WaveNet

A Generative Model For Raw Audio

이다경

INDEX

1. Introduction

2. WaveNet

- Dilated Causal Convolution
- Softmax Distribution
- Gated Activation Units
- Rasidual and Skip Connection
- Conditional WaveNets

3. Experiments

- Multi-Speaker Speech Generation
- Text To Speech
- Music
- Speech Recognition

1. Introduction

- We show that WaveNets can generate raw speech signals with subjective naturalness never before reported in the field of text-to-speech (TTS), as assessed by human raters.

  speech signal 생성 모델. 보고 된 적 없는 TTS 기법

- In order to deal with long-range temporal dependencies needed for raw audio generation, we develop new architectures based on dilated causal convolutions, which exhibit very large receptive fields.

  receptive fields : input fields

- We show that a single model can be used to generate different voices, conditioned on a speaker identity.

  speaker 의 identity를 반영한 audio를 generate 할 수 있다.

- The same architecture shows strong results when tested on a small speech recognition dataset, and is promising when used to generate other audio modalities such as music.

  music도 생성 할 수 있다.

2. WaveNet

Time Step에 따른 $x_1, x_2, x_3, \ldots$ 이 주어졌을 때 x_t 추정

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

PixelCNN 기법

- input size = output size
- no pooling layers
- The model outputs a categorical distribution over the next value x_t with a softmax layer.

2. WaveNet

2.1 Dilated Causal Convolution

Causal Convolution

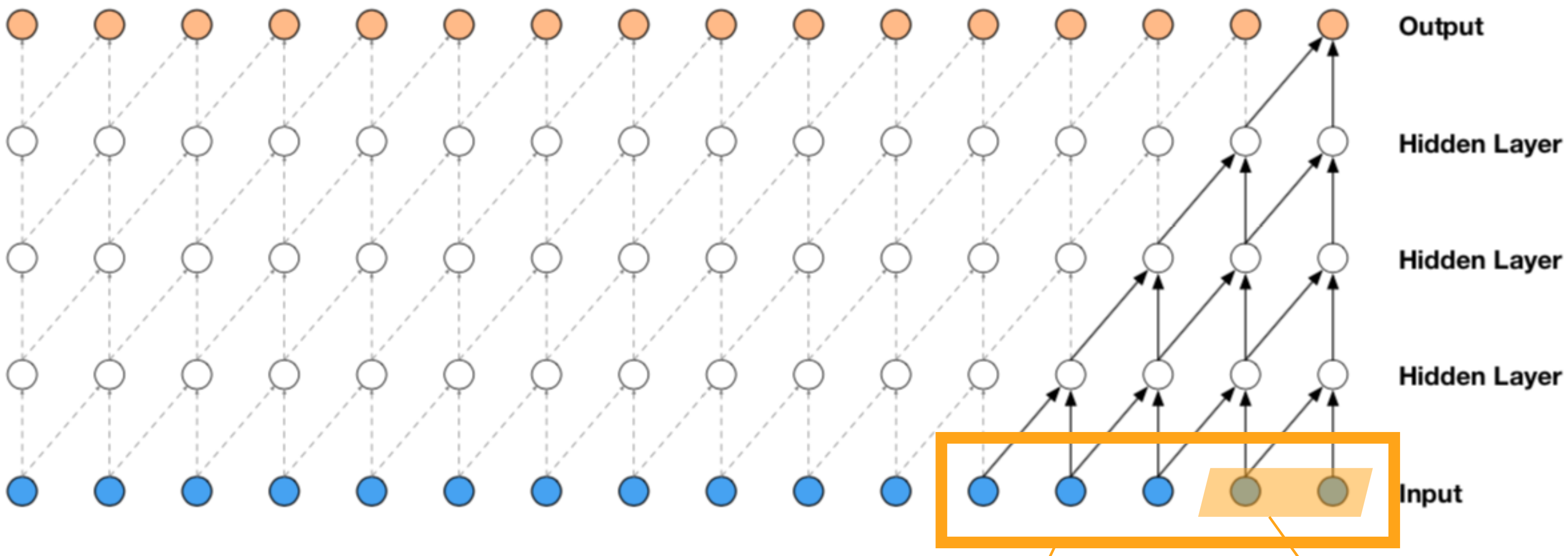


Figure 2: Visualization of a stack of causal convolutional layers.

size=2인 conv filter

receptive field

2. WaveNet

2.1 Dilated Causal Convolution

Causal Convolution

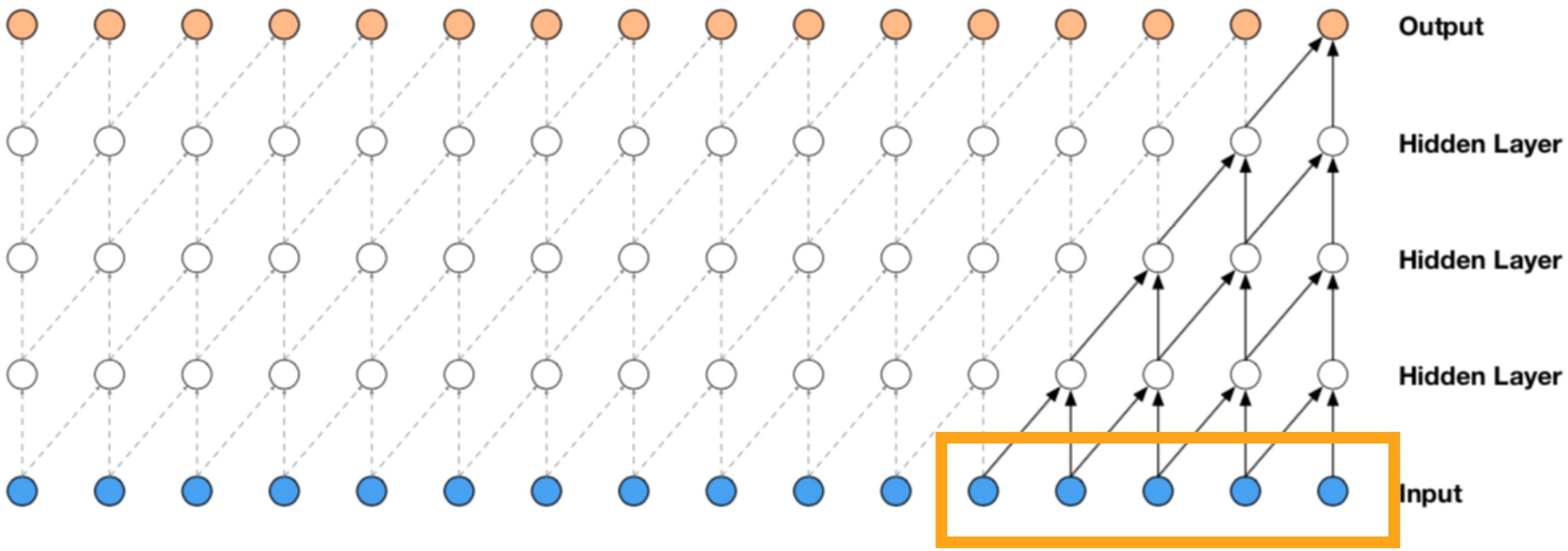


Figure 2: Visualization of a stack of causal convolutional layers.

But, 4개의 layer로 receptive filter 5개밖에..!
receptive filter 늘리려면 layer가 너무너무 깊어져..
→ Dilated Causal Convolution

2. WaveNet

2.1 Dilated Causal Convolution

Dilated Causal Convolution

* dilated : 넓히다
* Convolution with hole

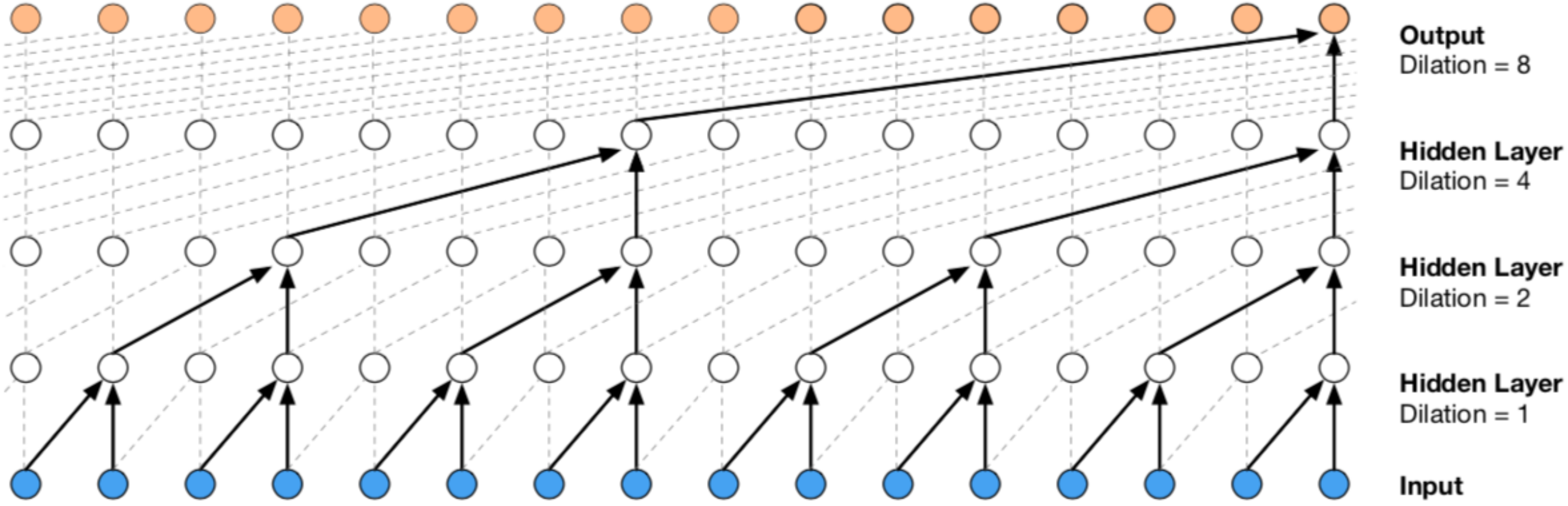


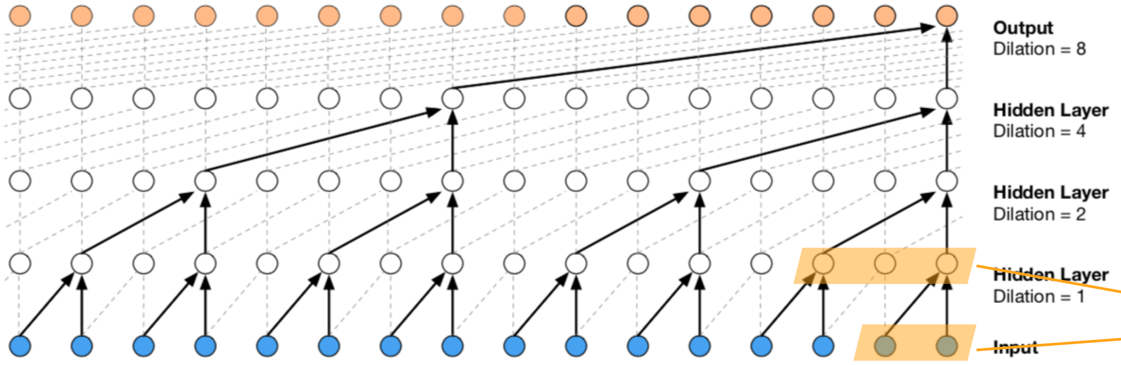
Figure 3: Visualization of a stack of *dilated* causal convolutional layers.

large receptive field with few layers

https://deepmind.com/blog/wavenet-generative-model-raw-audio/

2. WaveNet

2.1 Dilated Causal Convolution



마치 그냥 Causal에서 filter size가 늘어난 효과!
하지만 더 효율적

Figure 3: Visualization of a stack of *dilated* causal convolutional layers.

train할 때는 output 다시 input으로 넣을 필요 x (데이터가 원래 있으니까!)
실제 generating 할 땐 generate된 output이 다시 input으로!

2. WaveNet

2.1 Dilated Causal Convolution

$$\underline{1, 2, 4, \ldots, 512,}\ \underline{1, 2, 4, \ldots, 512,}\ \underline{1, 2, 4, \ldots, 512.}$$

10 layers, 1024 receptive field

↓

총 30 layers, large receptive field

2. WaveNet

2.2 Softmax Distribution

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

mixture density network (Bishop, 1994)

mixture of conditional Gaussian scale mixtures (MCGSM) (Theis & Bethge, 2015)

However, van den Oord et al. (2016a) showed that a softmax distribution tends to work better

reason 1. categorical distribution is more flexible
reason 2. can more easily model arbitrary distributions because it makes no assumptions about their shape.

shape에 대한 가정이 없기 때문에 분포를 쉽게 모델링 할 수 있다.

2. WaveNet

2.2 Softmax Distribution

raw audio는 일반적으로 16-bit integer values로 저장되어 있음(one per timestep)
→ softmaxt가 timestep 별 65536(= 2^{16}) output.
→ 사람도 실제 소리를 들을 때 저주파에 민감함을 반영. 비선형 양자화를 통해 256 output

$$f(x_t) = \operatorname{sign}(x_t)\frac{\ln(1 + \mu|x_t|)}{\ln(1 + \mu)}, \quad \text{where } -1 < x_t < 1 \text{ and } \mu = 255.$$

2. WaveNet

2.3 Gated Activation Units

그냥 CNN만 하는게 아니라 gate unit까지 * conv network를 통해 나온 결과를 얼마나 통과시킬지?

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

where
$*$ denotes a convolution operator,
$\odot$ denotes an element-wise multiplication operator,
σ(·) is a sigmoid function,
k is the layer index,
f and g denote filter and gate, respectively,
W is a learnable convolution filter.

2. WaveNet

2.4 Residual And Skip Connection

더 깊은 모델, 빨리 수렴하기 위해

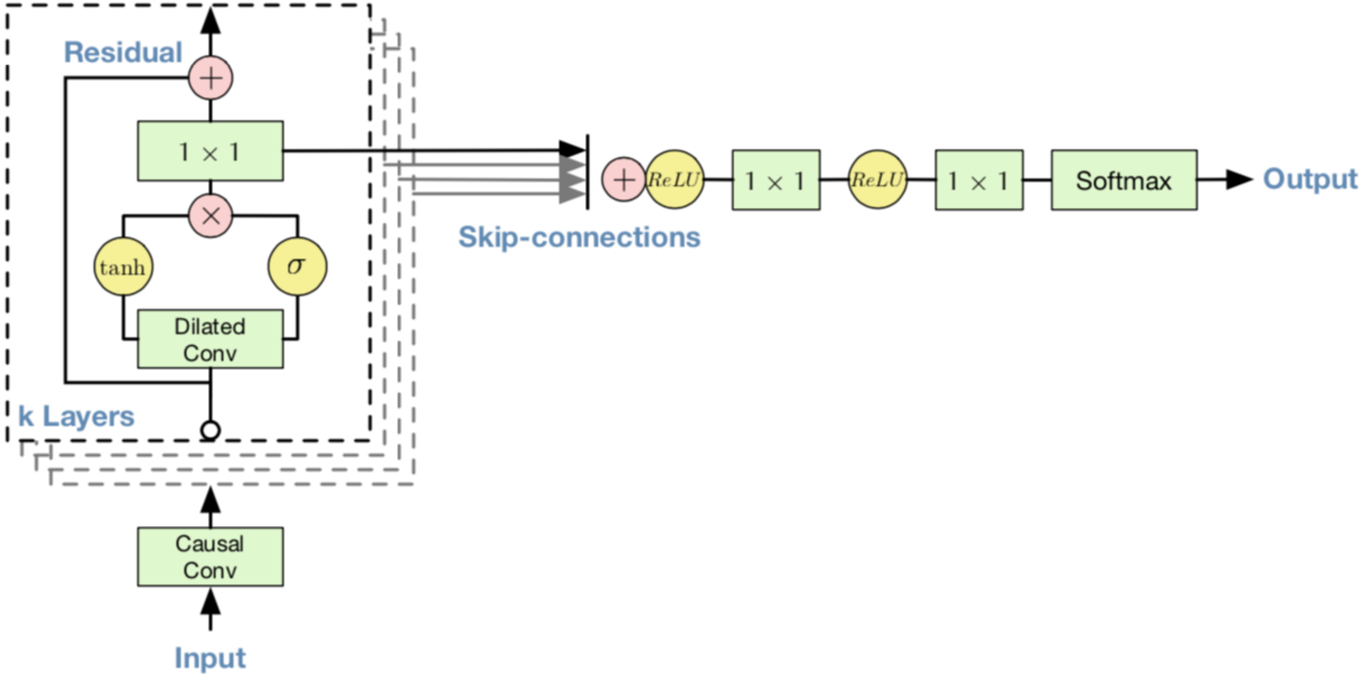
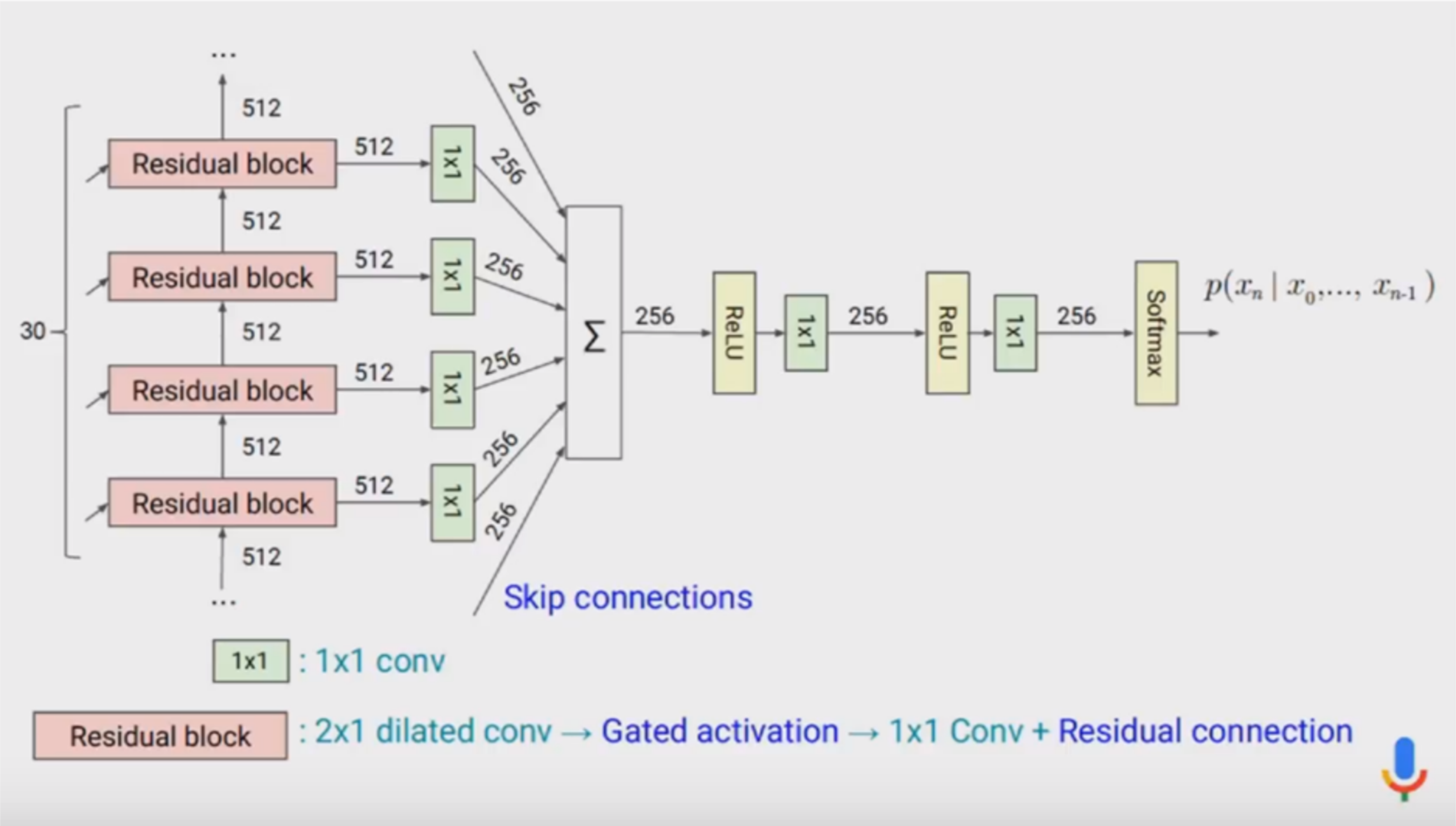


Figure 4: Overview of the residual block and the entire architecture.

2. WaveNet

2.4 Residual And Skip Connection



2. WaveNet

2.4 Residual And Skip Connection

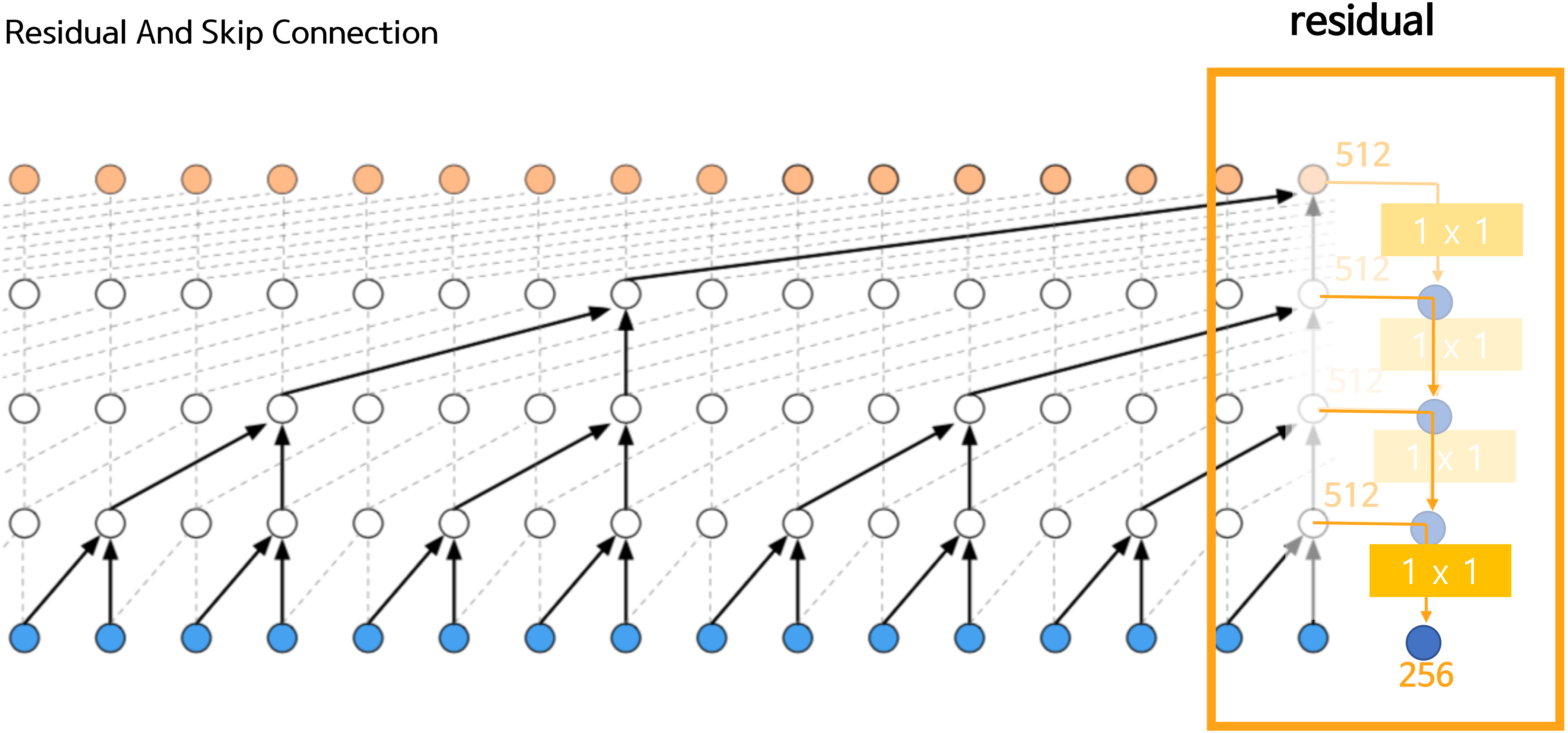


Figure 3: Visualization of a stack of *dilated* causal convolutional lay

2. WaveNet

2.4 Residual And Skip Connection

skip connection

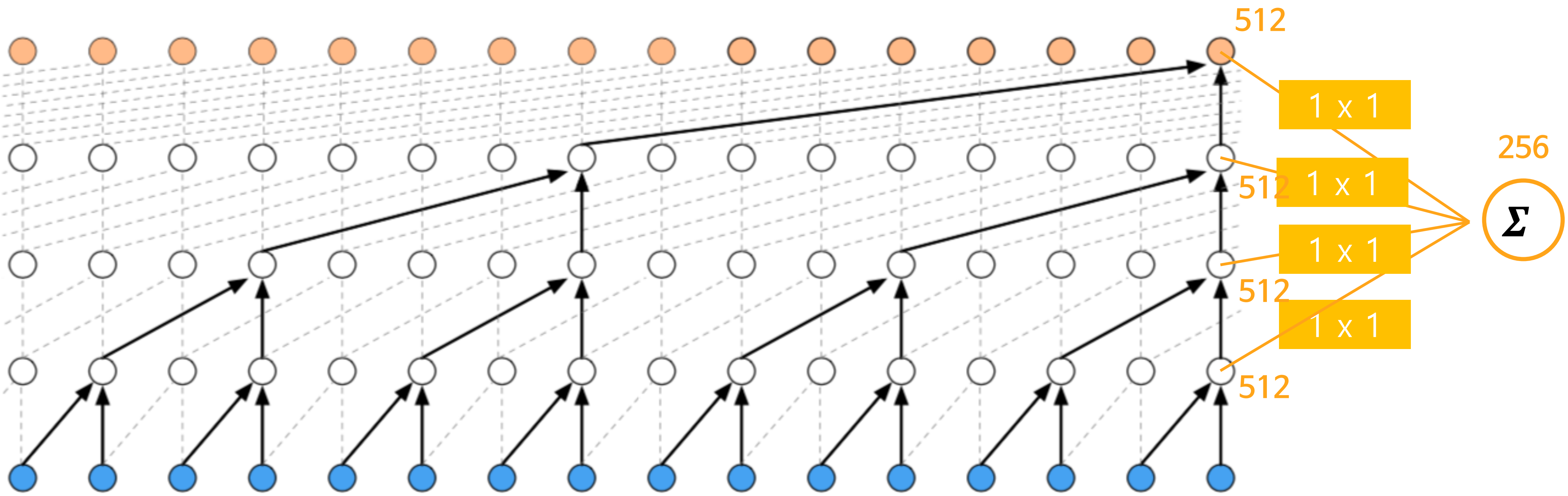


Figure 3: Visualization of a stack of *dilated* causal convolutional lay

2. WaveNet

2.5 Conditional WaveNets

Generate 하는 audio에 condition 줄 수 있음

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

Global conditioning

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

where $V_{*,k}$ is a learnable linear projection

all timestep에 같은 condition 적용
ex) speaker의 특징 (남, 여 등)

Local conditioning

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

where $V_{f,k} * y$ is now a 1×1 convolution.

audio 신호보다 낮은 샘플링 주파수를 갖는 second timeseries
ex) TTS의 텍스트

$y = f(h)$: audio signal과 같은 time series를 갖도록 upsampling하는 CNN network

3. Experiments

1. speaker speech generation
2. TTS
3. music audio modeling
4. speech recognition

https://deepmind.com/blog/wavenet-generative-model-raw-audio/

3. Experiments

3.1 Multi-Speaker Speech Generation

free-form speech generation (not conditioned on text)

DATA : http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html

CSTR VCTK Corpus 109 native speakers of English

Each speaker reads out about 400 sentences

Condition : speaker ID one-hot으로

text가 주어지지 않기 때문에 실제 존재하지는 않는 단어.
단지 사람의 언어와 비슷한 단어 생성.

speaker의 호흡, 입의 움직임 모방,
음향, 녹음품질까지 모방

3. Experiments

3.2 Text-To-Speech

DATA : single-speaker speech databases

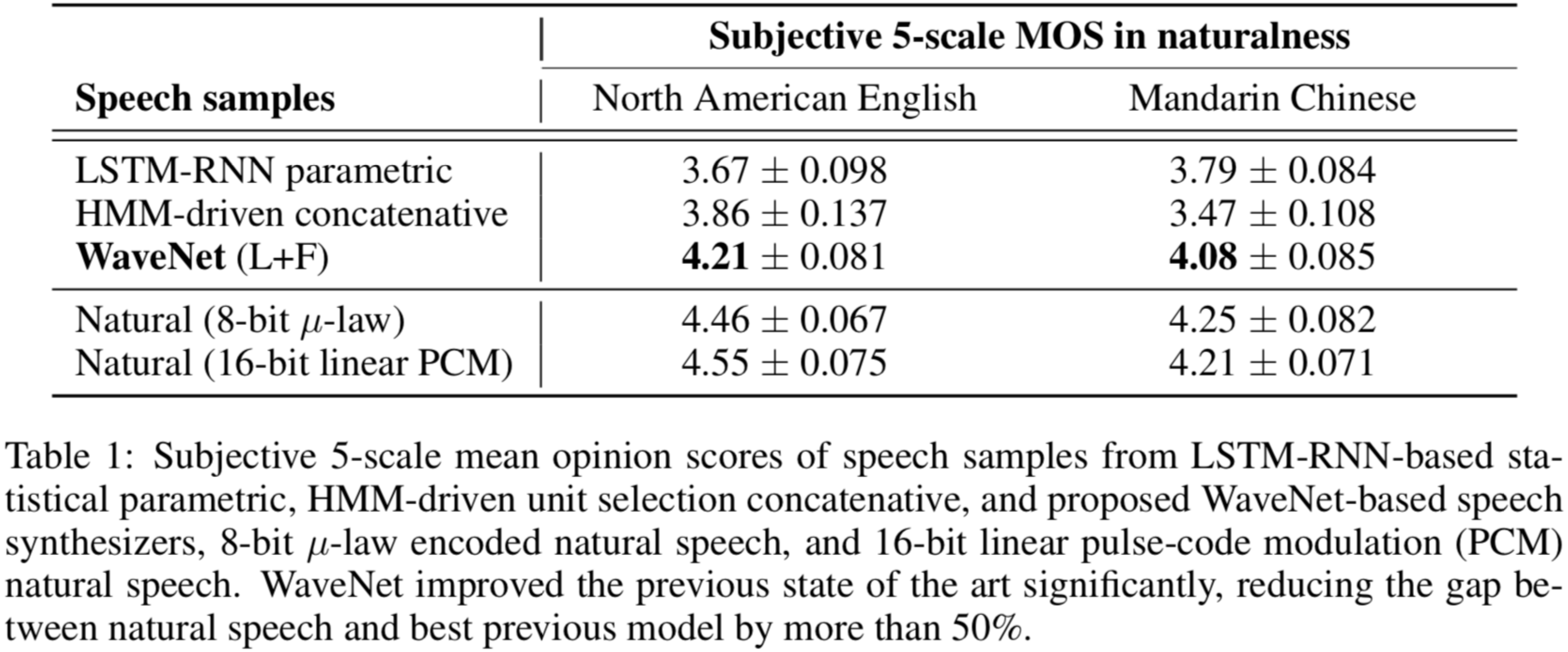
Google's North American English and Mandarin Chinese TTS systems are built. 자체 시스템 구축한듯

English dataset contains 24.6 hours Mandarin Chinese dataset contains 34.8 hours

Condition : Local conditioning

3. Experiments

3.2 Text-To-Speech



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)	**4.21** ± 0.081	**4.08** ± 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

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 μ-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%.

In the MOS tests, after listening to each stimulus,
the subjects were asked to rate the naturalness of the stimulus in a five-point Likert scale score
(1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent).

실험자가 소리를 듣고 1~5 점수

3. Experiments

3.3 Music

DATA：

- theMagnaTagATunedataset(Law&VonAhn,2009),whichconsistsofabout200hoursof music audio. Each 29-second clip is annotated with tags from a set of 188, which describe the genre, instrumentation, tempo, volume and mood of the music.
- the YouTube piano dataset, which consists of about 60 hours of solo piano music obtained from YouTube videos. Because it is constrained to a single instrument, it is considerably easier to model.

3. Experiments

3.4 Speech Recognition

Although WaveNet was designed as a generative model,
it can straightforwardly be adapted to discriminative audio tasks such as speech recognition.

Generating 모델이지만, discriminative 역할도

speech recognition research has largely focused on using log mel-filterbank energies or mel-frequency cepstral coefficients (MFCCs), but has been moving to raw audio recently
(Palaz et al., 2013; Tu¨ske et al., 2014; Hoshen et al., 2015; Sainath et al., 2015).

원래는 speech recognition의 feature가 주로 mel-filterbank나 MFCC였으나, 최근에 raw audio로

DATA : TIMIT https://catalog.ldc.upenn.edu/ldc93s1

dilated convolution 이후에 pooling layer 추가

two loss term - one to predict the next sample
one to classify the frame