

# Wavenet

WaveNet: A Generative Model for Raw Audio(Google DeepMind, 2016)

# WaveNet

## 보코더(Vocoder)의 한 종류

- 음성의 등장 확률을 학습하는 확률론적 모델
- 과거(t-1) 시점까지 음성데이터와 멜 스펙트로그램을 조건으로 현재(t) 시점의 특정 음성 등장 확률을 추출

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

# WaveNet 입출력

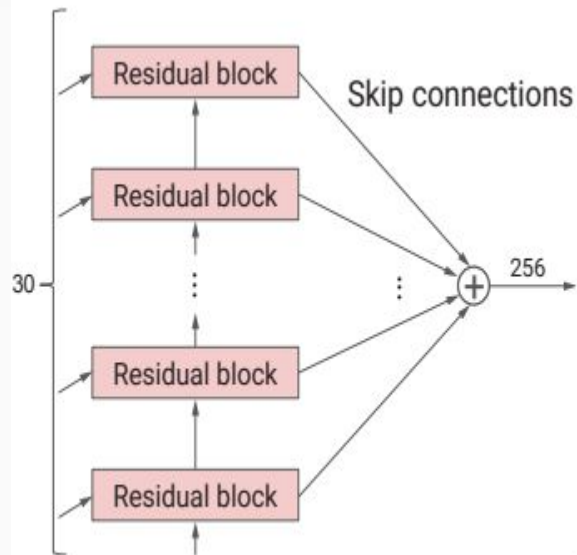
- 입력: 멜스펙트로그램 / 출력: 음성
- 일반적인 음성 데이터는 각 샘플을 16(bit) 정수 값으로 저장  
→  $-2^{15} \sim 2^{15}-1$  사이의 수로 표현  
→ 즉, **65,536개**의 숫자가 나올 확률(  $P(-2^{15}|x_1, \dots, x_{t-1}) \sim P(2^{15}-1|x_1, \dots, x_{t-1})$  )을 계산
- 따라서 이를 256개의 숫자로 변환(총 256개의 확률)  
→  $\mu$ -law Companding Transformation한 값이 WaveNet에서 입력으로 사용
$$f(x_t) = \text{sign}(x_t) \frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)}$$
- 출력 값 역시  $-128 \sim 127$ (256개) 범위의 정수로 최종적으로 이 정수를 이용해 음성 디지털 데이터로 변형

# WaveNet 구조

30개의 Residual Block으로 구성

각 Residual Block의 output이  
skip connection으로 합쳐져 최종 출력으로 활용

**Residual Block**  
= Dilated Casual Convolution + Gated Activation Units



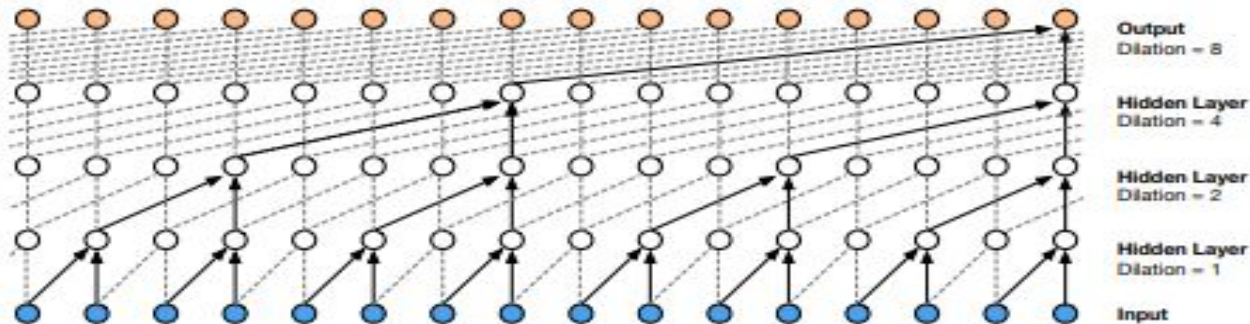
# WaveNet 구조

Dilated Casual Convolution: 과거 음성정보를 이용해 현재 시점의 정보를 생성

보통의 RNN/LSTM을 이용하게되면 비효율적(수용 범위가 넓으면 그 만큼 많은 연산)

Casual Convolution: 시간 순서를 고려하여 과거 정보만을 접근하여 정보를 추출

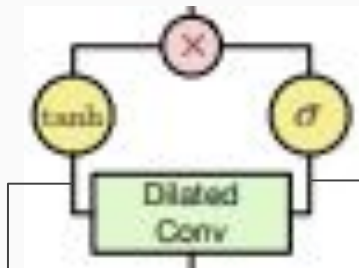
Dilated Convolution: 추출간격을 조절하여 Layer를 적게 쌓아도 **더 넓은 수용범위**를 갖음



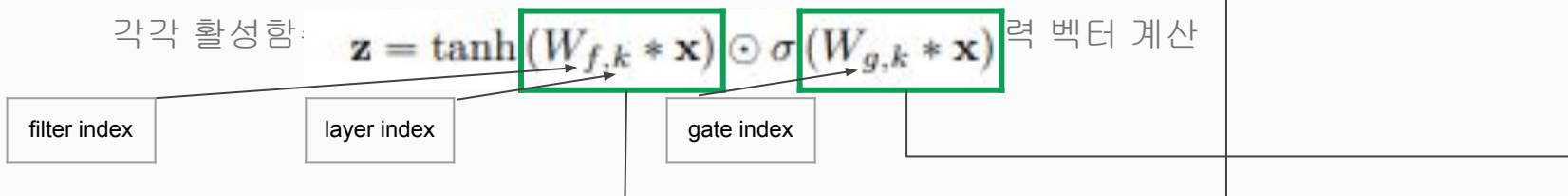
# WaveNet 구조

Gated Activation Units: Dilated Convolution에서 생성된 정보를 다음 Layer에 얼마나 전달할지 결정

filter 경로: Dilated Conv에서 생성된 정보를 가공 (**tanh** 부분)  
gate 경로: filter 경로에서 가공된 정보를 다음 Layer에 얼마나 전달할지 결정 (0~1) (**시그모이드** 부분)



Dilated Convolution된 값이 각각 convolution 연산을 거치고  
각각 활성화함:  $\mathbf{z} = \tanh(W_{f,k} * \mathbf{x}) \odot \sigma(W_{g,k} * \mathbf{x})$  력 벡터 계산

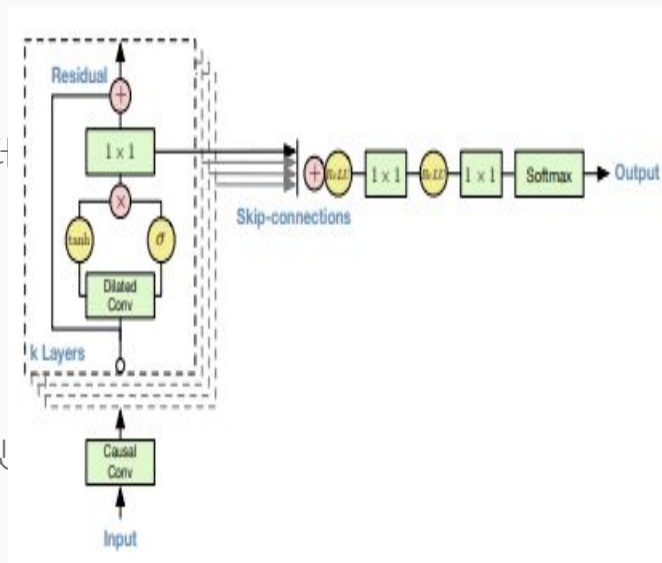


# WaveNet 구조

## Residual Block의 Output

→ 최종적으로 convolution과 gate를 통과하여 생성된 벡터  
Skip Connection으로 입력과 연결되어 최종 결과 생성

그레디언트 소멸 문제를 방지하여 많은 Layer를 쌓을 수 있



# Conditional WaveNet

## Conditional WaveNet 특징

- Conditional Modeling  $P(\mathbf{x} | \mathbf{h})$ 을 통해 특징을 추가하고 특징에 맞는 음성 생성
- Condition으로 전역적 특징(화자), 지역적 특징(멜스펙트로그램) 부여 가능
- 모델에 Condition을 추가함으로써 특별한 성질을 가진 오디오 생성 가능

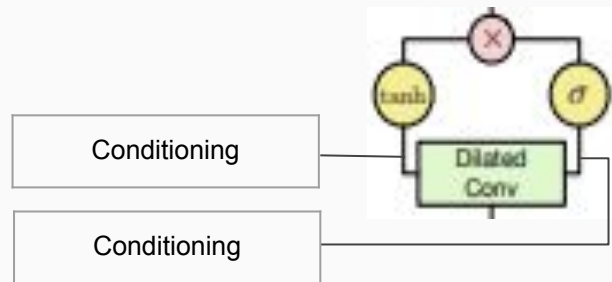
$$p(\mathbf{x} | \mathbf{h}) = \prod_{t=1}^T p(x_t | x_1, \dots, x_{t-1}, \mathbf{h})$$



# Conditional WaveNet

전역적 특징

- 출력 분포 전체에 영향을 미치는 요소(**h**)
- 화자의 목소리에 대한 특징이 해당될 수 있음



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

# Conditional WaveNet

## 지역적 특징

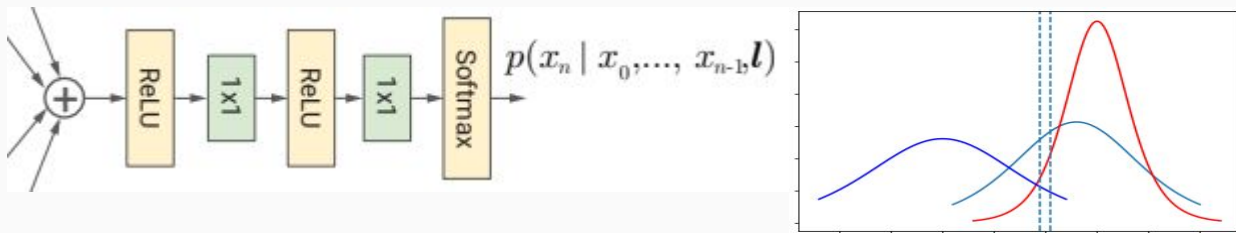
- 특정한 시간에 영향을 미치는 요소( $h_t$ )로 멜스펙트로그램이 해당될 수 있음
- $h_t$ 에 활성화함수를 적용하기 위해서 **오디오의 step수와 동일**하게 맞춰줘야 함  
즉, 멜스펙트로그램이 주어졌을 때 오디오의 step수와 동일하게 **upsampling**을 해줘야 함
- 스텝수를 맞춰준 정보를  $y = f(h)$ 라 하면 다음과 같이 연산

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

# Conditional WaveNet

## WaveNet Output

- 디지털 음성파형을 만들기 위해 **WaveNet**은  $-2^{15} \sim 2^{15}-1$  사이의 수로 표현
- 그러나 음성 생성 시 주변의 관계를 고려한 확률분포를 생성해야 함  
**Logistic Distribution**의 파라미터(평균, 분산(표준편차))를 추정
- 추정한 파라미터를 이용해서 **Mixture of Logistic Distribution(MoL)**을 구성
- 시점  $t$ 에서  $-2^{15} \sim 2^{15} - 1$ 의 값을 가질 수 있는 확률분포를 통해 출력을 생성



# Conditional WaveNet

## WaveNet 학습

- 오차함수: **Negative log-likelihood**(음의 로그우도)  
시점  $t$ 에서의 실제 음성 데이터(**True Label**)와 생성한 확률분포의 차이
- 데이터: **24.6시간**의 한 사람의 음성을 담은 **US English dataset**
- **MOS 지표**(전화 통화 성능 평가에 주로 사용되는 통계)를 이용

System	MOS
Parametric	$3.492 \pm 0.096$
Tacotron (Griffin-Lim)	$4.001 \pm 0.087$
Concatenative	$4.166 \pm 0.091$
WaveNet (Linguistic)	$4.341 \pm 0.051$
Ground truth	$4.582 \pm 0.053$
Tacotron 2 (this paper)	<b><math>4.526 \pm 0.066</math></b>

### **MOS 4~5: 고품질**

MOS 3.5~4: 중간품질

MOS 3~3.5: 대화는 잘 이루어지나 품질저하를 느낌

MOS 2.5~3: 군사품질, 대화 가능하지만 집중해야 함

→ Tacotron2로부터 생성된 음성이

**실제음성과 비슷한 평가를 받음**