# WAVENET
## :Audio 생성 AI

발표자: 이세영

2023. 02. 28

# 목차

1. Introduction
2. Wavenet
3. Experiment
4. Conclusion

# 01 Introduction

## WaveNet이란?

- 자연스러운 audio sample 생성
- dilated causal convolution 기반 새로운 아키텍처 제안
- 사람의 말소리 뿐만 아니라 음악 등 다양한 오디오 양식 생성에 사용 가능

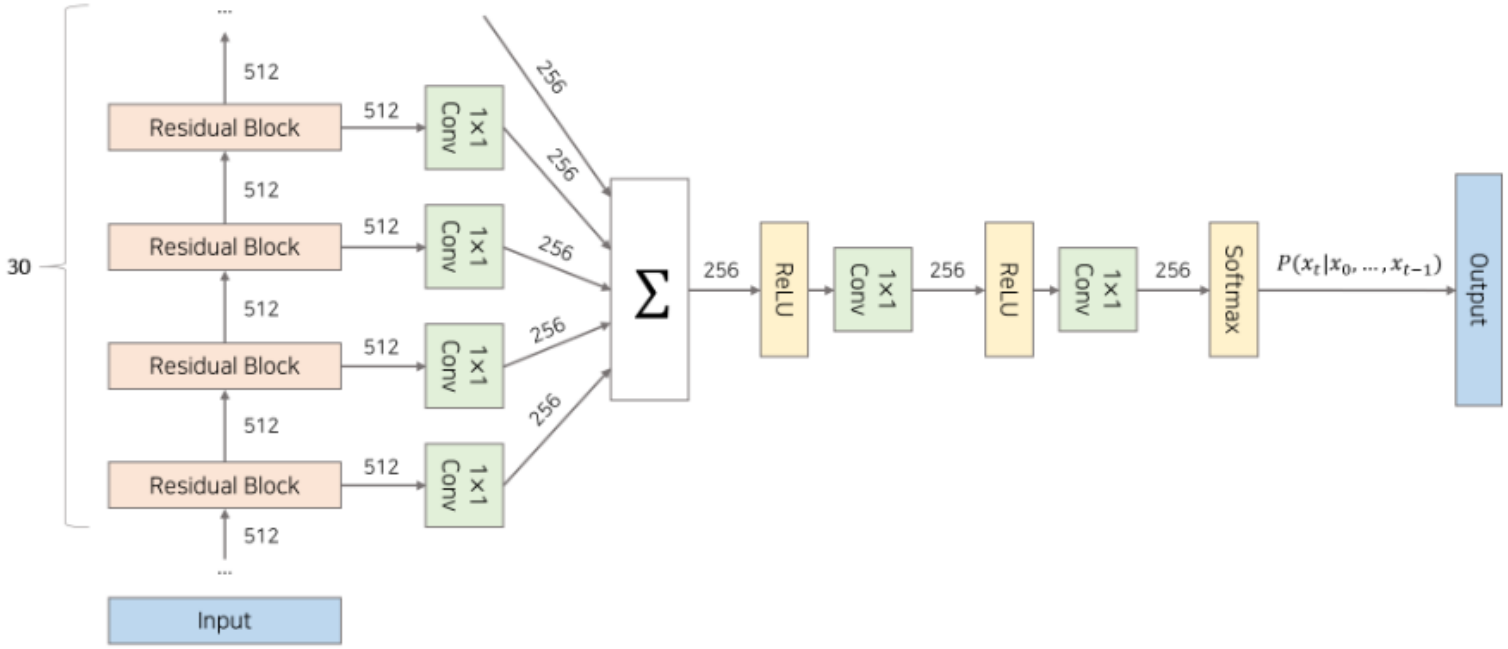



Figure 1 : WaveNet 전체구조

# 01 Introduction

## Audio Data

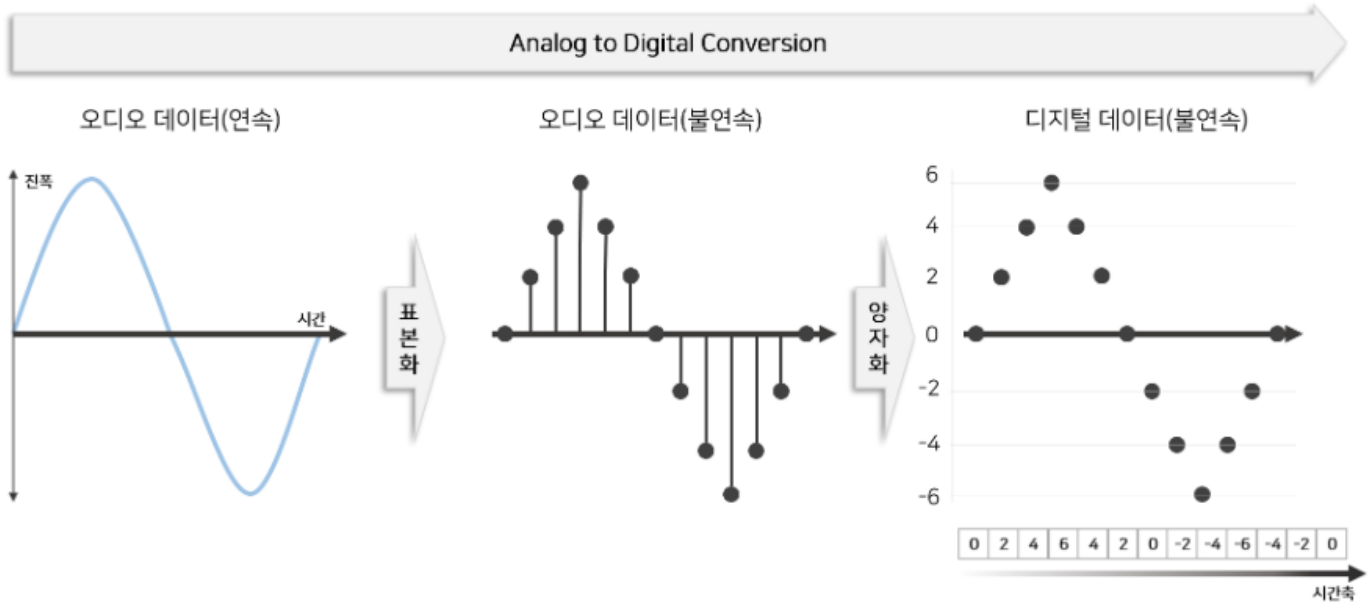



Figure 2 : 아날로그 음성 데이터로부터 디지털 데이터로 변환과정

Analog Digital Conversion

- 연속적인 음성 데이터를 정수배열로 만들어 디지털 데이터로 저장
- sampling rate: 초당 샘플링 횟수 (일반적으로 16000)

# 01 Introduction

## Audio Data

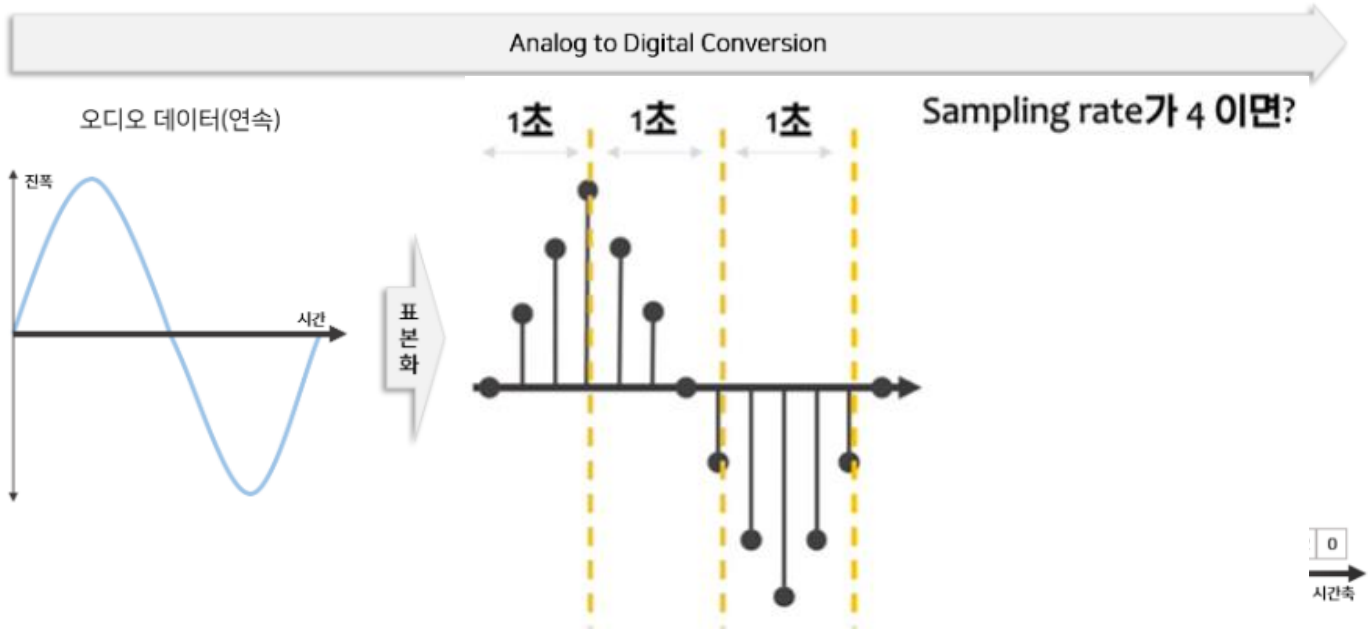



Figure 2 : 아날로그 음성 데이터로부터 디지털 데이터로 변환과정

Analog Digital Conversion

- 연속적인 음성 데이터를 정수배열로 만들어 디지털 데이터로 저장
- sampling rate: 초당 샘플링 횟수 (일반적으로 16000)

## Input Data - Softmax distribution

- 개별 오디오 샘플을 모델링하는 방법
- 음성데이터는 각 샘플을 16bit 정수 값으로 저장
- Analog Digital Conversion을 통해 생성된 정수배열의 정수는 $-2^{15}$~$2^{15}+1$ 사이의 숫자.
- 매 t 시점 특정 파형이 나올 확률 $P(x_t|x_1,\dots,x_t-1)$을 카테고리컬 분포로 가정하면,
  매 t 시점마다 $-2^{15}$~$2^{15}+1$ 사이의 숫자가 나올 확률을 계산 해야함. (총 65,536개의 확률)

**Audio: 읽기**

| 읽 | ? |
|---|---|
| $x_t-1$ | $x_t$ |

$P(x_t|x_1,\dots,x_t-1)$ => 기, 끼, 긱, ... 등

65,536개의 발음을 카테고리화한 것 중의 확률 계산

## Input Data - Softmax distribution

- 개별 오디오 샘플을 모델링하는 방법

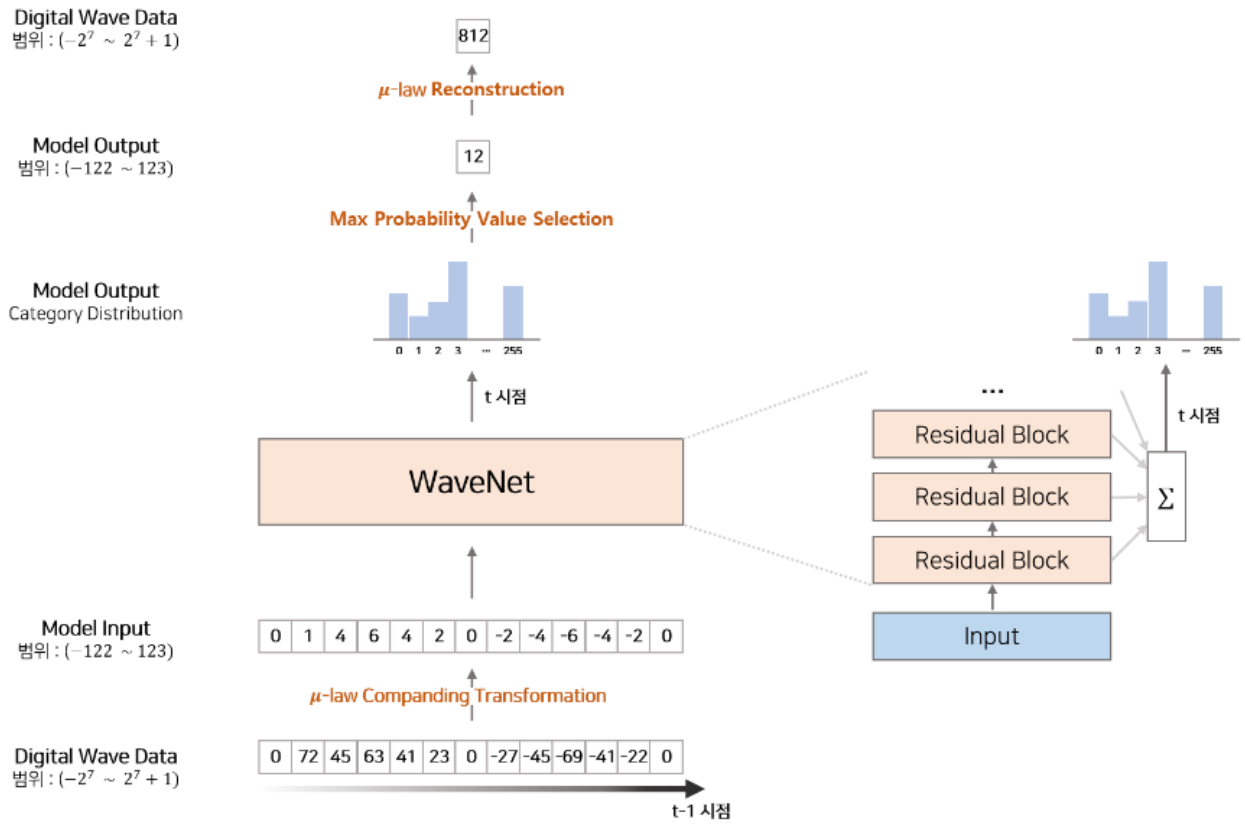



Figure 3 : 모델의 Input & Output 변환과정

**Input Data -** $\mu - law\ Companding\ Transformation$

- 65,536 개의 확률을 계산하는 것은 매우 어려움
- 따라서 이를 256개의 숫자로 변환하는 $\mu - law\ Companding\ Transformation$ 사용
- 사람의 귀가 소리가 작을 때는 작은 변화에 민감, 소리가 클 때는 변화에 민감 x

  =〉 값이 작은 것은 작게 자르고, 값이 큰 것은 크게 자르는 비선형적인 양자화 방식

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

# 02 WaveNet

## Overview

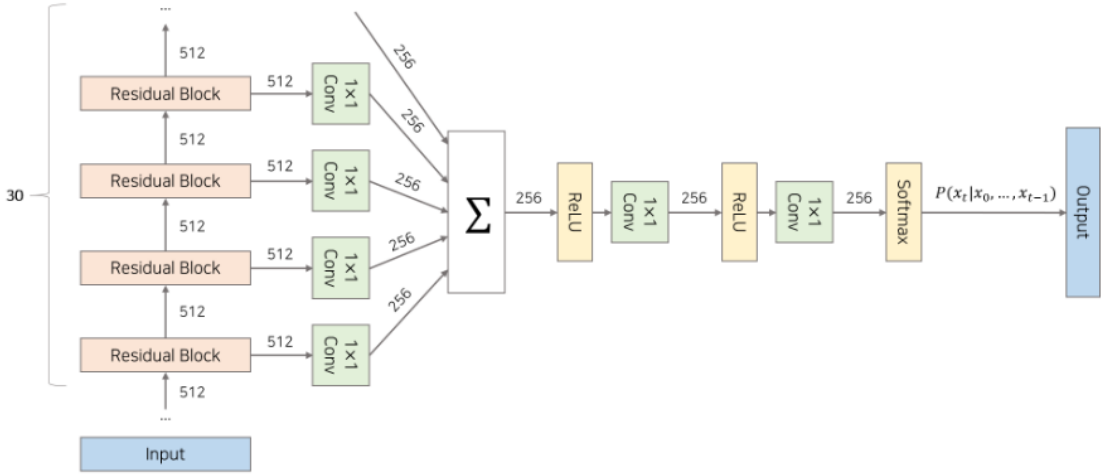



Figure 1 : WaveNet 전체구조

1. Dilated causal convolution
2. Residual connection & Gated activation units
3. Conditional Modeling

## Dilated Causal CNN

### Causal Convolution

- 시간 순서를 고려하여 Convolution Filter를 적용하는 변형 Convolution Layer
- RNN처럼 음성 데이터(시계열 데이터)를 모델링 가능
- 수용 범위(Receptive Field)를 넓히기 위해 많은 양의 Layer를 쌓아야 함.

### Dilated Convolution

- 추출 간격(Dilation)을 조절하여 더 넓은 수용 범위(Receptive Field)를 갖게 하는 변형 Convolution Layer
- Wavenet 에서는 총 30층의 Layer를 쌓아 모델을 구성함.

## Dilated Causal CNN

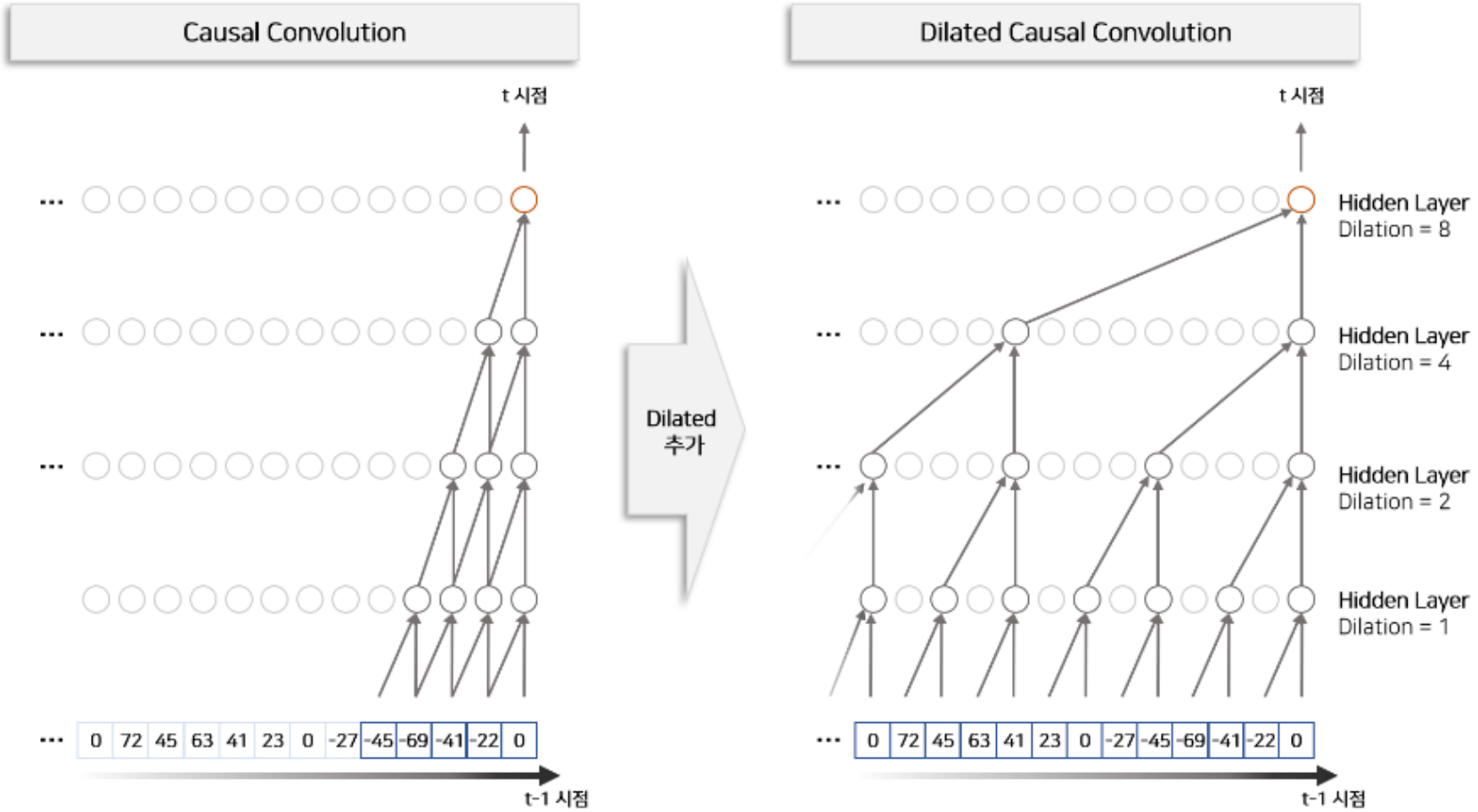



Figure 5 : Causal Convolutions VS Dilated Causal Convolutions

## Residual Connection & Gated activation units

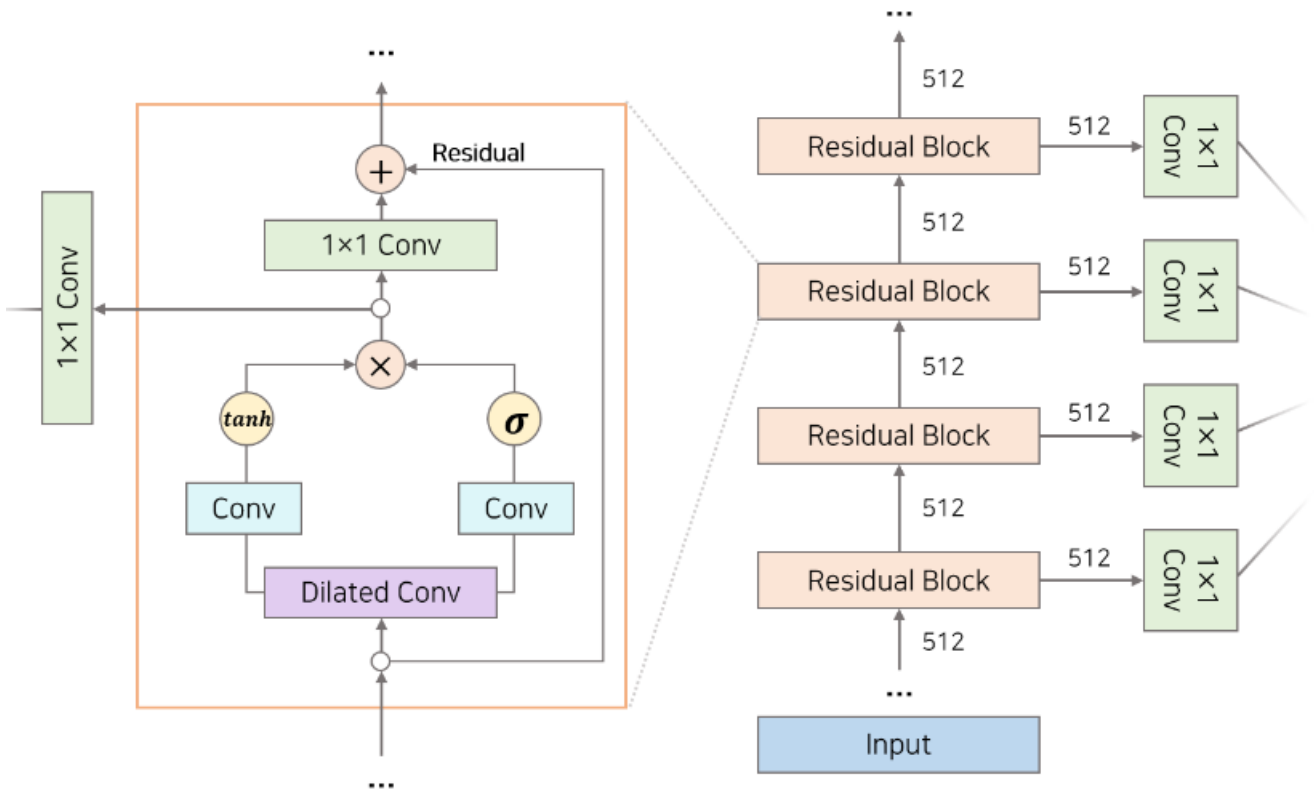
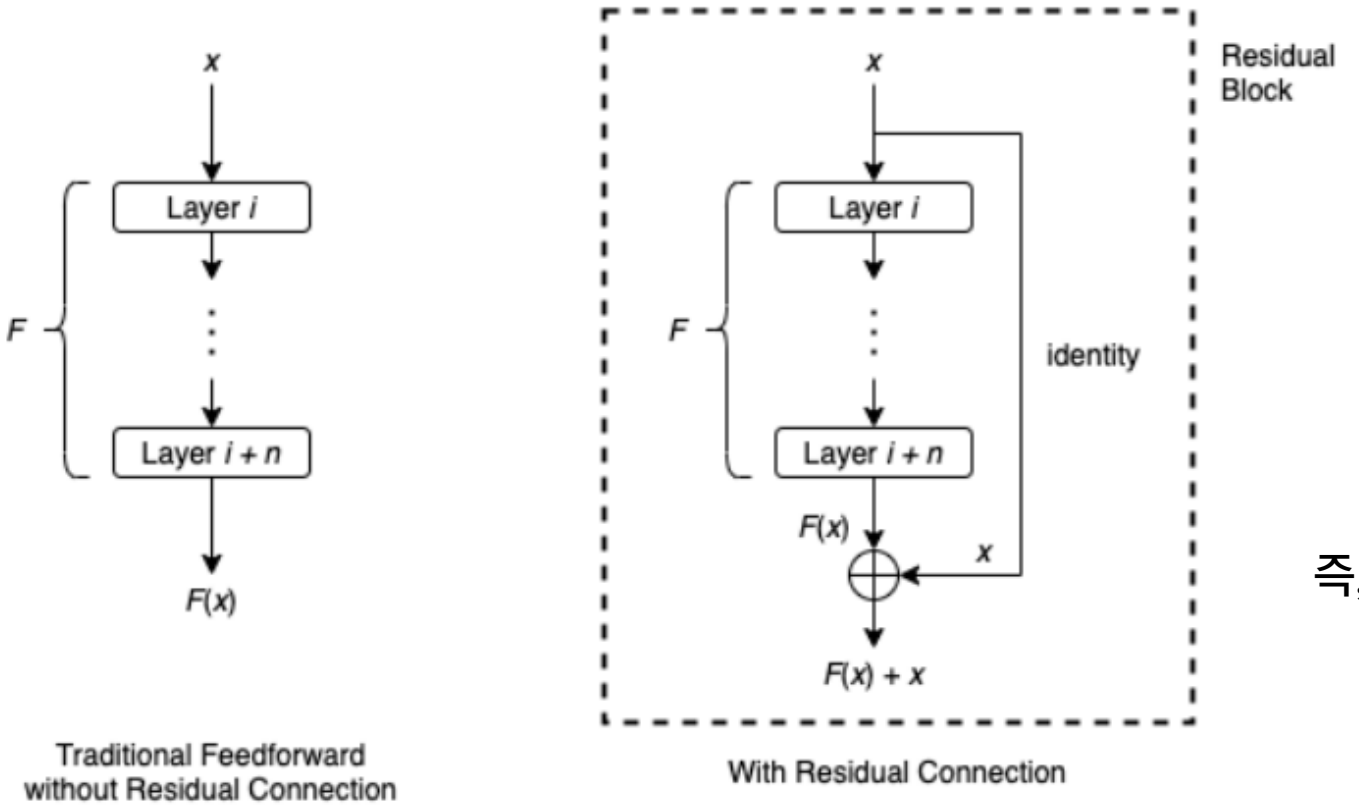



Figure 7 : Residaul Block 상세구조

## Residual Connection & Gated activation units




네트워크 출력 값이 X가 되도록

H(X)-x를 최소화하는 방향으로 학습

$$H(X)=y$$

$$F(X)=H(X)-X$$

즉, input과 output의 차이를 줄이는 방향으로 학습

## Residual Connection

출처 -https://towardsdatascience.com/what-is-residual-connection-efb07cab0d55

## Residual Connection & Gated activation units

### Gated Activation Units

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

$*$ : $Convolution$연산

$\odot$ : $Element - wise$곱셈

$\sigma()$ : $SigmoidFunction$

$W$ : 학습가능한$ConvolutionFilter$

$f$ : $filter$ $g$ : $gate$ $k$ : $layer$번호

- Gate의 기능: Local Feature를 필터(Filter)로 보고 이 필터의 정보를 다음 Layer에 얼만큼 전달해 줄지 정하는 것

## Residual Connection & Gated activation units

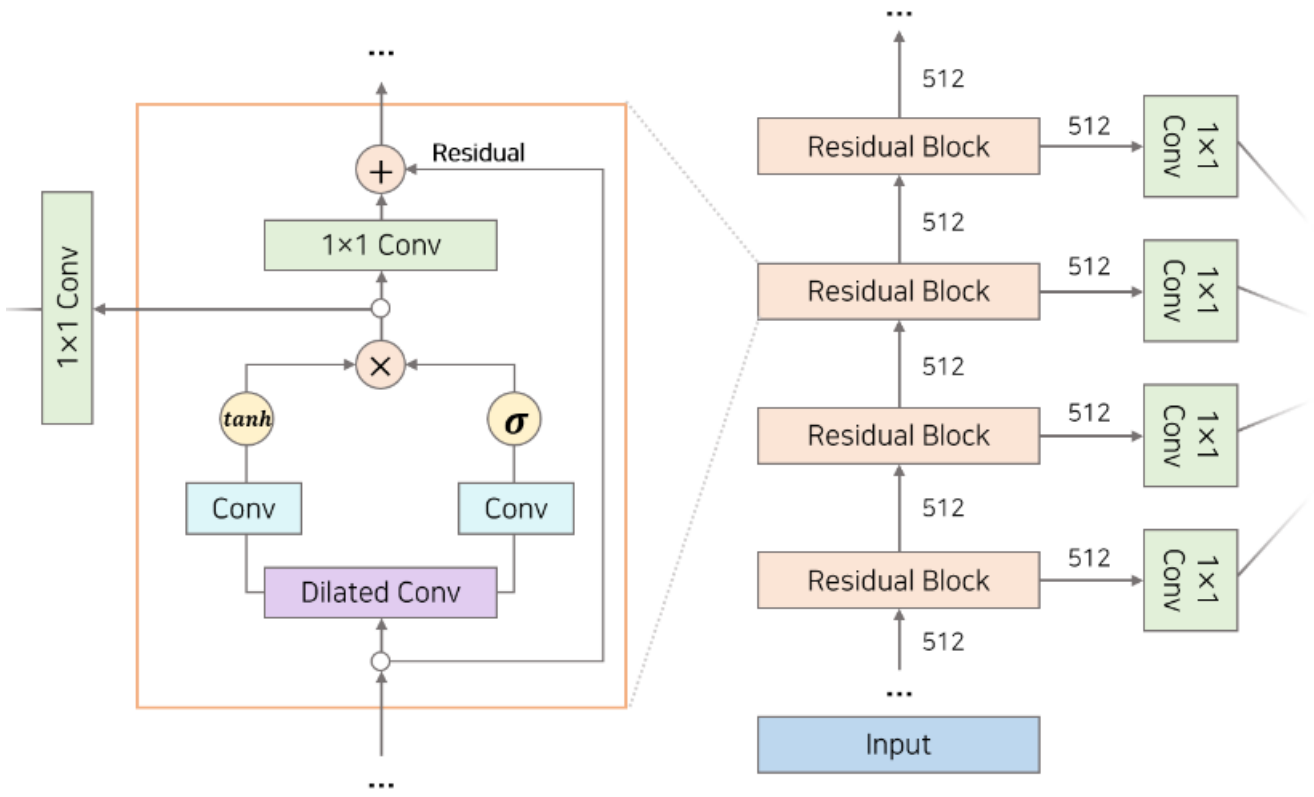



Figure 7 : Residaul Block 상세구조

# 02 WaveNet

## Conditional Modeling

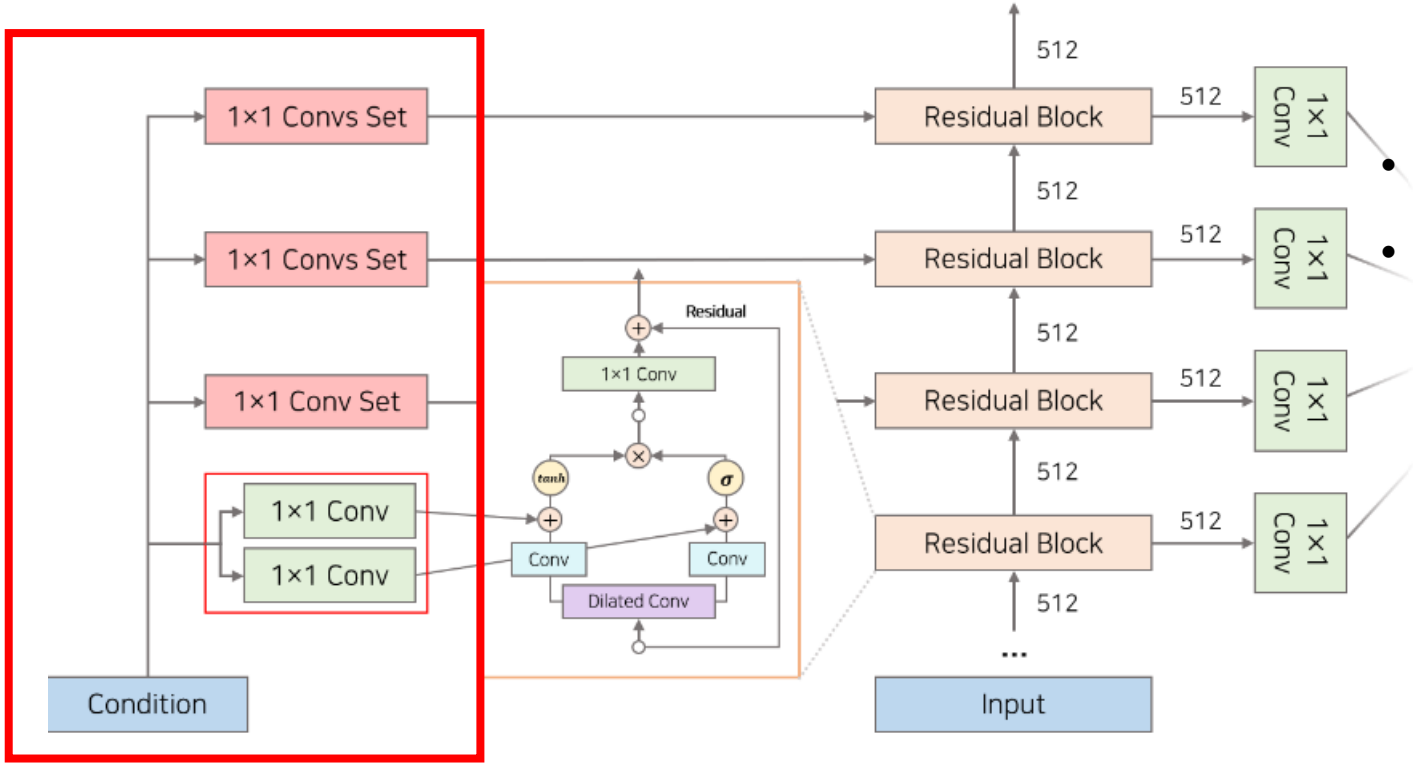



Figure 9 : 조건을 추가한 WaveNet 상세구조

- 특징(h)를 추가하여 특징에 맞는 음성 생성
- TTS(Text to Speech)의 경우, Text Embedding을 조건 정보로 추가하여 모델을 학습 -> Text 에 맞는 음성 생성

## Conditional Modeling

### Global Conditioning

- 시점에 따라 변하지 않는 조건 정보
- 화자의 고유 특성 정보(화자 ID)등
- 모든 시점에서 동일하게 조건 정보 추가

$h$: 조건에 해당하는 벡터
$V_{f,k}^T$, $V_{g,k}^T$ : 각 선형 함수

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

### Local Conditioning

- 시점에 따라 변하는 조건 정보
- TTS의 경우 Linguistic Feature 또는 Text Embedding 과 같은 정보 =〉 순서가 있는 일정 길이의 Sequence 벡터.
- Local 조건 정보를 음성 정보와 매칭시켜 시점에 따라 다르게 넣어주어야 함.

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

# 03 Experiment

## 실험 1. Multi-Speaker Speech Generation

- VCTK Dataset(English Multi-speaker corpus) 를 사용하여 다양한 화자 ID를 조건으로 추가하여 다양한 음성 생성

DeepMind - Wavenet
Speech Generation, Making Music
https://www.deepmind.com/blog/wavenet-a-generative-model-for-raw-audio

## 실험 2. Text to Speech

- TTS를 위해 음소, 음소길이, 기본주파수 등의 조건정보 추가
- 학습 할 때는 위 조건정보를 추가하여 학습, 생성할 때는 조건정보만을 이용하여 음성 생성

| 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 | $3.86 \pm 0.137$ | $3.47 \pm 0.108$ |
| **WaveNet** (L+F) | $\mathbf{4.21} \pm 0.081$ | $\mathbf{4.08} \pm 0.085$ |
| Natural (8-bit $\mu$-law) | $4.46 \pm 0.067$ | $4.25 \pm 0.082$ |
| Natural (16-bit linear PCM) | $4.55 \pm 0.075$ | $4.21 \pm 0.071$ |

Table 1: Subjective 5-scale mean opinion scores of speech samples from LSTM-RNN-based statistical parametric, HMM-driven unit selection concatenative, and proposed WaveNet-based speech synthesizers, 8-bit $\mu$-law encoded natural speech, and 16-bit linear pulse-code modulation (PCM) natural speech. WaveNet improved the previous state of the art significantly, reducing the gap between natural speech and best previous model by more than 50%.

L: Linguistic feature
F: 화자 특징
(로그 기본 주파수)

1. **Paired Comparison Test:** 두 개의 실험모델로부터 생성된 음성 중 더 좋은 음성 선택
2. **Mean Opinion Score(MOS) Test:** 생성된 음성에 1~5점의 품질 점수 부여, 평균을 구함.

# 04 Conclusion

## Contribution

- 인간 평가자가 평가했을 때 자연스러운 TTS(Text to Speech) sample 생성
- dilated causal convolution 기반 새로운 아키텍처 제안
- conditional feature 를 사용하여 조건에 따라 다른 audio 생성
- 사람의 말소리 뿐만 아니라 음악 등 다양한 오디오 양식 생성에 사용 가능

# 참고자료


- [논문리뷰] - WaveNet: A Generative Model for Raw Audio, Deep Mind (https://joungheekim.github.io/2020/09/17/paper-review/)
- [Paper Review] WaveNet: A Generative Model for Raw Audio- DSBA (https://www.youtube.com/watch?v=MNZepE1m-kI)
- [DL] Exploding & Vanishing Gradient 문제와 Residual Connection (https://heeya-stupidbutstudying.tistory.com/entry/DL-Exploding-Vanishing-gradient-%EB%AC%B8%EC%A0%9C%EC%99%80-Residual-Connection%EC%9E%94%EC%B0%A8%EC%97%B0%EA%B2%B0)
- (논문리뷰) ResNet 설명 및 정리 (https://ganghee-lee.tistory.com/41)

# Q&A