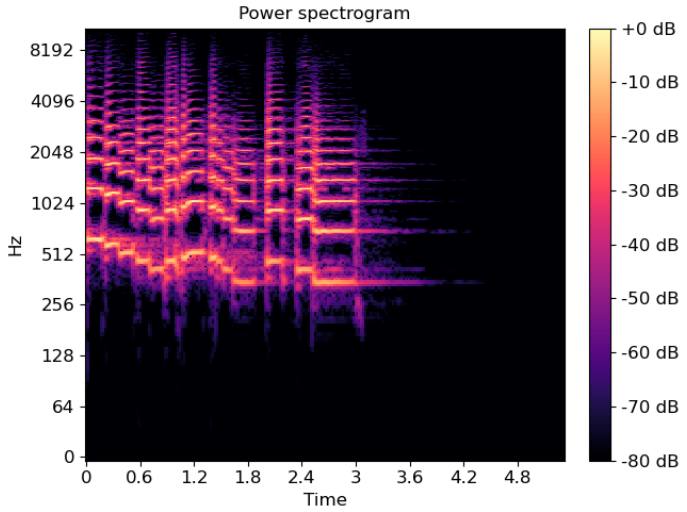
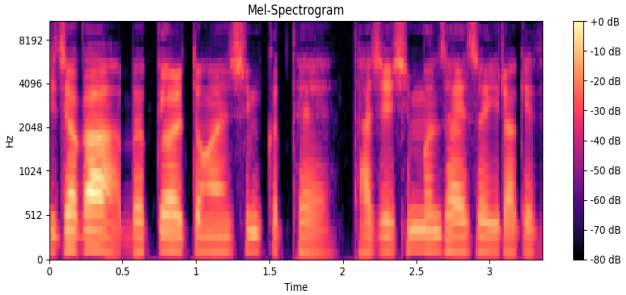
- X축- 시간(time)
- Y축 - 진폭(amplitude)

Sampling & Quantization



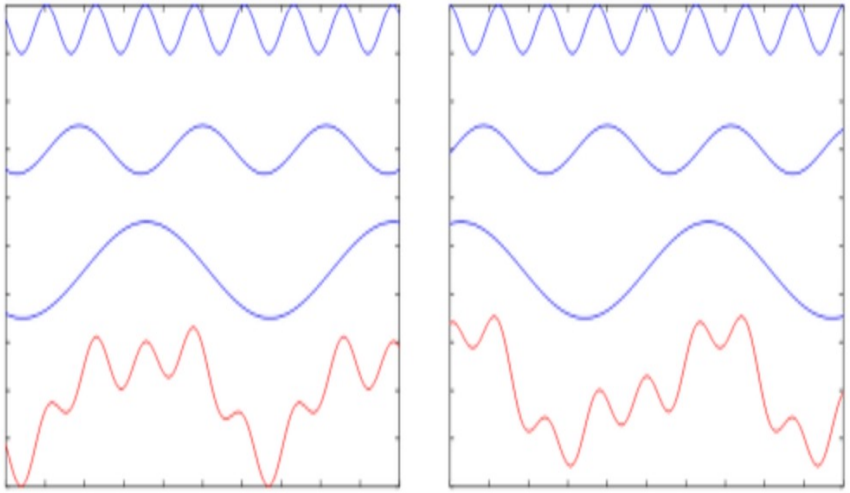
- Raw audio를 이산(discrete)적인 데이터로 바꾸는 작업
- Sampling: raw audio의 X축인 time을 일정 단위로 나눔
- Quantization: 특징값들을 추출

오디오 데이터 처리- **Feature extraction**

Fourier transform	Short time Fourier transform	Mel-Spectrogram 변환
		
• 주기 데이터인 음성 데이터를 주기함수인 sin함수, cos함수들의 집합으로 표현함 • 파도를 구성하는 다양한 요인으로 분해 • 특정 주파수에서 얼마나 진폭이 있었는지 파악할 수 있게 해주는 power spectrum	• 시간 정보가 담겨있는 power spectrum을 생성 • raw data가 (time , frequency(Hz), power(dB))의 3차원 tensor 로 변환됨	• 인간이 인식하는 주관적인 인지 기준에 맞춘 주파수 scale • 기존의 power spectrum 에 멜 변환(로그 변환) • dimension은 3차원으로 그대로 유지하면서 데이터의 해상도 높이기

Vocoder

위상 정보가 필요함



위상(파란색) 이 달라,
이들을 조합한 원본 음성(빨간색)의 모습이 상이함

Vocoder 과정

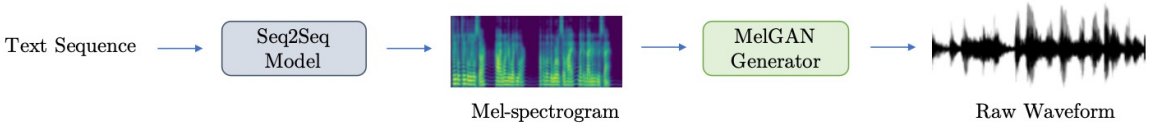
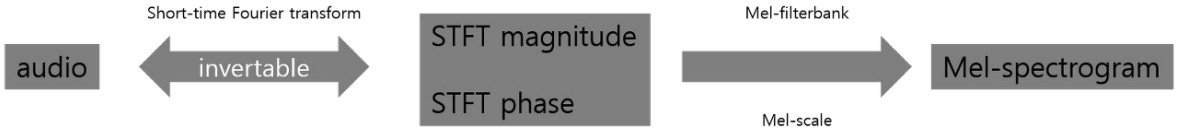


Figure 2: Text-to-speech pipeline.



- mel-spectrogram이라는 3차원 텐서를 생성함
- 이 텐서로 raw audio를 합성함
- STFT magnitude, STFT phase 정보가 필요함
- 초기 위상 정보를 임의로 설정한 뒤 STFT magnitude를 예측,
  오차를 최소화하며 학습함

TTS

Speech Synthesis

[KISTI과학향기]옛 가수들, AI 기술로 부활?

발행일 : 2021-01-25 07:00 지면 : 2021-01-25 17면

최근 세상을 떠난 옛 가수들을 인공지능(AI) 기술로 재현하는 TV 프로그램이 방영돼 화제가 되고 있다. 지난해 12월, 케이블 음악채널 엠넷은 '다시 한 번'이라는 프로그램을 통해 혼성 그룹 거북이의 리더였던 터틀맨과 가수 김현식의 목소리와 모습을 복원해 새로운 곡과 무대를 선보였다. 이미 세상을 떠난 가수들이 살아 돌아온 듯한 착각을 불러일으킬 정도로 목소리와 표정, 몸짓이 생생하게 구현됐다. 이 무대가 가능할 수 있었던 건 바로 AI의 음성합성 기술과 영상합성 기술의 발전 덕분이다.

<터틀맨의 목소리와 모습을 AI로 재현한 모습. (출처 : Mnet official 유튜브 캡처)>

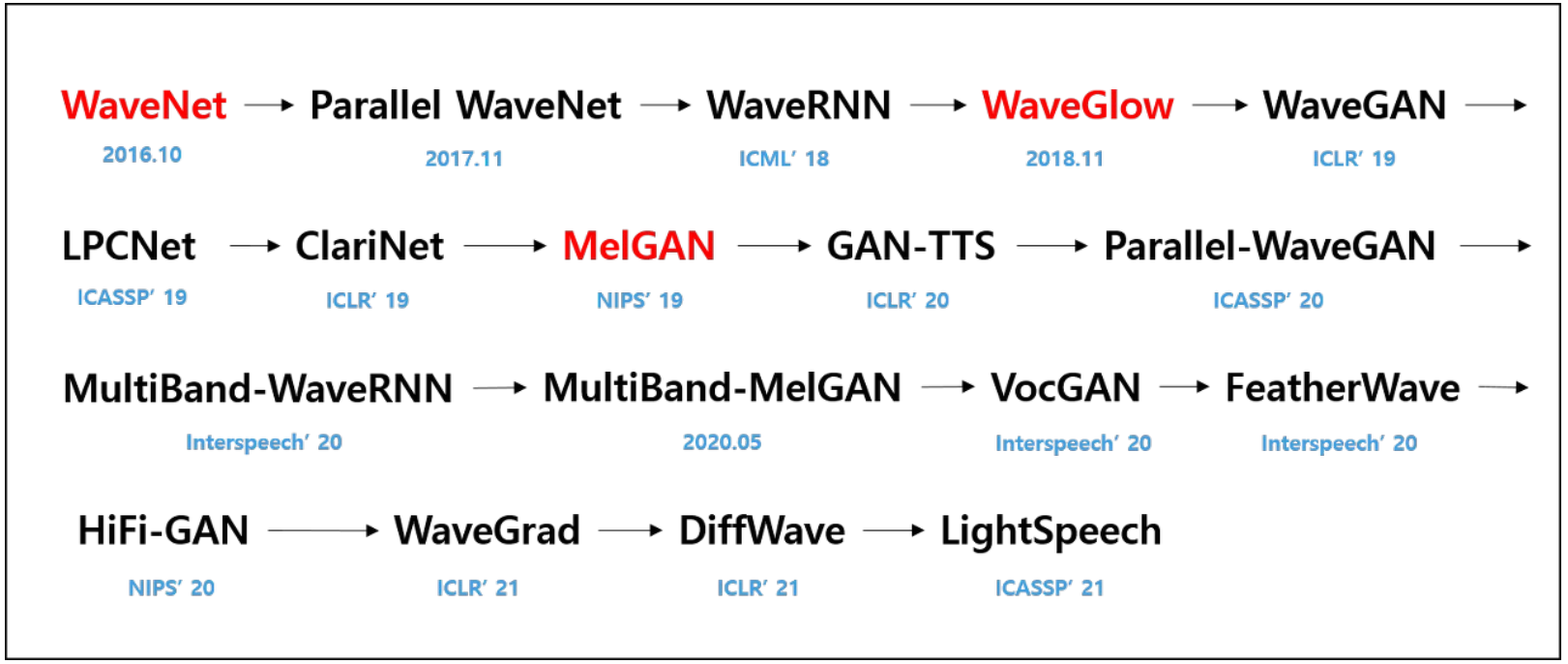
https://www.youtube.com/watch?v=NxQSxM0OkkY

https://www.youtube.com/watch?v=uflTK8c4w0c

https://www.youtube.com/watch?v=tnHop6WSIk8

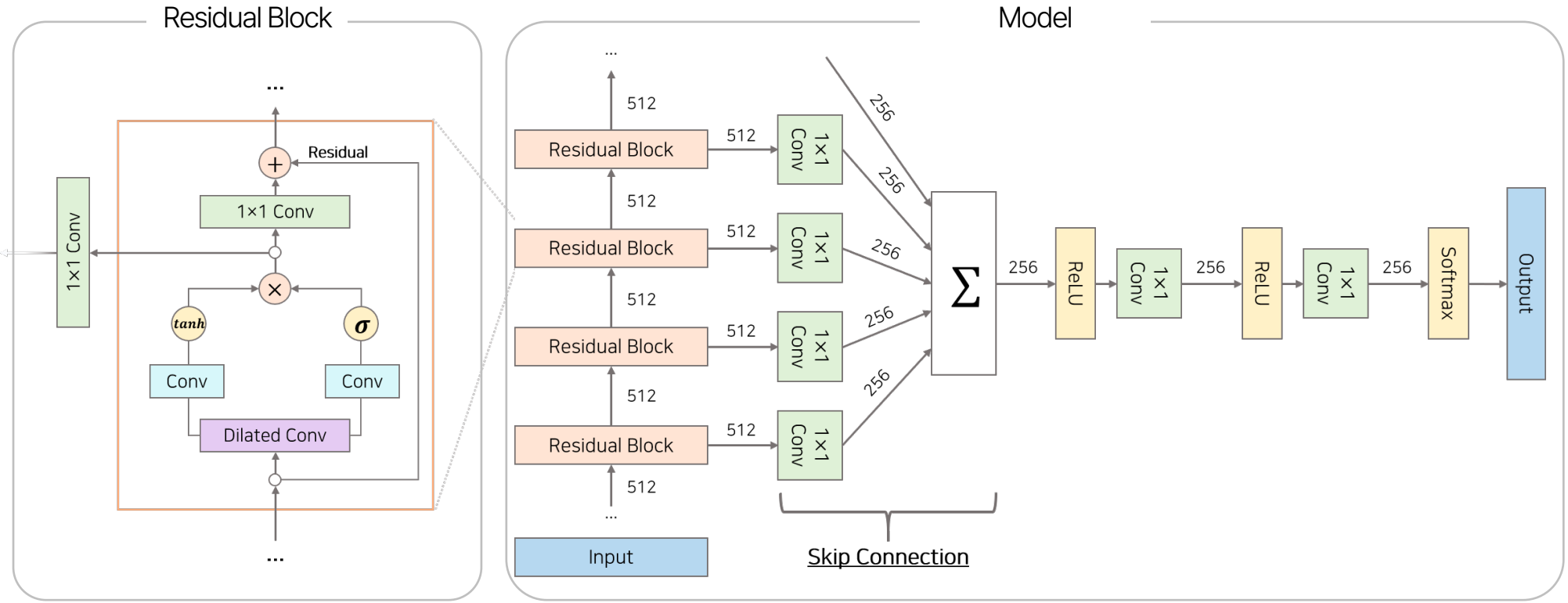
Vocoder : Wavenet~

Wavenet: A generative model for raw audio



Vocoder : Wavenet

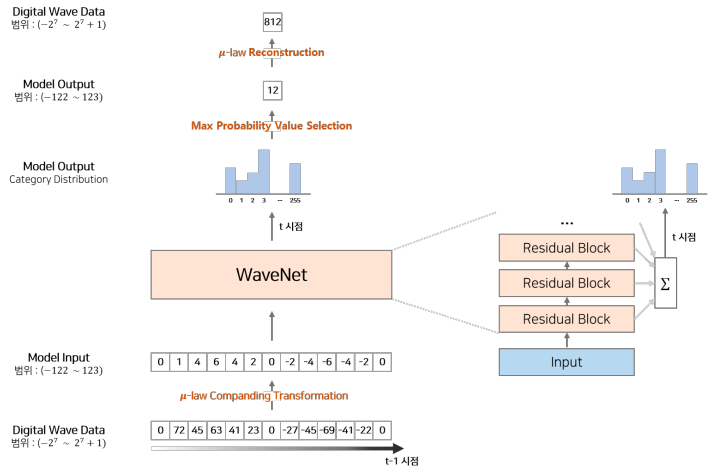
Wavenet: A generative model for raw audio



Vocoder : Wavenet

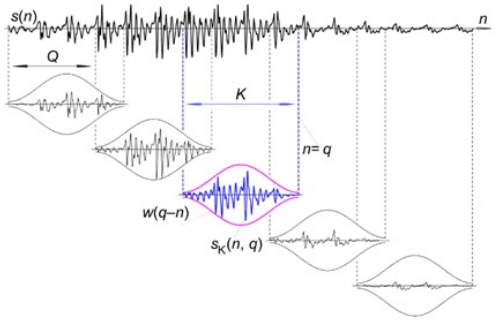
u-law Reconstruction:

데이터 벡터의 값들을 [-122, 123] 사이의 정수값들로 맵핑



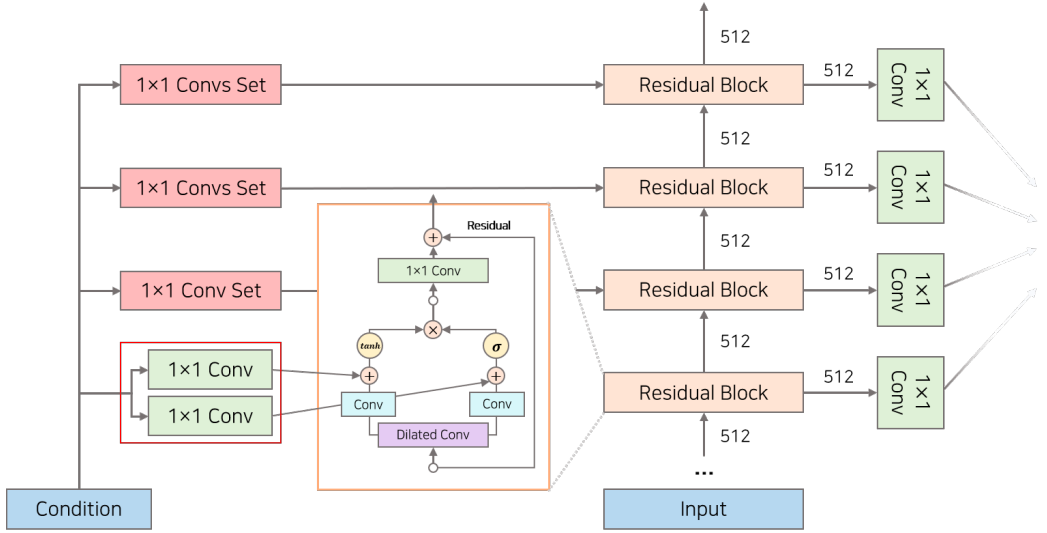
Overlapping:

안정적인 모델 추출을 위해
신호 겹치기



Conditional Wavenet

To add feature : 목소리, 발음, 띄어쓰기...



Global Condition	Local Condition
• 시간 불변 (데이터 끝까지) • 사람 목소리, 사투리	• character embedding • 음절별 정보 (발음 등)

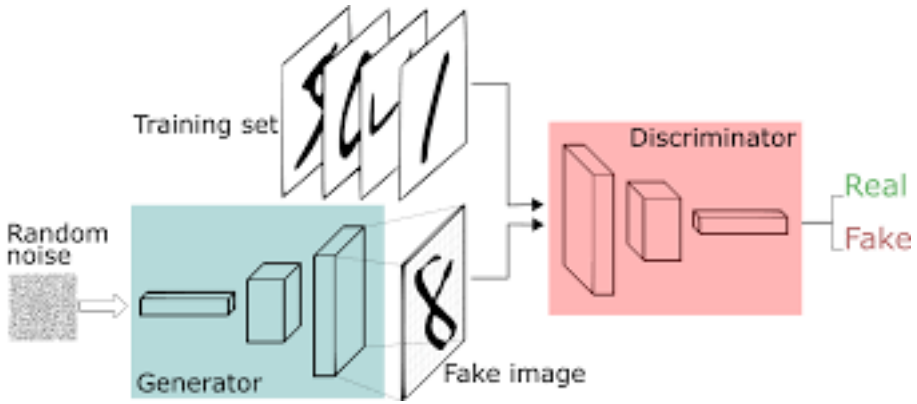
Vocoder : Wavenet

Wavenet: A generative model for raw audio

신호처리적 접근	Autoregressive (자기회귀 모델)	Non autoregressive (비자기회귀 모델)
the Griffin-Lim the WORLD vocoder	WaveNet SampleRNN , WaveRNN	Parallel Wavenet, Clarinet MelGAN
• Efficiently decode an STFT sequence • WORLD vocoder: attention-based recurrent neural network • Make too strong, robotic artifacts	• neural-networks-based • Produces realistic samples • Generate audio samples sequentially-> slow and inefficient • not suited way for real-time applications	• Parallelizable • faster than auto-regressive model • MelGAN: non-autoregressive feed-forward convolutional architecture

MelGAN

GAN을 활용한 음성 합성 모델



GAN

- Generative Adversarial Networks
- Generator vs Discriminator

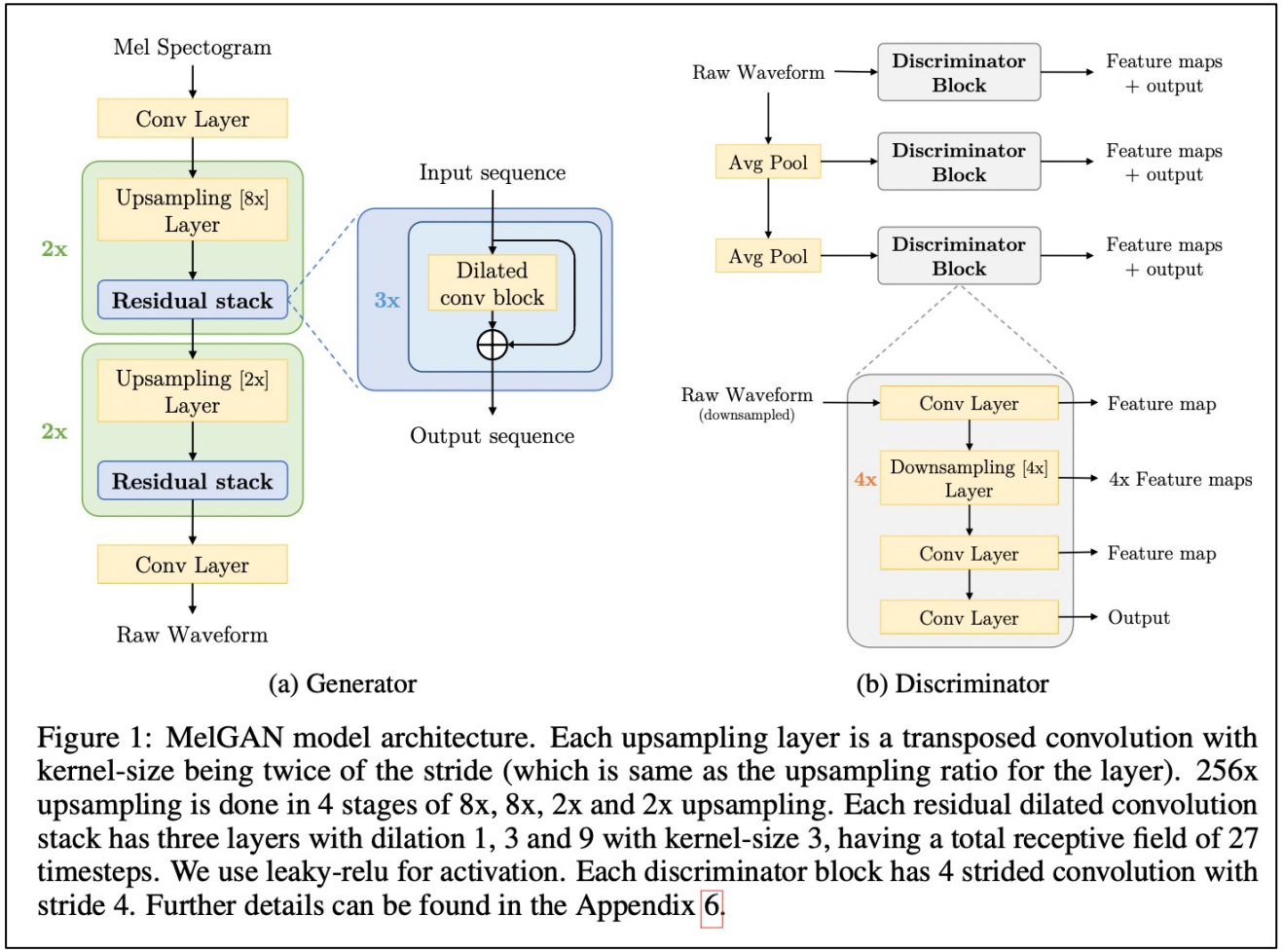


Figure 1: MelGAN model architecture. Each upsampling layer is a transposed convolution with kernel-size being twice of the stride (which is same as the upsampling ratio for the layer). 256x upsampling is done in 4 stages of 8x, 8x, 2x and 2x upsampling. Each residual dilated convolution stack has three layers with dilation 1, 3 and 9 with kernel-size 3, having a total receptive field of 27 timesteps. We use leaky-relu for activation. Each discriminator block has 4 strided convolution with stride 4. Further details can be found in the Appendix 6.

MelGAN

타 모델과의 성능 비교

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**

TTS Pipeline

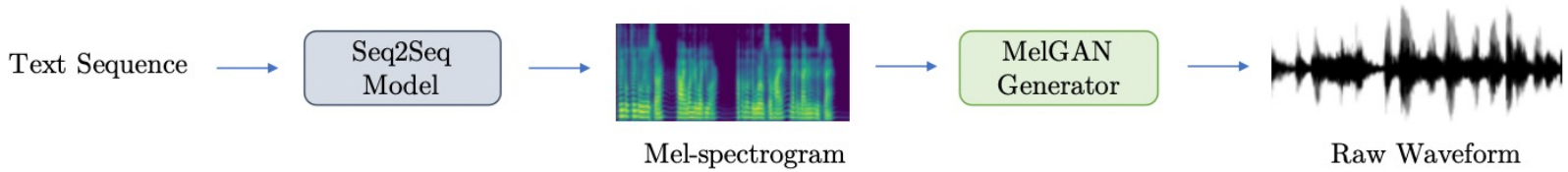


Figure 2: Text-to-speech pipeline.

MelGAN

http://swpark.me/melgan/

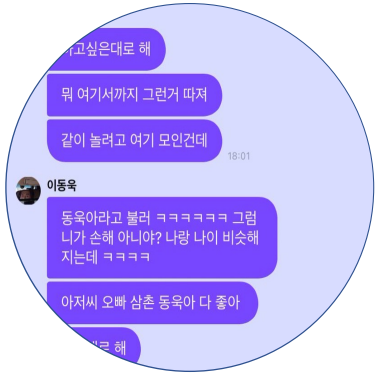
- Epoch 400
- Epoch 800
- Epoch 6400



Our project will be...

네비게이션

강의 영상

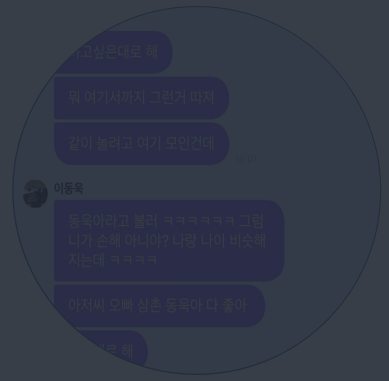


챗봇

Our project will be...

네비게이션

강의 영상



챗봇

감사합니다.