



# DL Seminar

## WaveNet

A Generative Model For Raw Audio



**한양대학교**  
HANYANG UNIVERSITY

인공지능 연구실  
김지성

# Introduction

## Google DeepMind WaveNet

지금 바로 텍스트를 음성으로 변환해보세요.

원하는 내용을 입력하고 언어를 선택한 다음 'Speak It(음성 변환)'을 클릭하세요.

Text to speak:

Google Cloud Text-to-Speech enables developers to synthesize natural-sounding speech with 100+ voices, available in multiple languages and variants. It applies DeepMind's groundbreaking research in WaveNet and Google's powerful neural networks to deliver the highest fidelity possible. As an easy-to-use API, you can create lifelike interactions with your users, across many applications and devices.

text [ssml](#)

Language / locale

English (United States)

Voice type

WaveNet

Voice name

en-US-Wavenet-D

Audio device profile

Default

Speed:

1.00

Pitch:

0.00

[Show JSON](#) ▼

▶ SPEAK IT

# Introduction

## Sound Data



A second of Generated speech

- 음성데이터 : 시계열 데이터  
사실적인 소리를 위해 1초에 16000bit(16k bps)를 사용
- Raw Audio : 압축되지 않은 오디오  
Ex) 이미지에서 비트맵 이미지

# Introduction

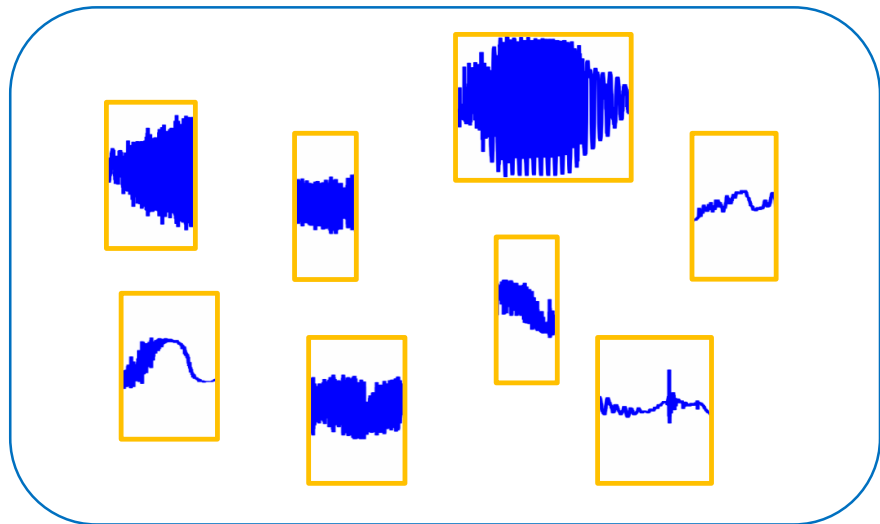
---

## Realistic Sound Data

- 32k bps : AM 품질
- 96k bps : FM 품질
- 192k bps : DAB (디지털 오디오 방송) 품질
- 224 ~ 320k bps : CD 품질
- 96 ~ 640k bps : 손실 데이터 압축 중 돌비 디지털 (AC3) 규격의 비트레이트 범위
- 1,536k bps : DTS, CD 디지털 오디오의 PCM 소리 포맷
- 6,000k bps : 손실 데이터 압축 중 DTS-HD High Resolution AUDIO 규격의 최대 비트레이트
- 18,000k bps : 무손실 데이터 압축 -> Dolby TRUE HD 규격의 최대 비트레이트 (VBR)
- 24,500k bps : 무손실 데이터 압축 -> DTS-HD MASTER AUDIO 규격의 최대 비트레이트 (VBR)

# Introduction

## Classical speech generation method



녹음된 음소정보

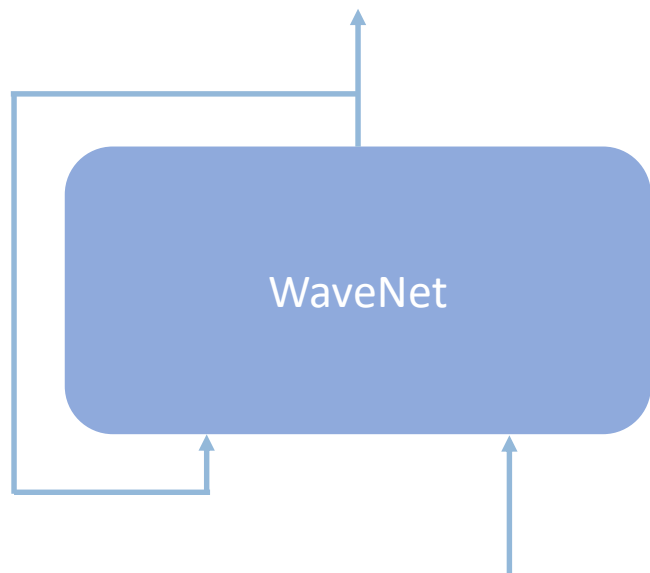


조합된 발음

- 파편화된 음성샘플을 조합하여 발음을 완성한다.
- 자연스러운 조합이 매우 어려워 일반적으로 부자연스러운 발음이 나타난다.

# Introduction

## Objective



Linguistic features

- CNN을 사용해서 사실적인 시계열데이터를 생성하고 싶다
- 장기간의 시간 의존성을 유지하고 싶다.
- 다양한 Condition에 대해 유연하게 대처하고 싶다.

# Method

## Causal Convolution Layers



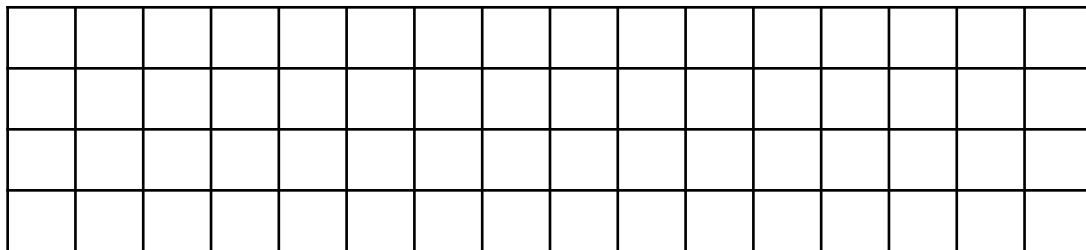
16bit의 경우의 수(65536개)

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

$\mu$ -law companding ( $\mu=255, -1 < x_t < 1$ )



8bit의 경우의 수(255개)



SoftMax를 위한 One-Hot 인코딩

## Method

---

### WaveNet

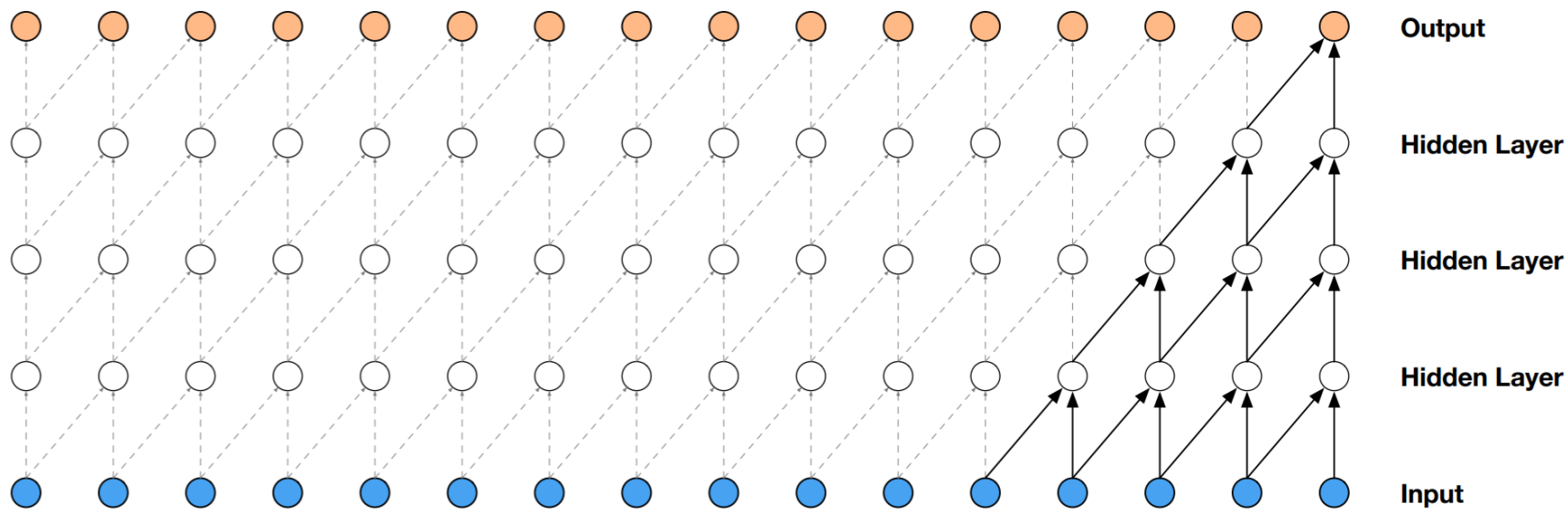
$$p(\mathbf{x}) = \prod_{t=1}^T p(x_t \mid x_1, \dots, x_{t-1})$$

- 음성  $X = \{x_1, \dots, x_{t-1}\}$   
오디오 샘플  $x_t$ 는  $x_{t-1}$ 까지의 샘플에 대해 조건화됨
- 조건부 확률 분포를 Stack of Convolution Layer로 모델링  
PixelCNN을 이용



# Method

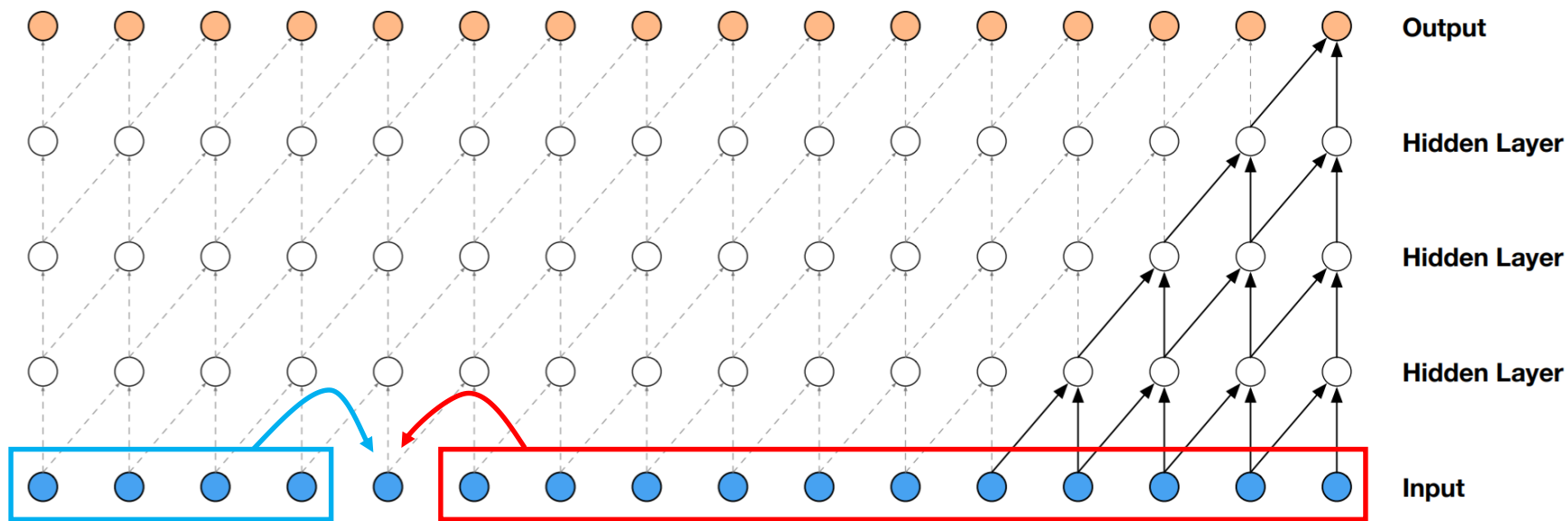
## Causal Convolution Layers



Visualization of a stack of **causal** convolutional layers

# Method

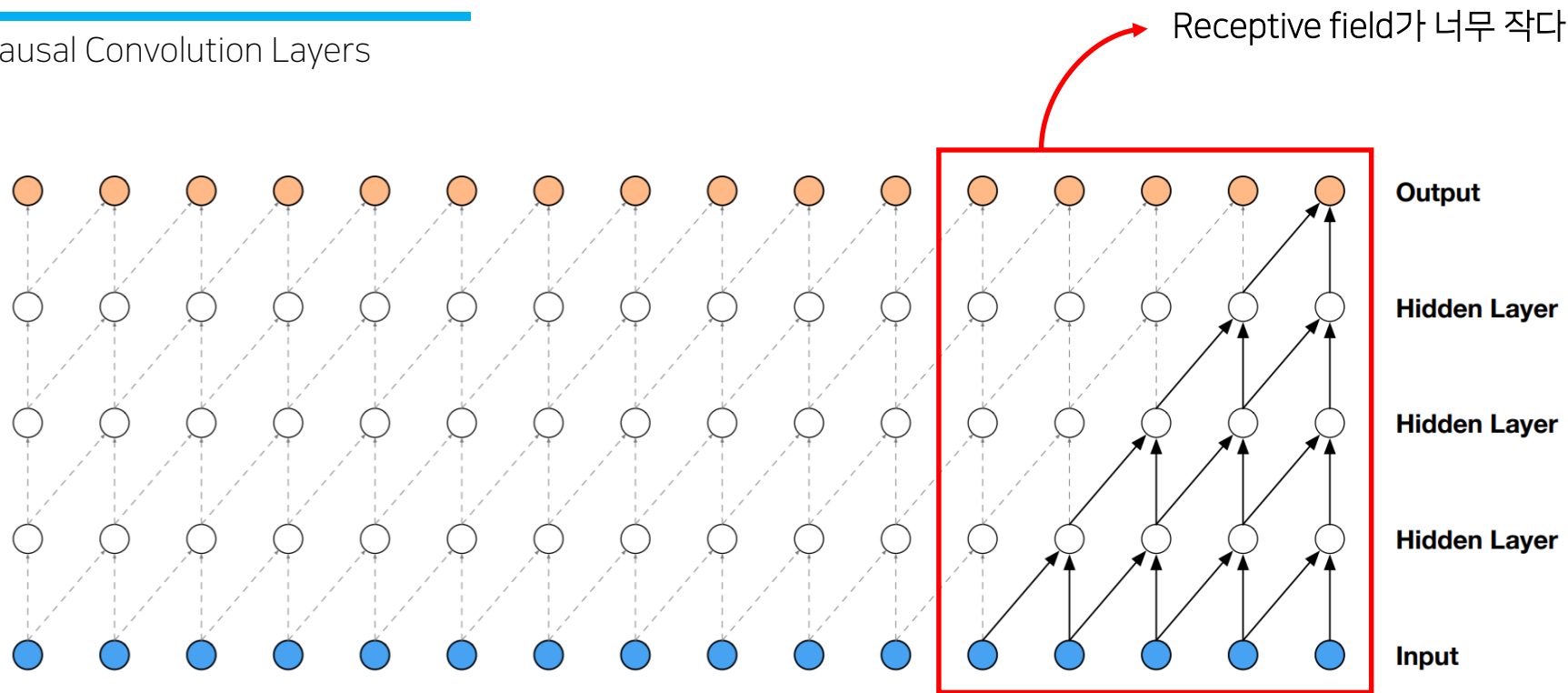
## Causal Convolution Layers



Visualization of a stack of **Non causal** convolutional layers

# Method

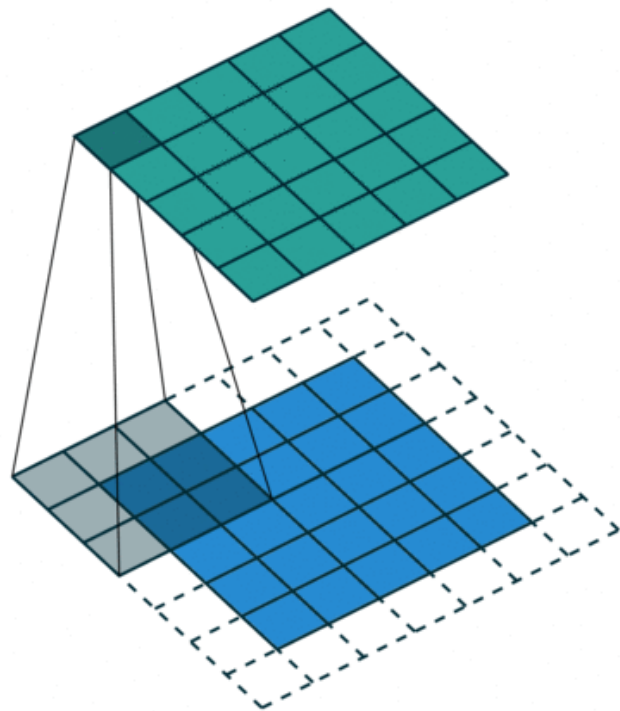
## Causal Convolution Layers



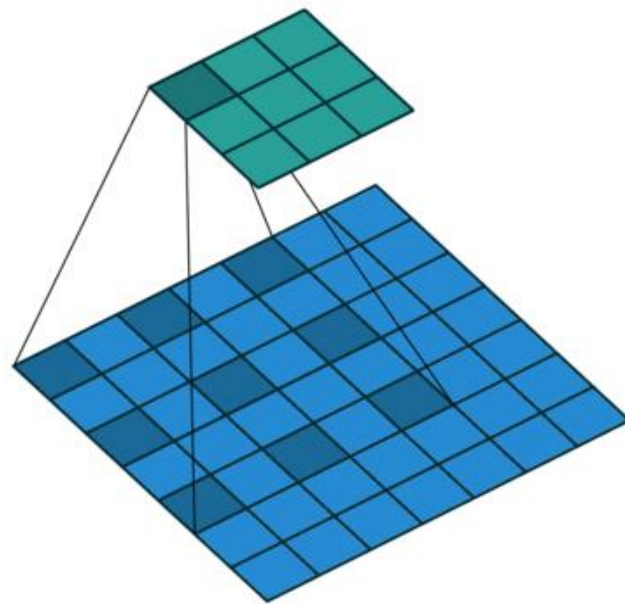
Visualization of a stack of causal convolutional layers

# Method

## Dilated Convolution



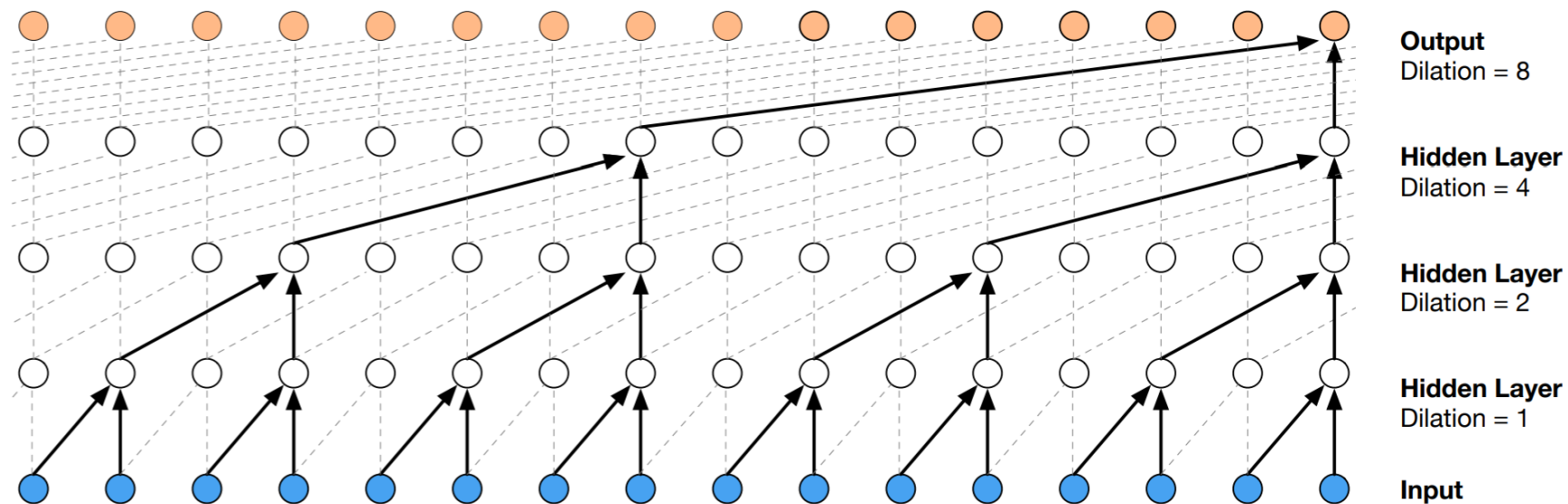
Convolutions



Dilated Convolutions

# Method

## Dilated causal Convolution Layers



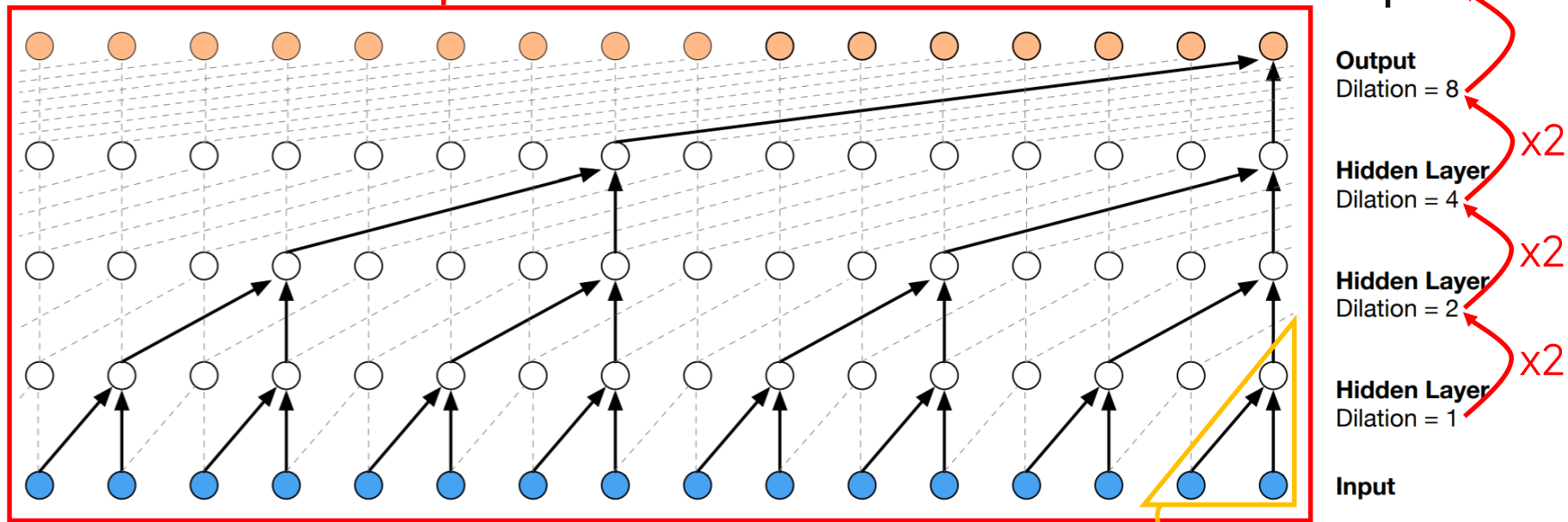
Visualization of a stack of **dilated** causal convolutional layers

# Method

Dilated causal Convolution Layers

Receptive field = Last dilation x 2

Dilation = 512

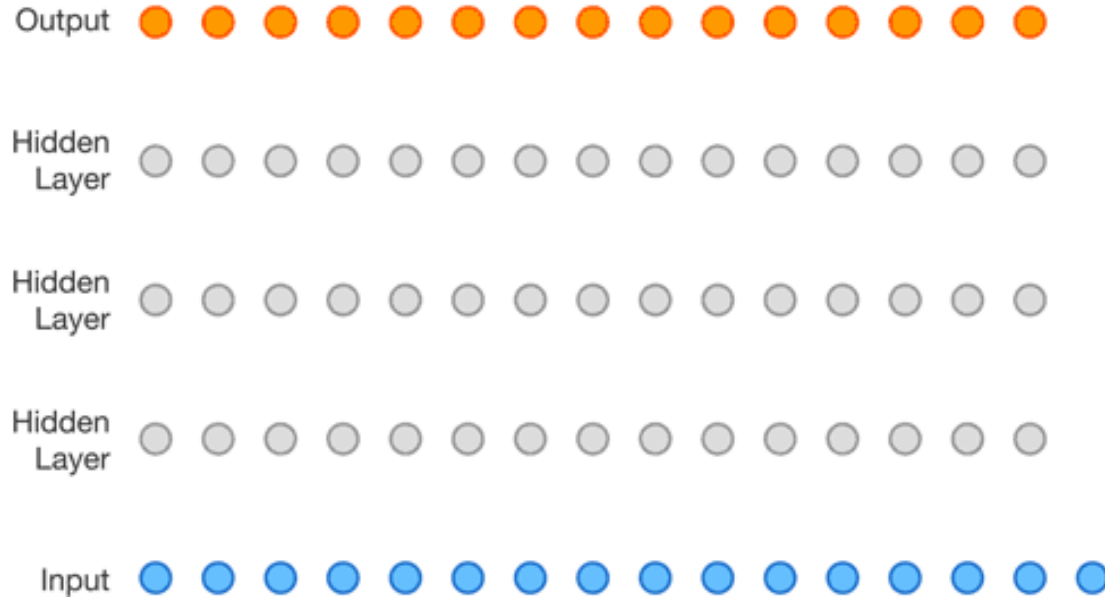


Visualization of a stack of **dilated** causal convolutional layers  $\tanh(W_{x,k} * x)$

1024 x 1 필터보다 효율적이다!

# Method


## Dilated causal Convolution Layers



Visualization of a stack of dilated causal convolutional layers

# Method

## Gated Activation Units

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


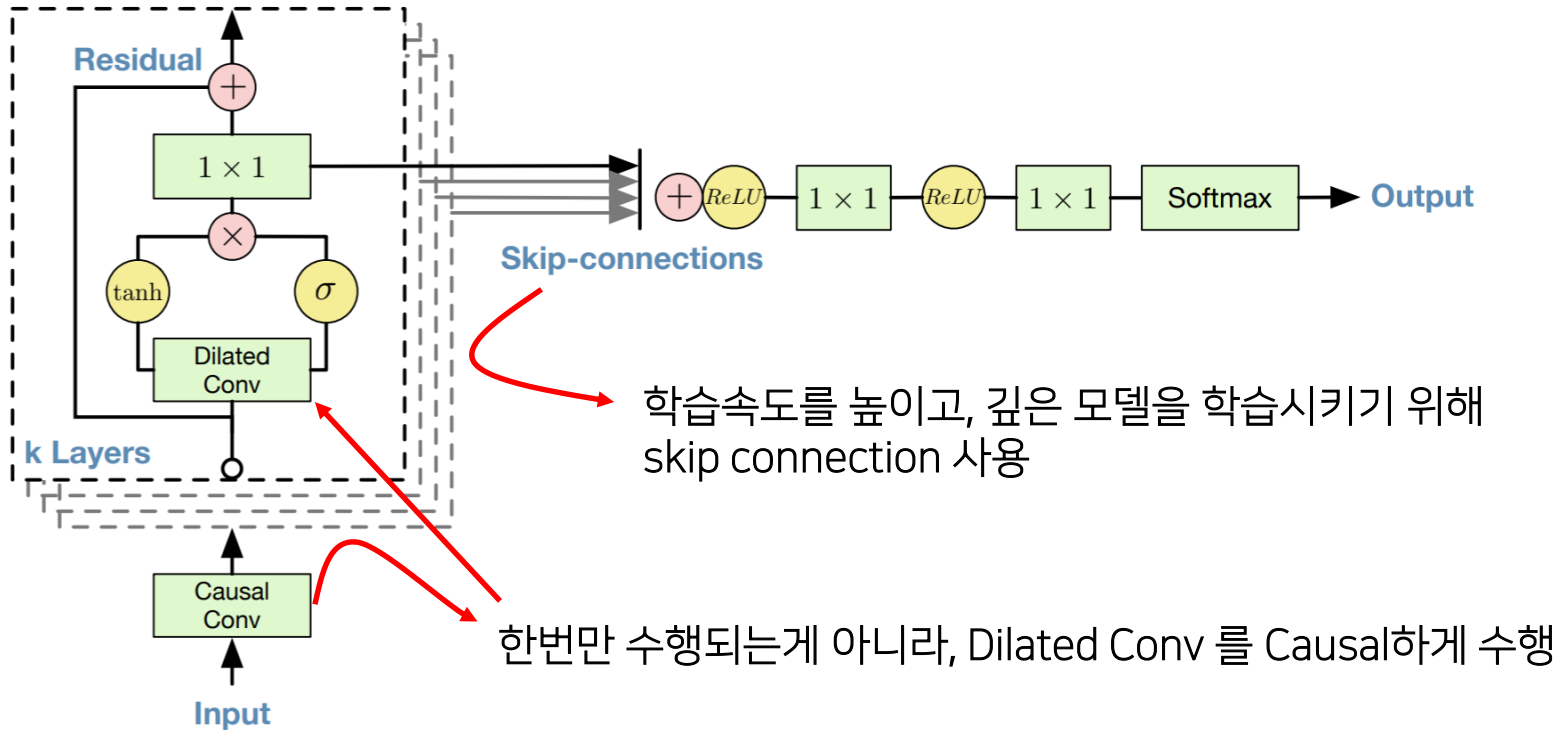
- $\mathbf{Z}$  : output data
- $\mathbf{W}_f$  : learnable convolution filter of filter (for generate)
- $\mathbf{W}_g$  : learnable convolution filter of gate
- $*$  : convolution operator
- $\mathbf{X}$  : input data
- $\odot$  : Element-wise multiplication
- $\sigma$  : sigmoid function
- $k$  : layer index

Convolution을 수행한 값을 얼마나 참여시킬지 결정하는 Gate Unit을 추가



# Method

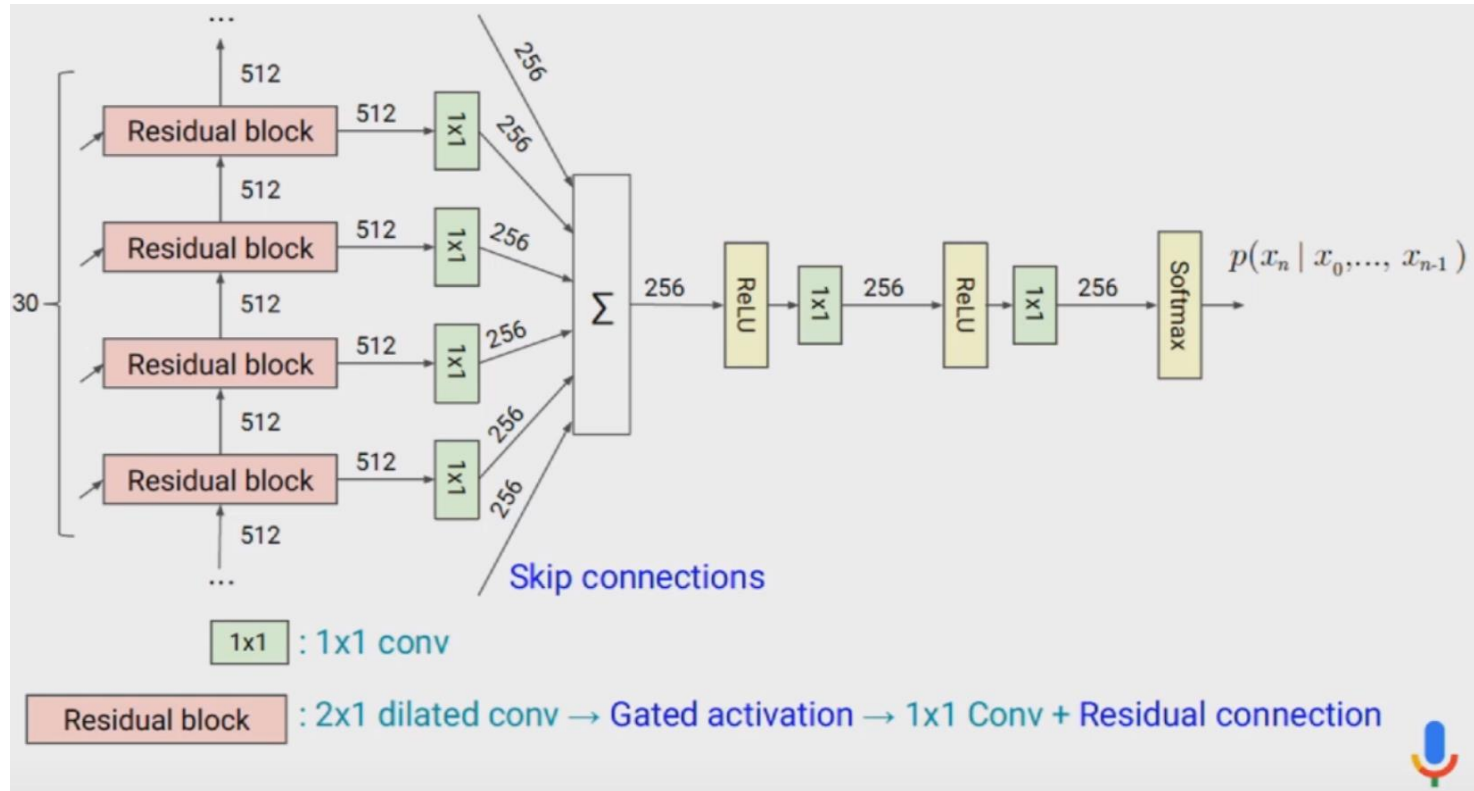
## Entire Architecture



Overview of the residual block and the entire architecture

# Method

## Entire Architecture



Overview of the residual block and the entire architecture

## Method

### Global Conditional WaveNets

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

새로운 조건  $\mathbf{h}$ 가 추가될 경우의 조건부 확률

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

- $V_{?,k}$ : learnable vector
- $V^T$ : T 번에 걸쳐 broadcast

남자 목소리, 여자 목소리 처럼 데이터 전역에 적용하고 싶은 조건일 때 T는 무시  
일부에만 적용하고 싶을 때 broadcast 시간 T 사용

## Method

### Local Conditional WaveNets

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

새로운 조건  $\mathbf{h}$ 가 추가될 경우의 조건부 확률

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

- $V_{?,k}$ : learnable vector
- $\mathbf{y}$ : 새로운 시계열 데이터
- $* \mathbf{y}$ :  $\mathbf{y}$ 를 1x1 convolution

실험해보니 Local보다는 Global Conditioning이 Wavenet에 적합했다.

# Experiments

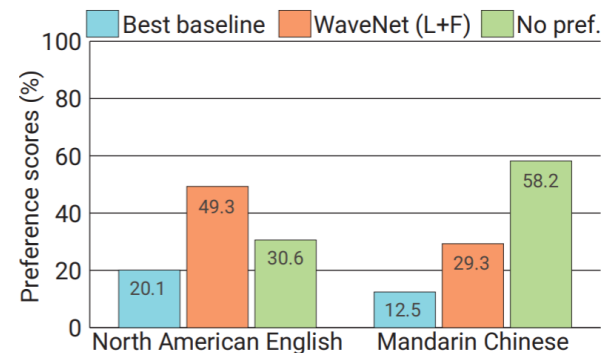
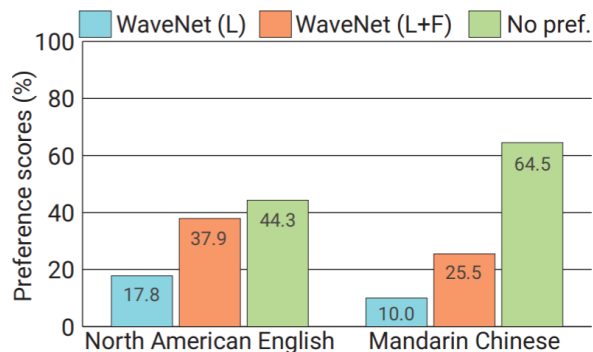
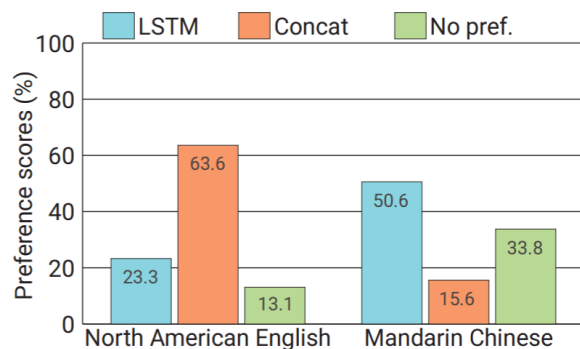
## Text To Speech

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

Subjective 5-scale mean opinion score of speech samples

# Experiments

## Music



## Conclusion

- CNN을 사용해서 시계열데이터 생성  
PixelCNN구조를 기반으로 오디오를 생성
- $\mu$ -law companding을 사용해도 사실적인 소리가 생성됨  
Softmax가 예측할 경우의 수 1/256 만큼 감소
- 장기간의 시간 의존성을 위해 Dilated Convolution 사용  
receptive field가 확장되며, filter size를 높이는 것 보다 효율적
- 하나의 모델이 다양한 목소리를 가질 수 있다.  
활성함수에 Condition을 추가할 수 있어 다양한 Condition에 유연한 학습이 가능

# 참고문헌

---

Deepmind - WaveNet : A generative Model for Raw Audio  
<https://deepmind.com/blog/wavenet-generative-model-raw-audio/>

Arxiv - WaveNet : A generative Model for Raw Audio  
<https://arxiv.org/pdf/1609.03499.pdf>

Youtube - WaveNet - A Generative Model for Raw Audio  
[https://www.youtube.com/watch?v=GyQnex\\_DK2k&t=1325s](https://www.youtube.com/watch?v=GyQnex_DK2k&t=1325s)

Youtube - Toward WaveNet speech synthesis  
<https://www.youtube.com/watch?v=m2A9g6Xu91I&t=174s>

Blog - WaveNet : A generative Model For Raw Audio  
<https://computer-nerd.tistory.com/71>

Github - WaveNet : A generative Model For Raw Audio(2016)  
[https://github.com/hwkim94/hwkim94.github.io/wiki/WAVENET:-A-GENERATIVE-MODEL-FOR-RAW-AUDIO\(2016\)](https://github.com/hwkim94/hwkim94.github.io/wiki/WAVENET:-A-GENERATIVE-MODEL-FOR-RAW-AUDIO(2016))

Github - Pixel CNN & WaveNet Review  
<https://yangyangii.github.io/2017/12/30/PixelCNN-WaveNet.html>

Wikipedia - WaveNet  
<https://en.wikipedia.org/wiki/WaveNet>

Github - tensorflow-wavenet  
<https://github.com/ibab/tensorflow-wavenet>

Slid - An implementation of WaveNet  
<https://slideplayer.com/slide/13025889/>

Medium - WaveNet: Increasing reception field using dilated convolution  
<https://medium.com/@kion.kim/wavenet-a-network-good-to-know-7caaae735435>

감사합니다.