WaveNet

A Generative Model for Raw Audio

2018-10-05

곽민구

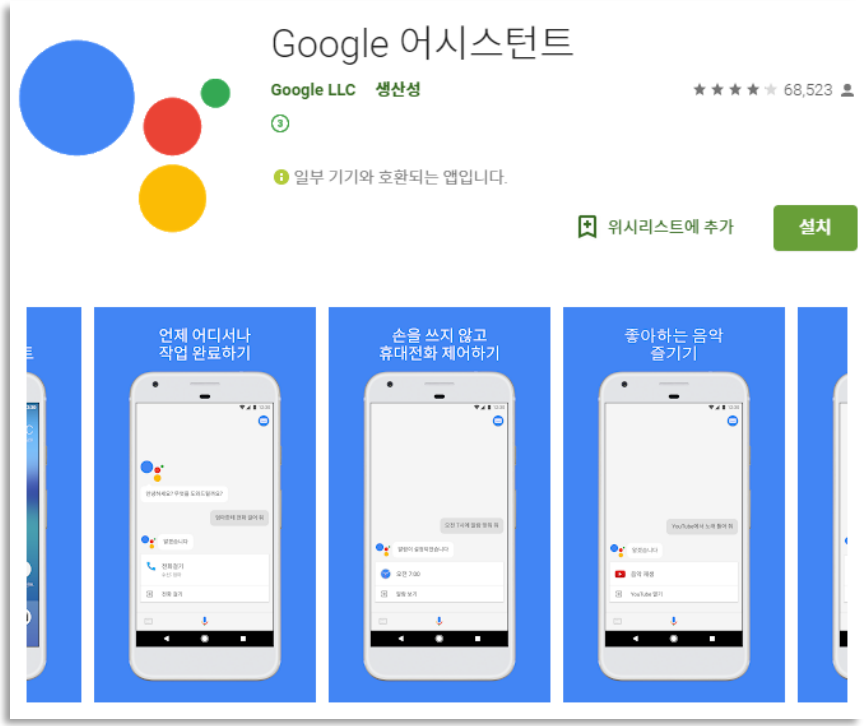
Contents

Contents

01 Google Assistant

- 2016년에 발표한 인공지능 비서 서비스
- 2018년 5월 구글 I/O 개발자 컨퍼런스: 추가적인 명령어 없이도 사람과 비슷한 억양, 톤으로 미용실 예약을 수행하기도 함

(https://www.youtube.com/watch?v=wMWdo8cVZ_U&feature=youtu.be)



- ❖ 현지 정보 확인: 기상, 음식, 길안내
- ❖ 커뮤니케이션: 통화, SMS, 이메일
- ❖ Google에 문의: 계산, 사전, 번역, 이미지 검색
- ❖ 스마트 홈 제어: 조명, 온도, 휴대전화 제어

구글 어시스턴트의 양방향 대화 기능 중,
Text-to-Speech에서 음성을 생성하는
알고리즘이 WaveNet

01 양방향 대화

- 양방향 대화가 가능하기 위해서는 기본적으로 2가지 기능이 필요

① Automatic Speech Recognition (ASR): 음성 인식. 음성 언어를 컴퓨터가 해석해 문자 데이터로 전환하는 처리.

* Speech-to-Text (STT)라고 부르기도 한다

→ "I want to make an appointment on Tuesday morning."

② Text-to-Speech Synthesis (TTS): 음성 합성. 말소리의 음파를 기계가 자동으로 만들어내는 기술.

"I want to make an appointment on Tuesday morning." →

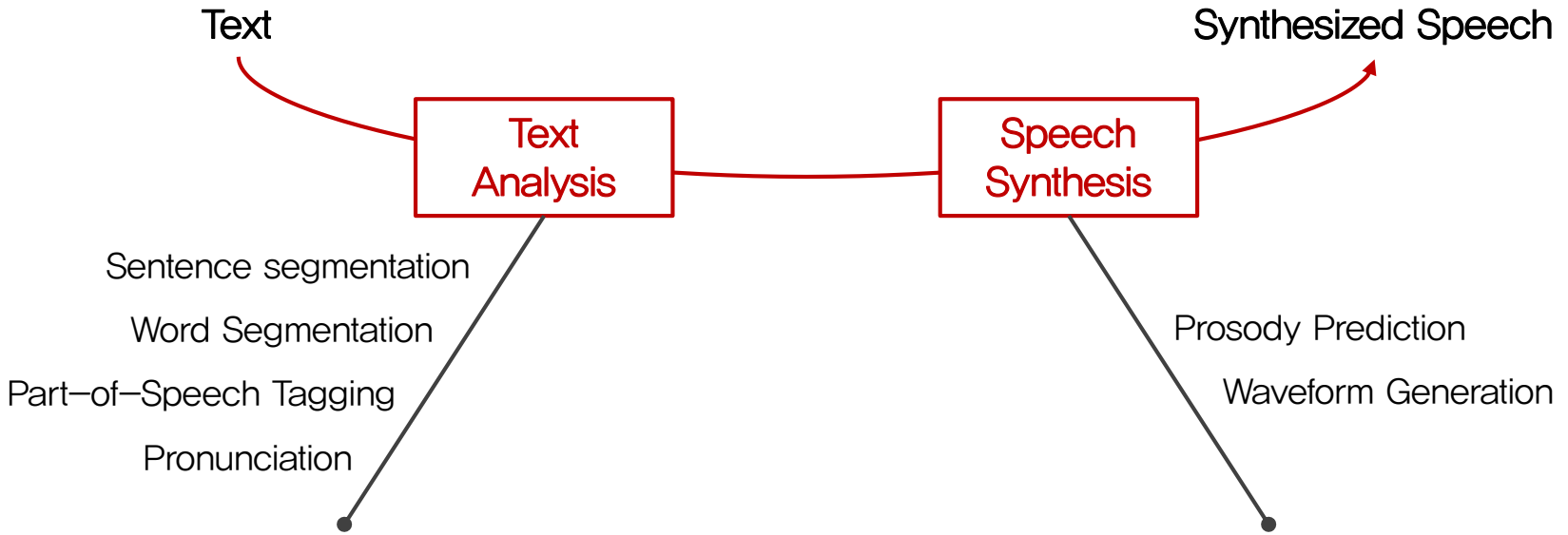
01 | Text-to-Speech

- Text-to-Speech는 2단계의 작업으로 구분이 가능

① Text Analysis

✓ ② Speech Synthesis

"I want to make an appointment on Tuesday morning." →



01 | WaveNet: Generative Synthesis

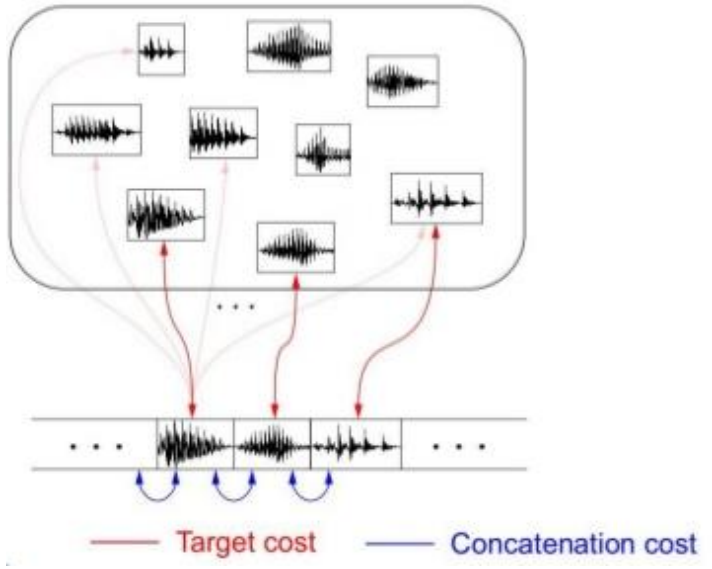
- Speech synthesis에 대한 3가지 접근 방법

① Rule-based, formant synthesis

〉〉 음성의 주파수 특징을 분석하여, 역으로 음성을 생성해내는 기법

② Sample-based, concatenative synthesis

〉〉 음성 데이터 집합 (pool)으로부터 샘플들을 추출하여, 이어 붙이는 기법



③ Model-based, generative synthesis

〉〉 음성 데이터의 분포를 학습하여, 분포로부터 샘플을 추출하는 기법

$p(\text{speech} =$  $|\ \text{text} =$ "I want to make an appointment on Tuesday morning.")

Contents

02 | WaveNet의 주요 포인트

1. Autoregressive Model
2. Dilated causal convolutions
3. Output – categorical/softmax distribution
4. Gated activation units
5. Residual and skip connections

WaveNet

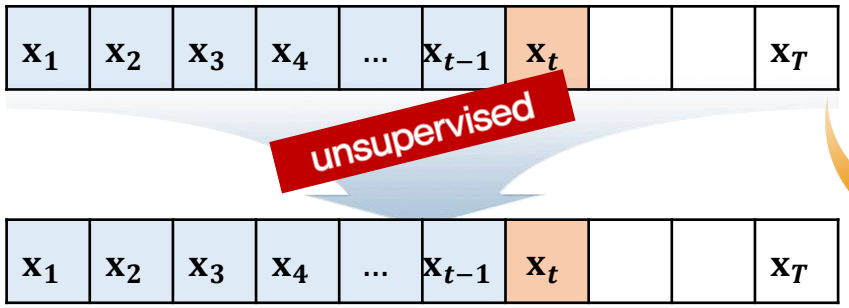
Input: audio → Output: audio

6. Conditional WaveNet → Input: audio & text information → Output: audio

(Text-to-Speech)

02 | 1. Autoregressive Model

- Audio 스스로의 분포를 학습하는 unsupervised learning
- Audio-sequence data의 statistical dependency를 반영하여 모델링을 해야 함
- T 시점까지의 audio waveform: $x = \{x_1, \dots, x_T\}$
- $p(x_t)$는 x_1부터 x_{t-1}까지의 데이터가 주어졌을 때의 conditional probability로 표현할 수 있음: autoregressive
  → $p(x_t) = p(x_t|x_1, x_2, \dots, x_{t-1})$
- Joint probability is factorized as a product of conditional probabilities



- $p(x_1)$
- $p(x_1, x_2) = p(x_1)p(x_2|x_1)$
- $p(x_1, x_2, x_3) = p(x_1, x_2)p(x_3|x_1, x_2)$

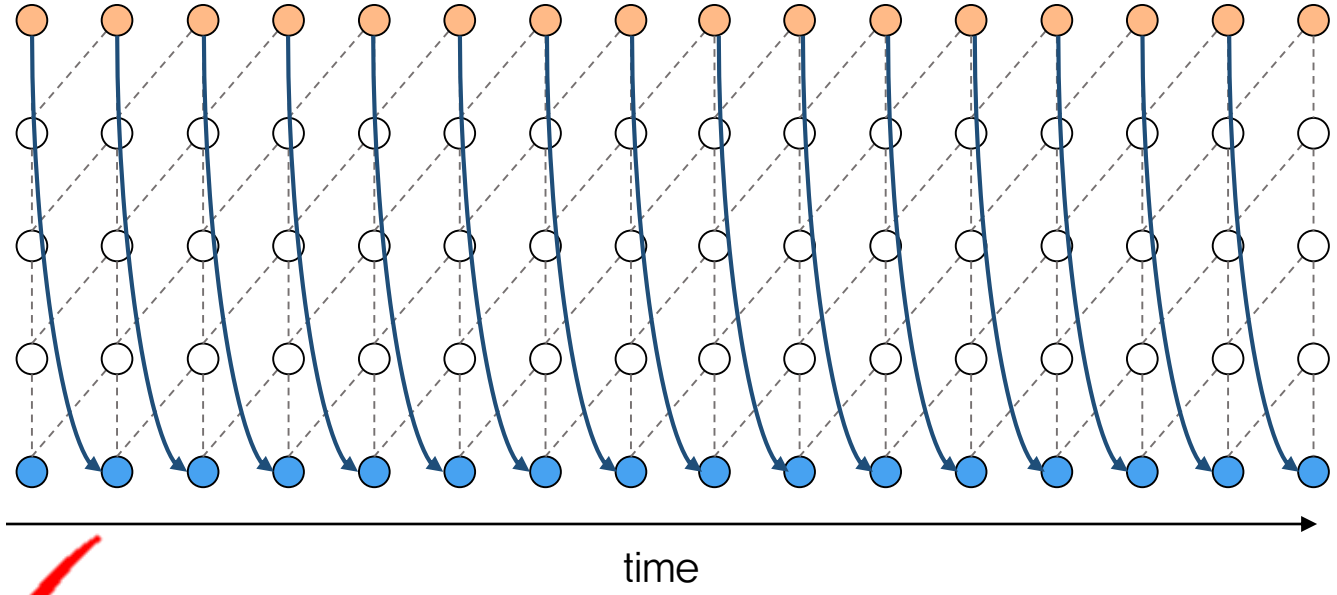
autoregressive

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

Stack of Convolution Layers로 표현하는 것 ⟩⟩ WaveNet

02 | 2. Dilated causal convolutions

- Autoregressive model with convolution

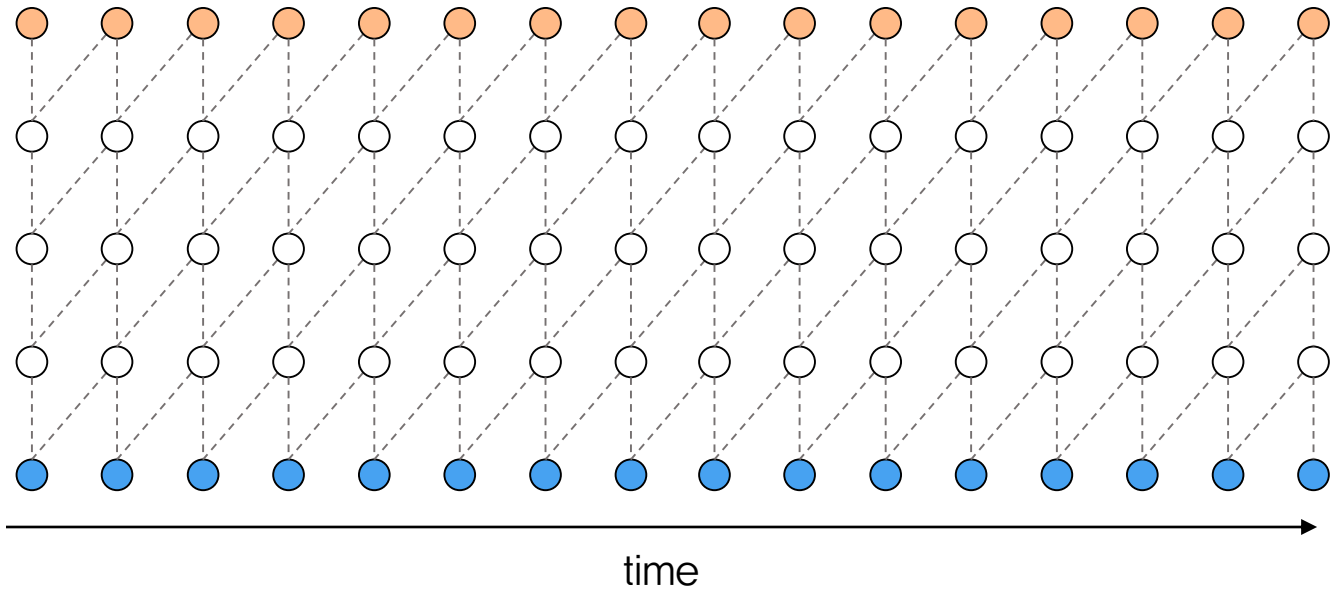


- Autoregressive model:
  $t-1$ 시점에서 얻은 output은 t 시점의 input으로 사용이 된다
- Input, hidden, output size = 16
  →pooling layer를 사용하지 않고,
  zero-padding을 통해 size를 유지함

- 왜 recurrent 대신 convolution을 사용하는가?
  → 예를 들어, 16kHz에서 100ms는 1,600 time steps이다. RNN, LSTM을 사용하여 특징을 학습하기에는 너무 길다.

02 | 2. Dilated causal convolutions

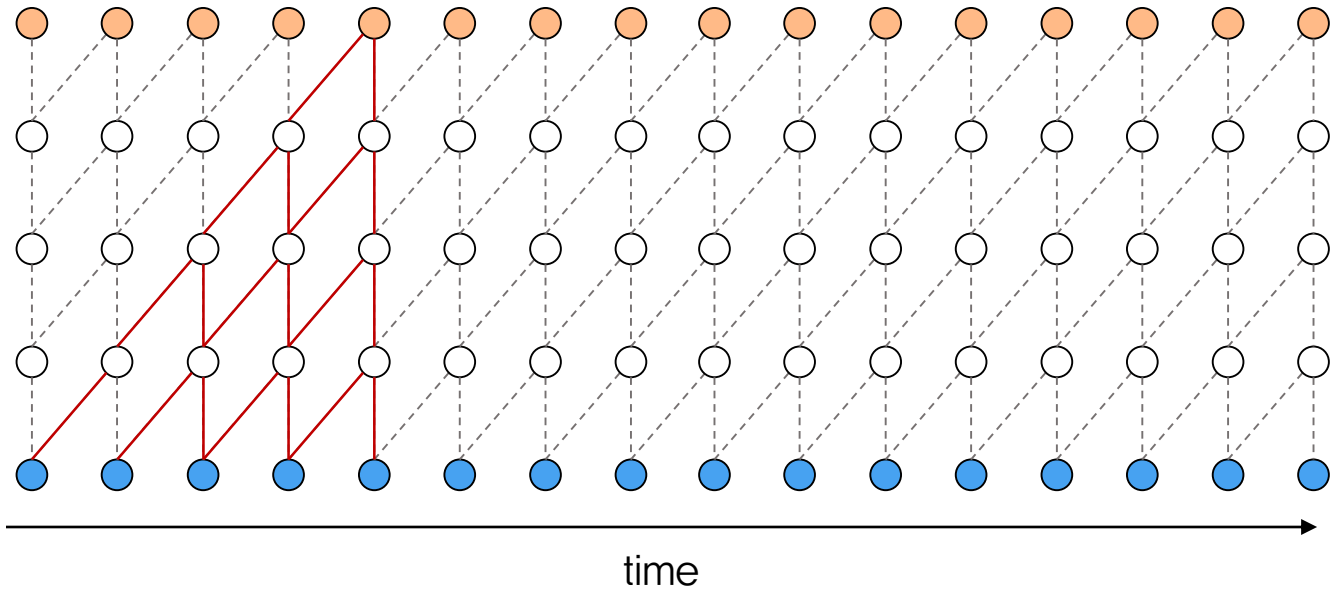
- Causal convolution layers



- # hidden layers = 3
- filter size = (2, 1)

02 | 2. Dilated causal convolutions

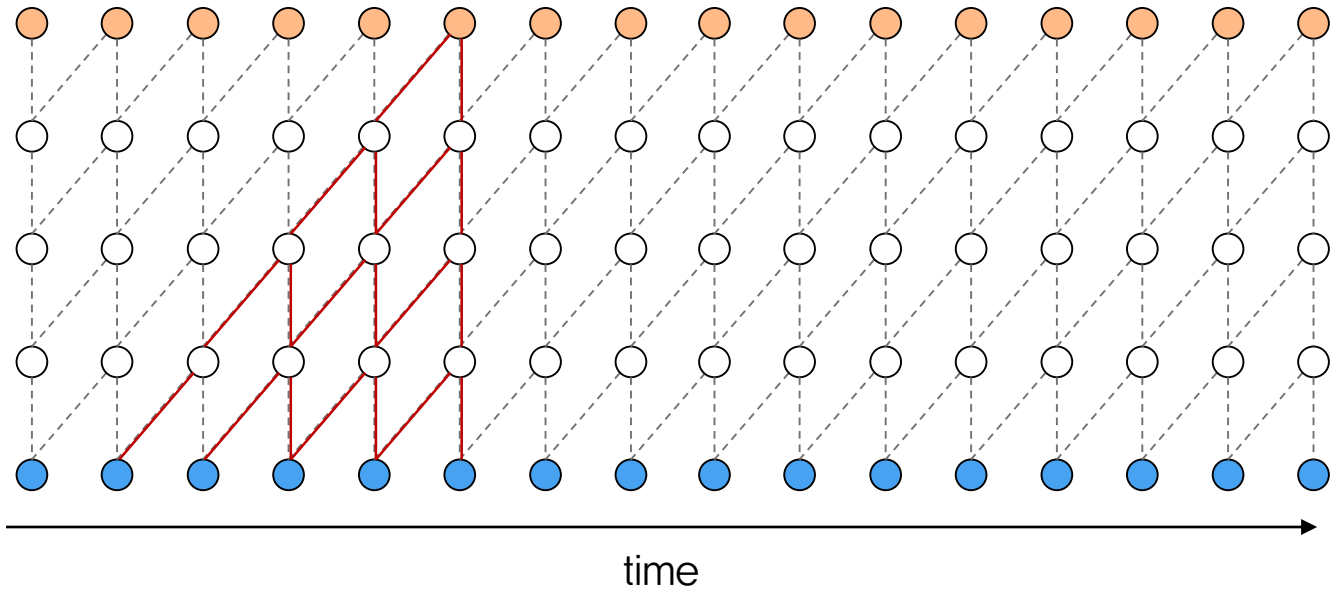
- Causal convolution layers



- 일반적인 convolution을 학습할 때,
  각 filter는 모든 input space를 지나간다

02 | 2. Dilated causal convolutions

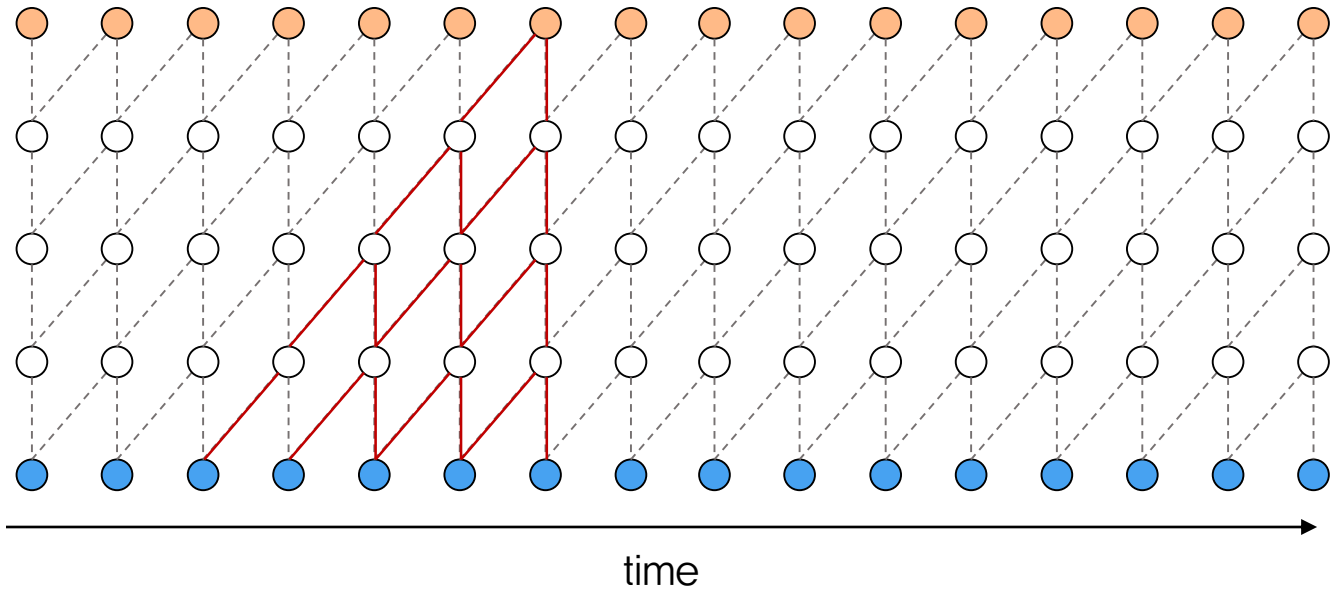
- Causal convolution layers



- 일반적인 convolution을 학습할 때,
  각 filter는 모든 input space를 지나간다

02 | 2. Dilated causal convolutions

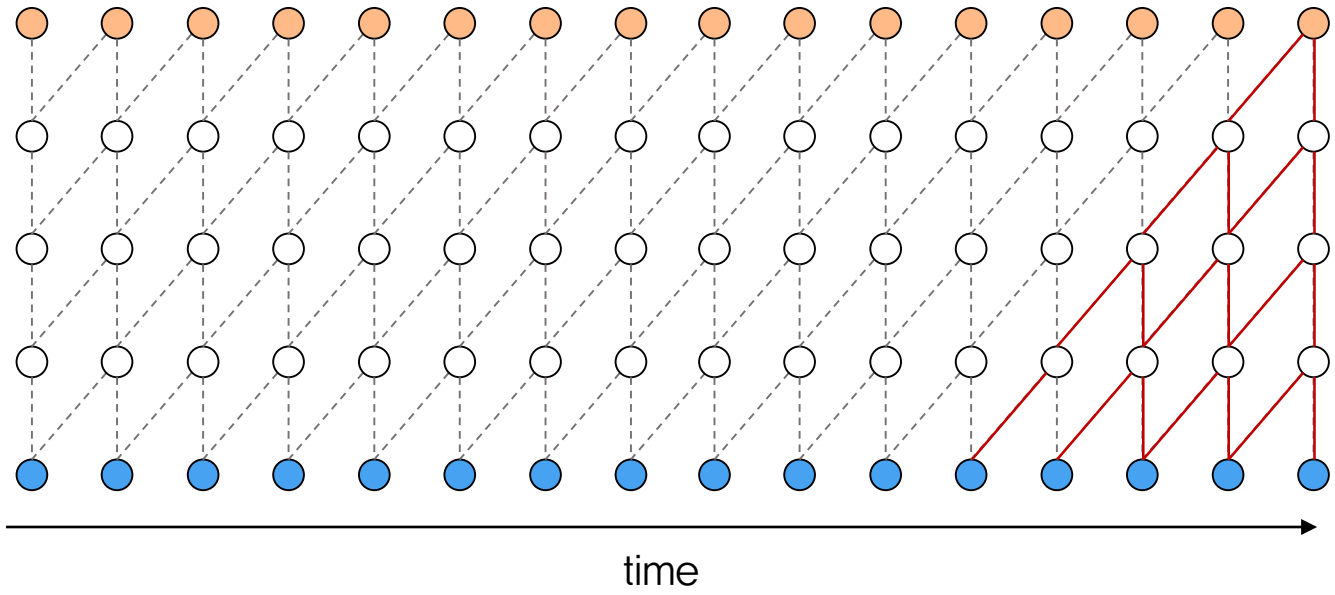
- Causal convolution layers



- 일반적인 convolution을 학습할 때,
  각 filter는 모든 input space를 지나간다

02 | 2. Dilated causal convolutions

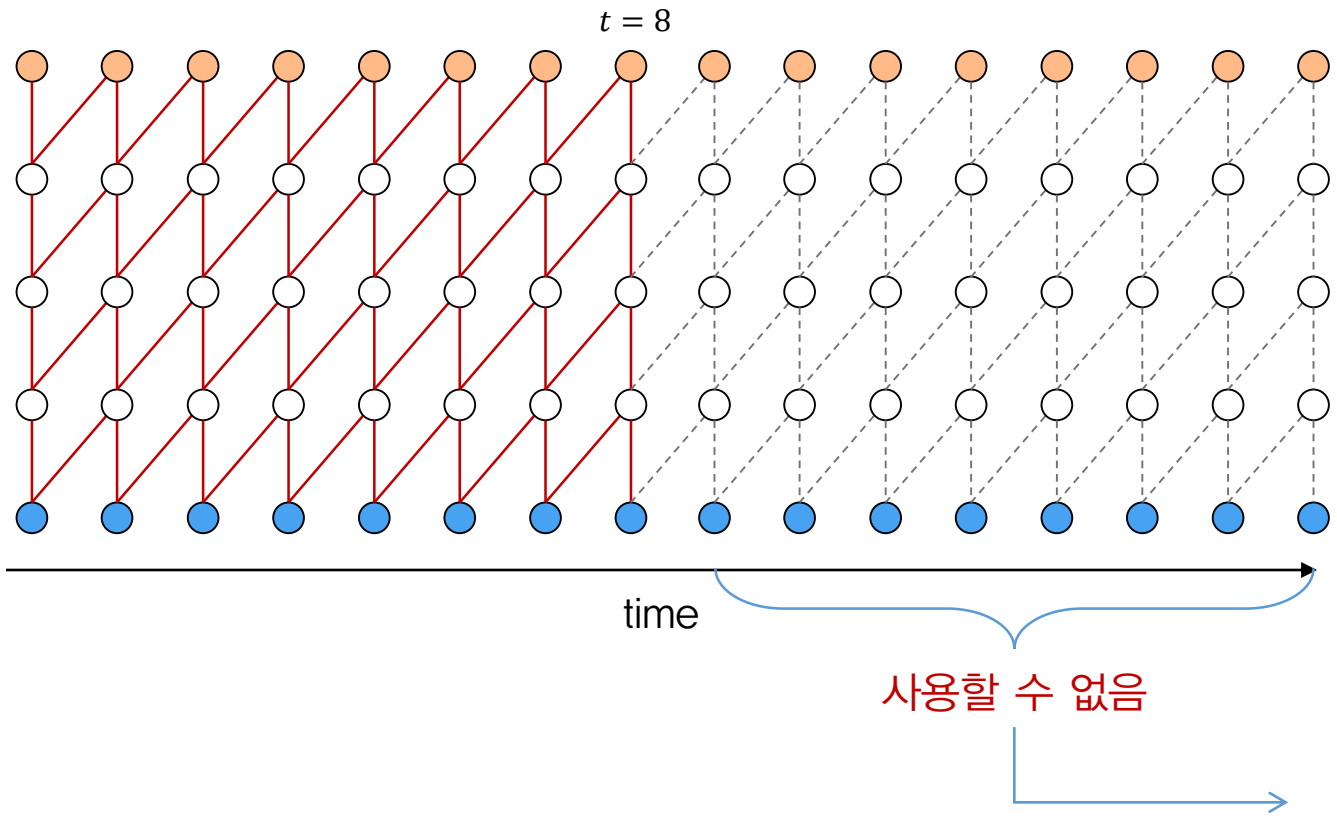
- Causal convolution layers



- 일반적인 convolution을 학습할 때,
  각 filter는 모든 input space를 지나간다

02 | 2. Dilated causal convolutions

- Causal convolution layers

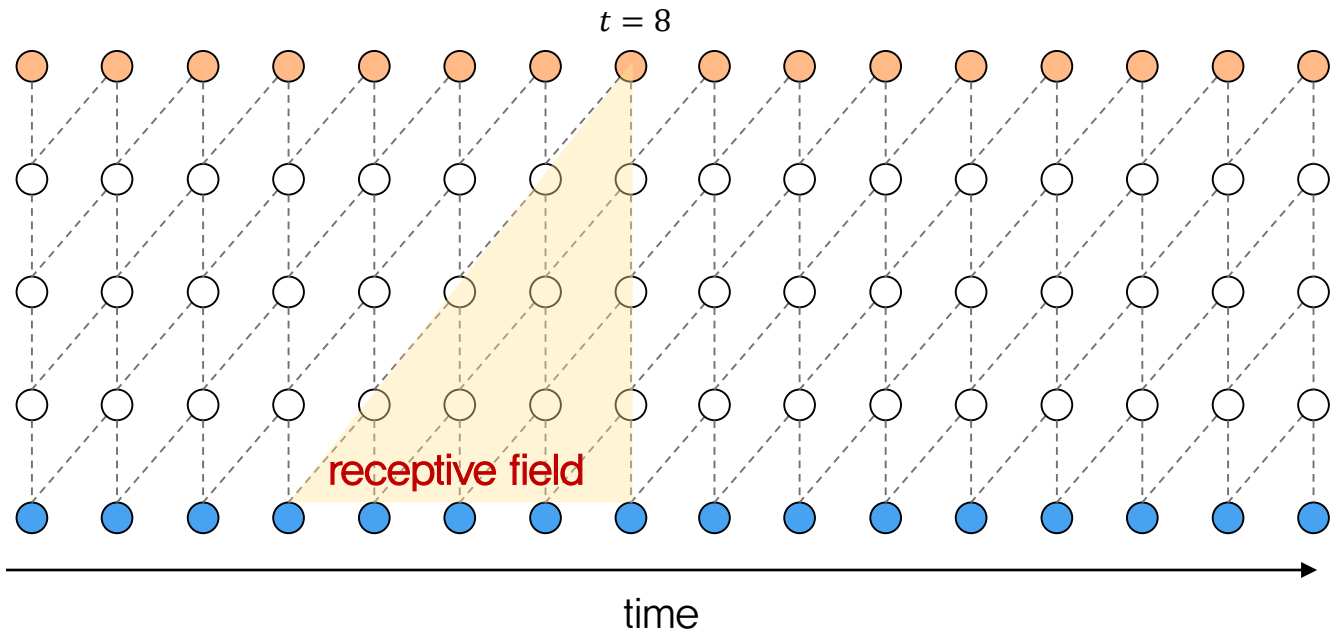


- 일반적인 convolution을 학습할 때,
  각 filter는 모든 input space를 지나간다
- 하지만, 미래의 값을 알 수 없기 때문에
  t 시점 이후의 값들은 사용할 수 없다
  (t 시점의 input은 $t-1$ 시점의 output)

time-based masking: causal filter

02 | 2. Dilated causal convolutions

- Dilated causal convolution layers



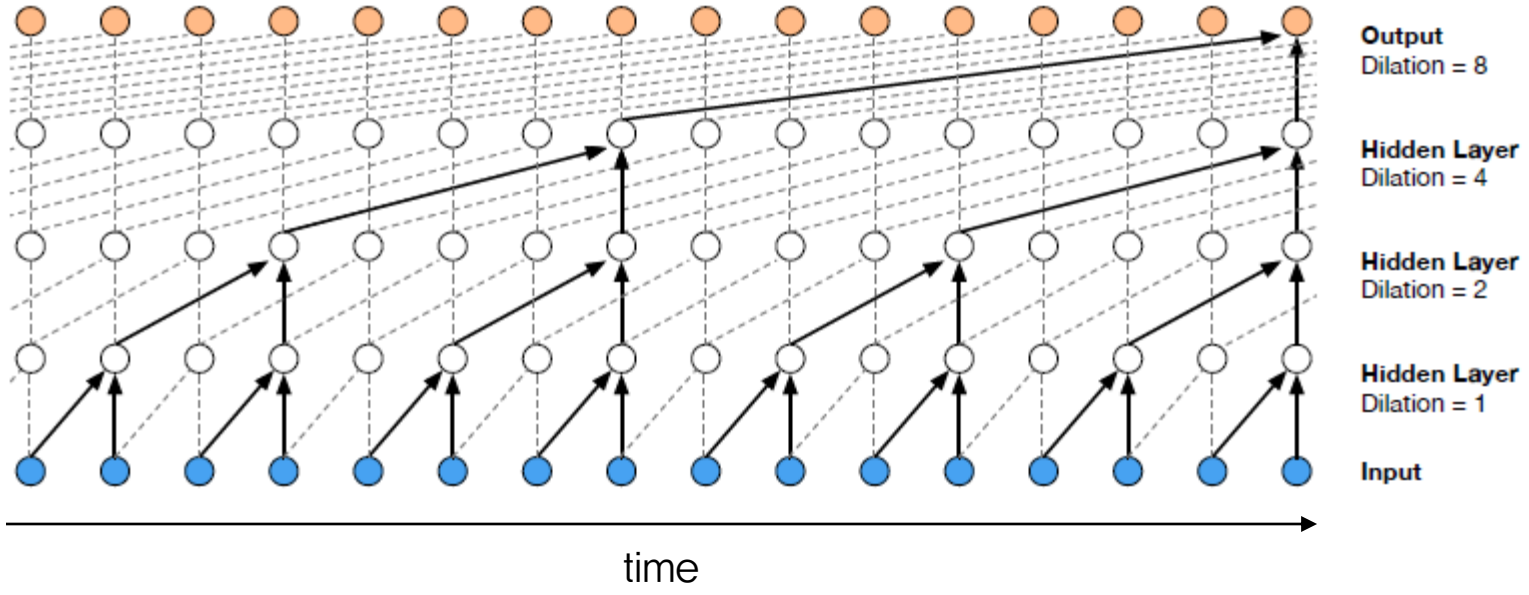
- 문제점: 작은 receptive field
  (참조하게 되는 input space의 크기)
- RNN은 이전 모든 시점들의 값을 고려하기 때문에, 이와 유사한 효과를 가지게 아키텍쳐를 구성하는 것이 필요함

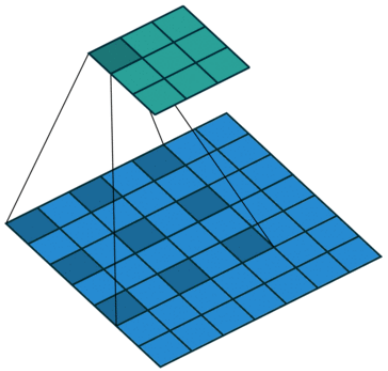
- Convolution의 receptive field를 증가시키기 위해서는
  → layer를 많이 쌓거나, filter size를 키워야 함

- Computational cost가 매우 높아 비효율적임

02 | 2. Dilated causal convolutions

- Dilated causal convolution layers

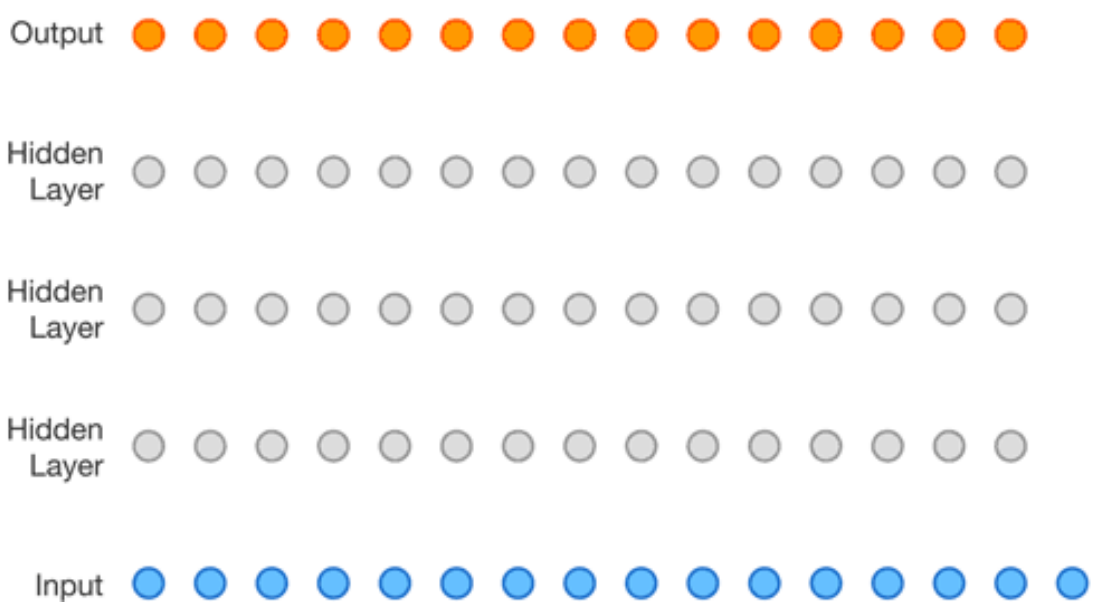


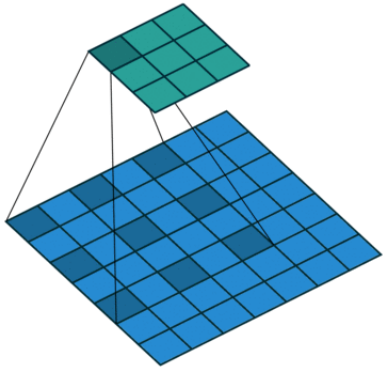


- Dilated convolution을 이용해, 적은 계산량으로 큰 receptive field를 얻음
- Dilation을 1, 2, 4, ⋯, 512, 1, 2, 4, ⋯, 512, 1, 2, 4, ⋯, 512 총 30개의 layer를 구축: receptive field = 1024
  (1x1024 convolution보다 훨씬 효율적)

02 | 2. Dilated causal convolutions

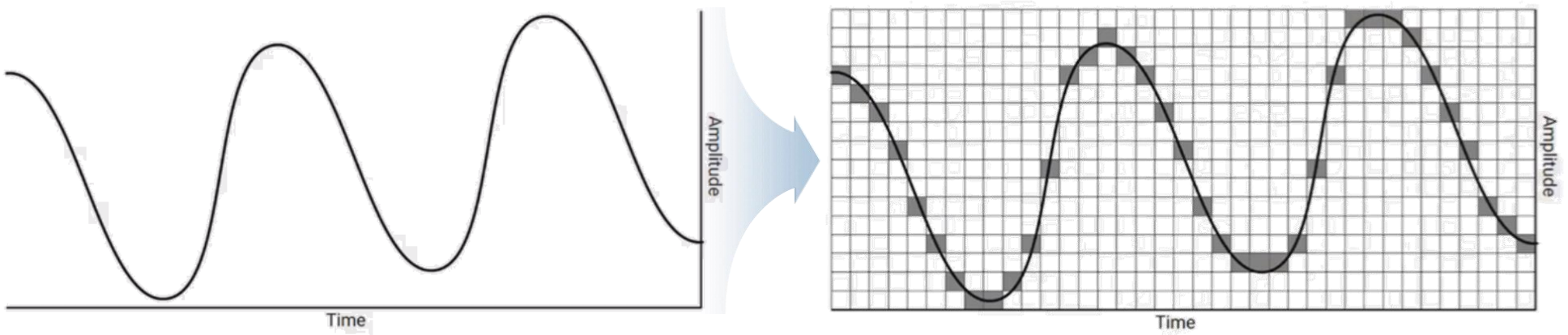
- 예시: Dilated causal convolution layers를 이용한 순차적 예측





02 | 3. Softmax distributions

- 일반적으로 audio의 distribution을 표현하기 위해, Gaussian Mixture Model 등을 사용 (continuous)
- PixelRNN, PixelCNN과 유사하게, 각 amplitude 값을 하나의 class로 가정하여 softmax distribution을 사용함 (discrete)
- 음성 데이터를 256개 클래스로 묶어 regression 문제를 multi-class classification 문제로 변환



analogue: waveform

digital(computer): quantization

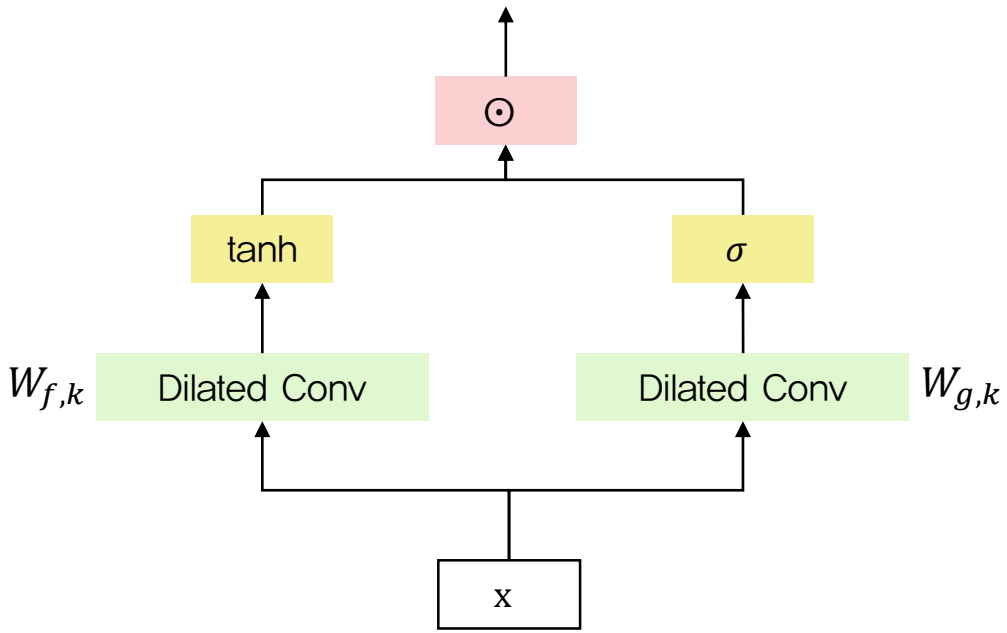
02 | 4. Gated Activation Units

- LSTM: gates에서 사용되는 여러 번의 연산들을 통해 complex interactions를 모델링에 반영할 수 있음
- 이 방식을 사용하여 ReLU activation 대신 Gated Activation을 사용

Gated Activation

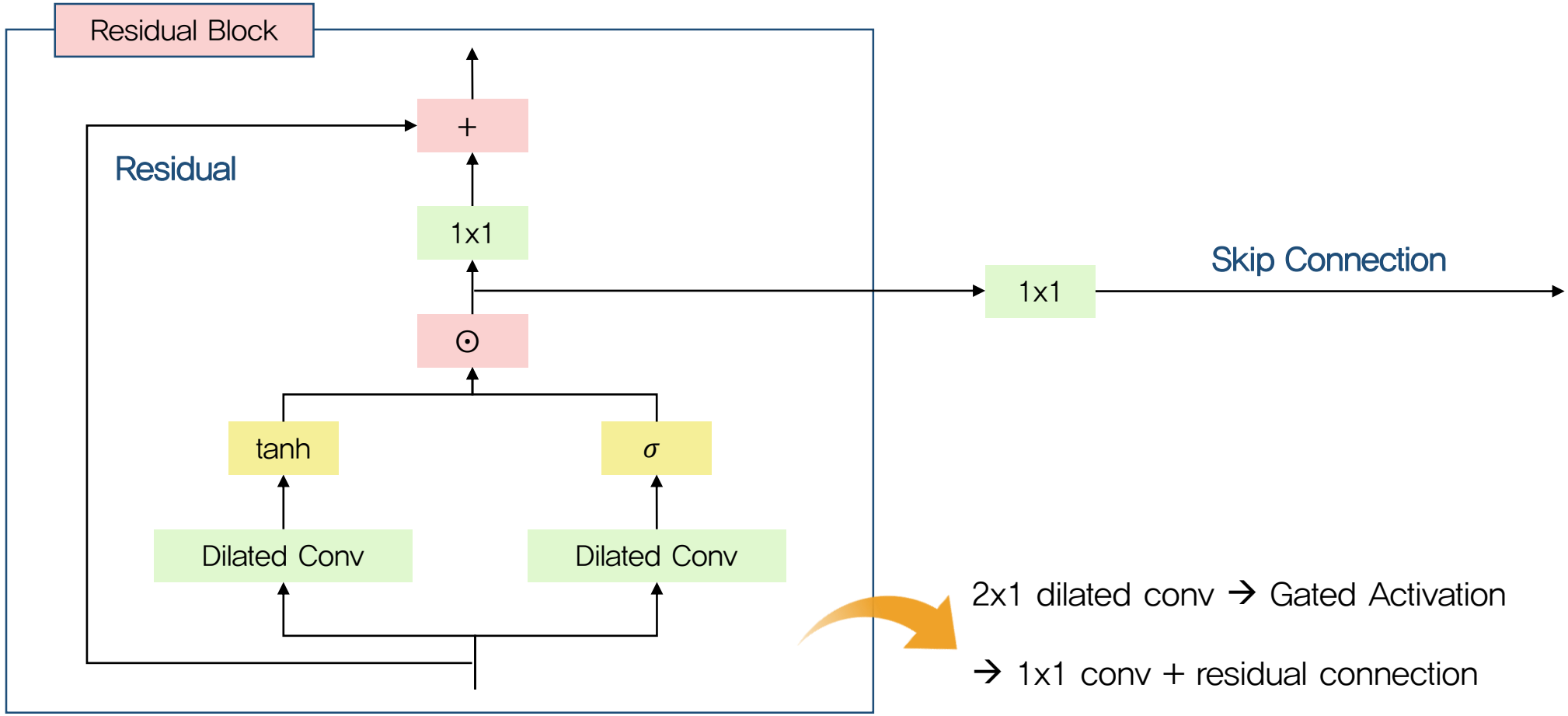
$$z = \tanh(W_{f,k} * \mathrm{x}) \odot \sigma(W_{g,k} * \mathrm{x})$$

- $*$: convolution operator
- $\odot$: element-wise multiplication operator
- k: layer index
- f, g: filter, gate → W is learnable convolution filter



02 | 5. Residual and Skip Connections

- Residual and parameterized skip connections: speed up convergence & enable deeper models

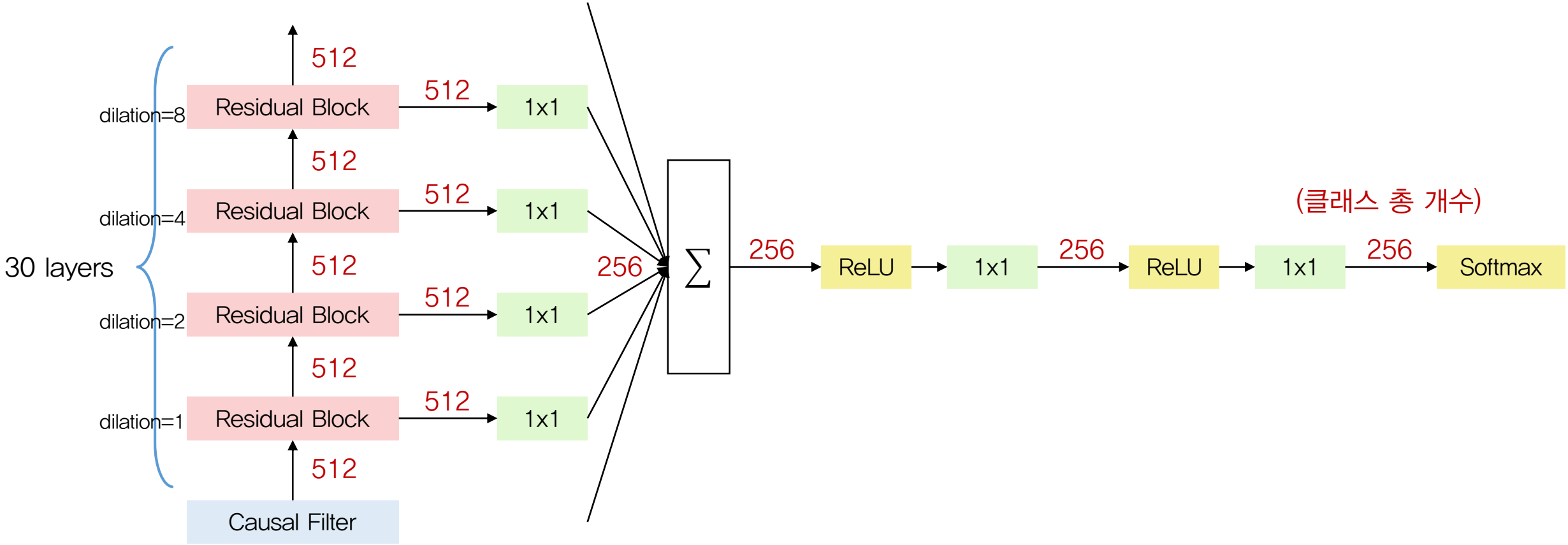


2x1 dilated conv → Gated Activation

→ 1x1 conv + residual connection

02 | 5. Residual and Skip Connections

- WaveNet을 펼쳤을 때의 모습 (한 timestep에서)



02 | 6. Conditional WaveNet

- Guide WaveNet's generation to produce audio with the required characteristics
- Global condition: multi-speaker setting에서 화자의 특징을 표현
- Local condition: text-to-speech에서 text의 특징을 표현

WaveNet

- $p(\mathbf{x}) = \prod_{t=1}^{T} p(\mathrm{x}_t|\mathrm{x}_1, \ldots, \mathrm{x}_{t-1})$
- $z = \tanh(W_{f,k} * \mathrm{x}) \odot \sigma(W_{g,k} * \mathrm{x})$

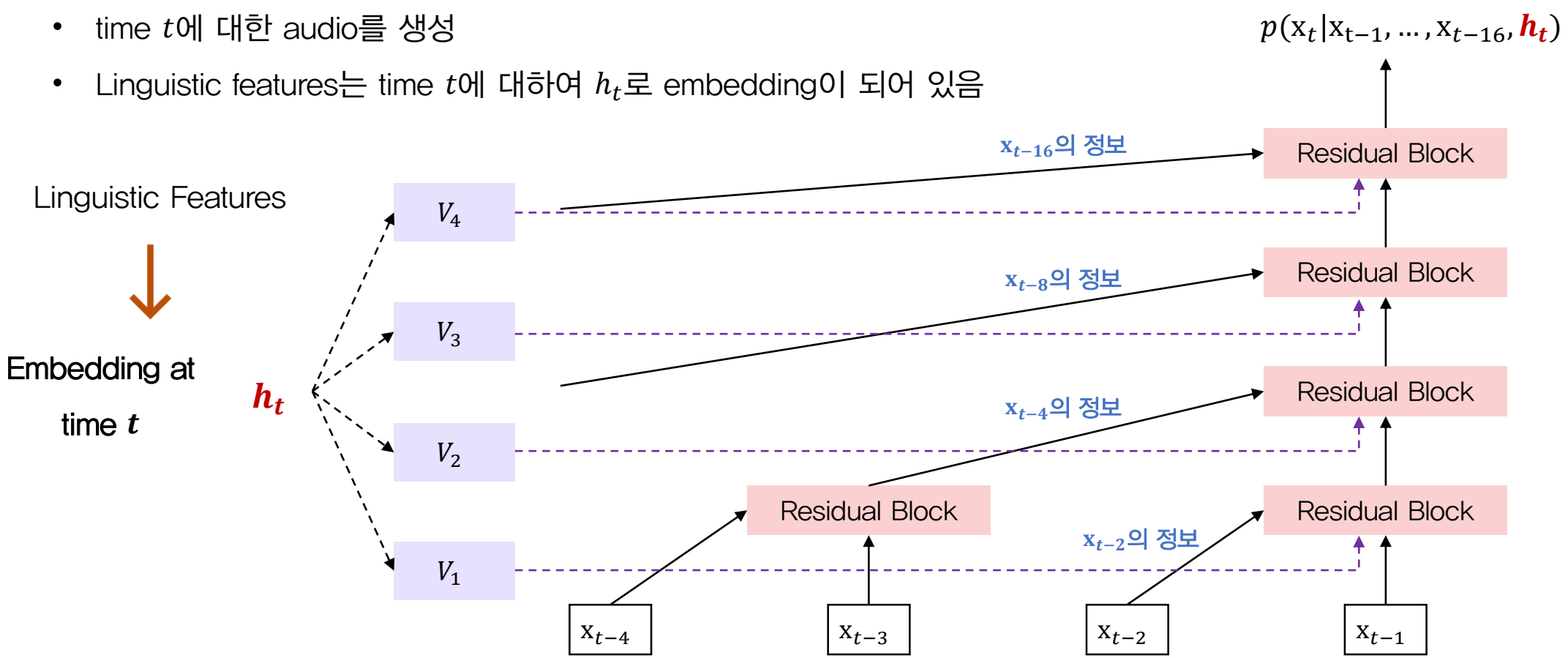
Conditional WaveNet

- $p(\mathbf{x}|\boldsymbol{h}) = \prod_{t=1}^{T} p(\mathrm{x}_t|\mathrm{x}_1, \ldots, \mathrm{x}_{t-1}, \boldsymbol{h})$
- $z = \tanh(W_{f,k} * \mathrm{x} + \mathrm{V}_{f,k}^{T} h) \odot \sigma(W_{g,k} * \mathrm{x} + V_{g,k}^{T} h)$

- V: learnable linear projection
- h가 시점에 따라 달라지는 함수 형태면 local condition,
  전체적으로 같으면 global condition으로 구분한다

02 | 6. Conditional WaveNet

- Local condition 예시: 1x2 dilation으로 4개의 layer를 쌓았을 때
- time t에 대한 audio를 생성
- Linguistic features는 time t에 대하여 h_t로 embedding이 되어 있음



02 | 6. Conditional WaveNet

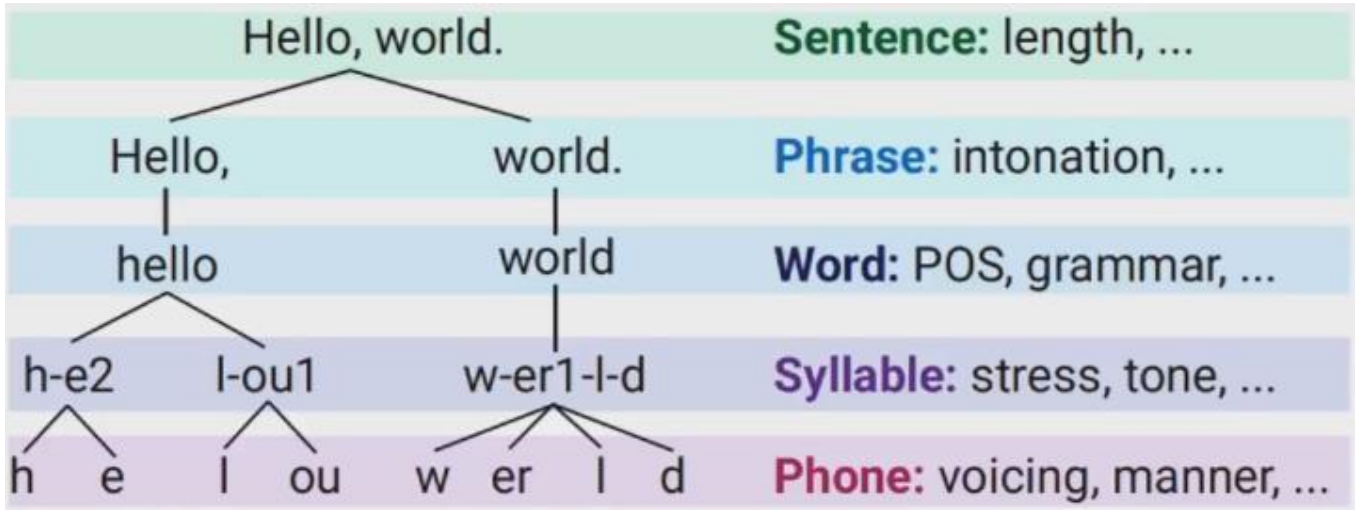
- Embedded linguistic features: sentence, phrase, word, syllable, phone 등 다양한 정보를 함축적으로 표현

Linguistic Features

↓

Embedding at time t

h_t



Contents

03 Experiments

- 총 3가지 종류의 실험을 진행하여 WaveNet의 generation 성능을 평가

1. Multi-Speaker Speech Generation
2. Text-to-Speech
3. Music

03 | 1. Multi-Speaker Speech

- 실험 세팅
  1. Free-form speech generation – not conditioned on text
  2. 109명의 다른 화자의 음성 데이터셋을 사용하고, 화자의 ID를 one-hot vector h로 인코딩하여 사용
  3. Globally conditioned WaveNet을 사용

- 실험 결과
  1. Text (linguistic) 정보가 없기 때문에, 실제로 존재하는 단어로 이루어진 음성이 생성되지 않음
  2. 하지만, 듣기에는 human language-like words로 구성되어 있음
     (Image domain: 사진처럼 보이지만 실제로는 존재하지 않는 물체)
  3. 하나의 WaveNet만으로도 각 화자의 특징을 반영할 수 있음

03 | 2. Text-to-Speech

- 실험 세팅
  1. 사용 언어: North American English & Mandarin Chinese
  2. Single speaker
  3. Locally conditioned WaveNet

- 평가 모델
  1. Statistical parametric: LSTM-RNN
  2. Concatenative: HMM-driven
  3. WaveNet (L): linguistic feature
  4. WaveNet (L+F): linguistic feature + $\log F_0$

     → logarithmic fundamental frequency: linguistic feature로부터 값을 얻는 모델이 추가적으로 존재

03 | 2. Text-to-Speech

- 평가 방법
  1. Subjective paired comparison test

     → 평가 모델을 각각 짝지어서, 더 나은 모델을 선택하거나, 아예 선택하지 않음 (no preference)

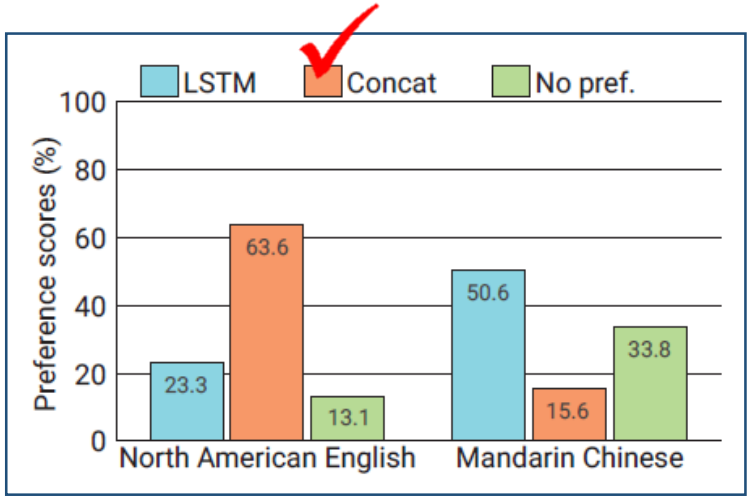
     → 선택된 비율로 모델의 성능을 평가

  2. Mean opinion score (MOS) test

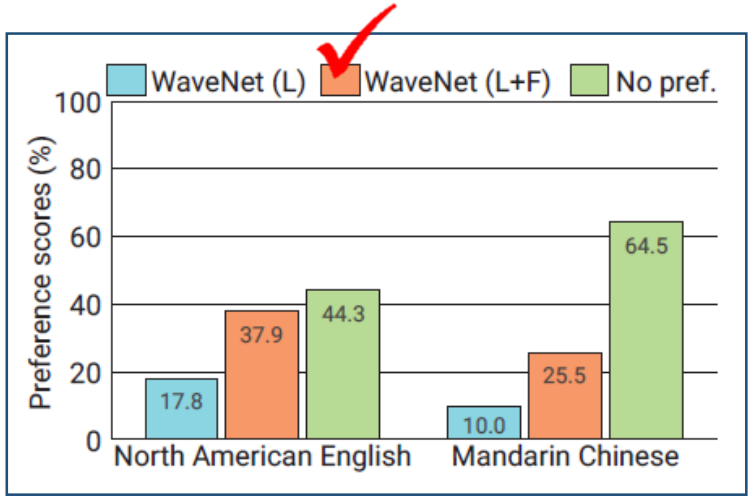
     → 각 샘플을 들은 후, 5점 평가 (1: bad, 2: poor, 3: fair, 4: good 5: excellent)

03 | 2. Text-to-Speech

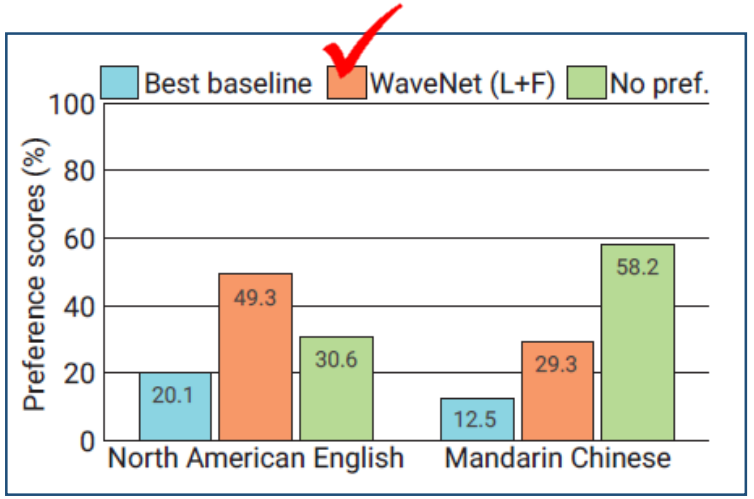
- 평가 결과
  1. Subjective paired comparison test



Baseline 모델 평가



F_0에 대한 평가



최종 평가

03 | 2. Text-to-Speech

- 평가 결과

2. MOS test

Speech samples	Subjective 5-scale MOS in naturalness	
	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

- WaveNet이 다른 모델들에 비해 높은 점수를 얻었으며,
  실제 음성과 비교해서도 좋은 점수를 얻었다

03 | 3. Music

- 실험 세팅
  1. 사용 데이터
     a. MagnaTagATune dataset: 29초 음악 샘플. 장르, 악기, 템포, 분위기 등에 대한 tag가 있음.
     b. YouTube piano dataset: 60시간 피아노 독주.
  2. Unconditional WaveNet
  3. Globally conditional WaveNet

- 평가 방법 및 결과
  1. 주관적인 평가를 수행: often harmonic and aesthetically pleasing
  2. 장르, 악기 등에 대한 conditional generation도 원활히 작동

03 Conclusions

- WaveNet: a deep generative model of audio data that operates directly at the waveform level

- Autoregressive and combine causal filters with dilated convolutions
  → exponentially increase receptive fields
  → learn long-range temporal dependencies in audio signal

- Text-to-Speech application: outperform the current best TTS systems


Thank
you