투빅스 정규 세미나

투빅스 9기 김유리안나

음성 인식과 음성 합성

Contents

투빅스 정규 세미나

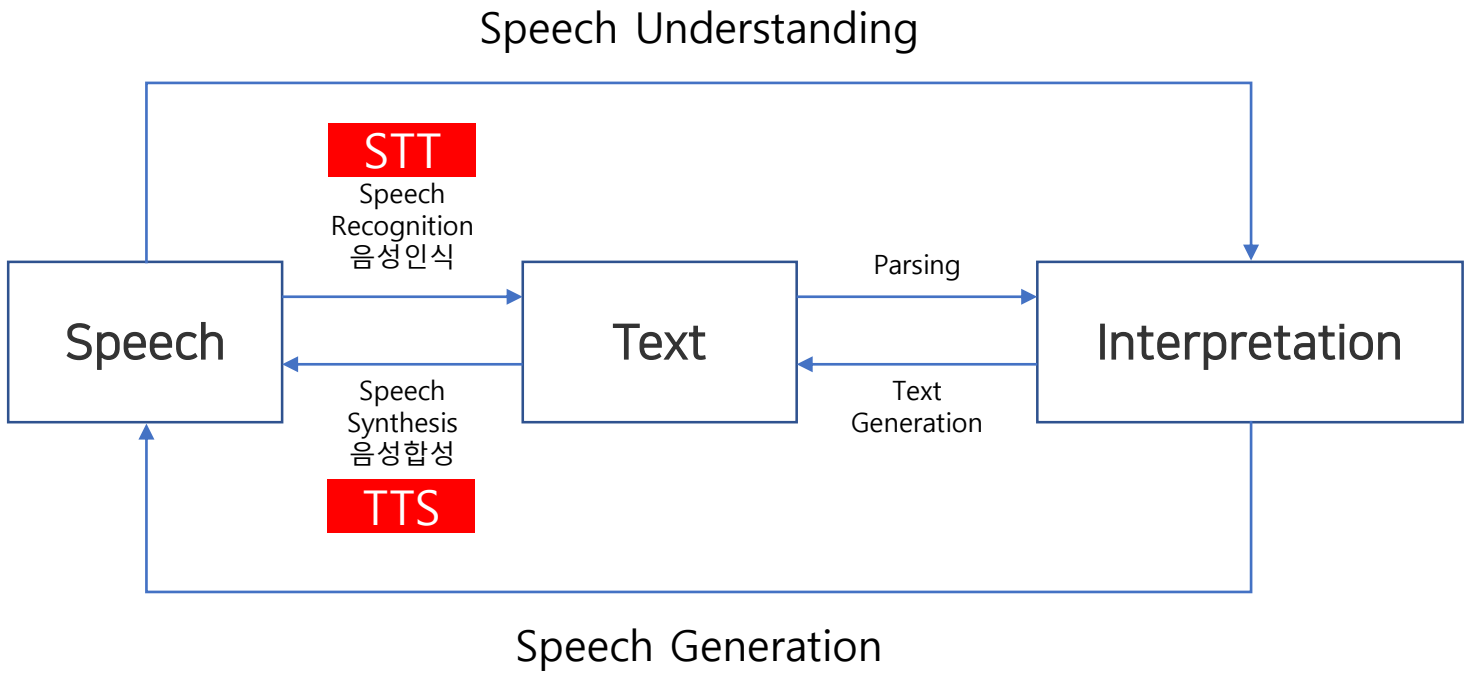
투빅스 9기 김유리안나

서론

Unit 01 | 음성/언어 분야 AI 기술 개요 및 동향

NLP의 다양한 분야 → 음성 분야 기술 개요

정보 검색
정보 추출
단어 분류
구문 분석
문장/문서 분류
감정 분석
의미역 결정
음성 인식
음성 합성
기계 번역
대화 처리
자동 대화 시스템(QA)
...



Unit 01 | 음성/언어 분야 AI 기술 개요 및 동향

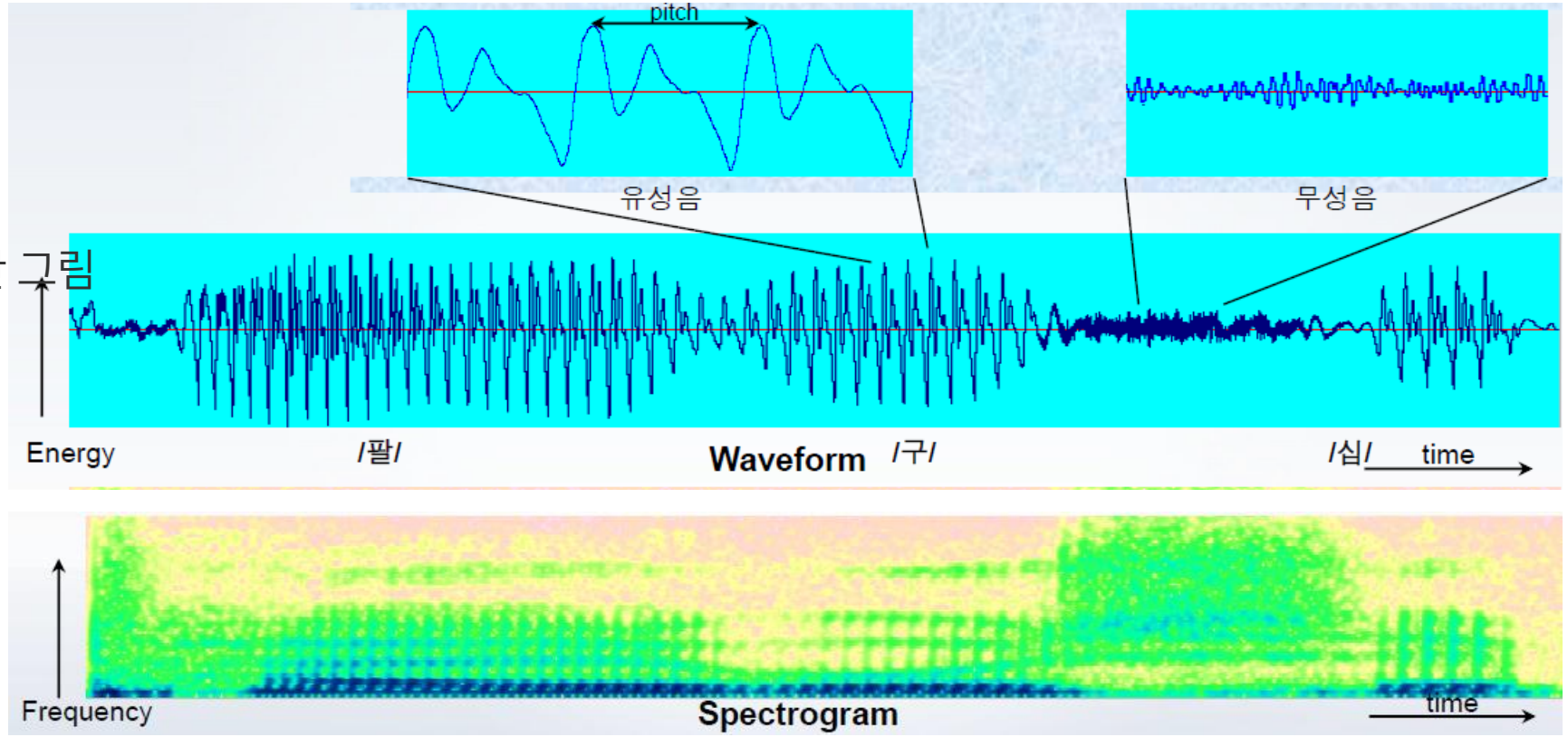
들어가기 전에…

1) 음성의 구조

- Waveform

- Spectrogram

-> 음성을 숫자로 표현하기 위한 그림

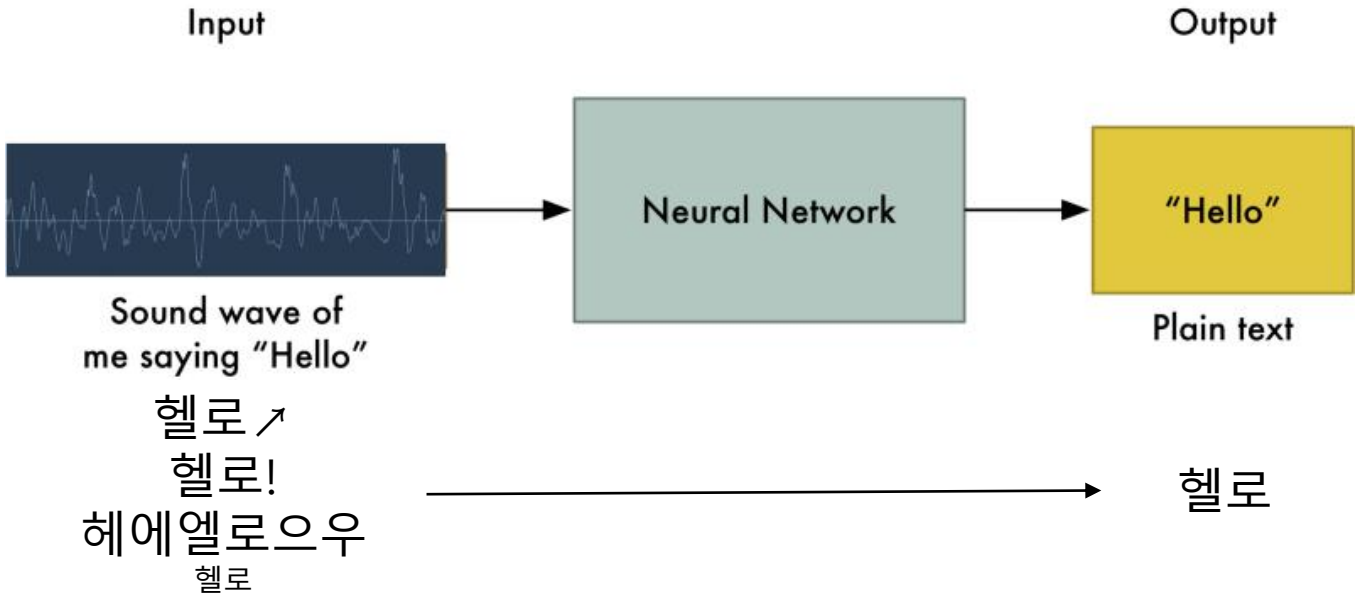


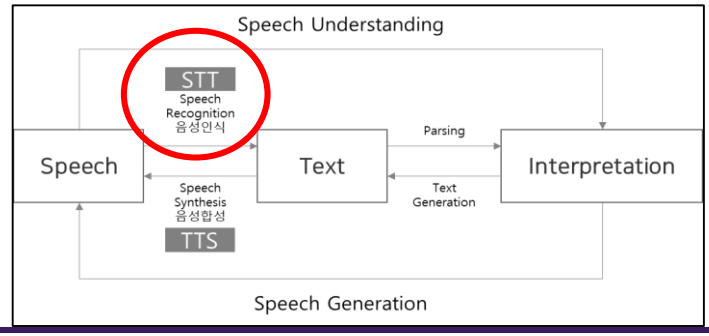
Unit 01 | 음성/언어 분야 AI 기술 개요 및 동향

1. 음성인식 (STT(speech-to-text), Speech Recognition)

1) 음성인식의 어려움

- 동일한 화자/다양한 변이 : 음의 높낮이, 발성 속도, 주변 잡음의 영향
- 동일한 단어/다양한 화자 : 음색, 발음, 강세 등
- 다양한 문맥/다양한 발성 : 음운 변화 (자음접변, 구개음화, 경음화 등)



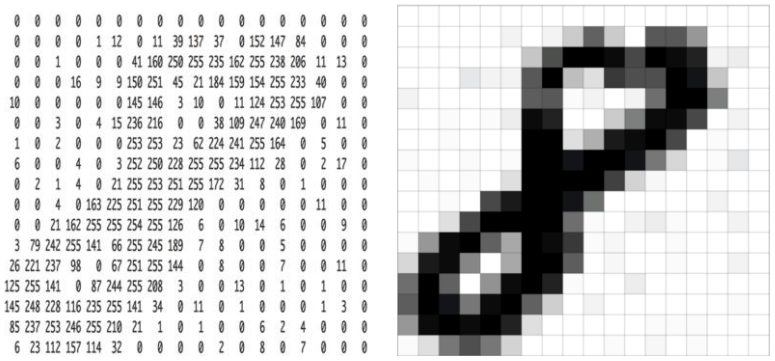


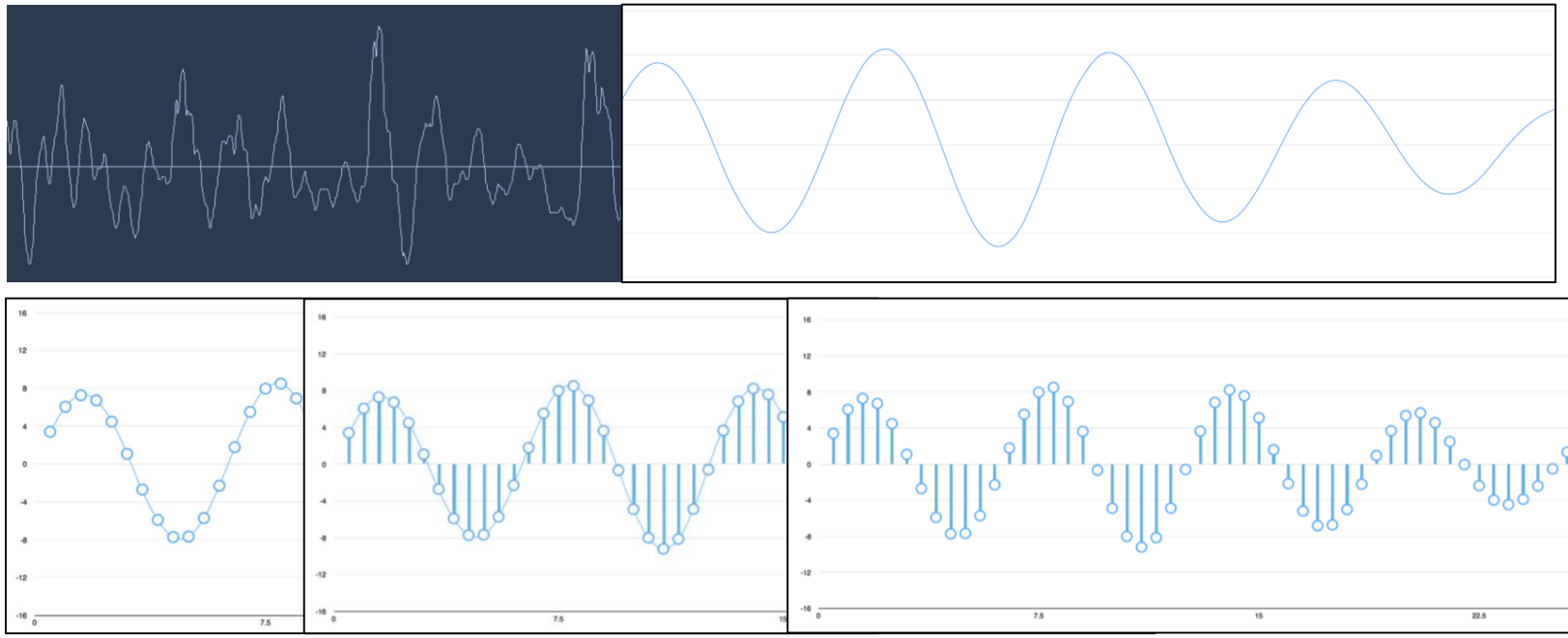
Unit 01 | 음성/언어 분야 AI 기술 개요 및 동향

1. 음성인식 (STT(speech-to-text), Speech Recognition)

2) 샘플링

- 16khz (초당 16,000번 추출)의 샘플링 속도로도 인간의 음성 주파수 범위를 커버





이미지를 숫자로 바꾸듯

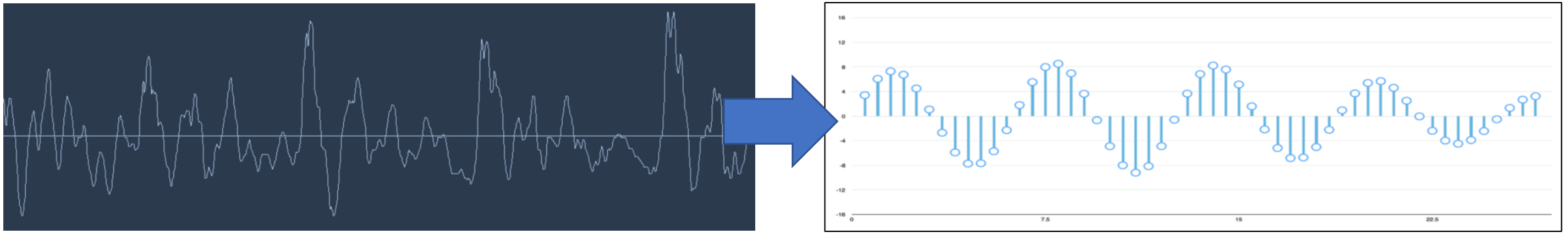
음파의 진폭을 숫자로 바꿈

```
[-1274, -1252, -1160, -986, -792, -692, -614, -429, -286, -134,
, -397, -212, 193, 114, -17, -110, 128, 261, 198, 390, 461, 772,
 4820, 4353, 3611, 2740, 2004, 1349, 1178, 1085, 901, 301, -262,
1648, -970, -364, 13, 260, 494, 788, 1011, 938, 717, 507, 323, 3
```

Unit 01 | 음성/언어 분야 AI 기술 개요 및 동향

1. 음성인식 (STT(speech-to-text), Speech Recognition)

2) 샘플링
- 16khz (초당 16,000번 추출)의 샘플링 속도로도 인간의 음성 주파수 범위를 커버



샘플링 과정에서 원본과 차이가 있지만 Nyquist–Shannon sampling 정리 덕분에,
가장 높은 주파수의 최소 두 배 빠르게 샘플을 추출한다면,
간격이 생긴 샘플로부터 원래의 음파를 수학적으로 재구성해 사용 가능

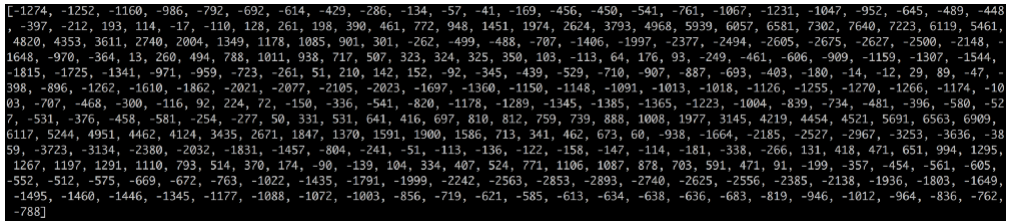
=> 우리가 아는 고해상도 음악은 샘플링이 매우 높게 추출된 음성. 굳이 엄청 고해상도 음성이 아니어도 괜찮다는 의미

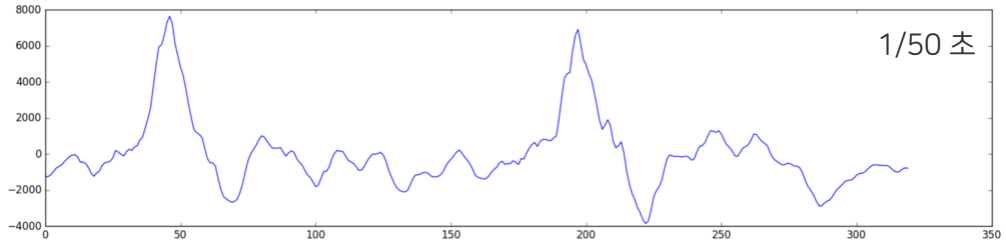
Unit 01 | 음성/언어 분야 AI 기술 개요 및 동향

1. 음성인식 (STT(speech-to-text), Speech Recognition)

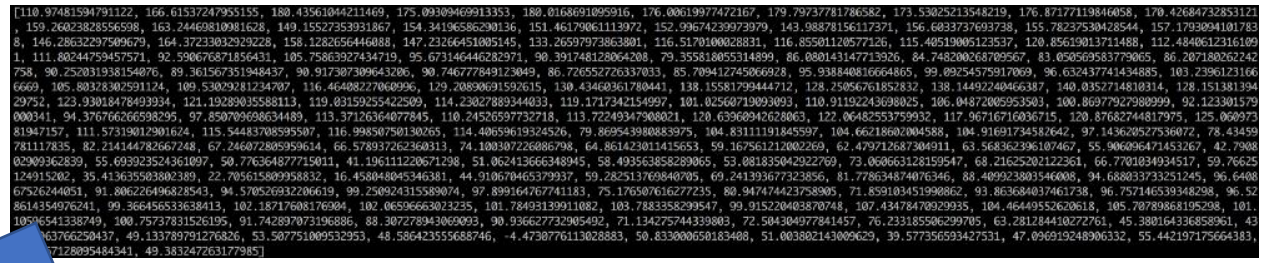
3) 샘플링된 데이터 전처리

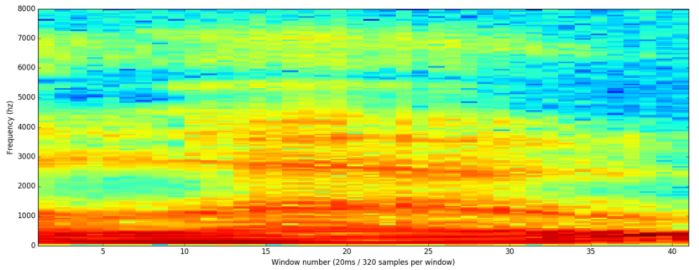
- 샘플링한 오디오를 20 밀리 초 길이로 잘라 그룹화





푸리에 변환 (Fourier transform)을 통해
복잡한 음파를 단순한 음파로 분해
각각의 음파에 얼마나 많은 에너지가 포함되어 있는지
합산하여 오른쪽 위 점수로 표현.





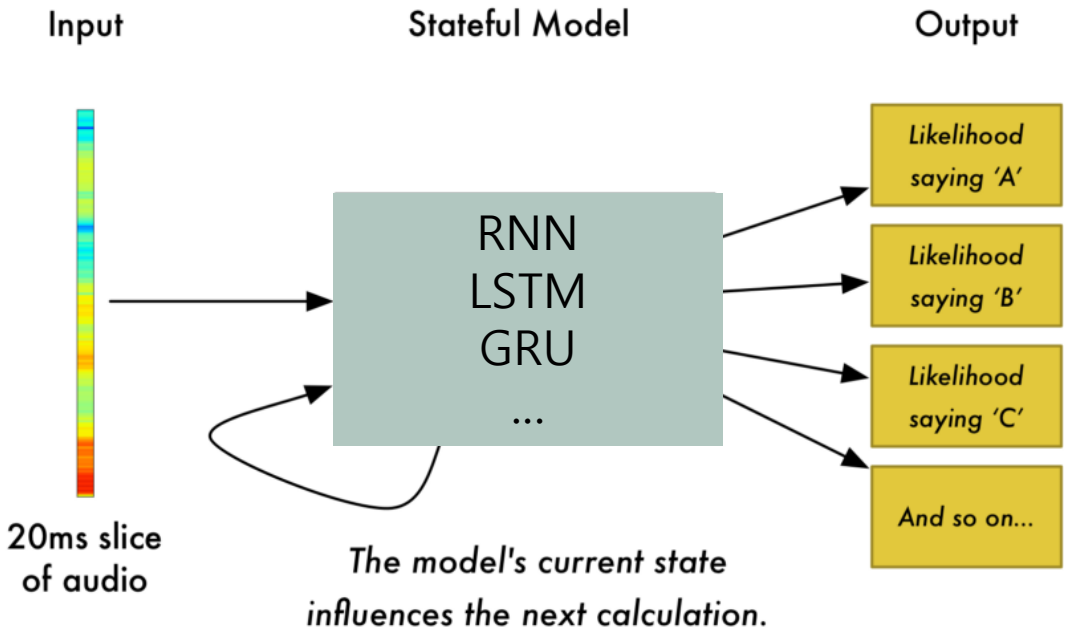
20 밀리초의 오디오 클립
각 50hz 대역별로 얼마나
많은 에너지가 있었는지 표현

-> ‘Hello’의 Spectrogram

Unit 01 | 음성/언어 분야 AI 기술 개요 및 동향

1. 음성인식 (STT(speech-to-text), Speech Recognition)

4) 전처리된 데이터로 text 생성
- 이후 이미지 처리와 동일. Spectrogram 이미지를 모델에 넣어 Output 생성

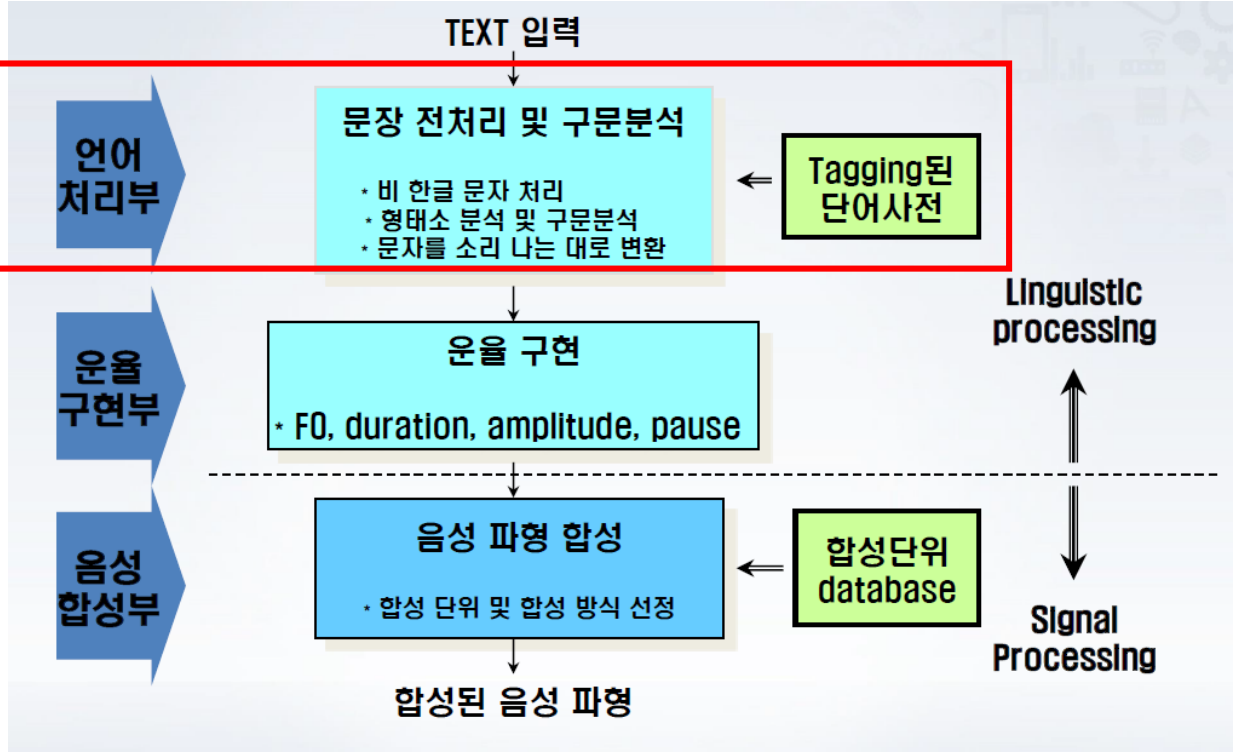


이후 반복문자 제거 후 공백 제거 => DB와 비교

Unit 01 | 음성/언어 분야 AI 기술 개요 및 동향

2. 음성합성 (TTS(text-to-speech), Speech Synthesis)

1) 언어 처리



- 비 한글 문자 처리

- 숫자: 123.45 -> 백이십삼점사오
- 특수 기호: % -> 퍼센트, & 앤드
- 영어 및 약자 : 사전 및 오토메타 이용

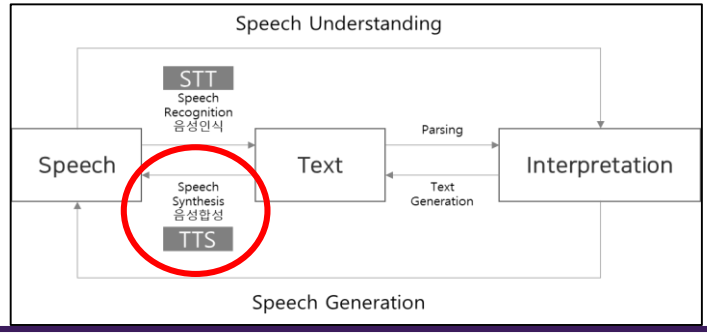
- 구문 분석

- 운율 구현을 위한 형태소 분석 및 구문 분석

- 문자를 소리나는 대로 변환

- 변환 Table 이용
- 불규칙 변환인 경우 형태소분석기 및 사전 이용

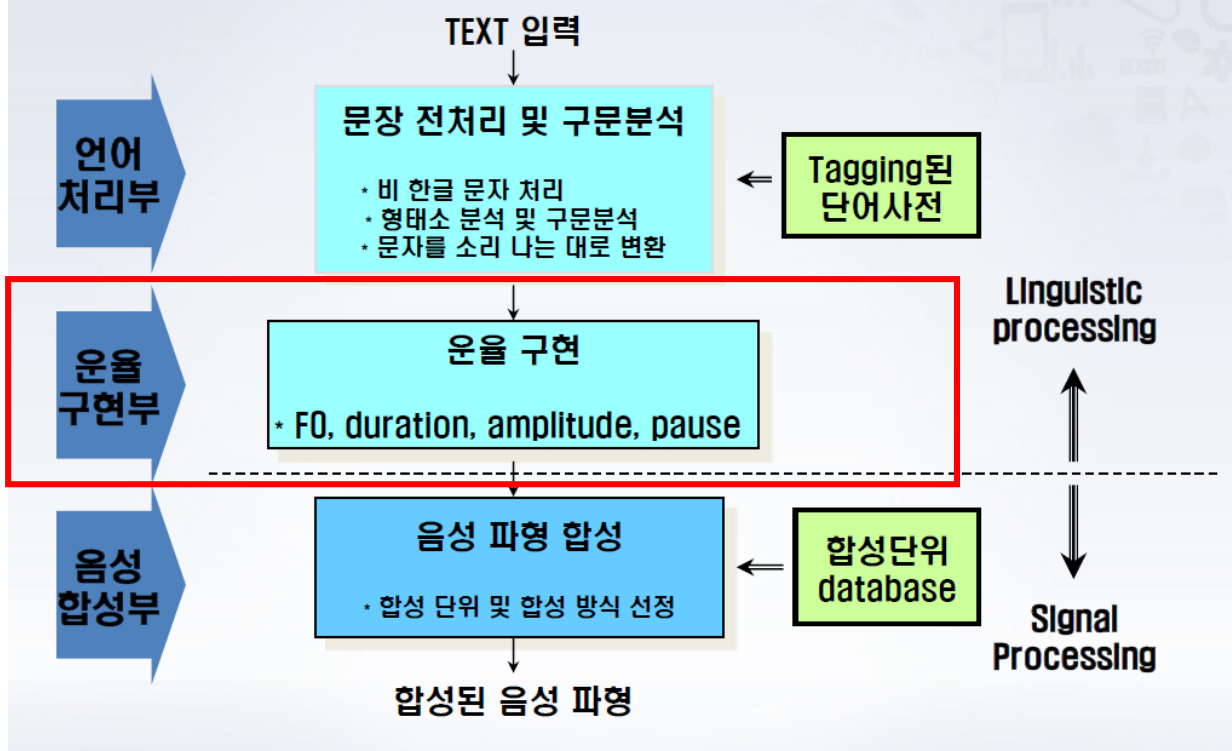
(예) 학교에 다녀와서 밥을 먹는다. -> 학꾜에 다녀와서 바블 멍는다.



Unit 01 | 음성/언어 분야 AI 기술 개요 및 동향

2. 음성합성 (TTS(text-to-speech), Speech Synthesis)

2) 운율 구현



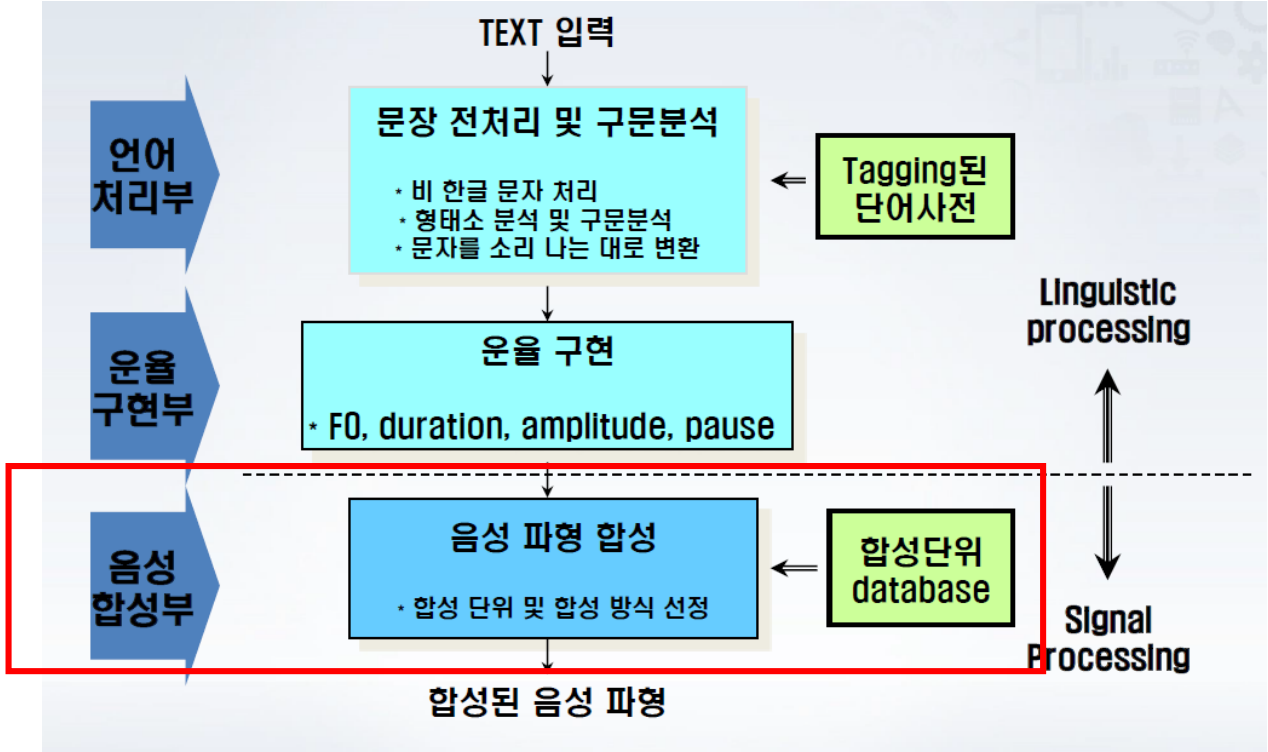
- 운율이란?
  - 발성시 나타나는 억양, 강세, 리듬 등의 특성
  - 기본 주파수 궤적(Intonation), 음소의 지속시간(duration), 음량(amplitude), 휴지구간 길이(pause length) 등에 의해 결정

- 운율에 영향을 미치는 요소는?
  - 말을 구성하는 음소들의 특성
  - 문장의 계층적 구조

Unit 01 | 음성/언어 분야 AI 기술 개요 및 동향

2. 음성합성 (TTS(text-to-speech), Speech Synthesis)

3) 음성 합성



- 1세대 : 고정 합성 단위 설계(Fixed Length Unit)
- 단어(word), 음절(syllable), 음소(phoneme)
- Formant, LPC 정보 등을 이용한 파형 생성
- (소용량 DB를 이용한 음성합성이 가능, 음성품질이 나쁨)

- 2세대 : 가변 합성 단위 연결 방식(Corpus based TTS)
- 음성파형을 최소한의 가공을 통해 연결
- (대용량 DB를 이용한 고품질 음성생성 가능)

- 3세대 : HTS (HMM based TTS system)
- 음성인식을 위한 음향모델링에 주로 사용하는 HMM 방식을 적용
- HMM을 이용하여 스펙트럼정보, 여기신호정보, 음성지속시간 등을 동시에 모델링하여 context dependent HMM을 생성하며, 이를 이용하여 음성을 합성
- 적절한 크기의 데이터베이스를 이용한 고품질 음성합성이 가능

Unit 01 | 음성/언어 분야 AI 기술 개요 및 동향

2. 음성합성 (TTS(text-to-speech), Speech Synthesis)

4) 음성 합성의 난제

- Emotional TTS

- 합성음에 감정을 구현하는 기술
- 음의 높낮이, 세기, 음색 등의 변화가 심해 음질이 저하되며 제어하기가 어려움

- Voice Conversion

- 특정인의 목소리로 변환하는 기술
- 특정인의 목소리를 나타내기 위해서는 음색 뿐만 아니라 발음, 억양 등 복합적인 요소가 작용

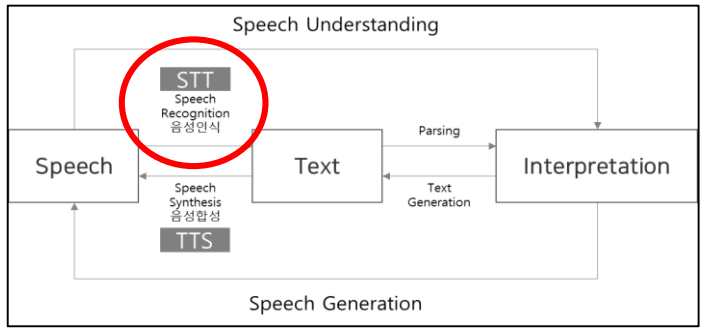
투빅스 정규 세미나

투빅스 9기 김유리안나

본론

음성인식

Unit 01 | 음성인식/합성 분야 기초 논문 소개 및 설명



(기존)
1) GMM(gaussian mixture model)으로 음소 모델링

2) HMM(Hidden markov model)으로 이들의 연속적 음성변화를 포착

=> 변화무상한 인간의 음성을 이해하기에 부족

(최근)
1) ~~GMM~~ -> DBN(deep belief network)같은 비지도학습 모델로 대체해 성능 개선

2) ~~HMM~~ -> RNN으로 대체시켜 End-to-End 학습을 달성했다.

대표논문: Speech recognition with deep recurent neural networks
End-to-end attention-based large vocabulary speech recognition

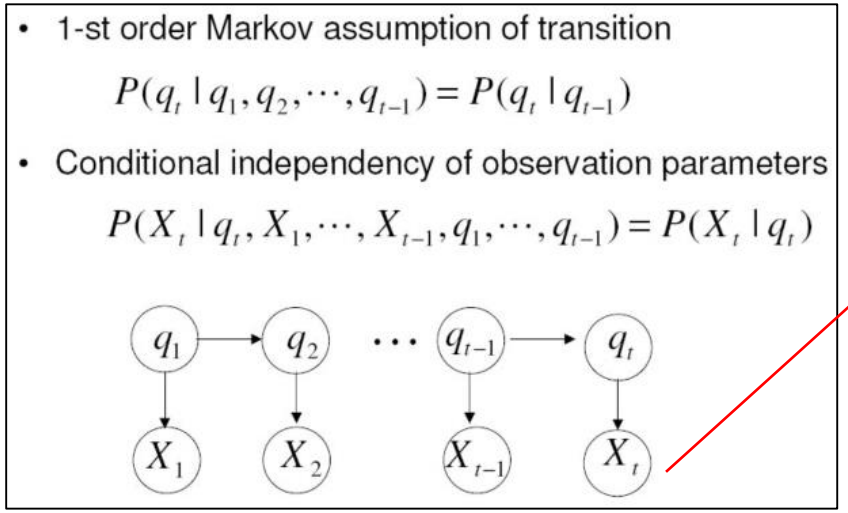
음성인식

Unit 01 | 음성인식/합성 분야 기초 논문 소개 및 설명

(기존)
1) GMM(gaussian mixture model)으로 음소 모델링

2) HMM(Hidden markov model)으로 이들의 연속적 음성변화(dynamics)를 포착

=> 변화무상한 인간의 음성을 이해하기에 부족



Gaussian Mixture Model (GMM)
각 state에서, 특징 벡터의 관측확률 분포를,
Gaussian들의 weighted summation으로 모델링

=> 수식 복잡

$$\sum_{j}^{k} \pi_j N(x_i|\mu_j, \Sigma_j)$$

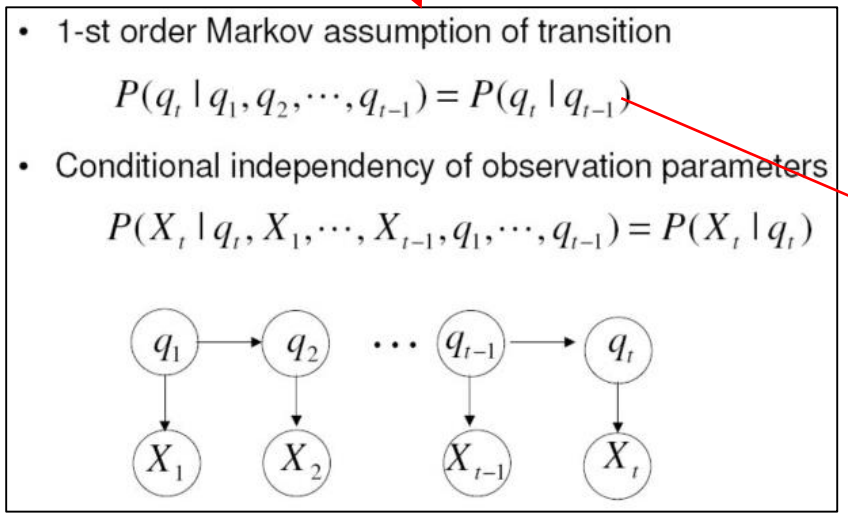
음성인식

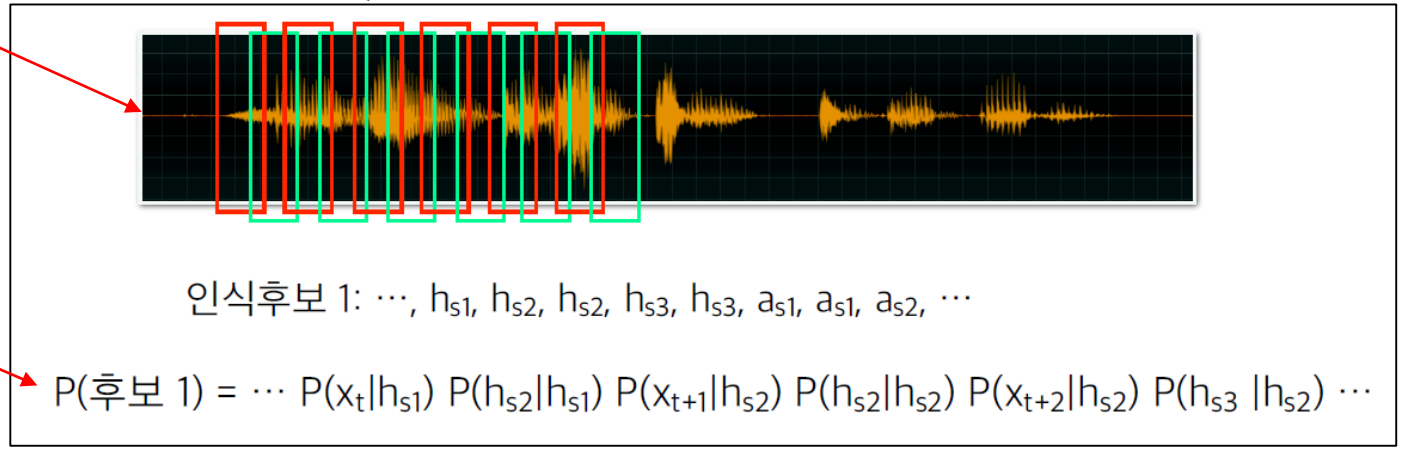
Unit 01 | 음성인식/합성 분야 기초 논문 소개 및 설명

(기존)
1) GMM(gaussian mixture model)으로 음소 모델링

2) HMM(Hidden markov model)으로 이들의 연속적 음성변화(dynamics)를 포착

=> 변화무상한 인간의 음성을 이해하기에 부족





음성신호의 dynamics는 HMM으로 모델링

각 state에서의 관측확률은 GMM으로 모델링

음성인식

Unit 01 | 음성인식/합성 분야 기초 논문 소개 및 설명

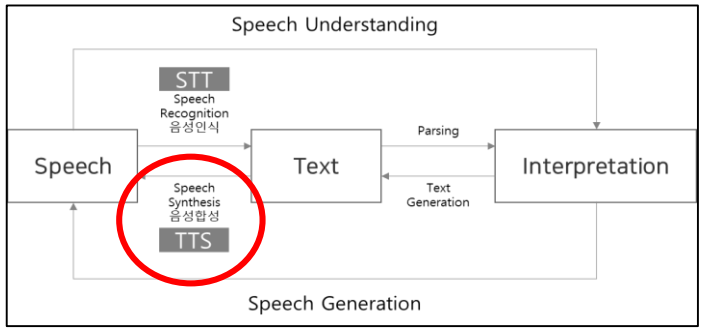
(최근)
1) ~~GMM~~ -> DBN(deep belief network)같은 비지도학습 모델로 대체해 성능 개선

2) ~~HMM~~ -> RNN/LSTM으로 대체시켜 End-to-End **학습**을 달성했다.

기존 hybrid 모델(DNN-HMM, LSTM-HMM 등)에서 HMM을 제거
전체 데이터를 모델로 한번에 학습 가능

음성합성

Unit 01 | 음성인식/합성 분야 기초 논문 소개 및 설명



- 대표적인 기본 논문/기술

1) 구글 딥마인드의 WaveNet
- 샘플을 순차적으로 생성하여 시간이 다소 걸림
- 고품질의 음성을 생성한다.
→ 기본 논문, 여기서 파생된 논문 다수: Fast WaveNet, Parallel WaveNet, WaveRNN

2) 구글의 TACOTRON
- 쉼표의 위치에 따라 문장 읽는 높낮이/속도가 달라져서, 같은 철자도 다르게 발음 (read의 현재/과거형 발음 구분 등) → Attention을 통해 가능해짐.
- 기존의 WaveNet보다 더 자연스러워 짐

음성합성

Unit 02 | WaveNet: A Generative Model for Raw Audio

1) 구글 딥마인드의 WaveNet

2 WAVENET

In this paper we introduce a new generative model operating directly on the raw audio waveform. The joint probability of a waveform $\mathbf{x} = \{x_1, \ldots, x_T\}$ is factorised as a product of conditional probabilities as follows:

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

Each audio sample x_t is therefore conditioned on the samples at all previous timesteps.

Similarly to PixelCNNs (van den Oord et al., 2016a;b), the conditional probability distribution is modelled by a stack of convolutional layers. There are no pooling layers in the network, and the output of the model has the same time dimensionality as the input. The model outputs a categorical distribution over the next value x_t with a softmax layer and it is optimized to maximize the log-likelihood of the data w.r.t. the parameters. Because log-likelihoods are tractable, we tune hyper-parameters on a validation set and can easily measure if the model is overfitting or underfitting.

p(x1)
p(x1,x2)=p(x1)p(x2|x1)
p(x1,x2,x3)=p(x1,x2)p(x3|x1,x2)
=p(x1)p(x2|x1)p(x3|x1,x2)
…

이 식을
Stack of Convolution Layer로 표현

음성합성

Unit 02 | WaveNet: A Generative Model for Raw Audio

1) 구글 딥마인드의 WaveNet

2.1 DILATED CAUSAL CONVOLUTIONS

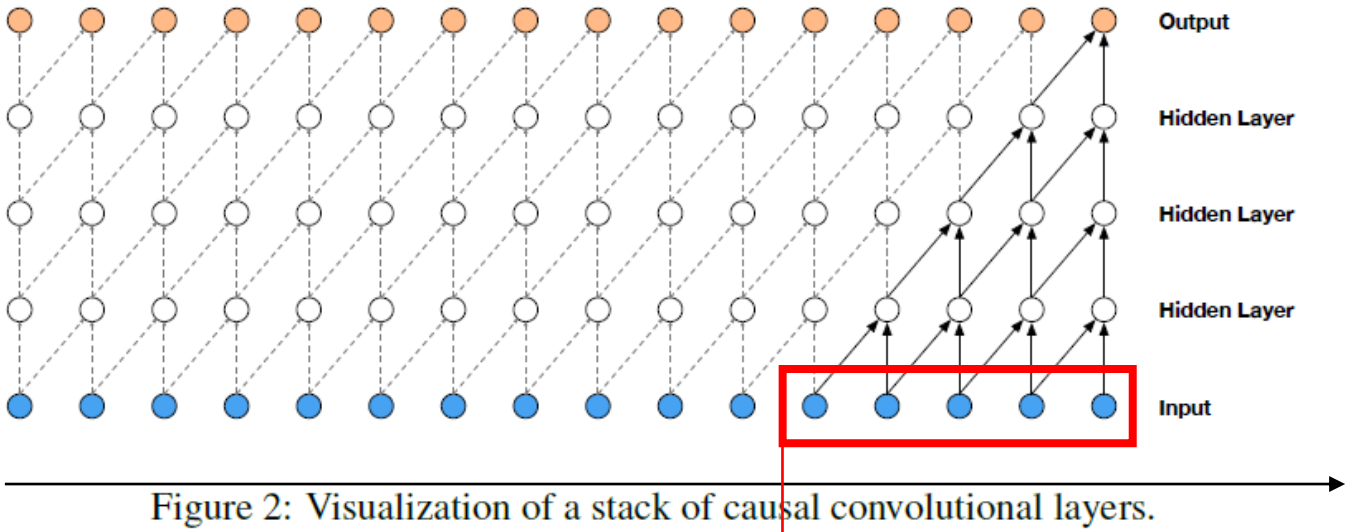


Figure 2: Visualization of a stack of causal convolutional layers.

Causal?
- 데이터를 만들기 위해 과거값만 가져와서 만들겠다는 개념
- Uncausal이면 오른쪽 데이터도 사용가능

하나의 아웃풋 생성을 위한
CNN의 Convoultion layer와 똑같

Receptive field= #layers + filter length -

Unit 02 | WaveNet: A Generative Model for Raw Audio

음성합성 1) 구글 딥마인드의 WaveNet

2.1 DILATED CAUSAL CONVOLUTIONS

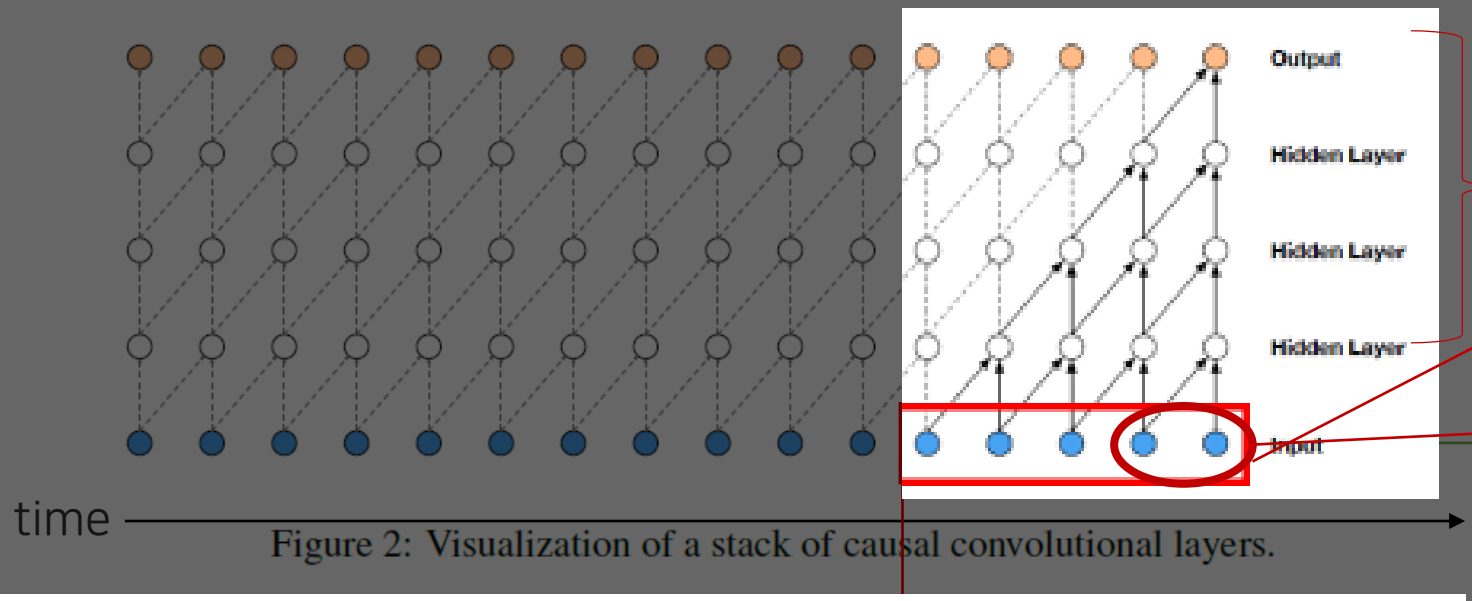


Figure 2: Visualization of a stack of causal convolutional layers.

Causal?
- 데이터를 만들기 위해 과거값만 가져와서 만들겠다는 개념
- Uncausal이면 오른쪽 데이터도 사용가능

#layers : 4

Receptive field : 5

Filter length : 2

식: Receptive field= #layers + Filter length -1
여기서 Receptive field가 5밖에 안되는데,
이것을 키우기 위해서는 너무 많은 Layer가 필요함.
와 똑같

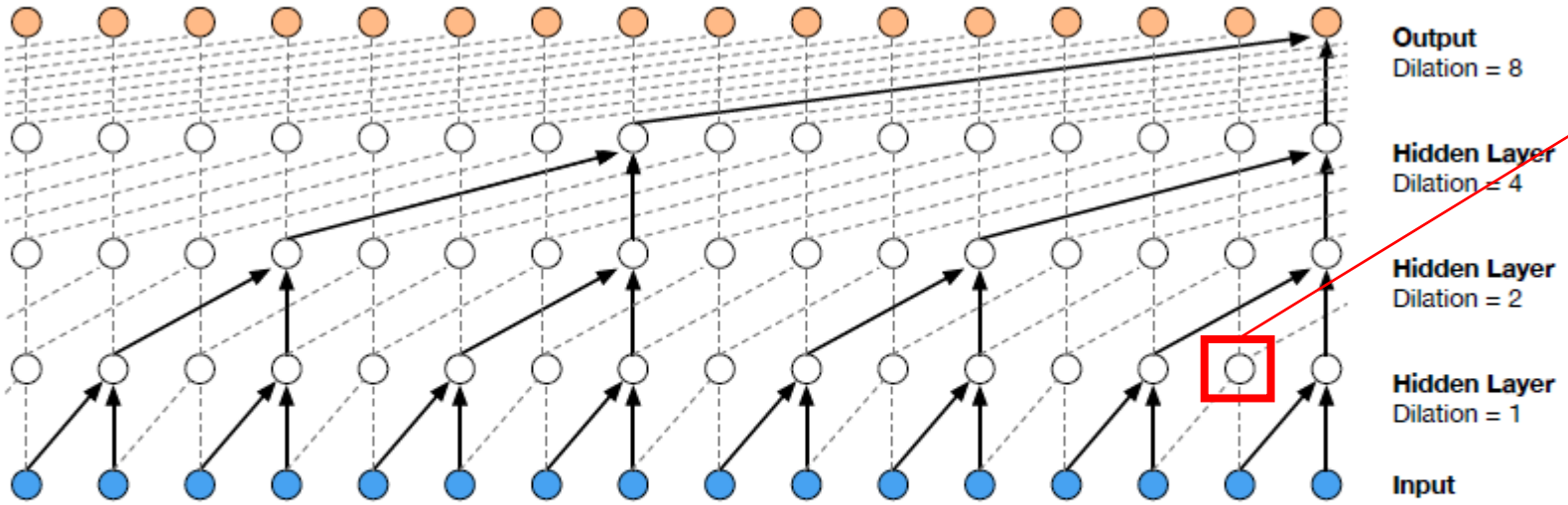
Because models with causal convolutions do not have recurrent connections, they are typically faster to train than RNNs, especially when applied to very long sequences. One of the problems of causal convolutions is that they require many layers, or large filters to increase the receptive field. For

음성합성

Unit 02 | WaveNet: A Generative Model for Raw Audio

1) 구글 딥마인드의 WaveNet

그래서 등장한 것이
Stack of Dilated Causal convolutional layers



Layer가 올라갈 때마다
중간의 값을 하나씩 안 쓰고
convolution 하는 것

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

Unit 02 | WaveNet: A Generative Model for Raw Audio

음성합성

1) 구글 딥마인드의 WaveNet

그래서 등장한 것이
Stack of Dilated Causal convolutional layers

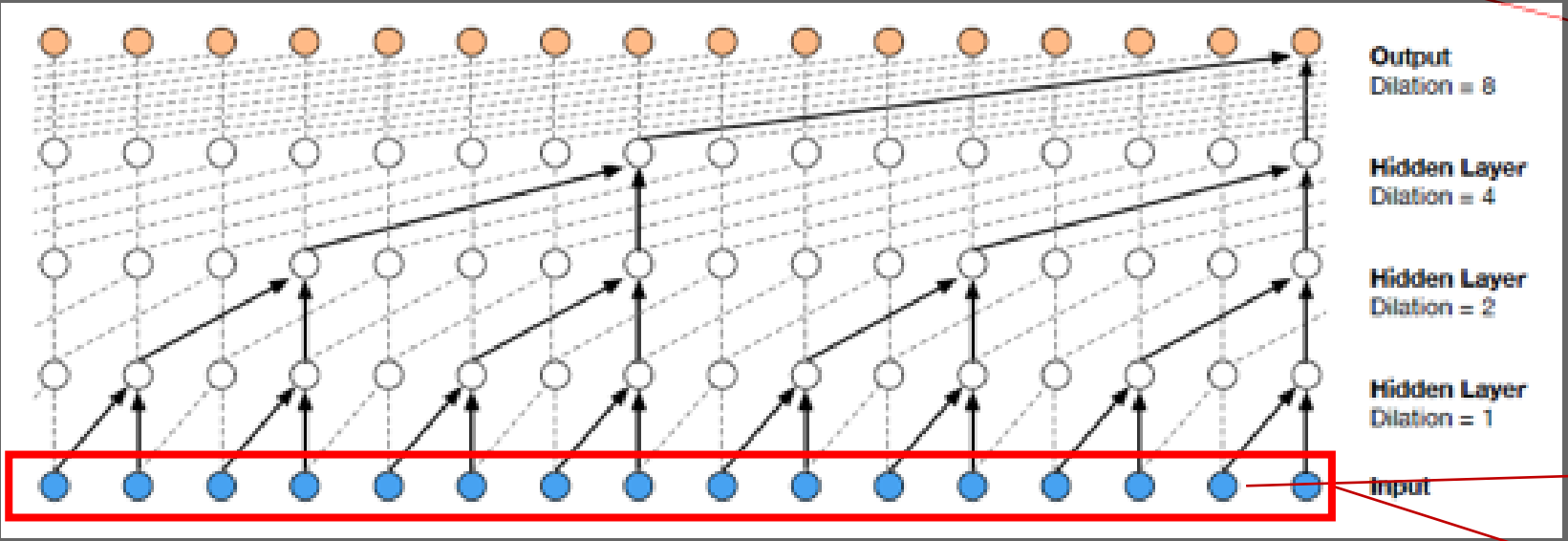


Figure 3: Visualization of a stack of *dilated* causal convolutional layers.

Layer가 올라갈 때마다
중간의 값을 하나씩 안 쓰고
convolution 하는 것

#layers : 4

Filter length : 2

Receptive field : 16

식: Receptive field= #layers + Filter length -1
Input 데이터 양이 많아짐☺

음성합성

Unit 02 | WaveNet: A Generative Model for Raw Audio

1) 구글 딥마인드의 WaveNet

Stacked dilated convolutions enable networks to have very large receptive fields with just a few layers, while preserving the input resolution throughout the network as well as computational efficiency. In this paper, the dilation is doubled for every layer up to a limit and then repeated: e.g.

$$1, 2, 4, \ldots, 512, 1, 2, 4, \ldots, 512, 1, 2, 4, \ldots, 512.$$

The intuition behind this configuration is two-fold. First, exponentially increasing the dilation factor results in exponential receptive field growth with depth (Yu & Koltun, 2016). For example each $1, 2, 4, \ldots, 512$ block has receptive field of size 1024, and can be seen as a more efficient and discriminative (non-linear) counterpart of a 1×1024 convolution. Second, stacking these blocks further increases the model capacity and the receptive field size.

그래서 이 논문은 총 2^0 ... 2^{10} 의 10개 layer를 3번 반복하여 30개의 layer 형성
=> 샘플 하나를 만들기 위해 과거 값을 많이 보는 것.

음성합성

Unit 02 | WaveNet: A Generative Model for Raw Audio

1) 구글 딥마인드의 WaveNet

이 논문에서 conditional probability를 modeling하는데 있어서, softmax distributions을 사용

Audio 신호는 대개 16bit로 Audio 정보를 softmax로 표현하려면 Sample마다 $2^{16} = 65536$ 개의 output이 필요

⇒ mu-law companding 기법을 사용.
⇒ 사람의 귀는 소리 크기가 작을 때는 작은 변화에 민감/ 소리 크기가 클 때는 큰 변화에도 둔감
⇒ quantization을 nonlinear하게 해줌.
⇒ 이렇게 하면 8bit($2^8 = 256$ outputs)로도 좋은 성능으로 encoding/decoding이 가능

음성합성

Unit 02 | WaveNet: A Generative Model for Raw Audio

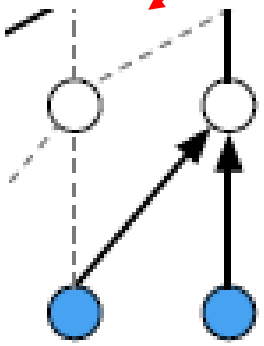
1) 구글 딥마인드의 WaveNet

2.3 GATED ACTIVATION UNITS

We use the same gated activation unit as used in the gated PixelCNN (van den Oord et al., 2016b):

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \sigma\left(W_{g,k} * \mathbf{x}\right) \quad (2)$$

where $*$ denotes a convolution operator, $\odot$ denotes an element-wise multiplication operator, $\sigma(\cdot)$ is a sigmoid function, k is the layer index, f and g denote filter and gate, respectively, and W is a learnable convolution filter. In our initial experiments, we observed that this non-linearity worked significantly better than the rectified linear activation function (Nair & Hinton, 2010) for modeling audio signals.



RNN/LSTM 당시 했던 gate units

-> 얼마나 통과시킬까를 결정.
-> 시그모이드로 0~1사이 값으로 반환

2.4 RESIDUAL AND SKIP CONNECTIONS

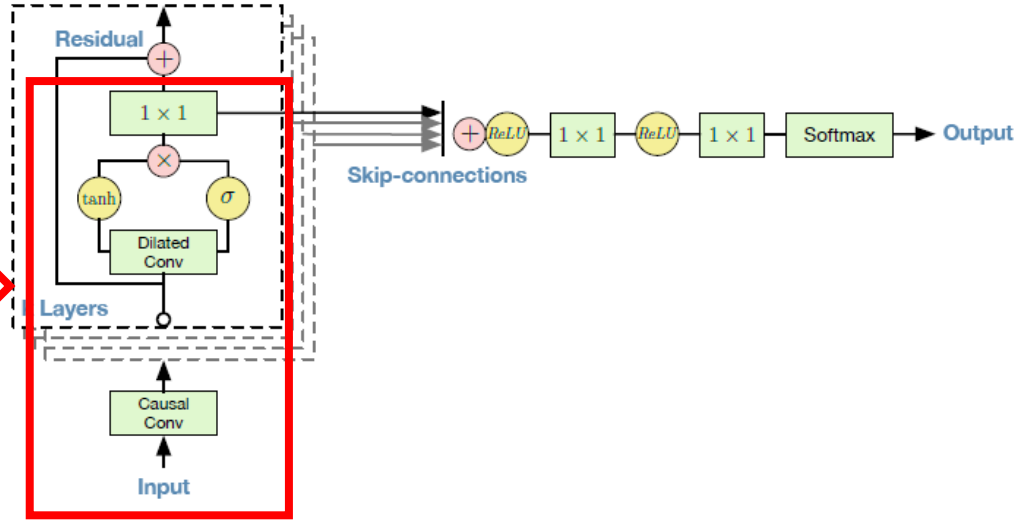


Figure 4: Overview of the residual block and the entire architecture.

1) 구글 딥마인드의 [WaveNet]

2.4 RESIDUAL AND SKIP CONNECTIONS

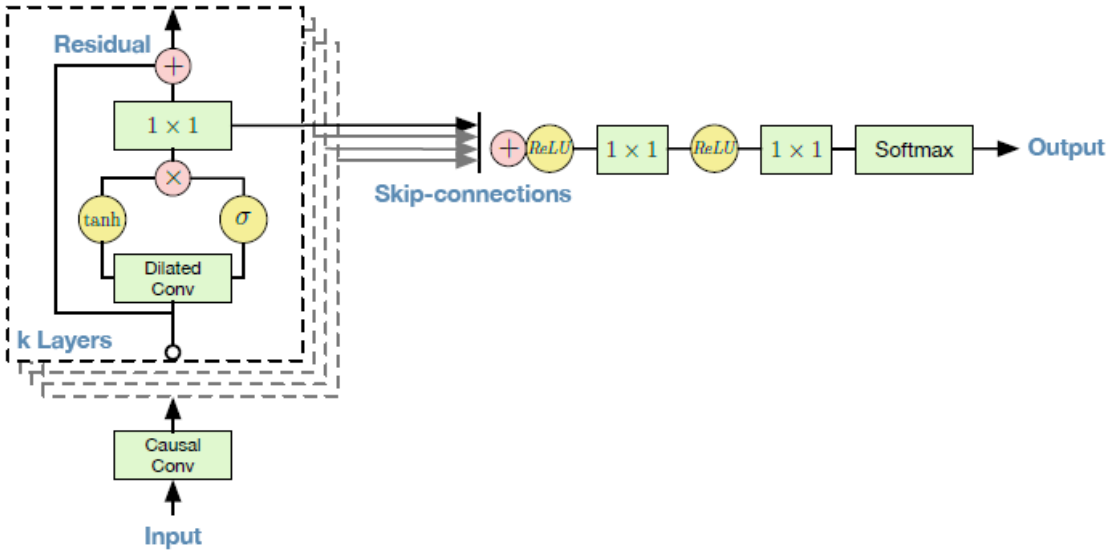
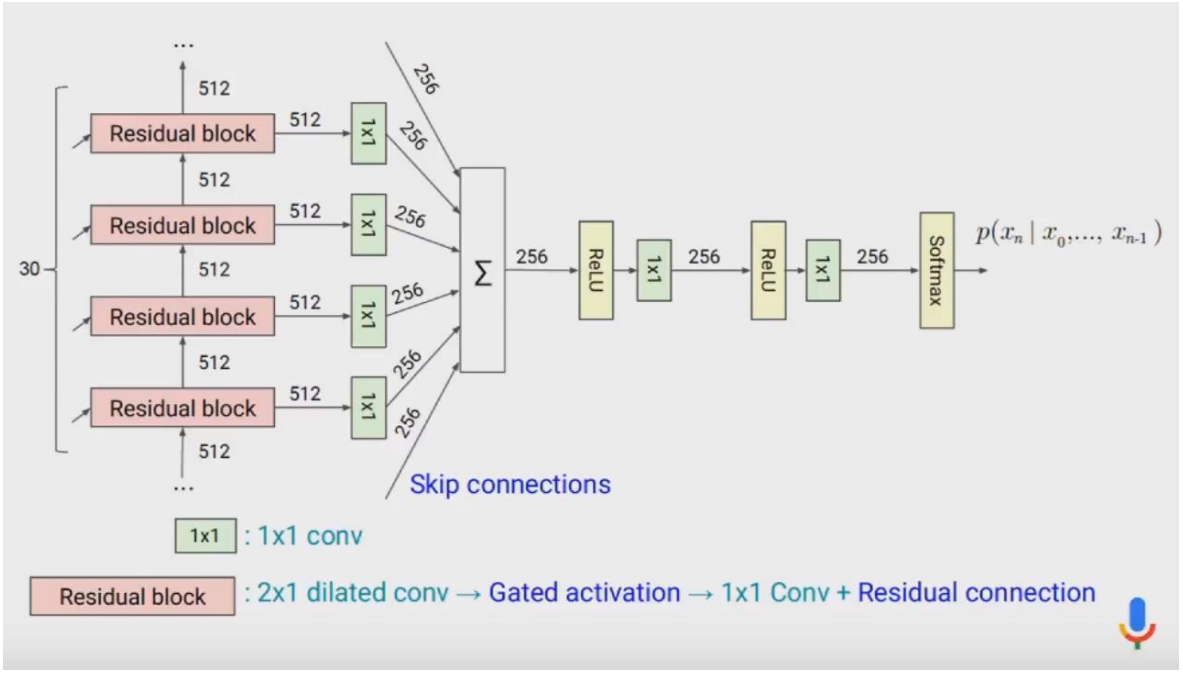


Figure 4: Overview of the residual block and the entire architecture.

논문 발표자료



논문 3저자 발표자료

음성합성

Unit 02 | WaveNet: A Generative Model for Raw Audio

1) 구글 딥마인드의 WaveNet

2.5 CONDITIONAL WAVENETS

Given an additional input $\mathbf{h}$, WaveNets can model the conditional distribution $p(\mathbf{x} \mid \mathbf{h})$ of the audio given this input. Eq. (1) now becomes

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h}). \quad (3)$$

By conditioning the model on other input variables, we can guide WaveNet's generation to produce audio with the required characteristics. For example, in a multi-speaker setting we can choose the speaker by feeding the speaker identity to the model as an extra input. Similarly, for TTS we need to feed information about the text as an extra input.

We condition the model on other inputs in two different ways: global conditioning and local conditioning. Global conditioning is characterised by a single latent representation $\mathbf{h}$ that influences the output distribution across all timesteps, e.g. a speaker embedding in a TTS model. The activation function from Eq. (2) now becomes:

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

마지막으로 WaveNets에
특정한 조건을 주고 오디오를 생성!

예시1)
화자가 여러명 일 때, 화자의 특성을 넣기
(남녀 대화시 남자 목소리만 넣기)

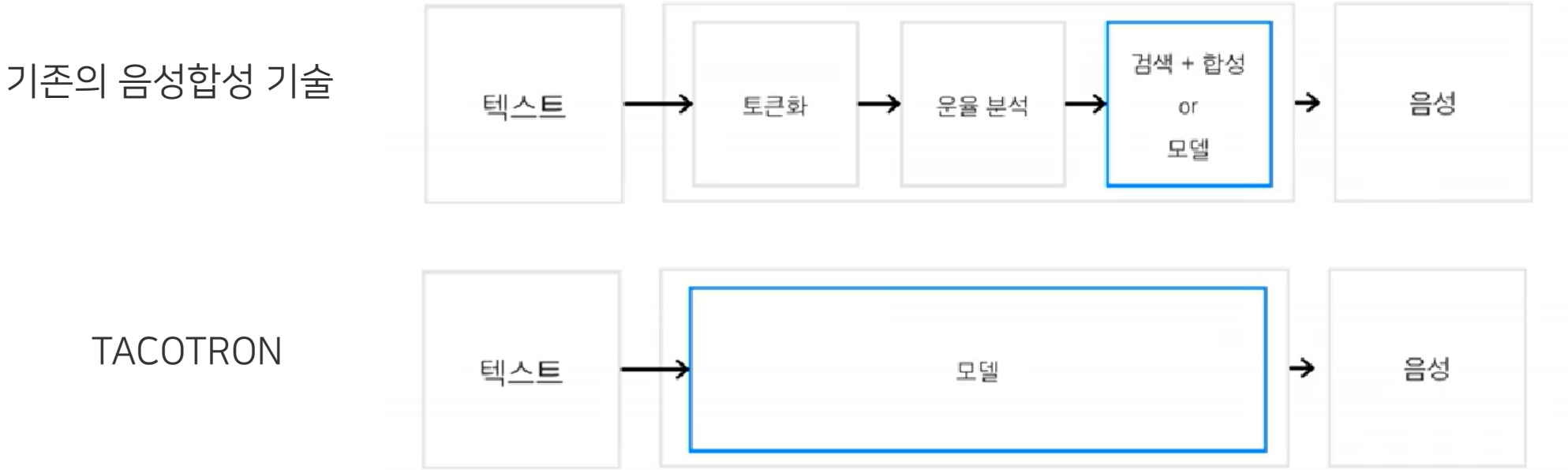
예시2) TTS : text

음성합성

Unit 03 | TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS

2) 구글의 TACOTRON

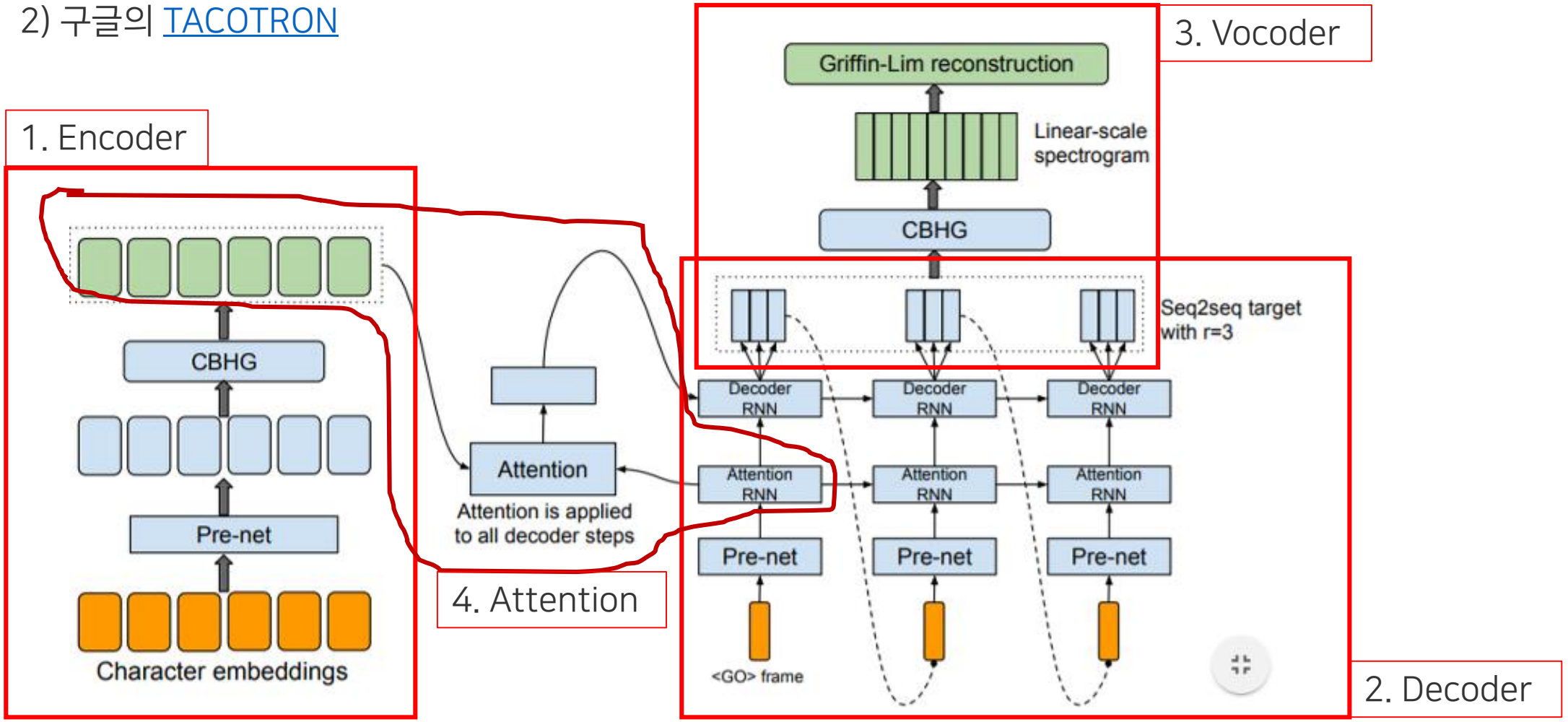
- 쉼표의 위치에 따라 문장 읽는 높낮이/속도가 달라져서, 같은 철자도 다르게 발음 (read의 현재/과거형 발음 구분 등)
- 기존의 WaveNet보다 더 자연스러워 짐



음성합성

Unit 03 | TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS

2) 구글의 TACOTRON



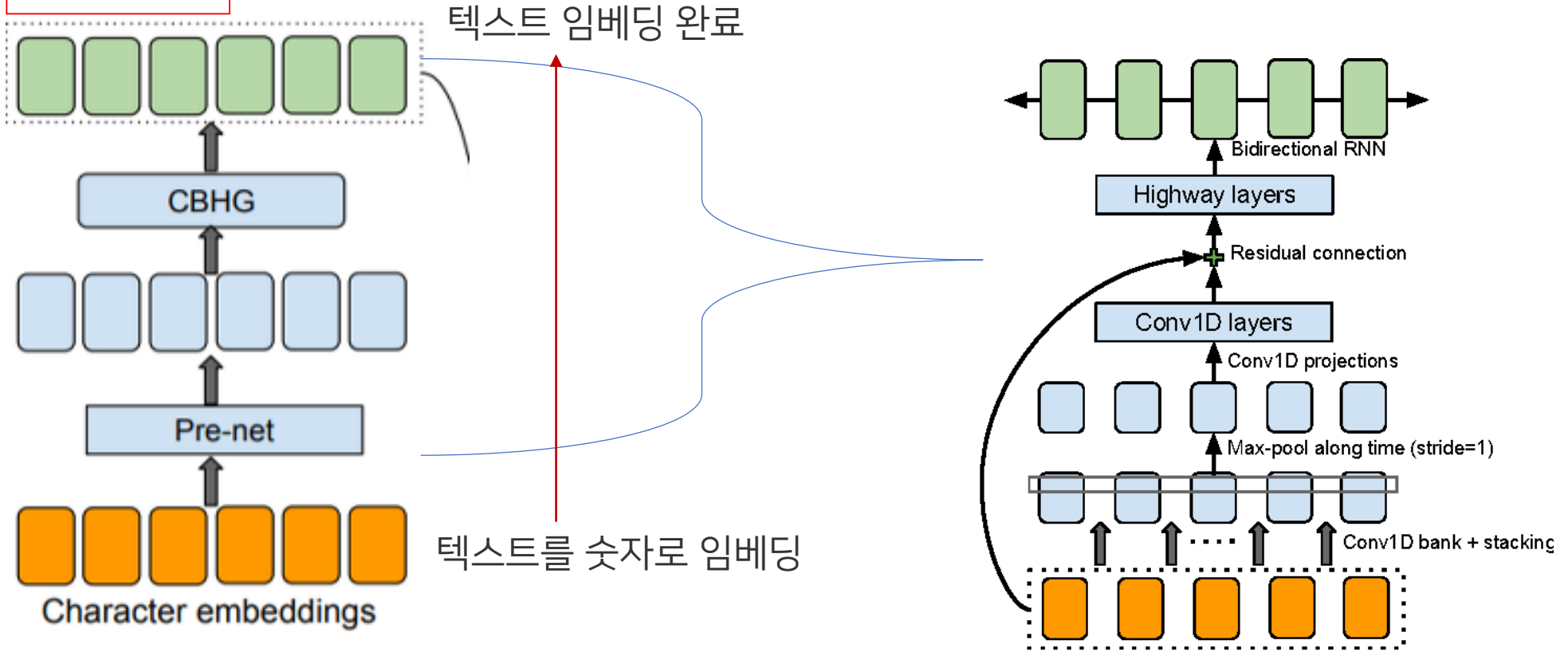
음성합성

Unit 03 | TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS

CBHG: **C**onvolutoin **B**ank + **H**ighway network + Bidirectional **G**RU

2) 구글의 TACOTRON

1. Encoder

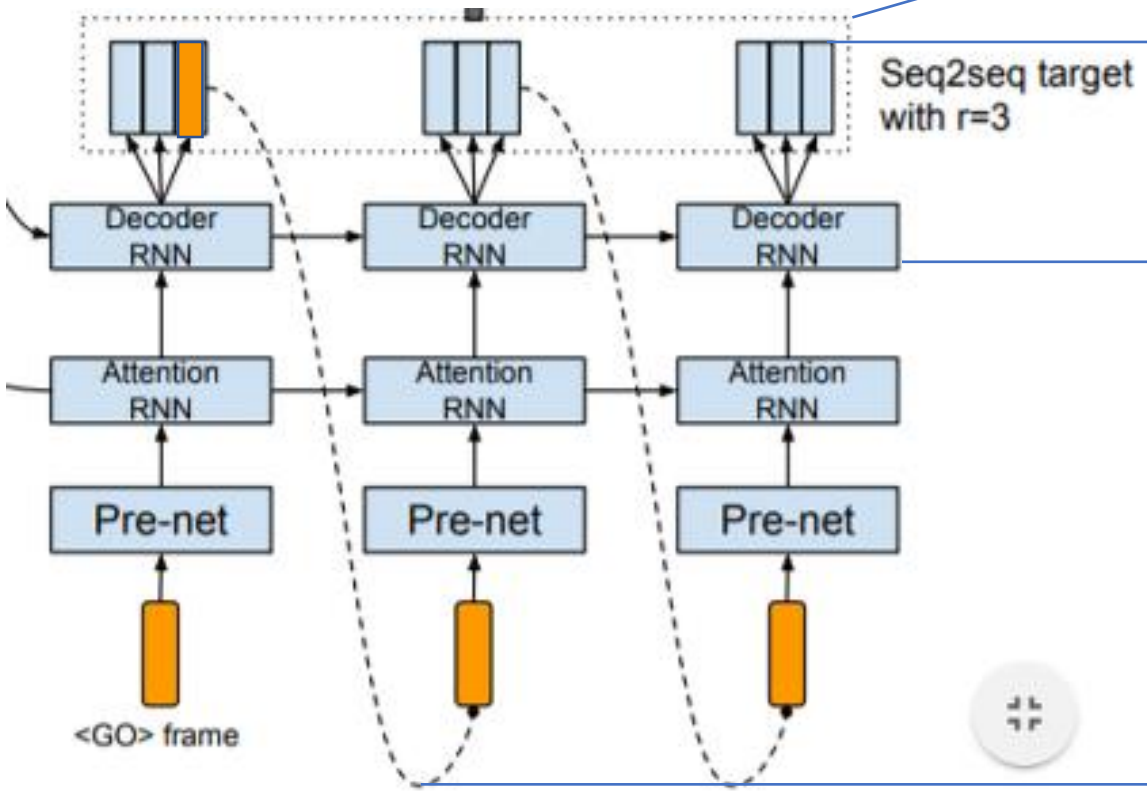


음성합성

Unit 03 | TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS

2) 구글의 TACOTRON

2. Decoder



5) 음성되기 직전의 스펙트로그램들!

2) 쭉 올라와서 N개의 스펙트로그램을 예측

4) 문장이 끝날때까지 반복

1) RNN과 똑같!

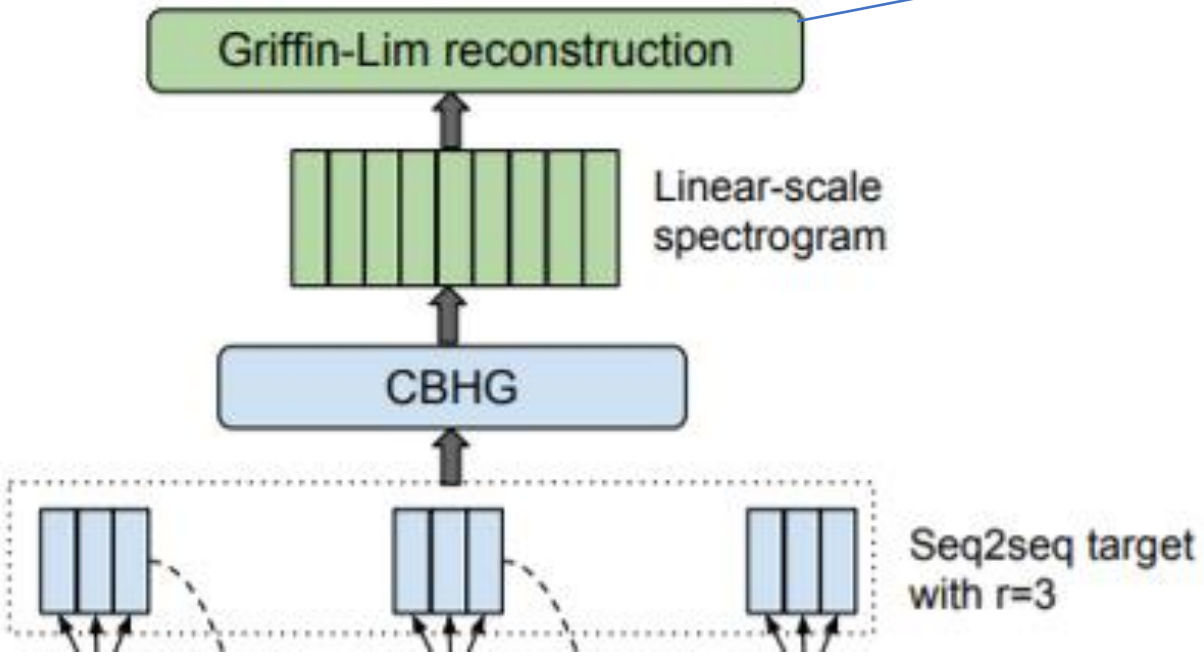
3) 마지막 스펙트로그램이
다음 입력값

음성합성

Unit 03 | TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS

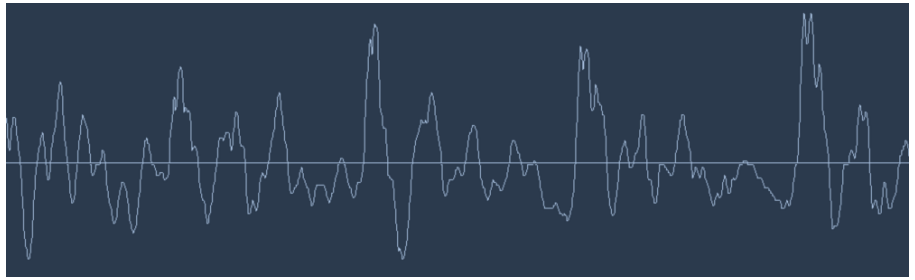
2) 구글의 TACOTRON

3. Vocoder



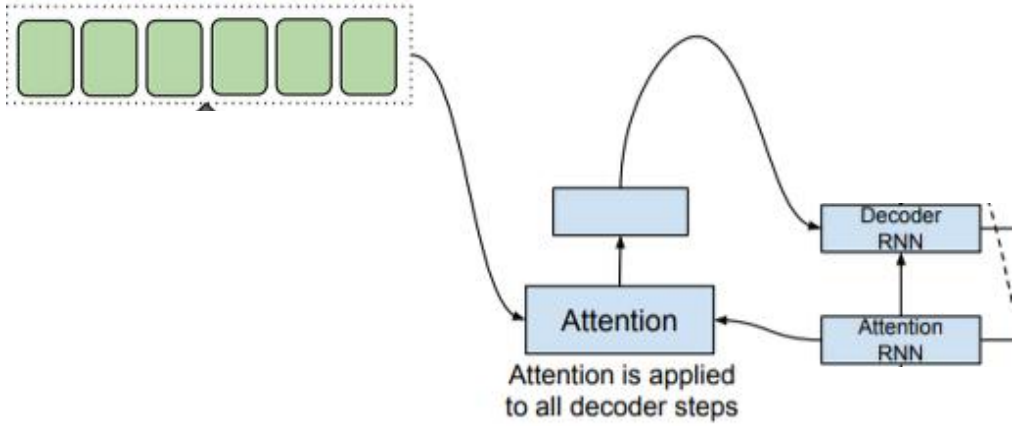
1) 스펙트로그램을 음성으로 만들어주는 알고리즘

2) 음성이 형성됨



2) 구글의 TACOTRON

4. Attention



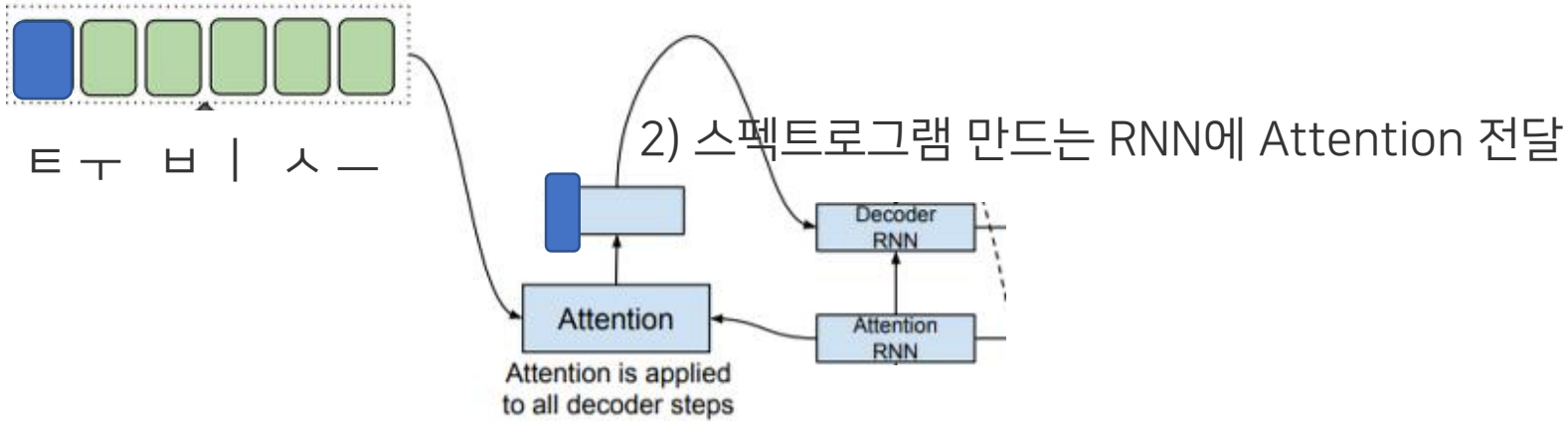
1) 어떤 글자에 집중할 것인지 계산하고

음성합성

Unit 03 | TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS

2) 구글의 TACOTRON

4. Attention



2) 스펙트로그램 만드는 RNN에 Attention 전달

1) 어떤 글자에 집중할 것인지 계산하고
(예를 들어 "ㅌ"에 집중)

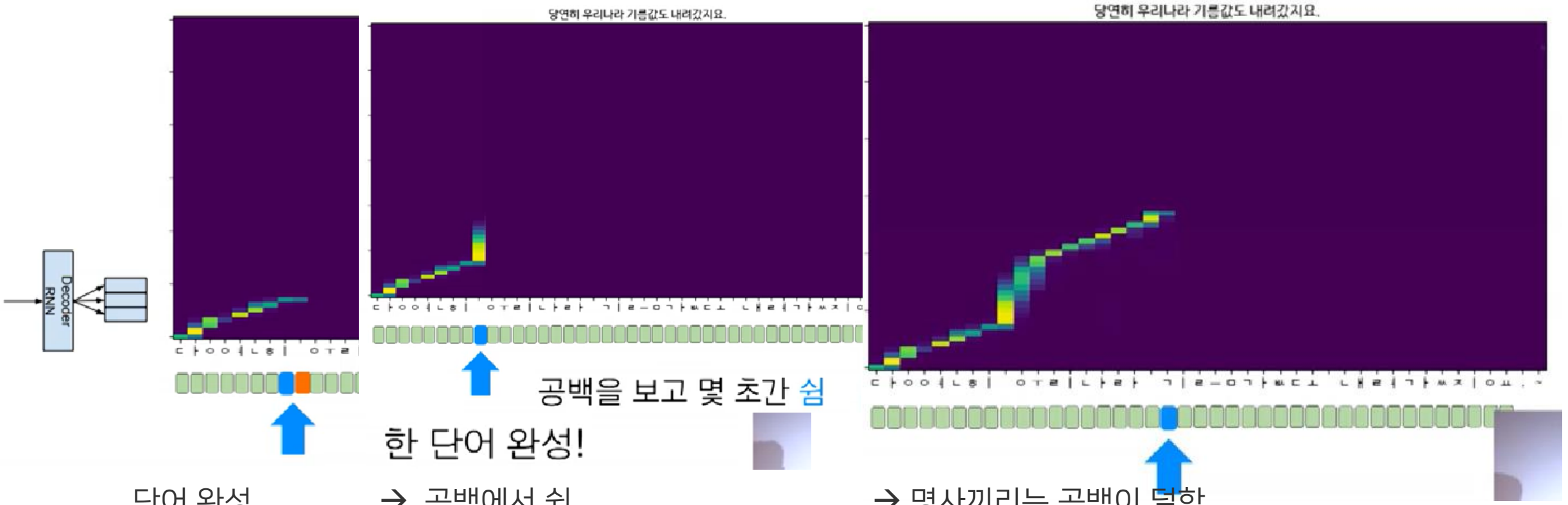
음성합성

Unit 03 | TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS

2) 구글의 TACOTRON

4. Attention

예시: 당연히 우리나라 기름값도 내려갔지요.



단어 완성 → 공백에서 쉼 → 명사끼리는 공백이 덜함

투빅스 정규 세미나

투빅스 9기 김유리안나

결론

Unit 01 | 심화 논문 소개 및 추가 논문 추천

심화 논문 소개

- Fast Wavenet Generation Algorithm (2016.11) -> Fast WaveNet
- SEGAN: Speech Enhancement Generative Adversarial Network (2017.06)
- Audio Super-Resolution using Neural Nets (2017.08)
- Synthesizing Audio with Generative Adversarial Networks(2018.02)
- Efficient Neural Audio Synthesis (2018.02) -> WaveRNN

추가 논문 추천

- Very Deep Convolutional Networks for Text Classification (2017.01)
- Neural audio synthesis of musical notes with WaveNet autoencoders (2017.04)
- Deep Voice 1,2,3
- Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions (2018.02) -> TACOTRON 2

출처

논문
https://arxiv.org/pdf/1508.04395.pdf
https://arxiv.org/pdf/1703.10135.pdf
https://arxiv.org/pdf/1303.5778.pdf

강의자료
https://medium.com/@ageitgey/machine-learning-is-fun-part-6-how-to-do-speech-recognition-with-deep-learning-28293c162f7a
https://brunch.co.kr/@kakao-it/65
https://deepmind.com/blog/wavenet-generative-model-raw-audio/
https://tacademy.sktechx.com/live/player/onlineLectureDetail.action?seq=110
https://github.com/buriburisuri/speech-to-text-wavenet
https://tv.naver.com/v/2292650

감사합니다