WaveNet:
A generative model for raw audio

Van Den Oord, Aäron, et al.

Deepmind

경상국립대학교 석사과정 오서영

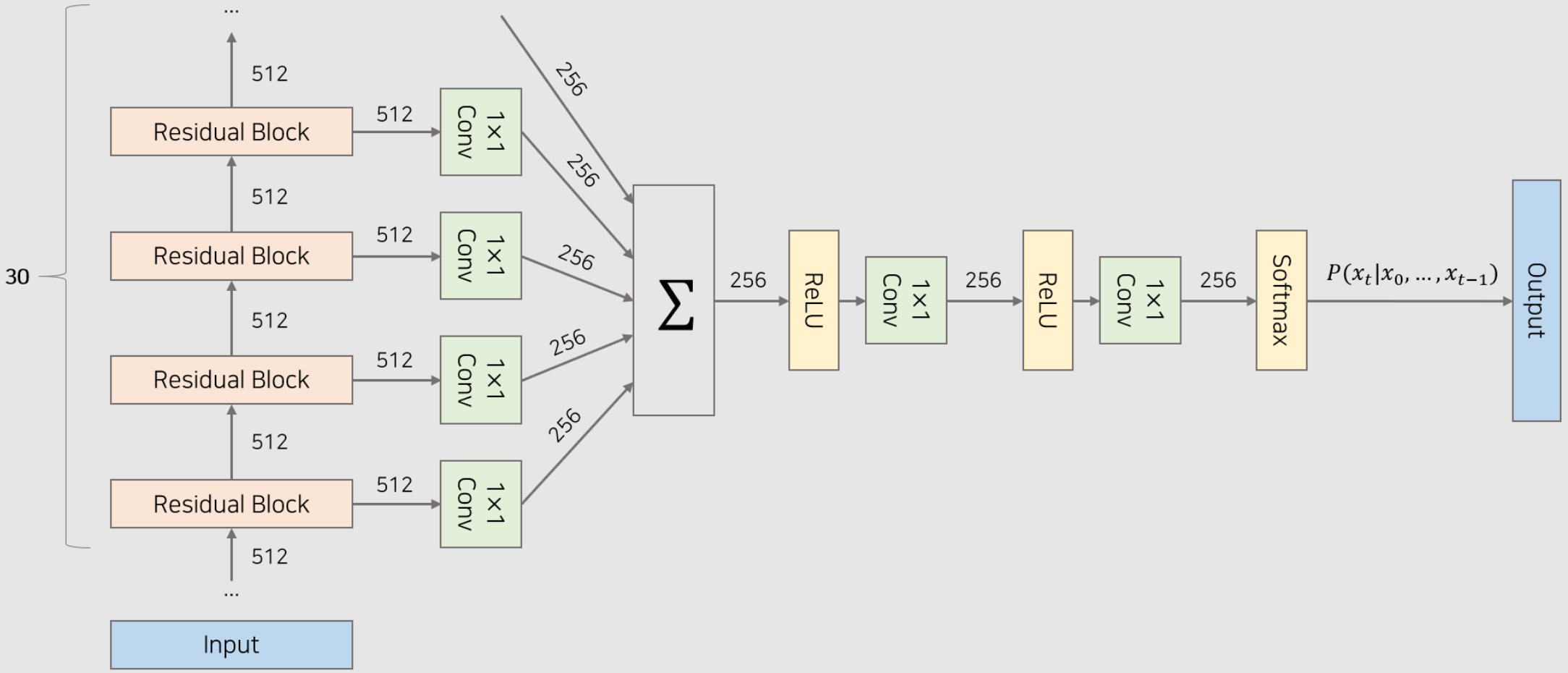
Introduction

Wavenet

- 조건부 분포의 산물로 neural 구조를 사용하여 픽셀 또는 단어에 대한 결합 확률을 모델링 함
- PixelCNN 아키텍쳐를 기반으로 하는 오디오 생성 모델
- 주요 네가지 기여:

1. TTS 분야에서 보고되지 않은 주관적이고 자연스러운 원시 음성 신호를 생성 할 수 있음
2. long-range temporal dependenc를 다루기 위해서, 매우 큰 receptive field를 나타내는 dilated causal convolution 기반 새 아키텍처를 개발함
3. 단일 모델을 사용하여 화자의 신원에 조정된 다른 음색을 생성하는데 사용가능
4. 음악과 같은 다른 오디오 양식을 생성하는데 유망함

Introduction



이미지 출처 : https://joungheekim.github.io/2020/09/17/paper-review/

Wavenet

파형의 결합확률 $x = \{x_1, \ldots, x_T\}$ 은 다음과 같이 조건부 확률의 곱으로 분해됨:

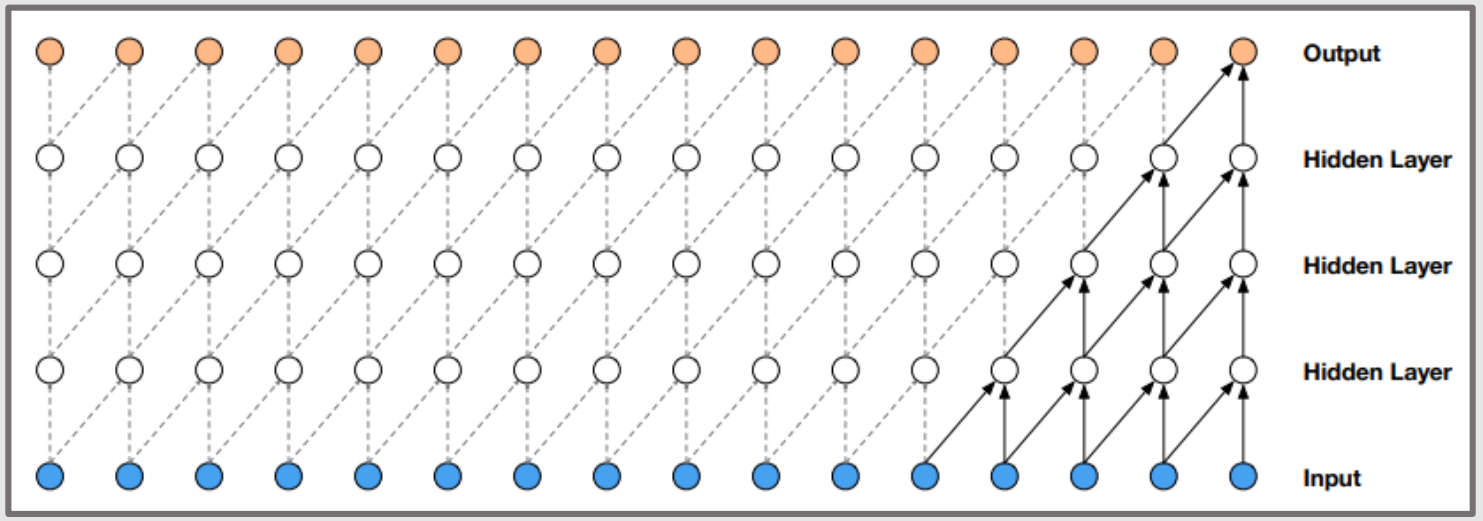
$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

- 따라서 각 오디오 샘플 x_t 는 이전의 이전 시간에 샘플에 따라 조정됨
- 조건부 확률 분포는 컨볼류션 레이어의 스택에 의해 모델링됨, 풀링 X
- 모델의 입력과 출력은 동일한 시간 차원
- 소프트 맥스 레이어를 사용하여 다음 값 x_t에 대한 범주형 분포를 출력함
- 모델은 log-likelihood를 최대화 하도록 최적화됨

Wavenet

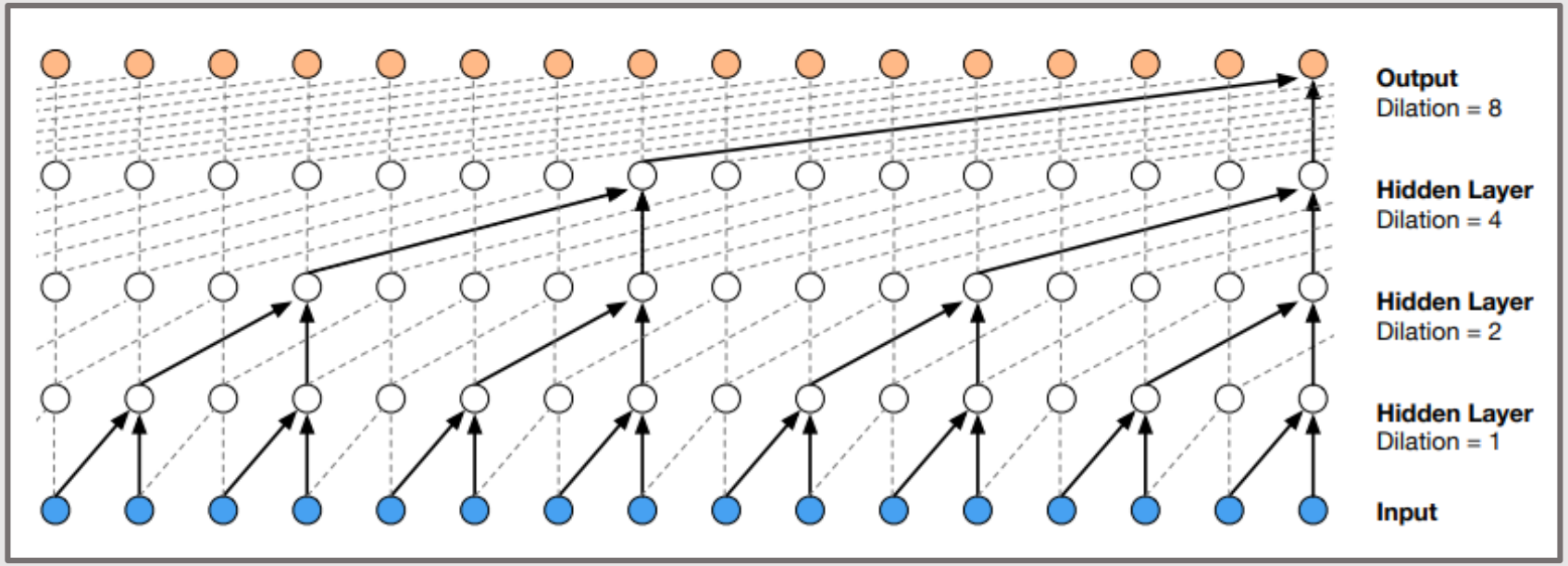
Dilated Causal Convolutions
- Dilated covolution layer + causal convolution layer

1. Causal convolution layer
- 모델이 예측을 모델링 하는 주문을 위반할 수 없는지 확인함
- time step t 에서 나온 $p(x_{t+1} \mid x_1, \ldots, x_t)$ 가 미래의 timestep $x_{t+1}, x_{t+2}, \ldots, x_T$.에 의존하지 않음
- 마스크 텐서를 구성하고 이를 적용하기 전에 이 요소를 컨볼류션 커널과 곱함
- 훈련시간에 모든 time step에 대한 조건부 예측은 실측값 x의 모든 time step이 알려져 있기 때문에 병렬로 이루어질 수 있음
- RNN보다 빠른 훈련, 하지만 많은 층이 필요하거나 receptive field를 확장하기 위해 큰 필터가 필요



Wavenet

2. Dilated causal convolution
- Dilated convolution은 특정단계로 입력값을 건너뛰어, 필터가 길이보다 큰 영역에 필터가 적용되는 컨볼루션
→ 이는 0으로 확장하여 원래 필터에서 파생된 더 큰 필터를 사용하는 것과 equivalent 하지만 더 효율적
- 이 논문에서는 특정 지점 (512)까지 모든 층에 대해 확장을 두 배로 한 다음 (1,2,4,...,512,1,2,4,...,512) 반복한다.



Wavenet

Softmax Distribution

- 개별 오디오 샘플을 통해 조건부 분포 $p(x_t \mid x_1, \ldots, x_{t-1})$ 를 모델링
- 범주형 분포는 형태에 대한 가정을 하지 않기 때문에 더 유연하고 임의의 분포를 더 쉽게 모델링 가능
- 원시 오디오는 일반적으로 16 비트 정수 값 (Timestep 당 하나)의 시퀀스로 저장되므로 SoftMax 레이어는 모든 가능한 값을 모델링하기 위해 타임 스텝 당 65,536 개의 확률을 출력해야합니다.

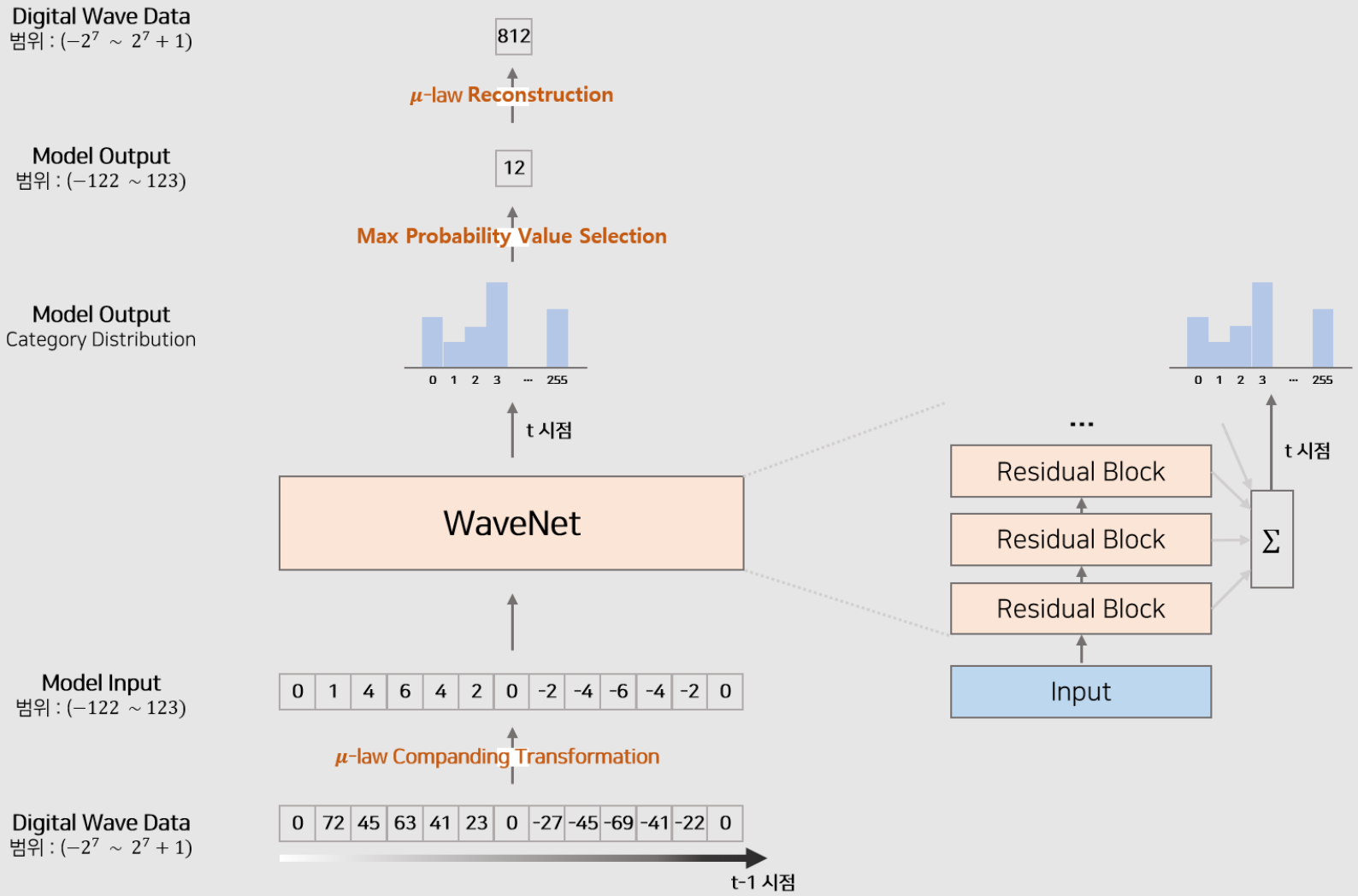
→ 이것을 더 다루기 쉽게 만들기 위해, 먼저 **μ-law companding transformation**을 데이터에 적용한 다음, 256개의 값으로 정량화한다.

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

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

- 이 비선형 양자화는 단순한 선형 양자화 체계보다 훨씬 더 나은 재구성을 생성함

Wavenet



이미지 출처 : https://joungheekim.github.io/2020/09/17/paper-review/

Wavenet

Gated activation units

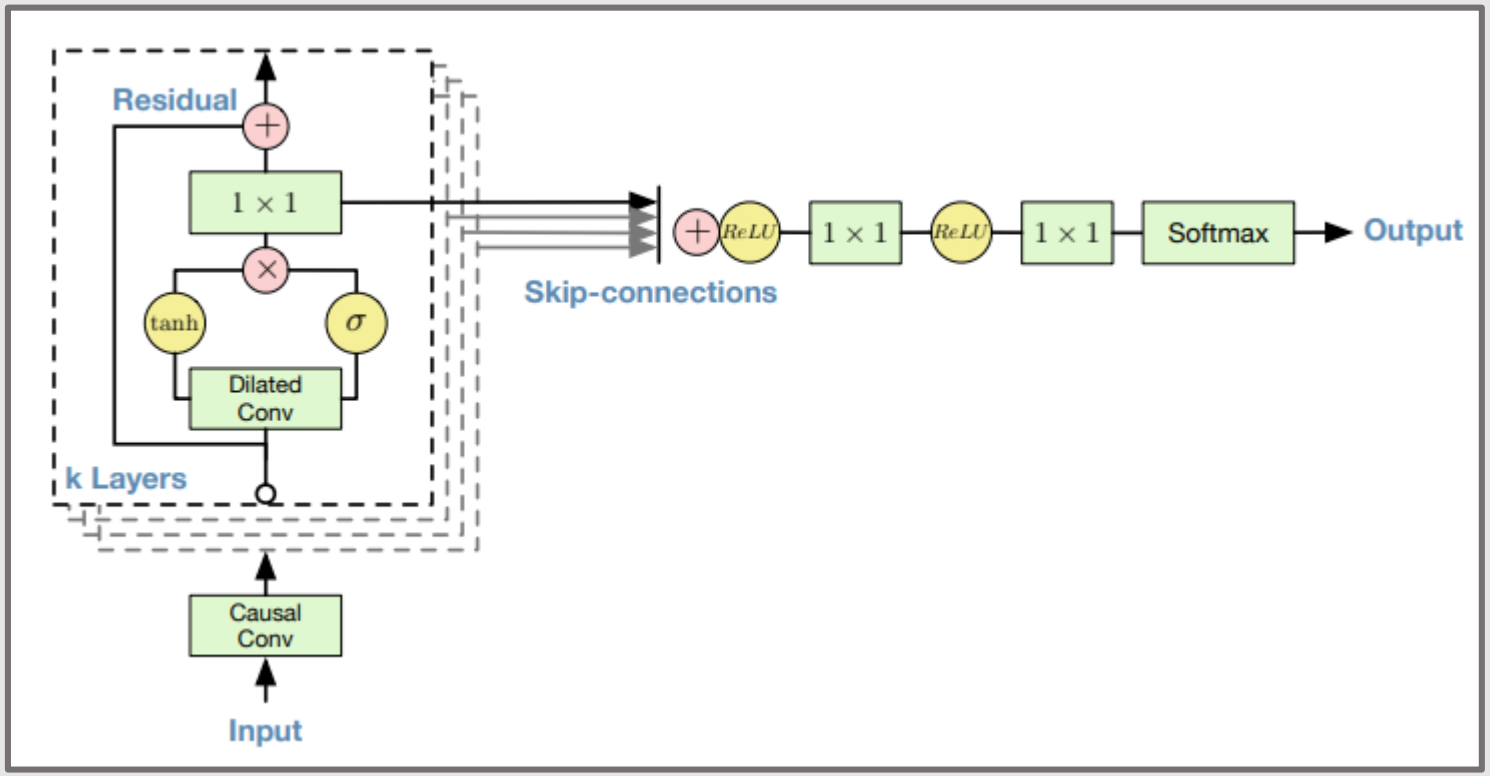
$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \sigma\left(W_{g,k} * \mathbf{x}\right),$$

* : 컨볼루션 연산자
σ : 시그모이드 함수
k : 층 인덱스
f : Filter
g : Gate

Gate : 특정 layer에서 생성한 local feature을 filter로 보고, 이 filter 의 정보를 다음 층에 얼마나 전달할지를 결정함

Wavenet

Residual and skip connections
- 네트워크 전체에서 사용되어 수렴 속도를 높이고 훨씬 더 깊은 모델을 학습가능



Residual block 과 전체구조

Wavenet

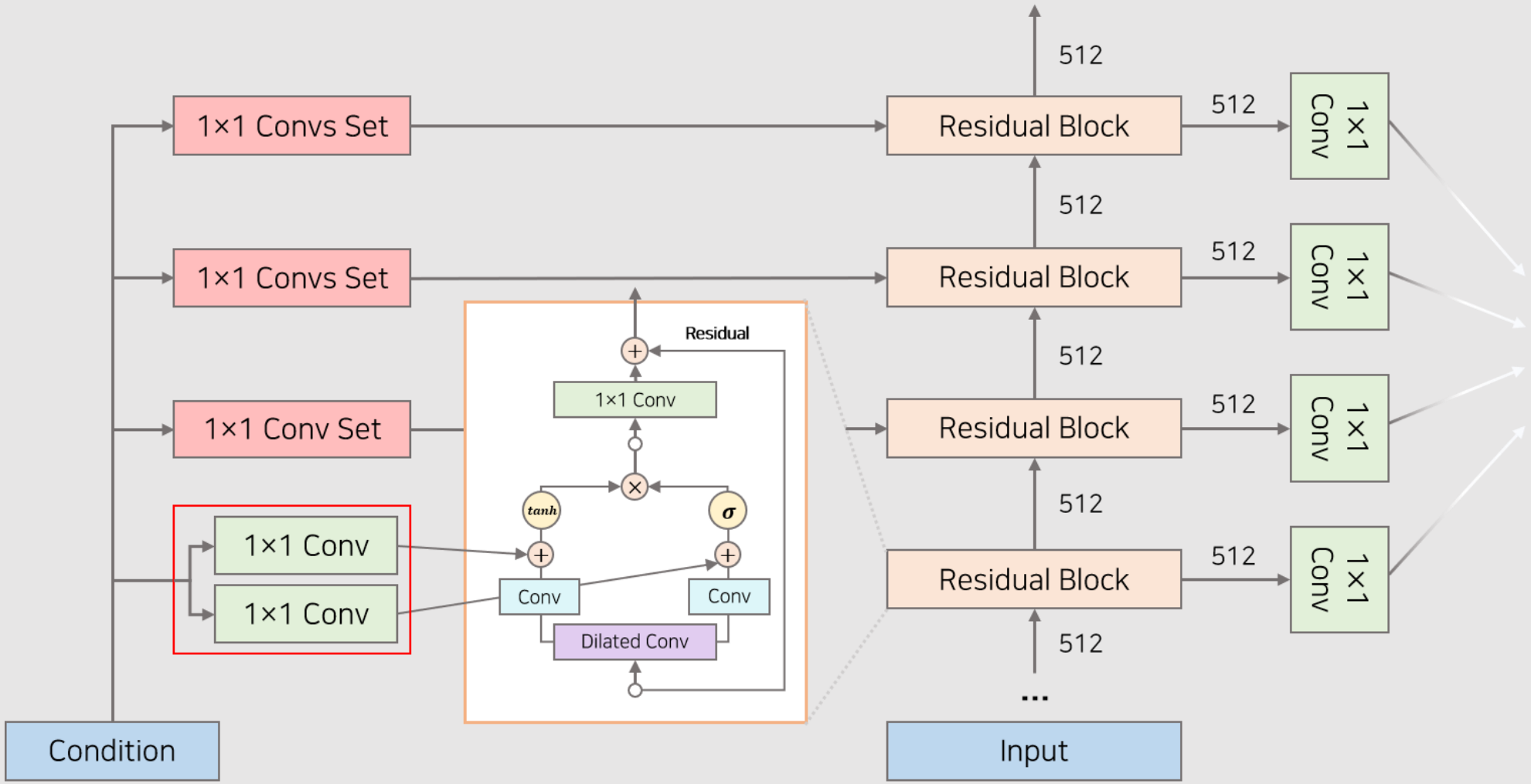
Conditional Wavenet

- 추가 입력 h가 주어지면, Wavenet은 오디오의 조건부 분포 p(x | h)를 모델링 가능

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

- 다른 입력 변수에서 모델을 조절함으로써 Wavenet의 생성을 가이드하여 필요한 특성을 가진 오디오를 생성할 수 있음
- 다중 화자 세팅에서는 추가 입력으로 모델에 화자 id를 공급하여 화자를 선택할 수 있음
- TTS의 경우 텍스트에 대한 정보를 추가 입력으로 제공

Wavenet



이미지 출처 : https://joungheekim.github.io/2020/09/17/paper-review/

Wavenet

우리는 두 가지 다른 방법으로 다른 입력에 모델을 조건화함

1. Global conditioning

- 모든 timestep에서 출력 분포에 영향을 미치는 단일 잠재적인 표현 **h**를 특성으로함.
- TTS 모델에 화자가 embedding

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

$V_{*,k}$: 학습 가능한 linear projection, $V^{T}_{*,k}$는 시간차원에 걸쳐 broadcast됨

2. local conditioning

- 오디오 신호보다 낮은 샘플링 주파수를 갖는 두 번째 시계열 ht를 가짐
(TTS 모델의 언어적인 특징)
- 먼저 오디오 신호와 동일한 해상도로 새로운 시계열 y = f(h)에 매핑하는 transposed convolutional network (학습된 업 샘플링)를 사용하여 이 시계열을 변환함
→ 그런 다음 아래와 같이 activation unit에 사용됨

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

$V_{f,k}$*y 는 1x1 convolution

Experiments

Multi-Speaker speech generation
- VCTK(Voice Cloning Toolkit) 데이터셋
- Conditioning은 화자 id를 원 핫 벡터 형태로 모델에 공급하여 적용됨
- 단일 WaveNet은 스피커의 단일 핫 인코딩을 조건으로하여 모든 스피커의 음성을 모델링가능
- 우리는 이 모델이 음성 자체와는 별도로 오디오의 다른 특징들도 잡아낼 수 있음을 관찰함

Text-to-speech
- TTS task의 WaveNets는 입력 텍스트에서 파생 된 언어 특징에 대해 local로 conditioning됨
- 언어 적 특징 외에도 대수 기본 주파수 (log F_0) 값에 conditioning 된 WaveNets을 훈련함.

Hidden Markov model (HMM), long short-term memory recurrent neural network (LSTM-RNN)
Wavenet(L+F) : Linguistic features + logF0

Experiments

TTS task의 WaveNets의 성능:

1. Subjective paired comparison tests

- 각 샘플의 쌍을 들은 후, 피실험자들은 그들이 선호하는 것을 선택하도록 요청 받음

2. Mean opinion score (MOS) tests

- 각각의 stimulus을 들은 후, 피실험자들은 five-point Likert scale score로 stimulus 의 자연스러움을 평가하도록 요청받음

	Subjective 5-scale MOS in naturalness	
Speech samples	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

WaveNet은 이전 최첨단 기술을 대폭 개선하여 자연스러운 화법과 최상의 이전 모델 간의 격차를 50% 이상 줄였습니다.

Conclusion

- WaveNet는 autoregressive 이며 dilated convolutions 으로 causal filters 를 결합하여 이들의 receptive field 가 깊이를 가지고 지수적으로 성장할 수 있도록 하는데, 이는 오디오 신호에서의 long-range temporal dependencies 을 모델링 하는데 중요한 역할을 함
- WaveNet은 global(화자의 신원) 또는 local(언어 특징)의 다른입력에 conditioning 될 수 있음