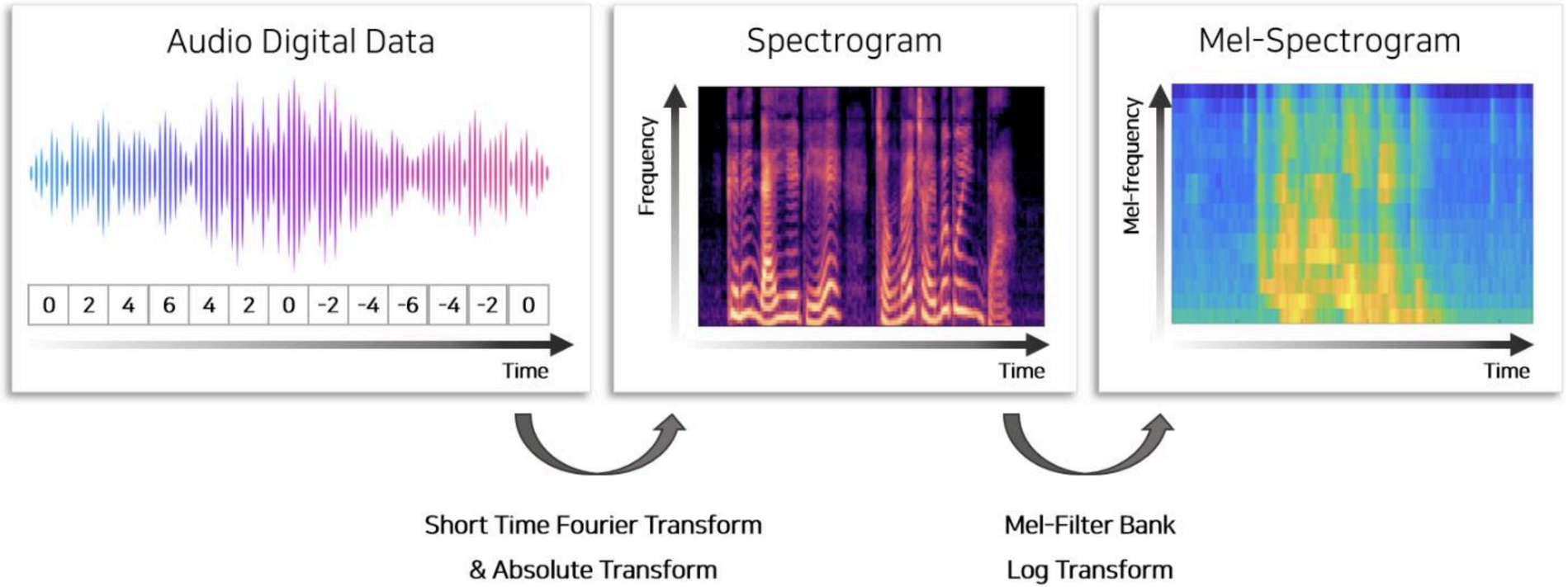
WaveNet : A Generative Model for Raw Audio

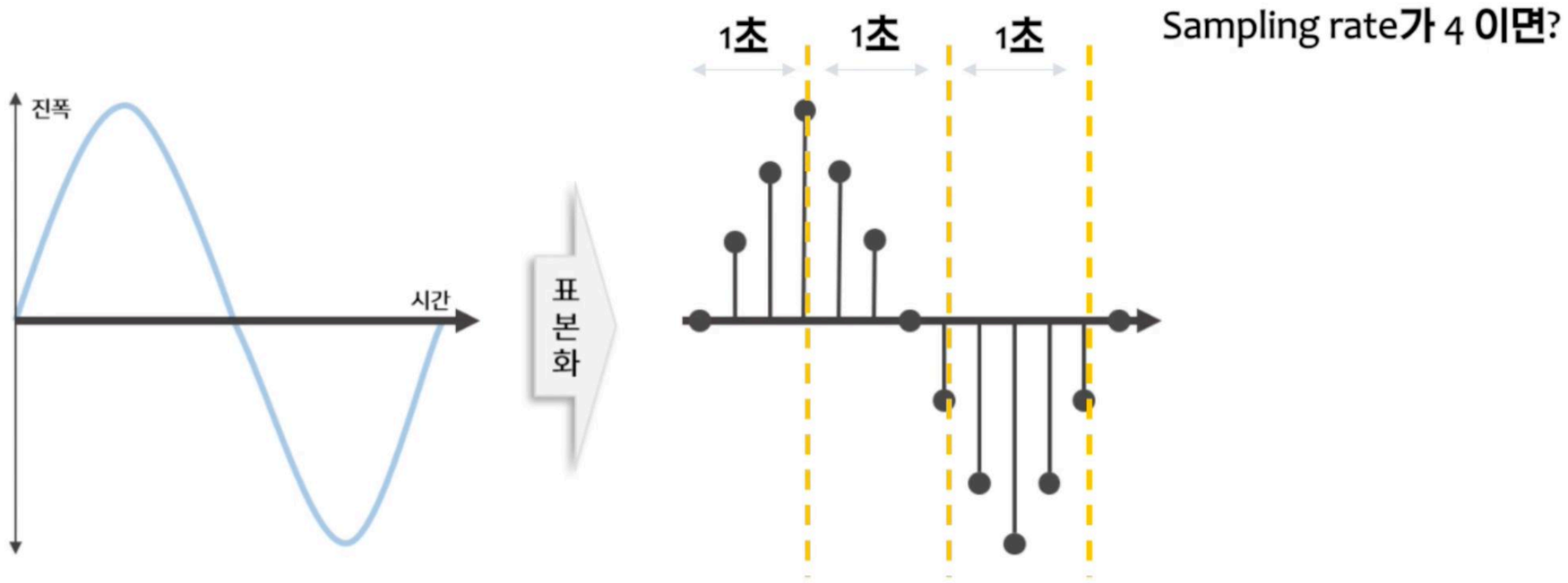
Deepshark Lab 소프트웨어융합학과 이민욱

WaveNet of Raw Waveform



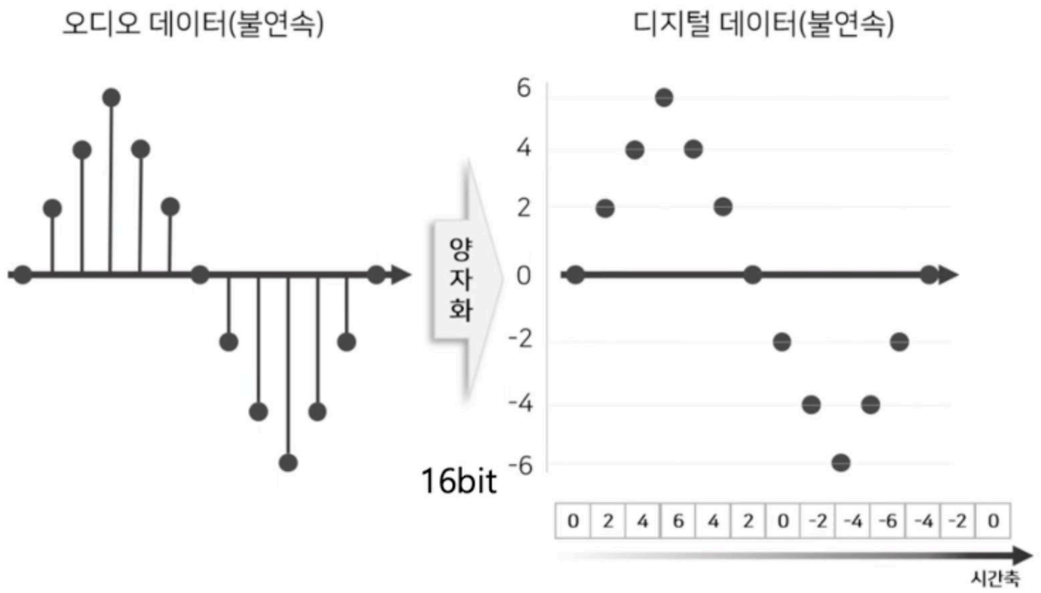
Amplitude sequence

- 아날로그 음성 데이터는 연속형(continuous) 데이터
- 연속형 데이터를 샘플링된 형태로 표현
  - 주로 초당 샘플링 횟수는 16KHz



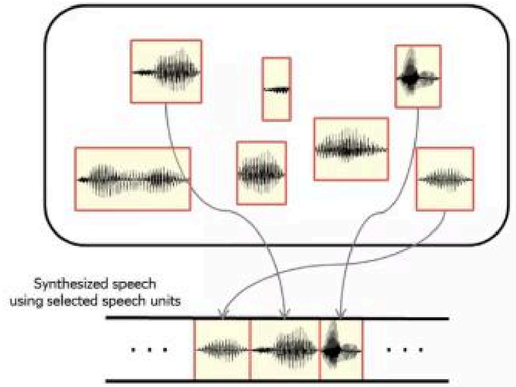
Amplitude sequence

- 진폭의 크기는 일반적으로 16bit
  - $(-2^{15}, 2^{15}-1)$



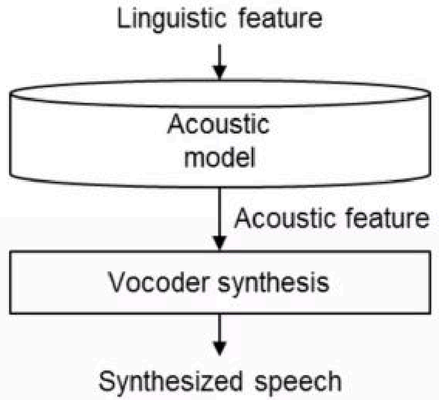
Classical TTS Model

Unit-selection speech synthesis [1]



- Concatenate tiny segment of real speech signal
- Pros) High quality
- Cons) Low controllability, large database

Statistical parametric speech synthesis (SPSS) [2]



- Parameterize speech signal using vocoder technique
- Pros) High controllability, small database
- Cons) Low quality

Generative model-based speech synthesis

WaveNet

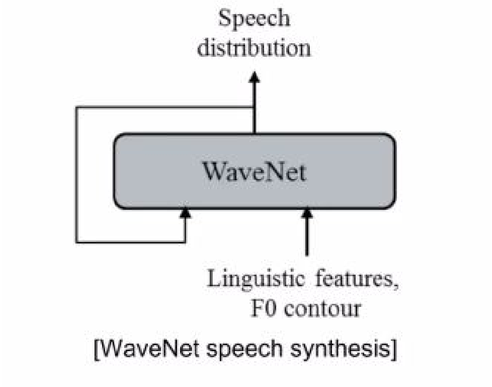
- Raw waveform을 사용한 최초의 generative model

$$p(\mathbf{x}) = \prod_{n=1}^{N} p(x_n \mid \mathbf{x}_{<n})$$

  - waveform sample의 확률 분포를 auto-regressive하기 예측
- 고품질의 오디오 / 음성 신호를 생성
  - TTS, voice conversion, music synthesis

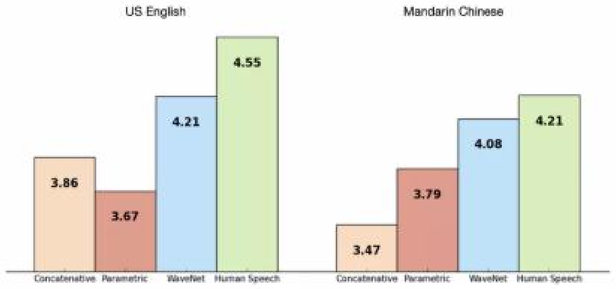
WaveNet in TTS task

- Utilize Linguistic feature and F0 contour as a conditional information



[WaveNet speech synthesis]

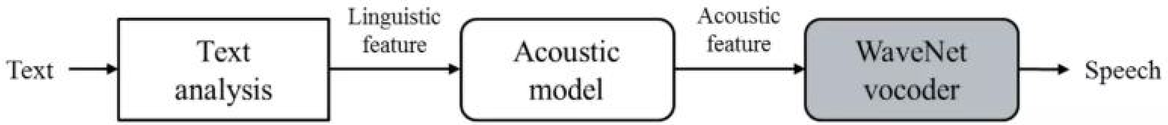
- Present higher quality than the conventional TTS Systems



[Mean opinion scores]

WaveNet Vocoder - Based Speech Synthesis

Utilize WaveNet as parametric vocoder



[WaveNet vocoder based parametric speech synthesis]

- **Acoustic feature을 conditional informationd으로 사용.**

Advantages

- **Higher quality Synthesis Speech than Conventional vocoder**
  - Don’t require hand-engineered processing pipeline
- **Higher controllability than the case of linguistic features**
  - Controlling acoustic features
- **Higher training efficiency than the case of linguistic features**
  - Linguistic feature: 25~35 hour database
  - Acousituc feature: 1 hour database

Basic of WaveNet

Auto- regressive generative model

$$p(\mathbf{x}) = \prod_{n=1}^{N} p(x_n \mid \mathbf{x}_{n-R:n-1})$$

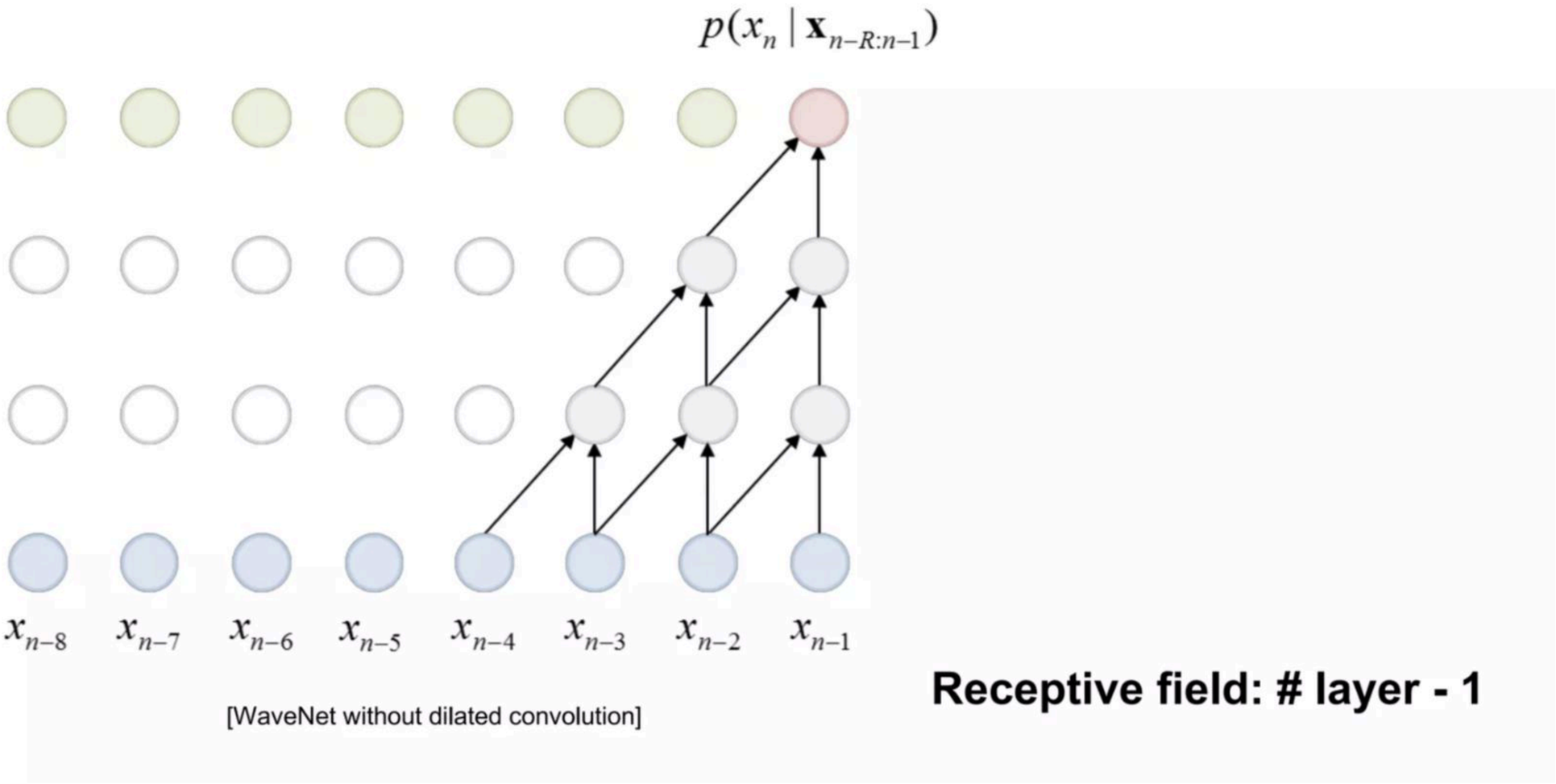
- WaveNet은 각 오디오 샘플을 이전 샘플들을 조건으로 하여 순차적으로 예측합니다.

Problem: Long - term depency nature of speech signal

- **시퀀스에서 멀리 떨어진 요소들 사이의 의존 관계를 학습하기 어려운 문제(vanila RNN)**
  - 화자의 일관성, 억양 패턴, 발음 스타일
- **LSTM**
  - **Highly correlated speech signal in high sampling rate, e.g. 16000 Hz**
  - **E.g. 1) 100Hz 음성을 표현하기 위해서는 적어도 160(16000/100) 음성 샘플**
  - **E.g. 2)평균적으로 한 문자를 표현하기 위해서는 6000개의 음성 샘플**
- **dilated causal CNN**
  - **다중 시간 스케일 처리에 효율적**
  - **지수적 receptive field 확장**

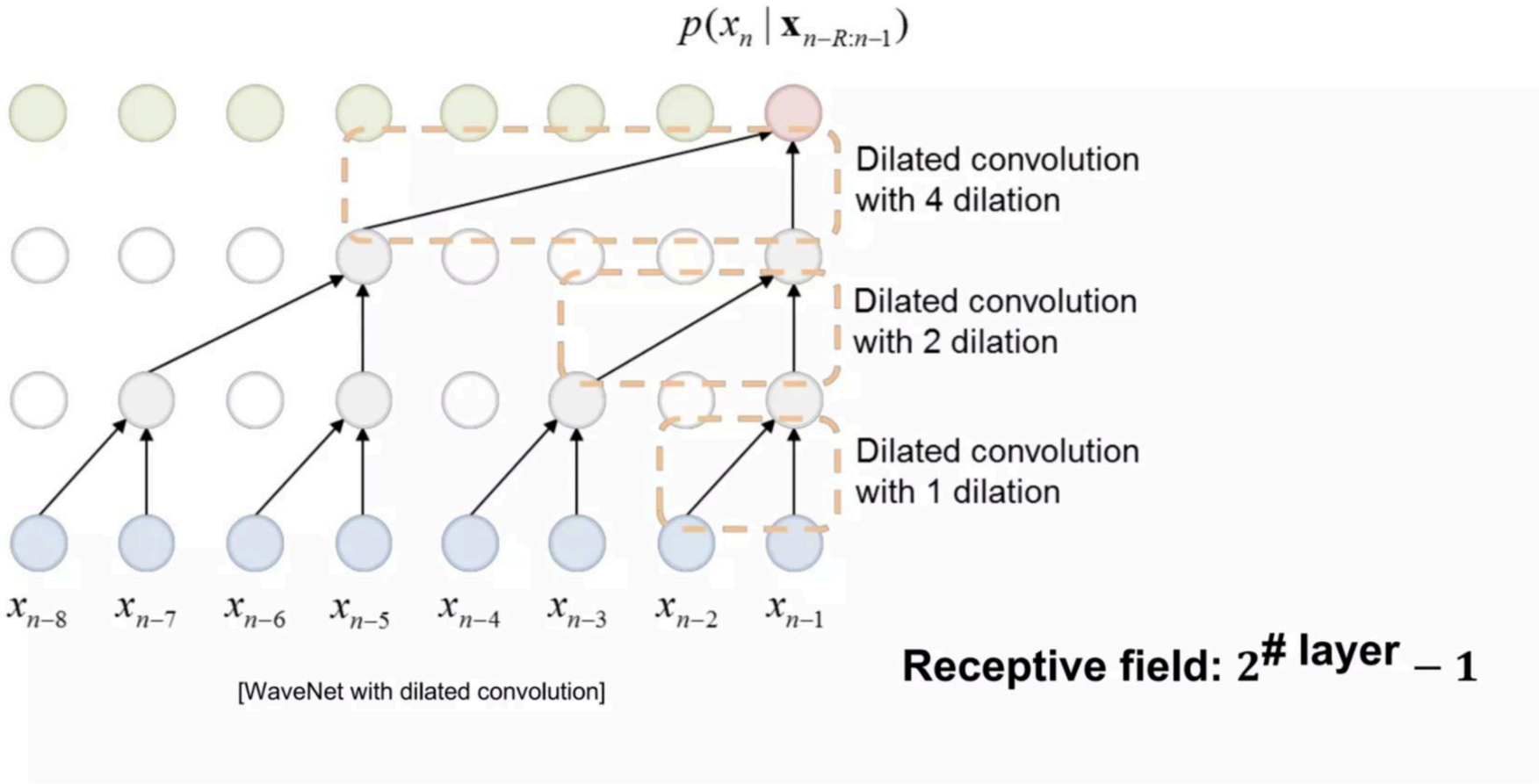
Basic of WaveNet

WaveNet without dilation



Basic of WaveNet

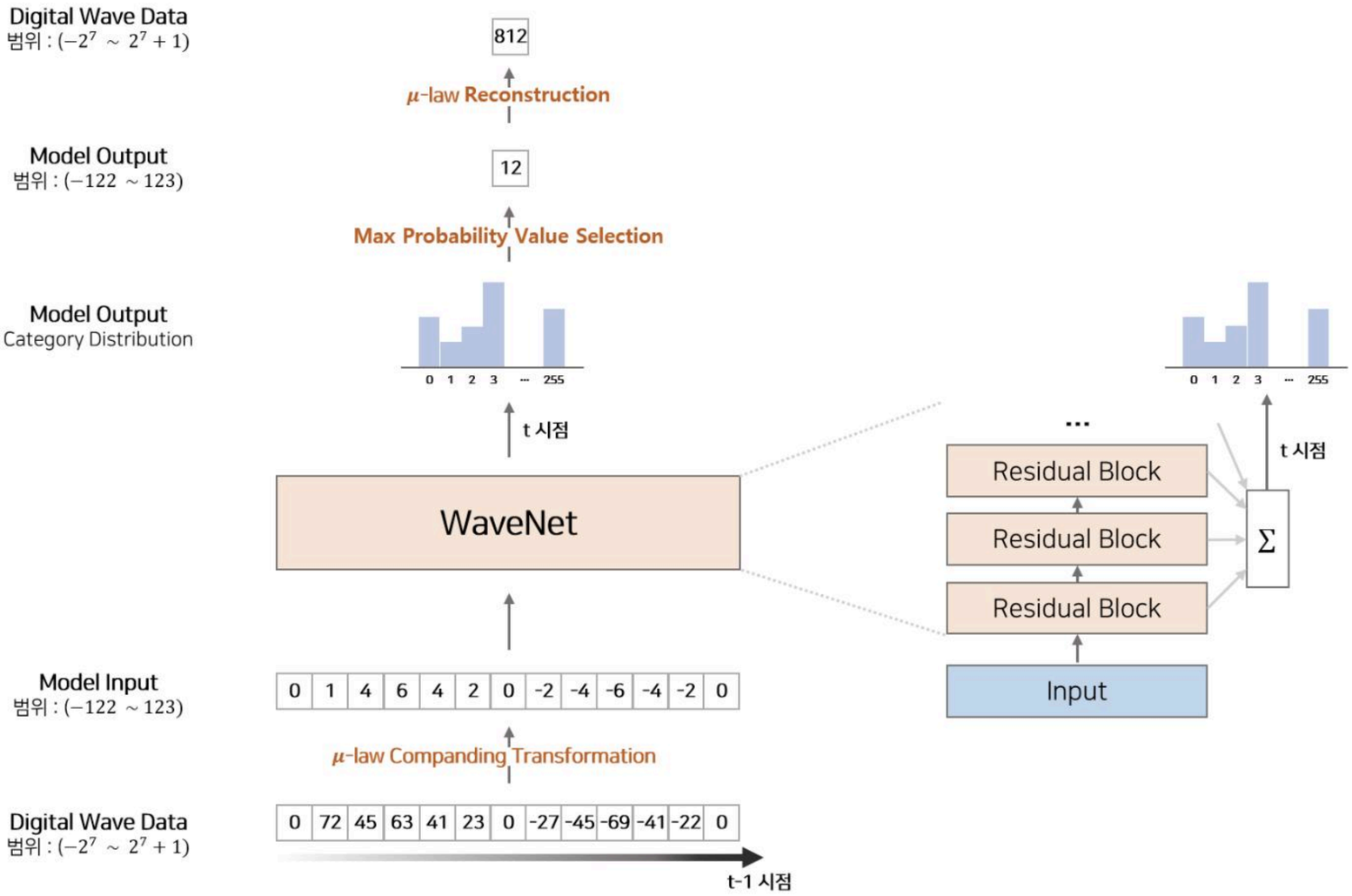
WaveNet with dilation



[WaveNet with dilated convolution]

Receptive field: $2^{\#\text{ layer}} - 1$

SoftMax Distribution

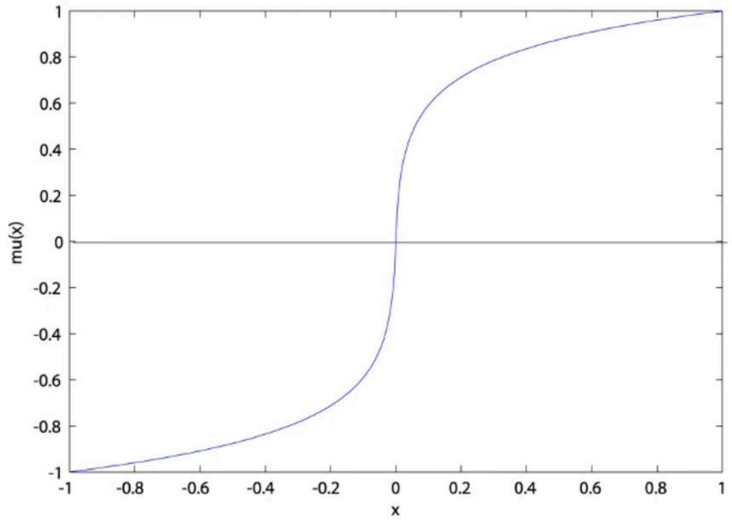


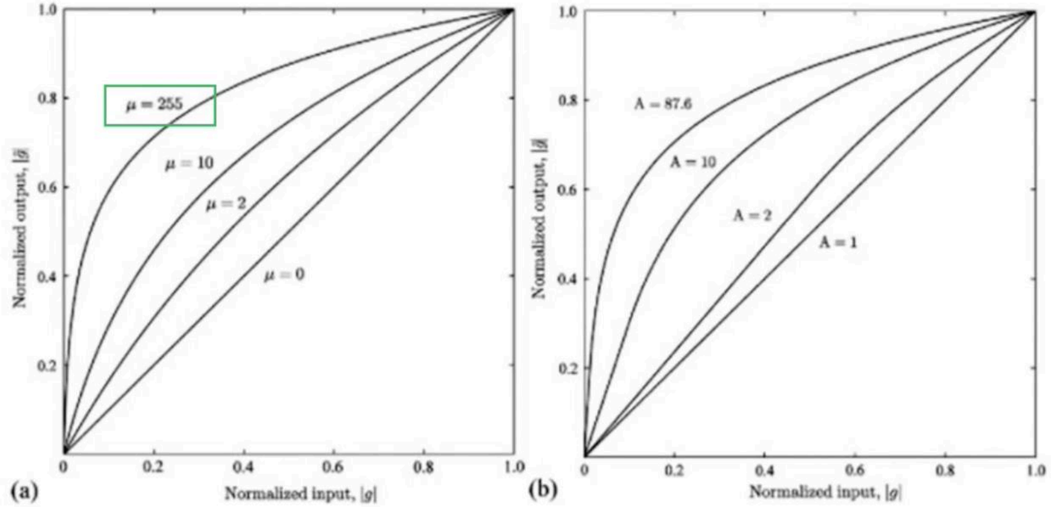
μ-law companding Transformation

- **bandwidth of speech signal**
  - **(-2^{15}, 2^{15} -1) = (-32768, 32767) 만큼의 discrete variable을 예측하기는 어렵다.**
    - **더 작은 법위의 정수배열로 표현 16bit → 8bit의 범위로 비선형함수를 이용하여 축소**
    - **(32767/256) = 127 with a remainder of 255**

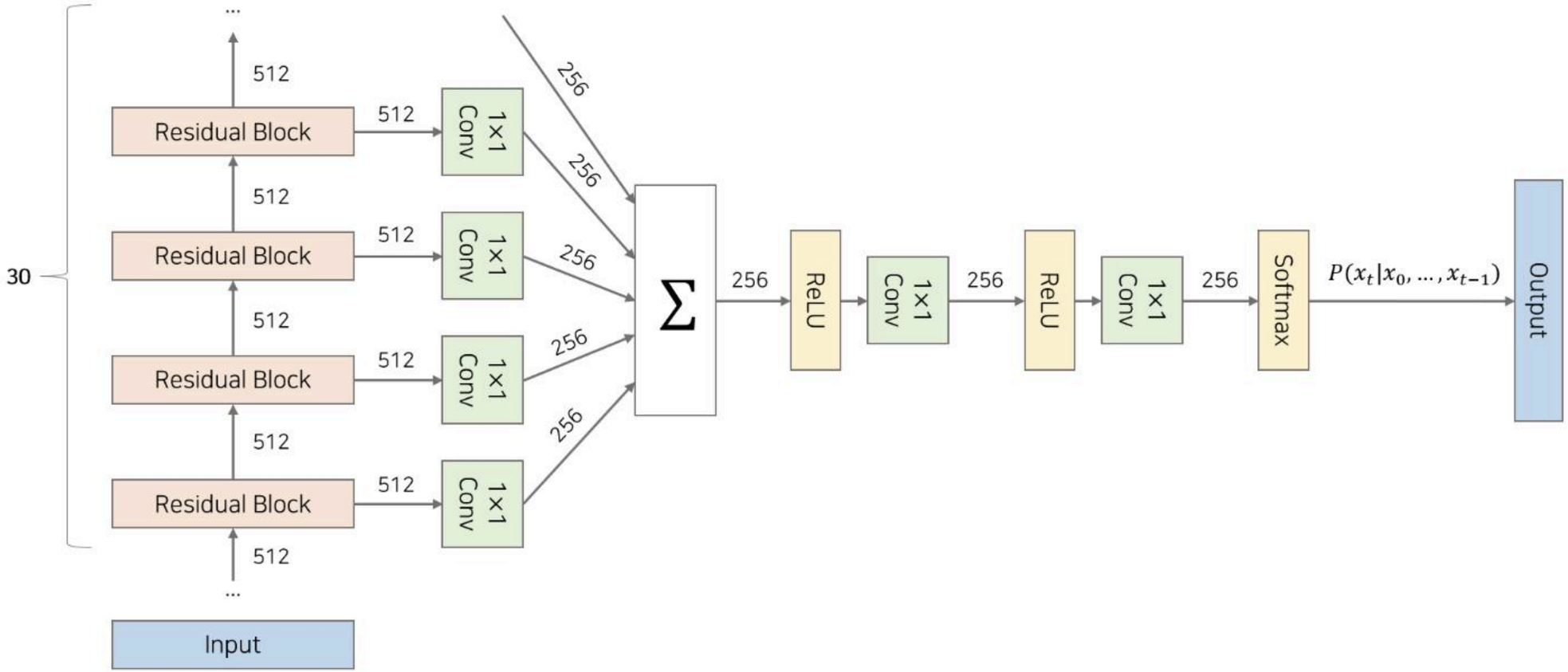
$$f(x_t) = sign(x_t)\frac{\ln(1+\mu|x_t|)}{\ln(1+\mu)}$$

If 16 -> 8bit
Then $\mu = 255$





Architecture of WaveNet

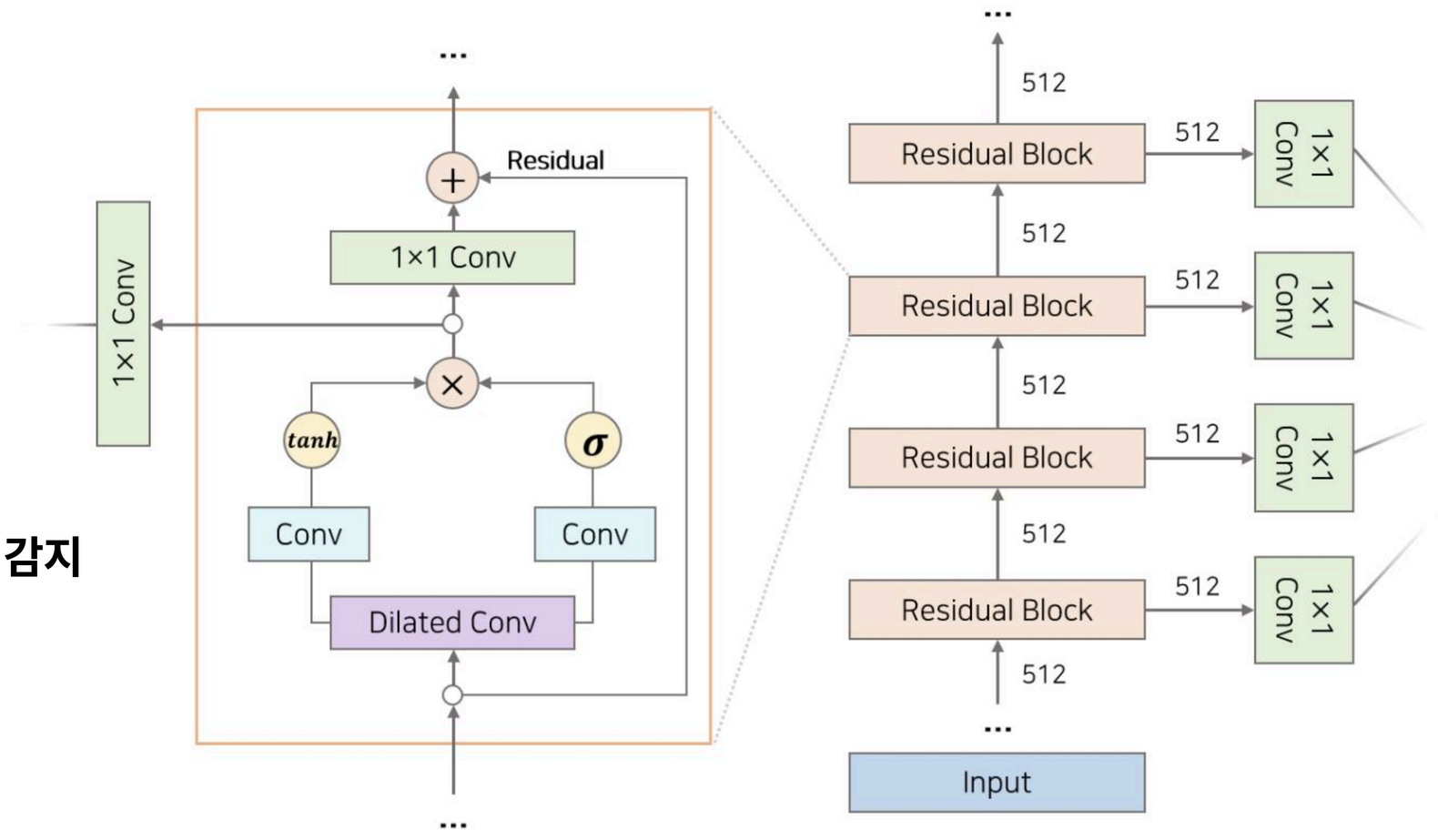


Residual Block, Gated Activation Units

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

$*$: $Convolution$연산
$\odot$: $Element-wise$곱셈
$\sigma()$: $SigmoidFunction$
W : 학습가능한$ConvolutionFilter$
f : $filter$ g : $gate$ k : $layer$번호

Filter(tanh) : positive한 feature 감지
Gate(sigmoid) : 일종의 가중치



Skip Connection

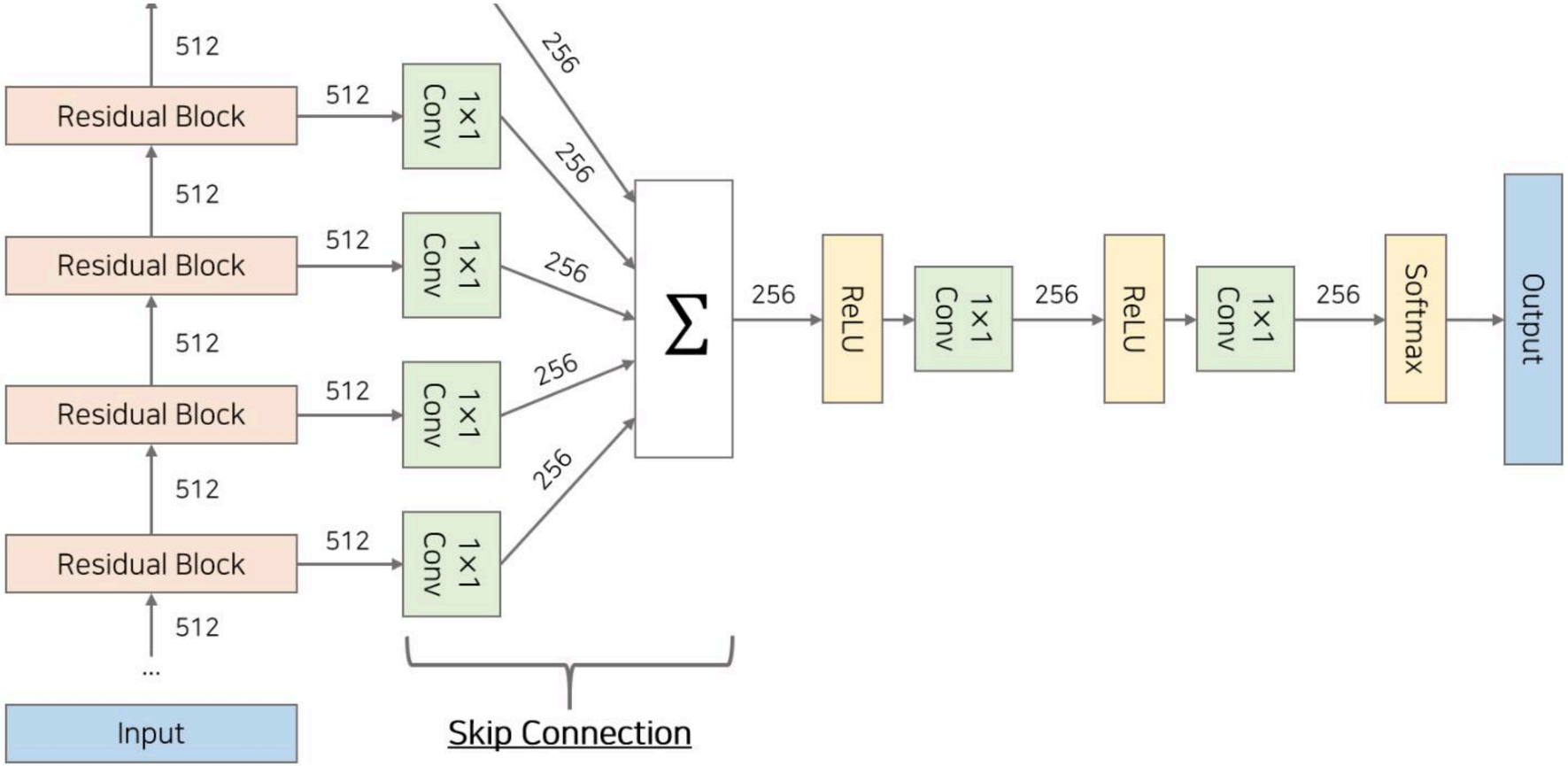
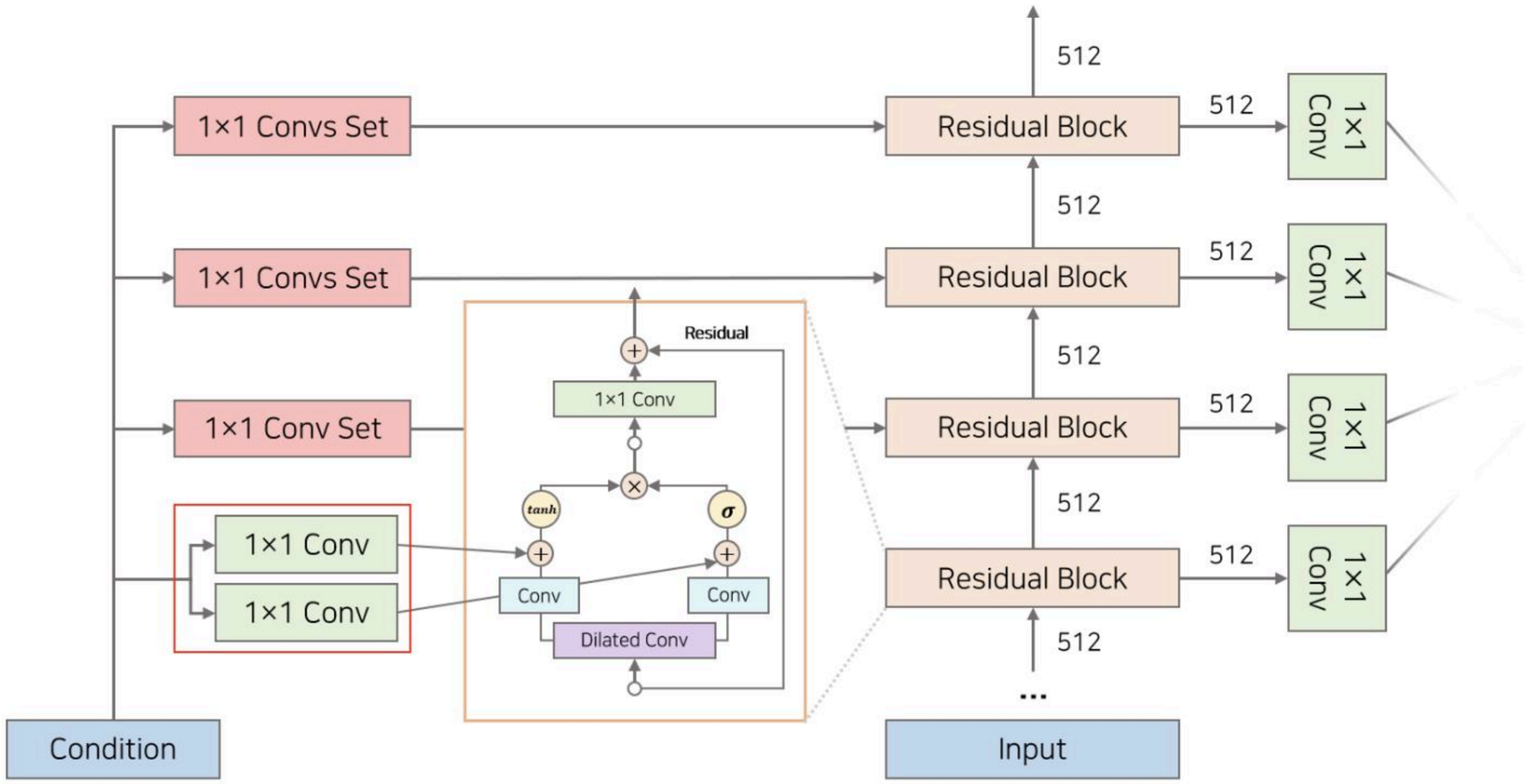


Figure 8 : Skip Connection 상세구조

Conditional WaveNet



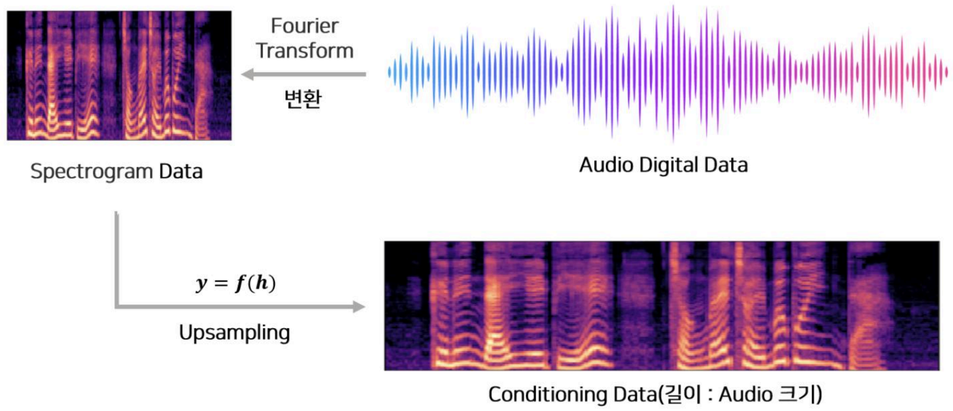
Conditional WaveNet

- **Global Condition**
  - **화자 정보에 대한 임베딩**
  - **화자의 ID를 One-Hot vector로 변환**

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

- **Local Condition**
  - **음성과 조건정보(e.g. Spectogram)**
  - **Transposed Convolution 사용(learnable parmeters)**

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



summary

- Uni - selection speech synthesis , Statistical speech synthesis
  - mel spectrogram, deterministic vocode
  - Linguistic feature
  - Low quality

- WaveNet(raw waveform)
  - WaveNet
  - Conditional WaveNet (acoustic feature)
  - Highly quality,

- Staitistical parmetric speech synthesis (WaveNet vocoder)