WaveNet

A Generative Model for Raw Audio

김민규

1. INTRODUCTION

- 이미지 및 텍스트와 같은 복잡한 분포를 모델링하는 신경 자기 회귀 생성 모델의 최근 발전에서 영감을 얻은 원시 오디오 생성 기술을 탐구

- 이 논문에서는 매우 높은 시간 분해능을 갖는 신호인 광대역 원시 오디오 파형을 생성하는 데 성공여부를 다룸

- PixelCNN 아키텍처를 기반으로 한 오디오 생성 모델인 WaveNet를 다룸

2. WAVENET

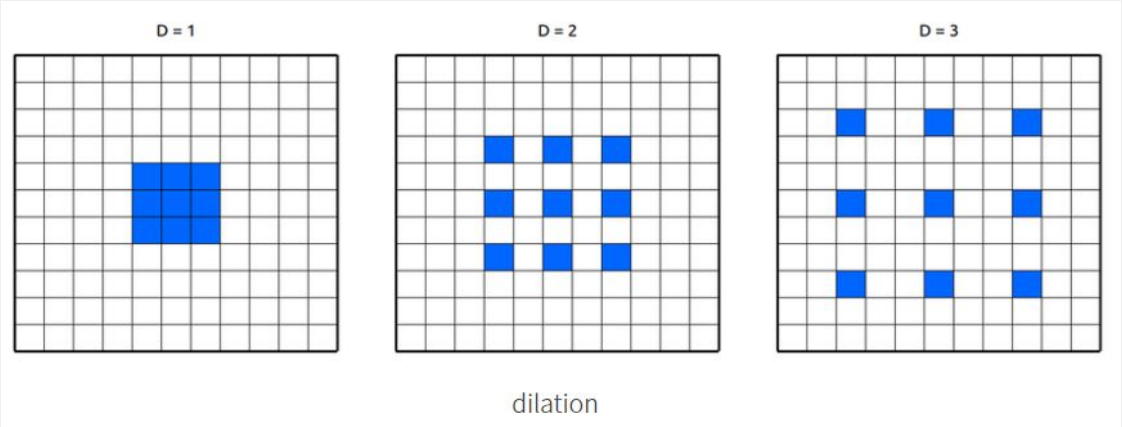
- WaveNets을 이용하여 텍스트 음성 변환(TTS) 분야에서 이전에 보고된 적이 없는 주관적인 자연성으로 원시 음성 신호를 생성
- WaveNets은 t 시점의 오디오 샘플을 t-1 시점까지의 샘플들의 조건부 분포로 모델링

$$p(x) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1})$$

- 네트워크에 pooling layer는 없고 input, output 차원이 동일
- Output에는 softmax를 취해 multinomial classification 문제를 다룸
- 최적화는 MLE을 사용

2.1 DILATED CAUSAL CONVOLUTION

- Dilated causal convolution은 dilation과 causal 두 개념이 같이 사용된 Conv 네트워크 구조
- dilation은 Conv의 receptive field는 넓히면서 연산량은 크게 증가시키지 않는 방법



2.2 SOFTMAX DISTRIBUTIONS

$$f(x_t) = sign(x_t)\frac{\ln(1+\mu|x_t|)}{(1+\mu)}$$

$$where, -1 < x_t < 1, \mu = 255$$

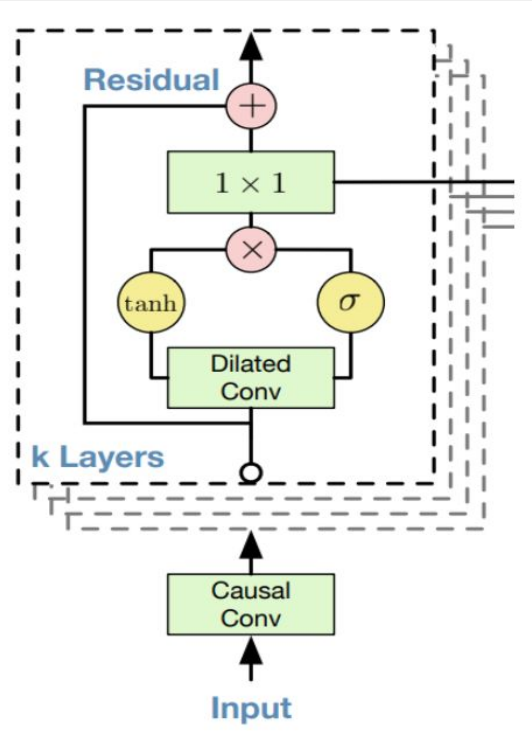
- Output을 모델링 하는데 있어서, softmax 분포를 사용
- Multinomial Logistic Regression 문제로 생각

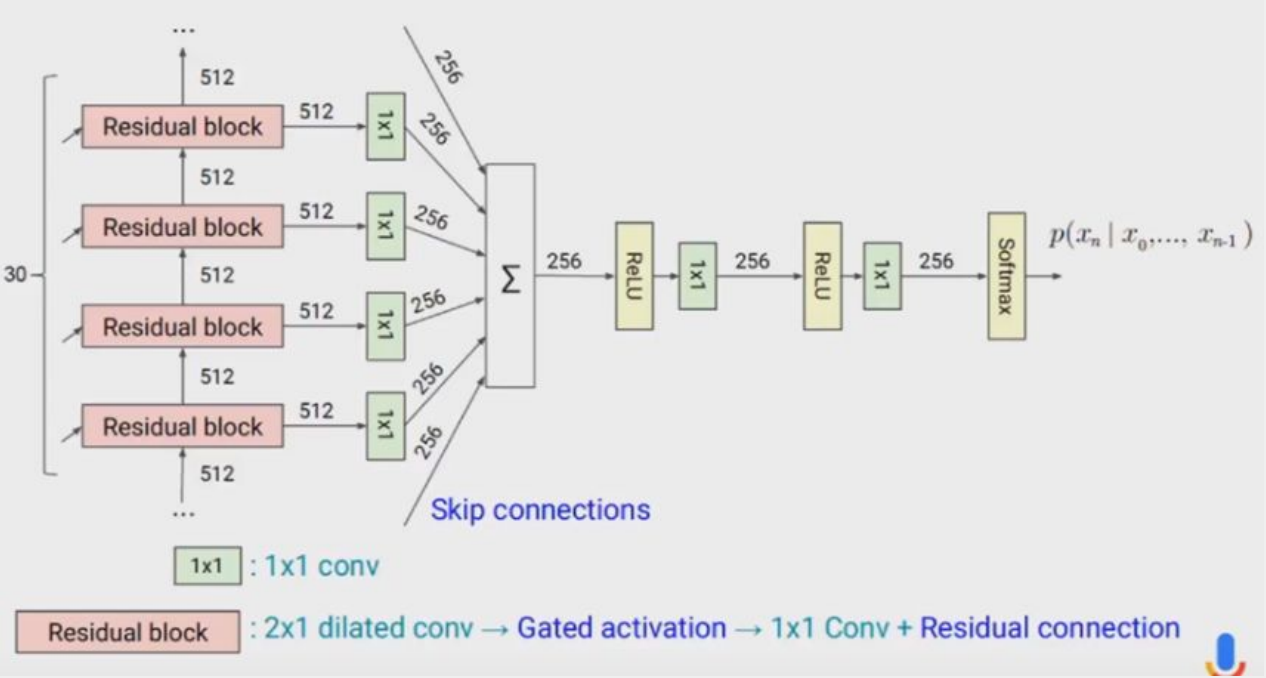
2.3 GATED ACTIVATION UNITS

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \sigma\left(W_{g,k} * \mathbf{x}\right)$$

- Dilated Conv 뒤에 gate network를 사용해서 다음 layer로 전달할 비율을 조절

2.4 RESIDUAL AND SKIP CONNECTIONS





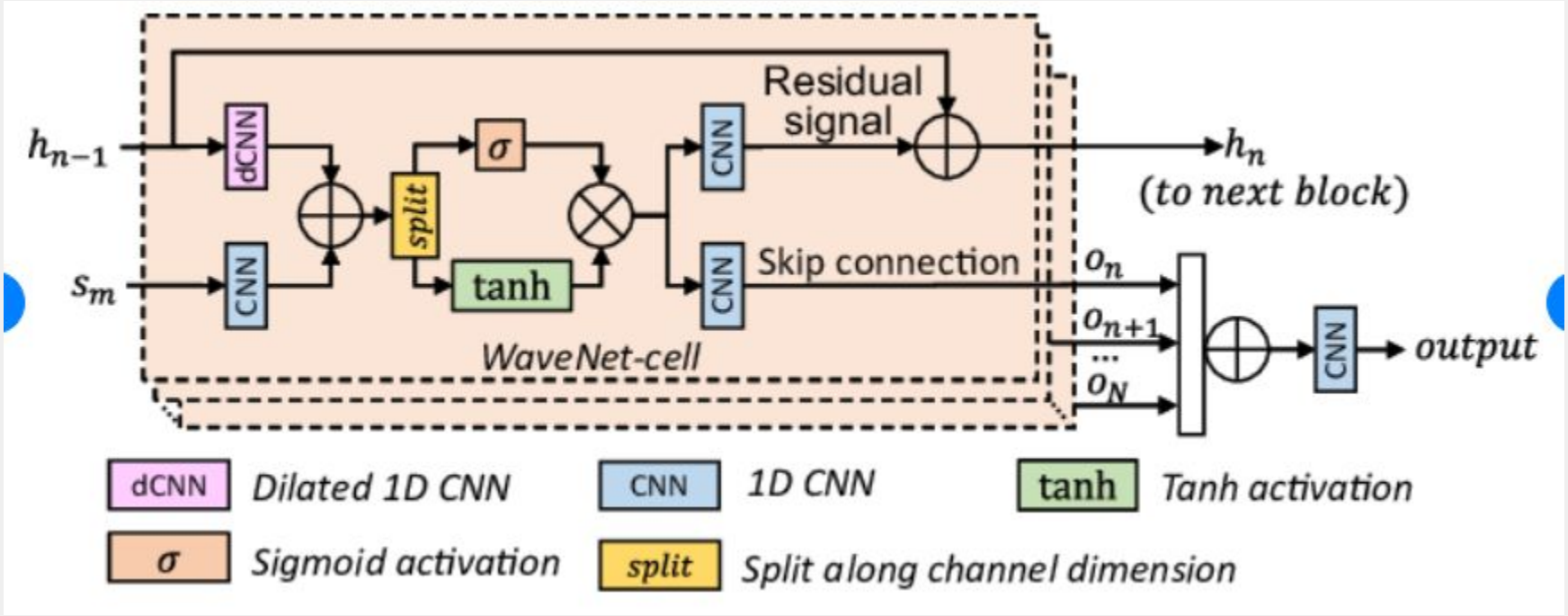
2.5 CONDITIONAL WAVENETS

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h}).$$

오디오의 조건부 분포 p(x | h)

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x} + V_{f,k}^T \mathbf{h}\right) \odot \sigma\left(W_{g,k} * \mathbf{x} + V_{g,k}^T \mathbf{h}\right).$$

2.6 CONTEXT STACKS



3. EXPERIMENTS

- WaveNet의 오디오 모델링 성능 측정을 위한 다중 스피커 음성 생성(텍스트에 구애받지 않음)

- TTS 및 음악 오디오 모델링의 세 가지 다른 작업에서 평가

3.1 MULTI-SPEAKER SPEECH GENERATION

- 단일 WaveNet은 스피커의 단일 한 인코딩으로 조건화함으로써 모든 스피커의 음성을 모델링 할 수 있음.

- 스피커를 추가하면 단일 스피커에 대한 교육보다 더 나은 감증 세트 성능을 얻을 수 있다는 것을 관찰.

- WaveNet의 내부 표현은 여러 스피커에서 공유.

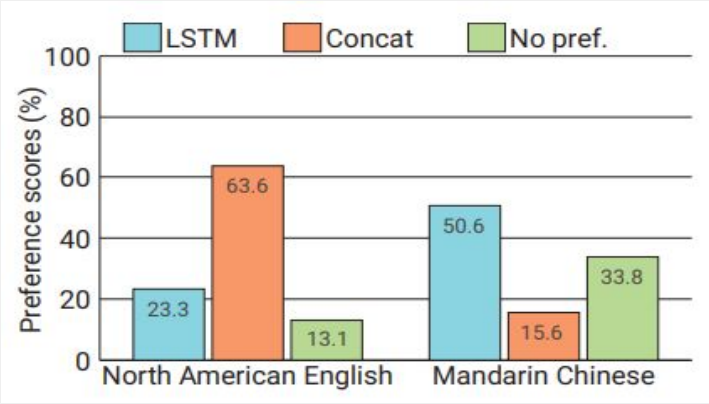
- 또한 모델이 음성 자체 이외에 스피커의 호흡, 입 움직임, 음향과 녹음 품질 등 오디오의 다른 특성도 포착할 수 있었음.

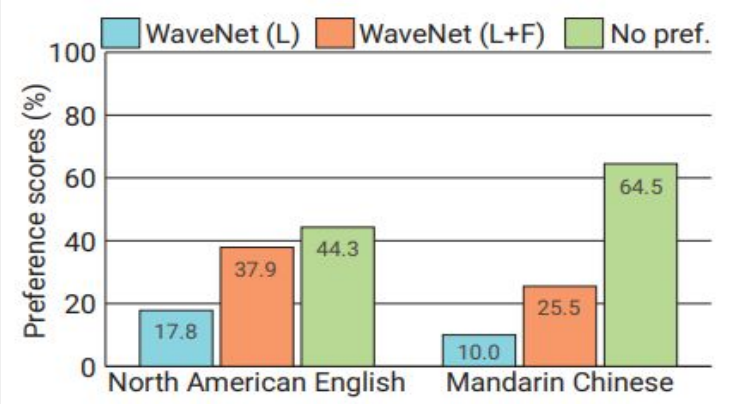
3.2 TEXT-TO-TEXT

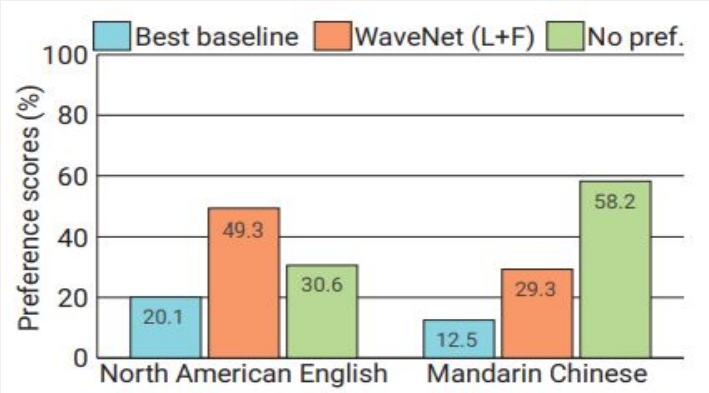
Speech samples	Subjective 5-scale MOS in naturalness	
	North American English	Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	$\mathbf{4.21} \pm 0.081$	$\mathbf{4.08} \pm 0.085$
Natural (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071

- LSTM-RNN 기반 통계 매개 변수, HMM 기반 단위 선택 연결 및 제안된 WaveNet 기반 음성 합성기, 8-bit u-law, 16-bit linear PCM 등에서 음성 샘플의 주관적인 5 스케일 평균 의견 점수
- WaveNet은 이전 기술을 크게 개선하여 자연 음성 및 최상의 이전 모델 간의 격차를 50% 이상 줄임

3.3 MUSIC


LSTM
Concat
No pref.
Preference scores (%)
100
80
60
40
20
0
63.6
23.3
13.1
50.6
15.6
33.8
North American English
Mandarin Chinese


WaveNet (L)
WaveNet (L+F)
No pref.
Preference scores (%)
100
80
60
40
20
0
17.8
37.9
44.3
10.0
25.5
64.5
North American English
Mandarin Chinese


Best baseline
WaveNet (L+F)
No pref.
Preference scores (%)
100
80
60
40
20
0
20.1
49.3
30.6
12.5
29.3
58.2
North American English
Mandarin Chinese