CLOVA

# Voice synthesis and applications

송은우 / HDTS, NAVER CLOVA

CLOVA

# Introduction

Text-to-speech (TTS) 란 기계가 사람 처럼 **텍스트를 읽어주는** 기술입니다

CLOVA

# Introduction

Text-to-speech (TTS) 란 기계가 사람 처럼 **텍스트를 읽어주는** 기술입니다


딥러닝 TTS 기술은 크게 acoustic model 과 vocoding model 두가지 모듈로 구성됩니다

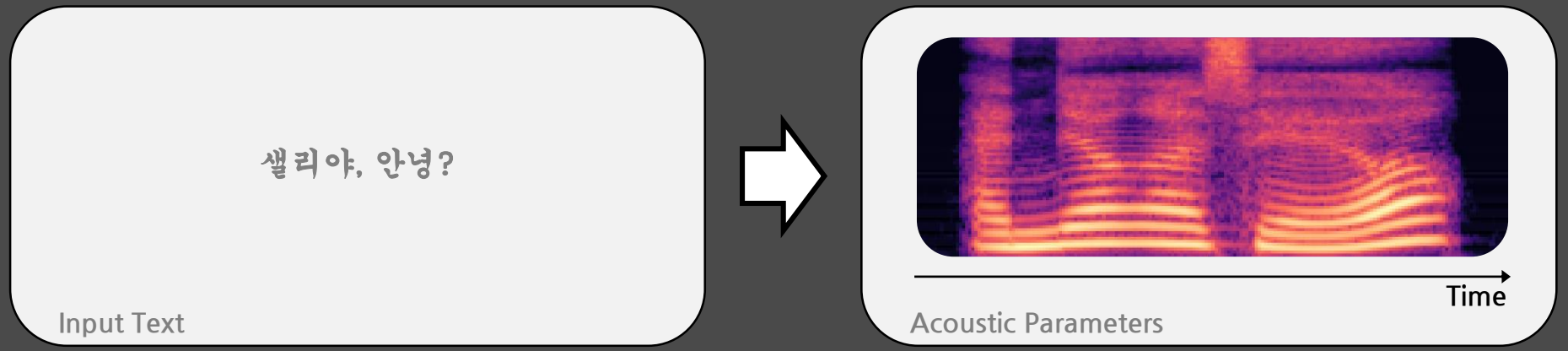
CLOVA

# Introduction

Text-to-speech (TTS) 란 기계가 사람 처럼 **텍스트를 읽어주는** 기술입니다

톤의 높낮이, 음색, 어조, 강세 등

**텍스트에서 Acoustic Parameter 를 추정**

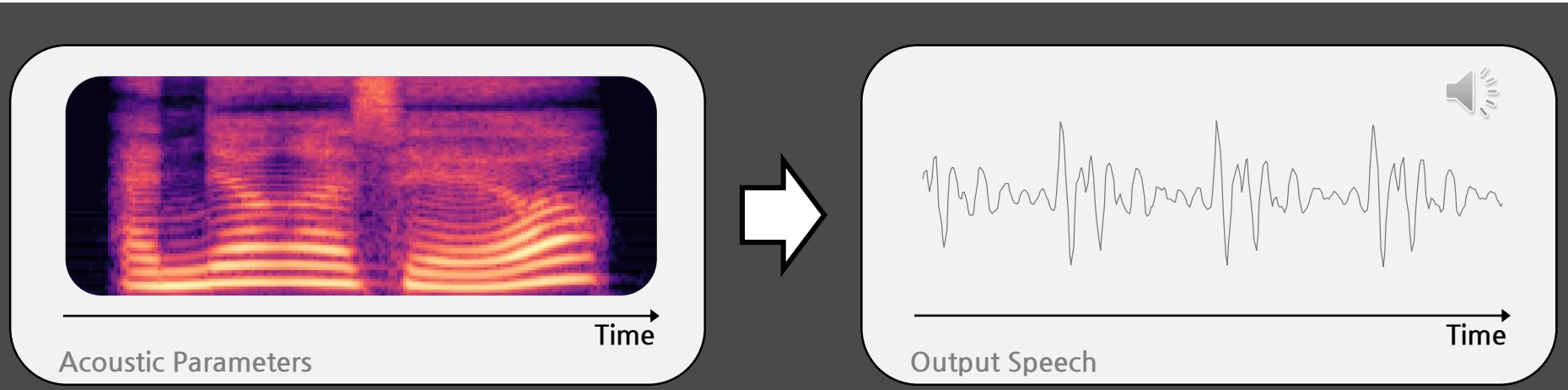



CLOVA

# Introduction

Text-to-speech (TTS) 란 기계가 사람 처럼 **텍스트를 읽어주는** 기술입니다

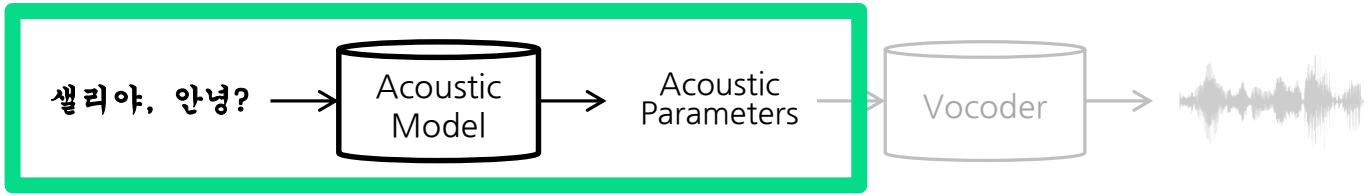
Acoustic Parameter 에서 음성 신호를 생성

CLOVA

# Introduction

Text-to-speech (TTS) 란 기계가 사람 처럼 **텍스트를 읽어주는** 기술입니다




본 발표에서는 TTS 엔진의 핵심 요소인

**Acoustic Model** 기술을 정리하고,

CLOVA

# Introduction

Text-to-speech (TTS) 란 기계가 사람 처럼 **텍스트를 읽어주는** 기술입니다

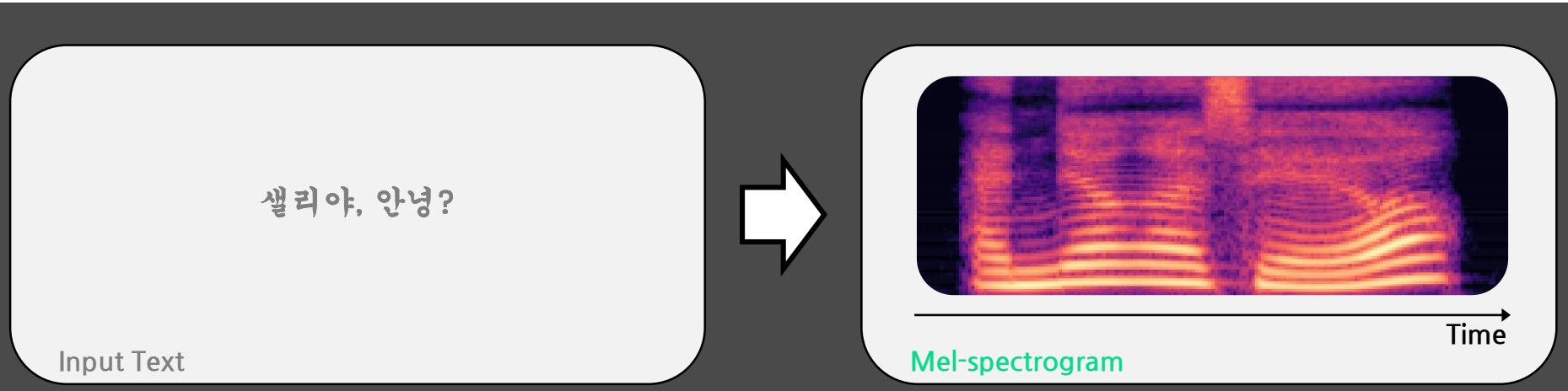

이어지는 발표에서는 TTS 엔진의 핵심 요소인

**Vocoder** 기술을 정리하고자 합니다.

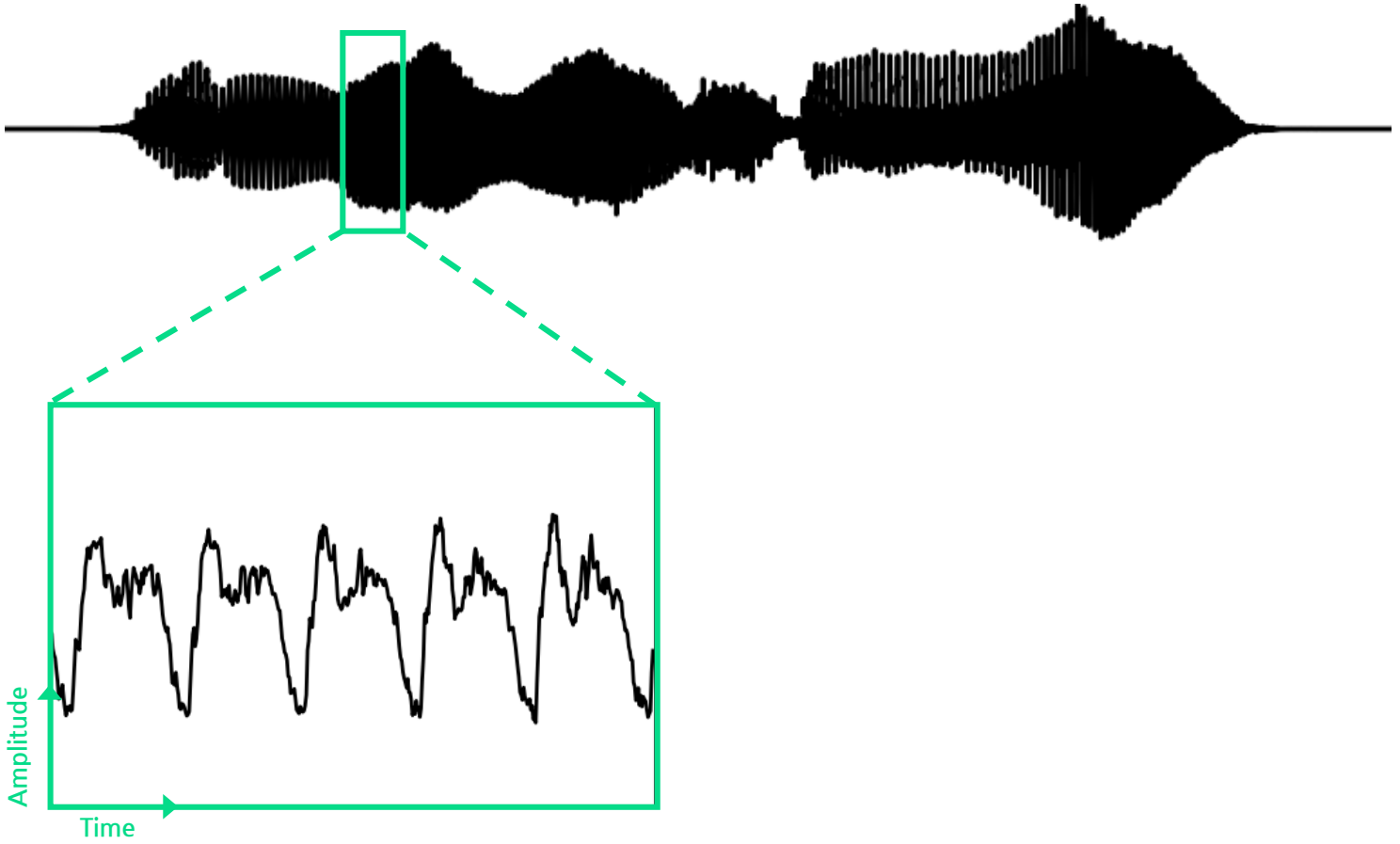
CLOVA

# Speech analysis

## Acoustic parameters: Mel-spectrogram

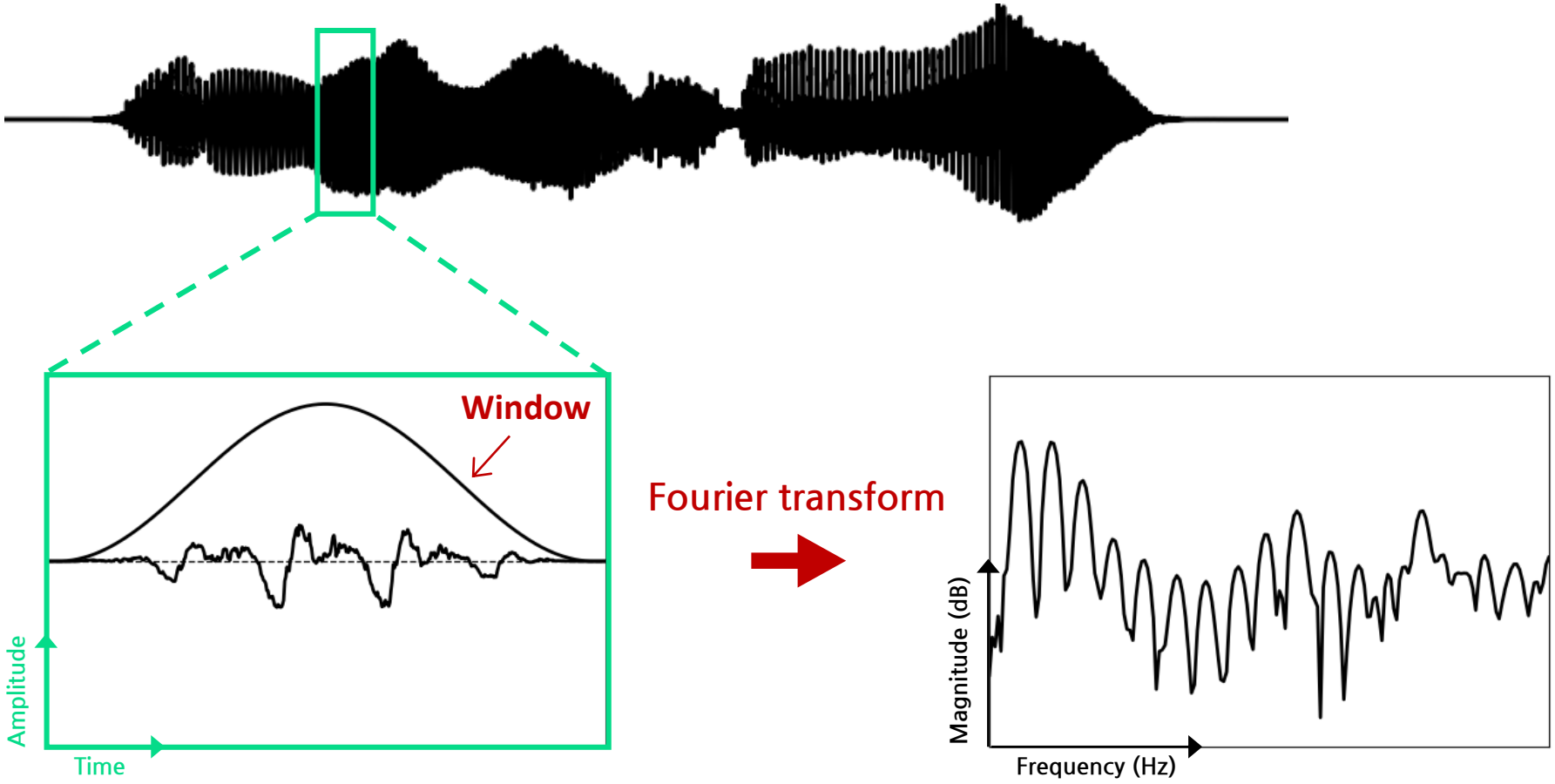


# Speech waveform




음성 신호는 시간 축에서 특정한 에너지를 갖는 파형의 형태로 존재합니다

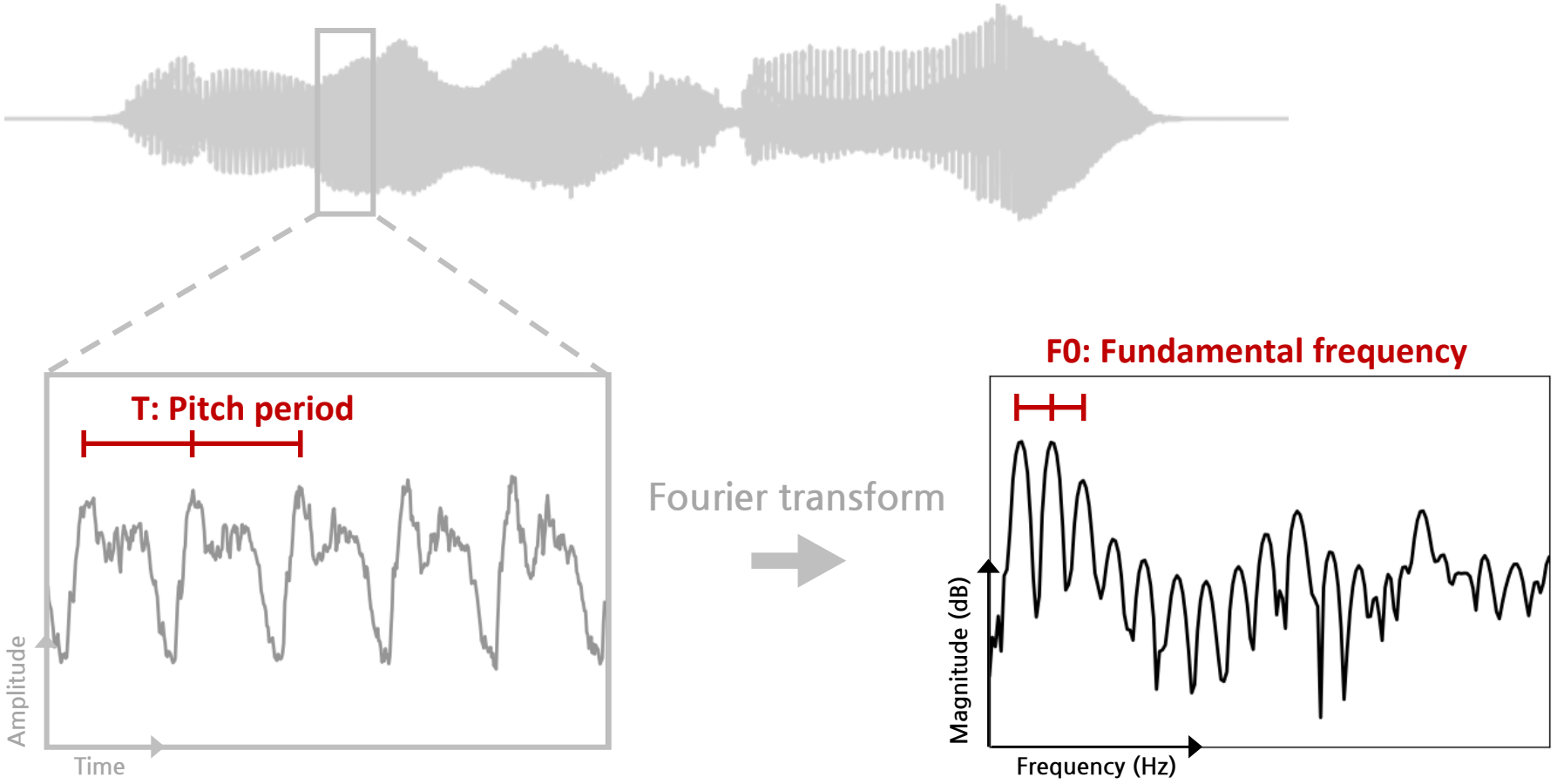
# Speech waveform




Fourier 변환을 통해 주파수 축에서 음성을 관찰할 수 있습니다

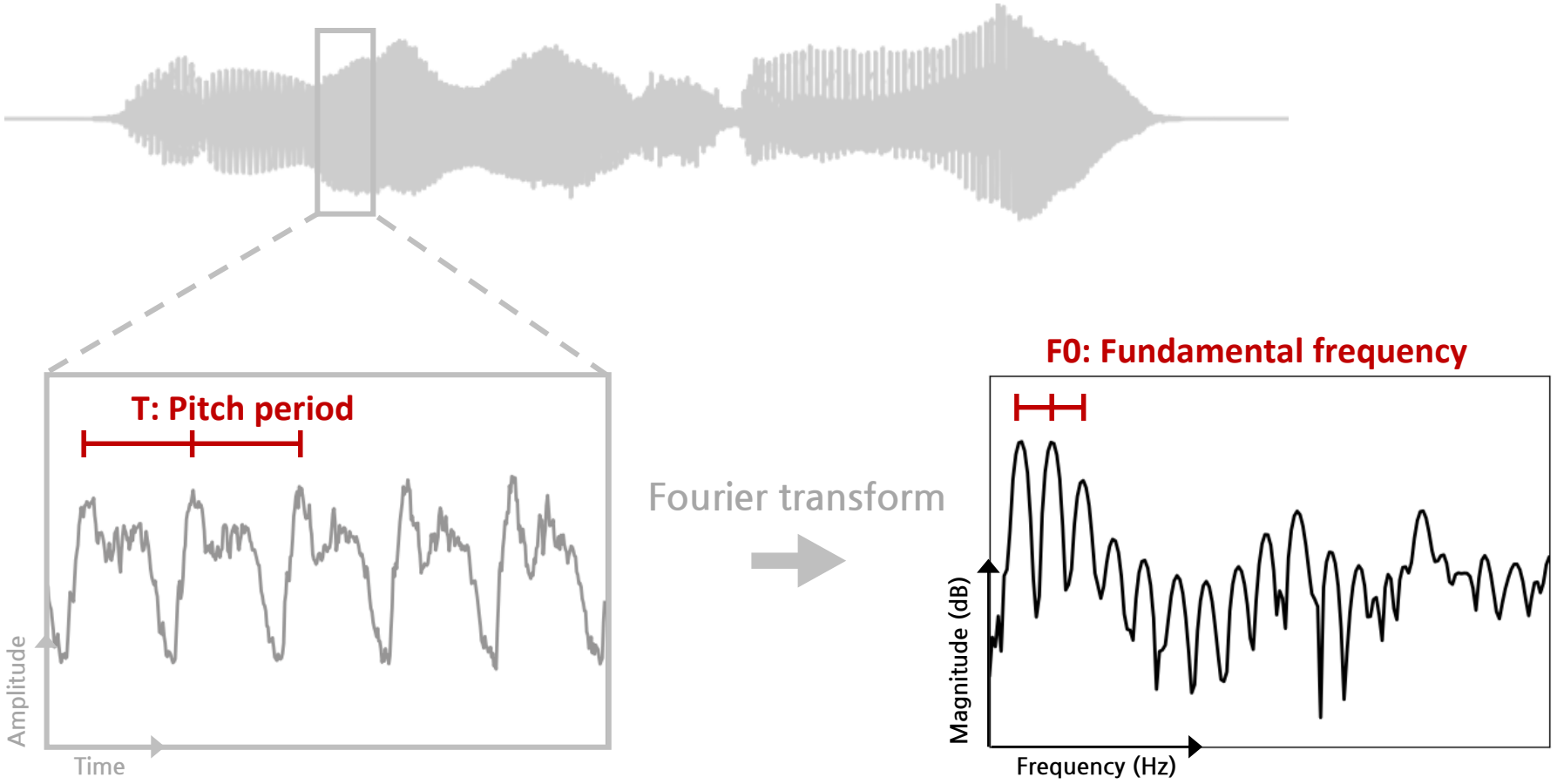
CLOVA

# Speech waveform




시간 축에서의 주기성을 pitch period, 주파수 축에서의 주기성을 fundamental frequency 라고 정의합니다

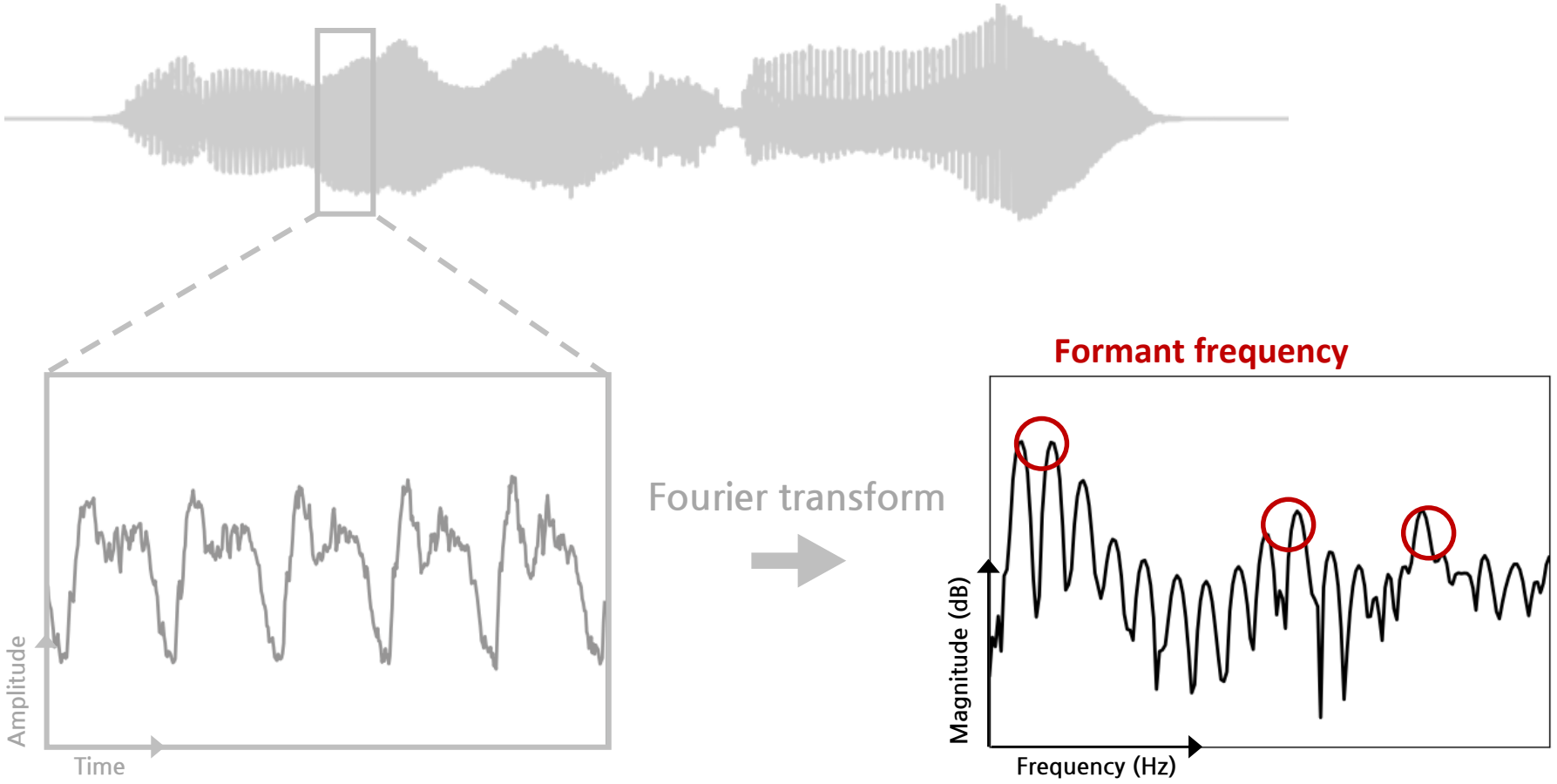
# Speech waveform




F0 의 높낮이에 따라 목소리의 톤이 결정됩니다 (아↘ 아↗)

CLOVA

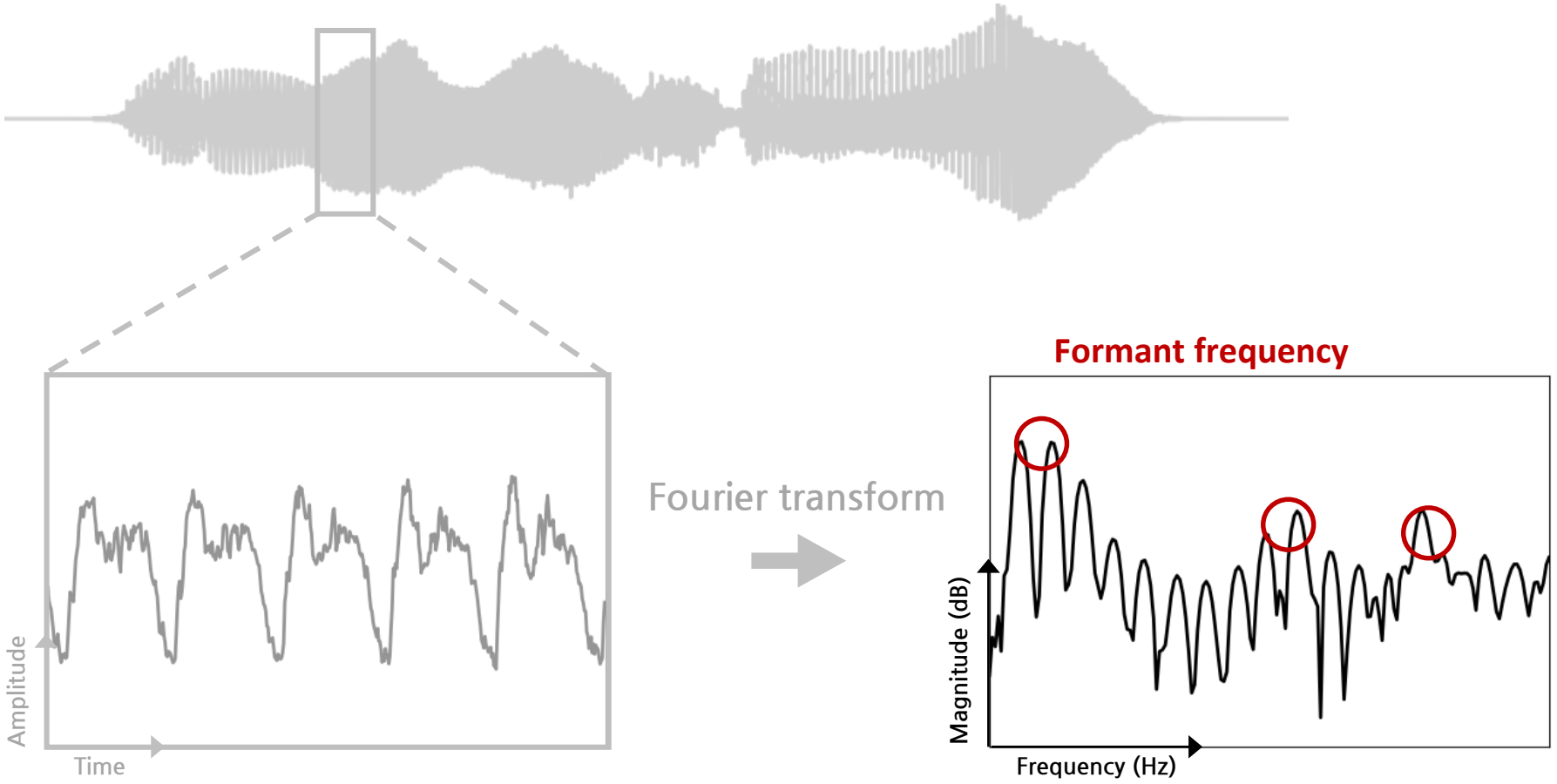
# Speech waveform




높은 에너지를 갖는 (spectral peak) 주파수 성분을 formant frequency 라고 정의합니다

CLOVA

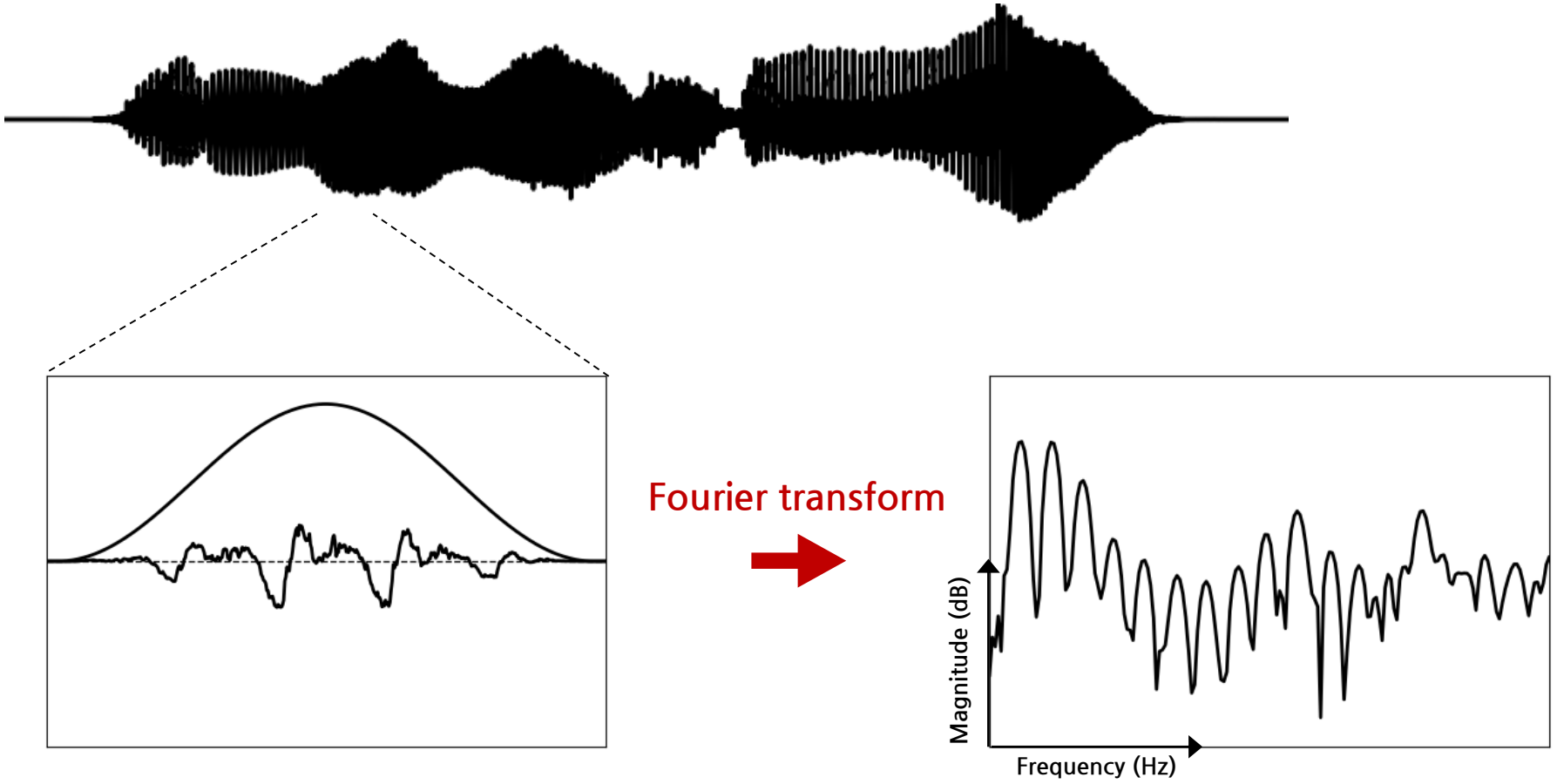
# Speech waveform



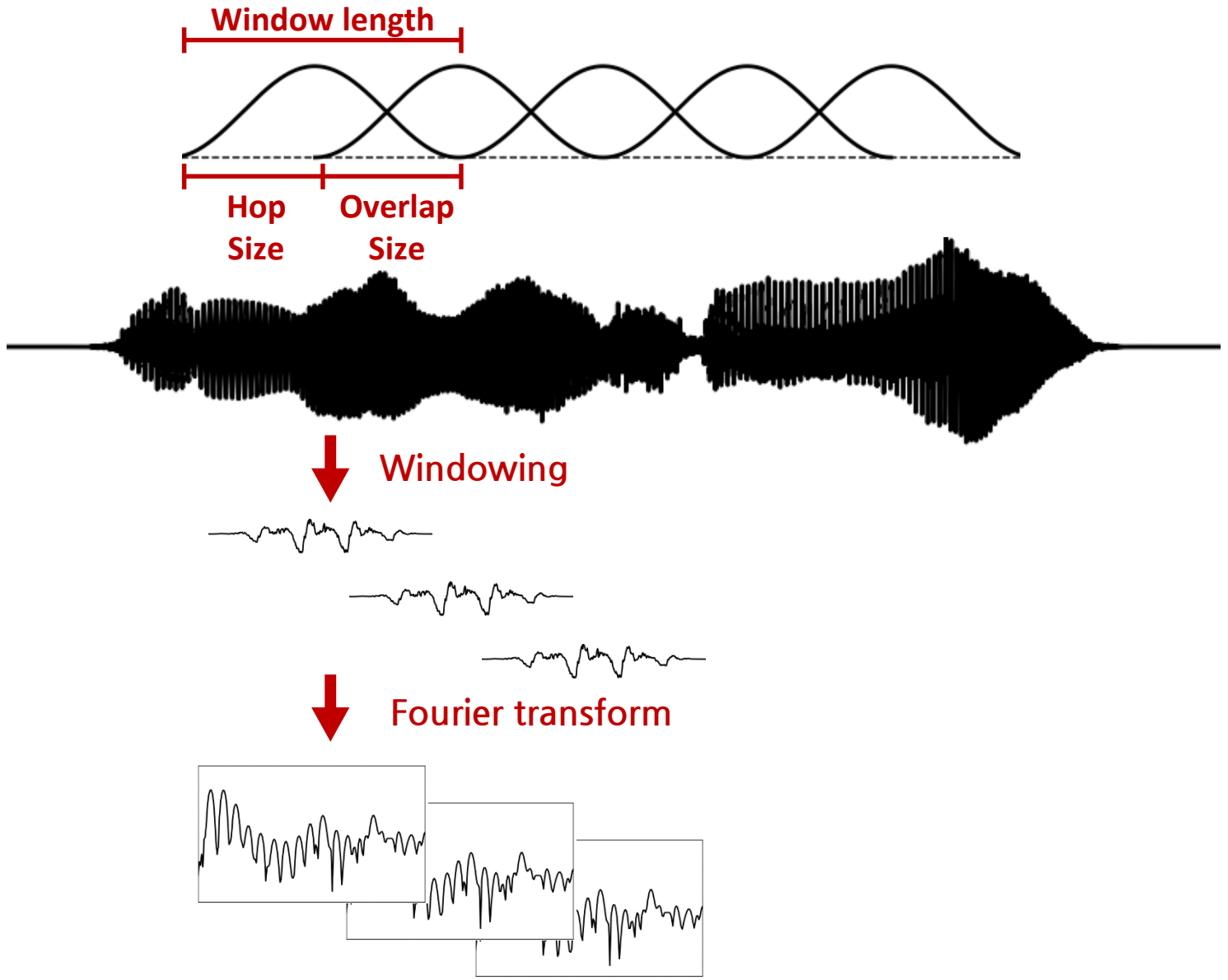

Formant 의 위치에 따라 발음이 결정됩니다 (아/에/이/오/우)

CLOVA

# 정리하자면..

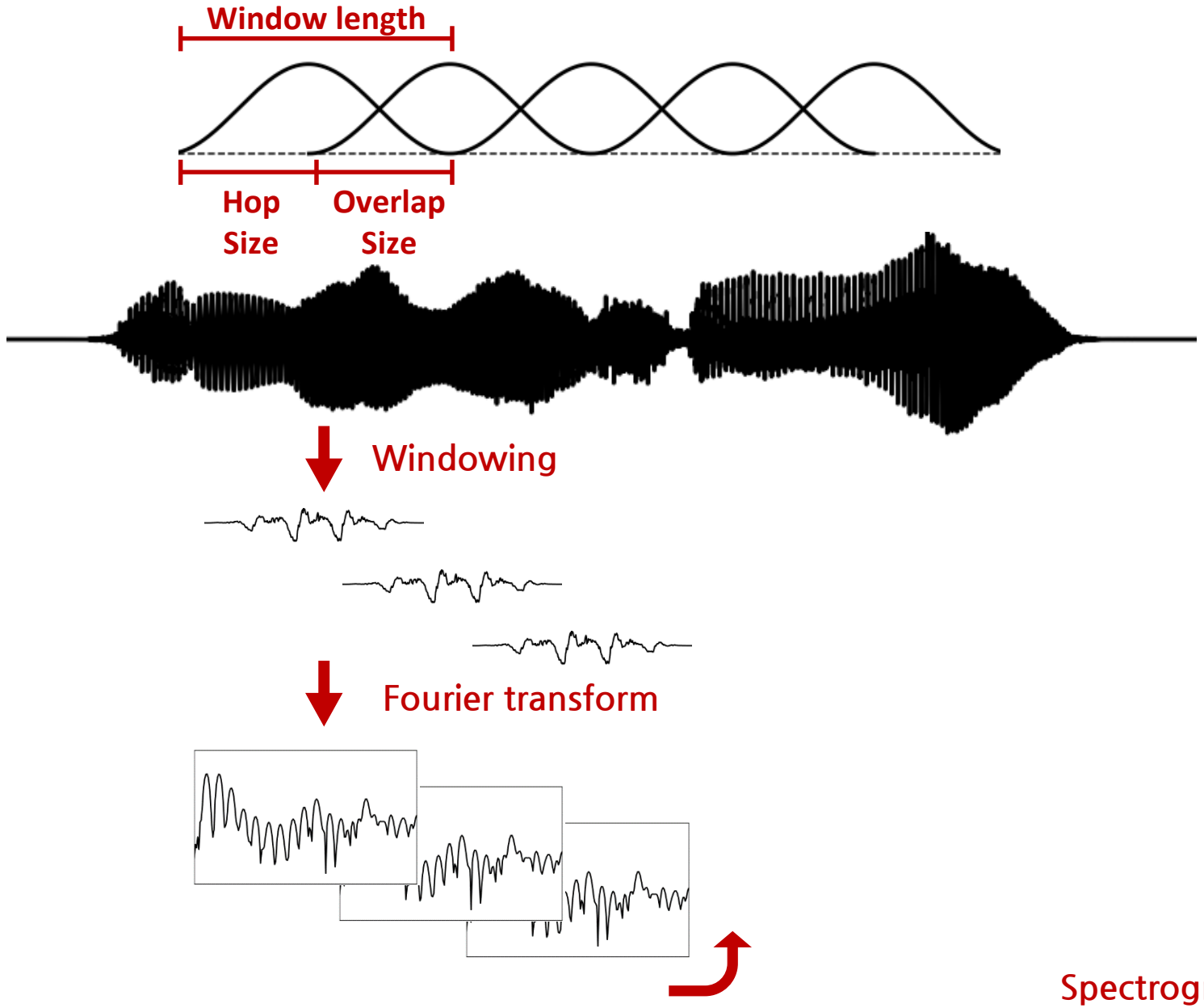



복잡해 보이는 시간 축 신호를 **주파수** 축에서 보면 음성을 **분석**하기 쉬워집니다




## Short-time Fourier transform (STFT)

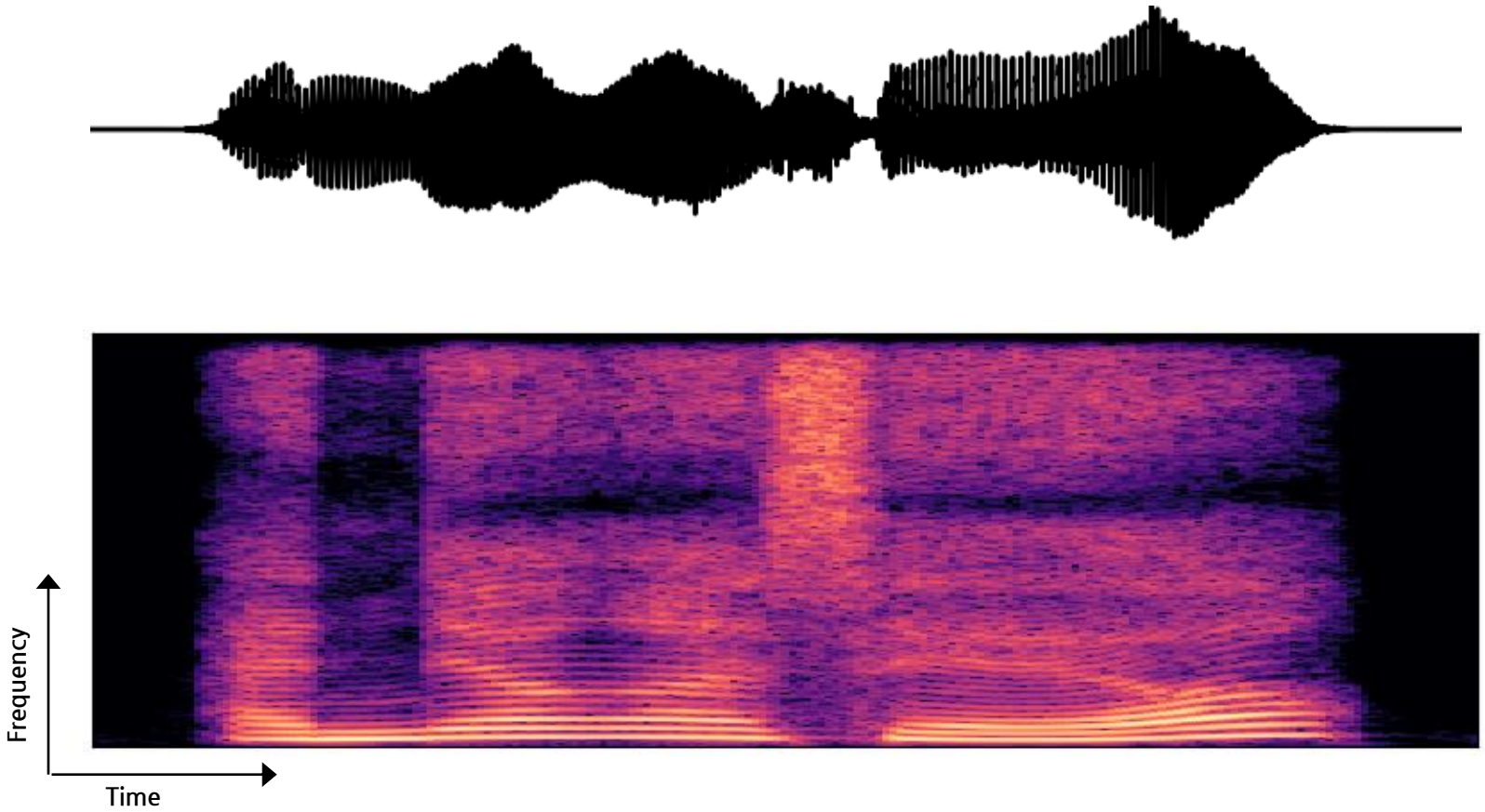
일정한 Hop 사이즈로 Fourier transform 적용합니다




Spectrogram

STFT 신호를 시간 축으로 붙인 2D 이미지

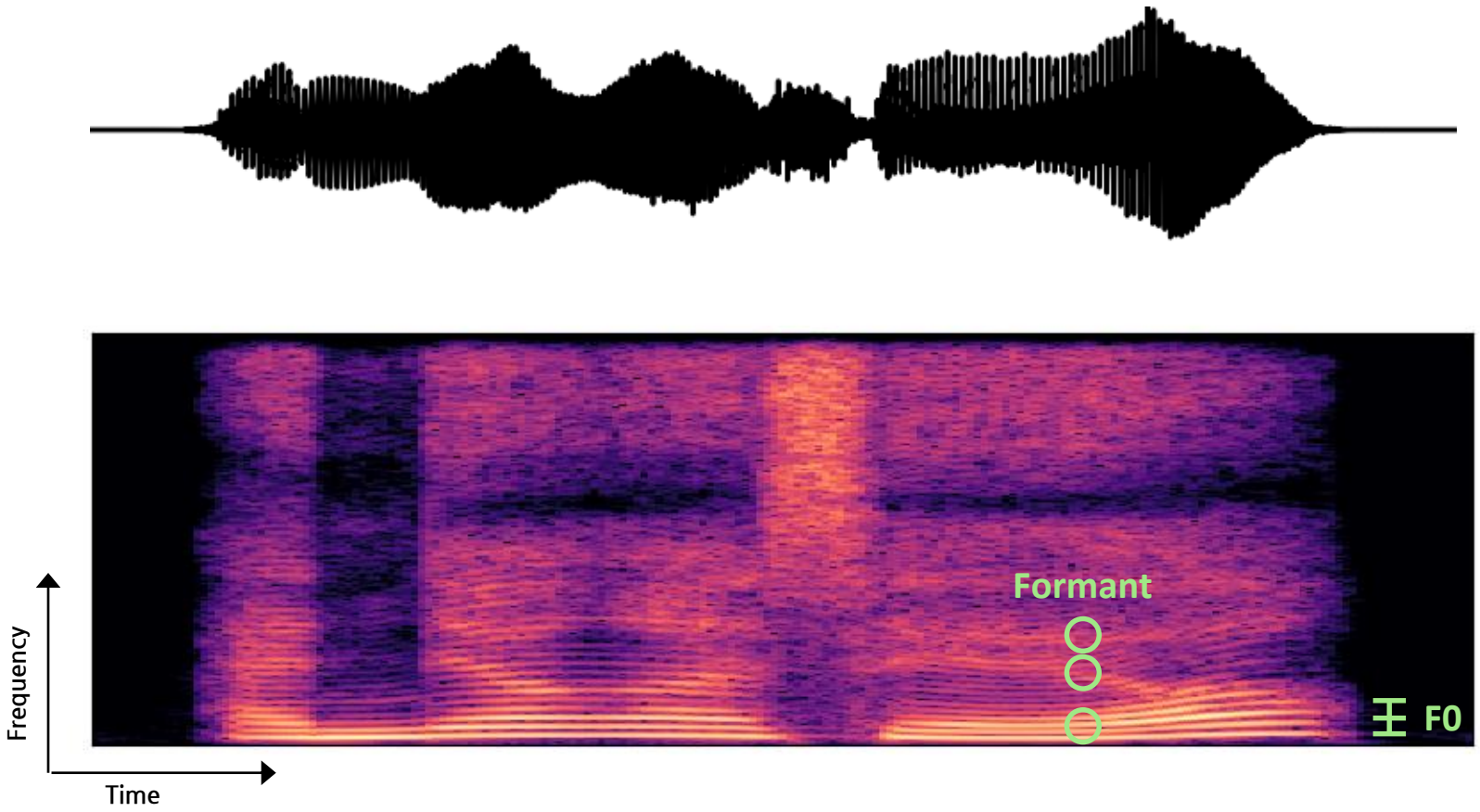
# Spectrogram




**Spectrogram**

STFT 신호를 시간 축으로 붙인 2D 이미지

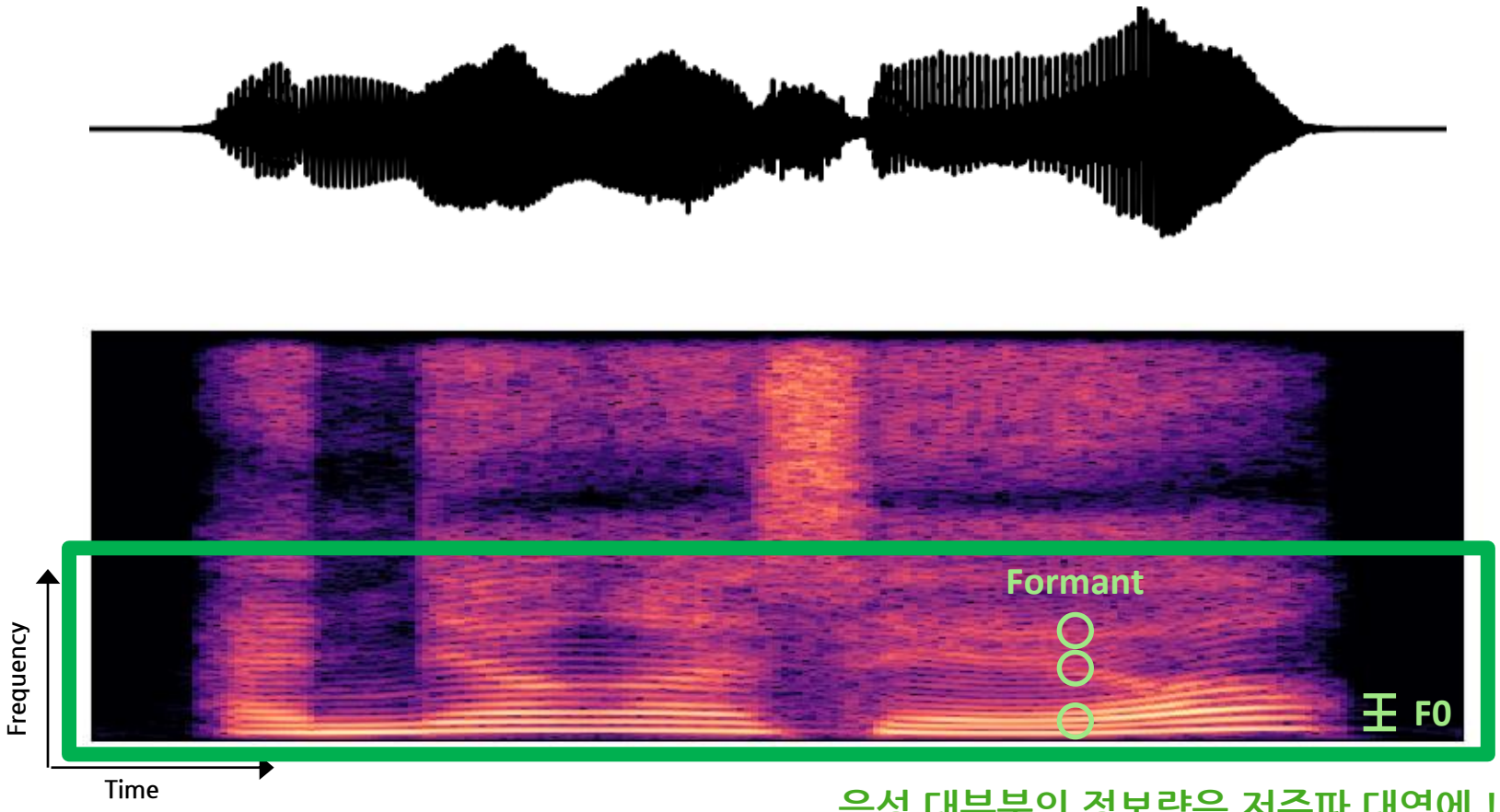
# Spectrogram




**Spectrogram**

음성 신호를 **시간-주파수** 축에서 **분석**할 수 있게 되었습니다

# Spectrogram

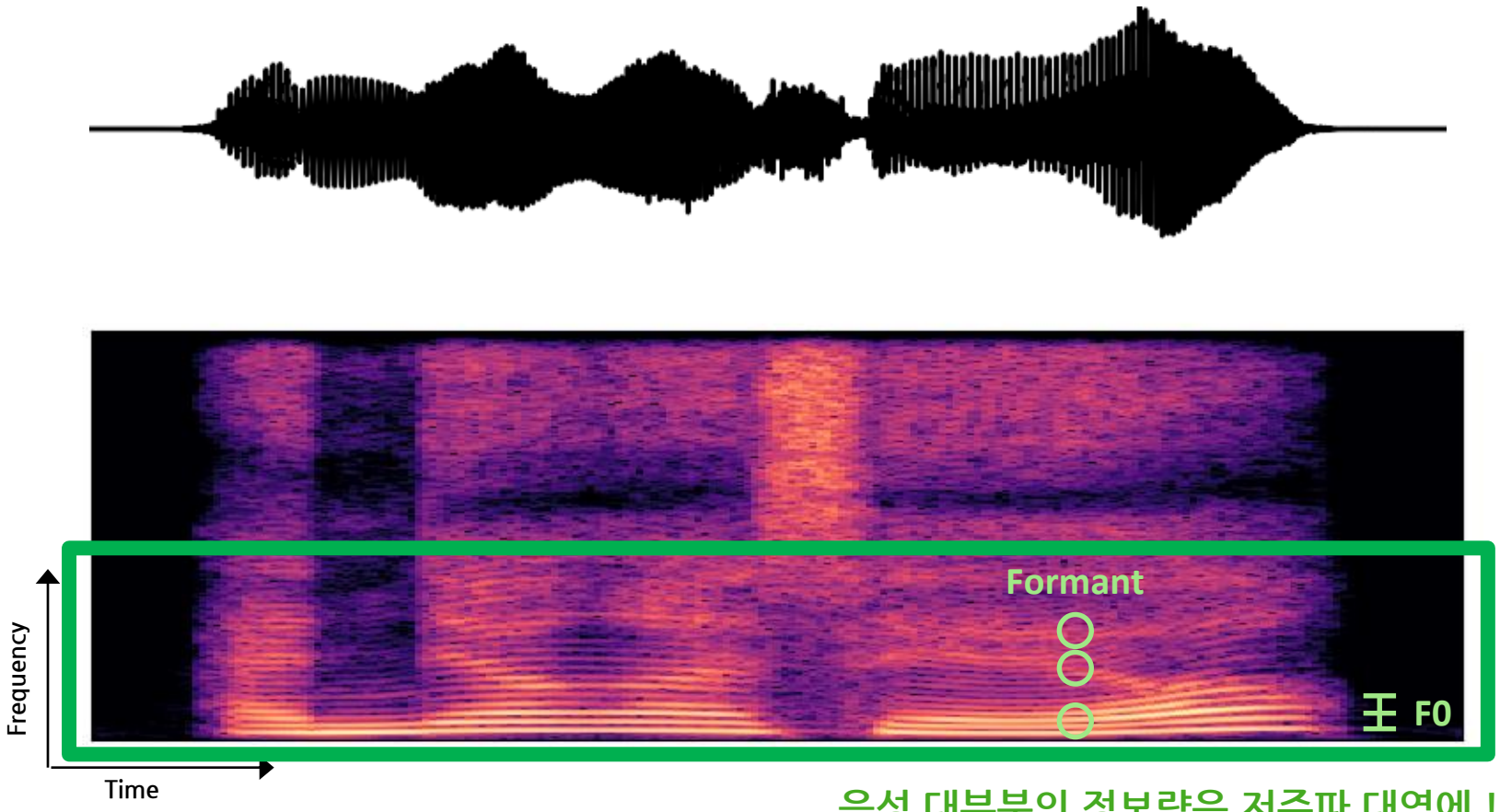



음성 대부분의 정보량은 저주파 대역에 !

Spectrogram

음성 신호를 시간-주파수 축에서 분석할 수 있게 되었습니다

# Spectrogram



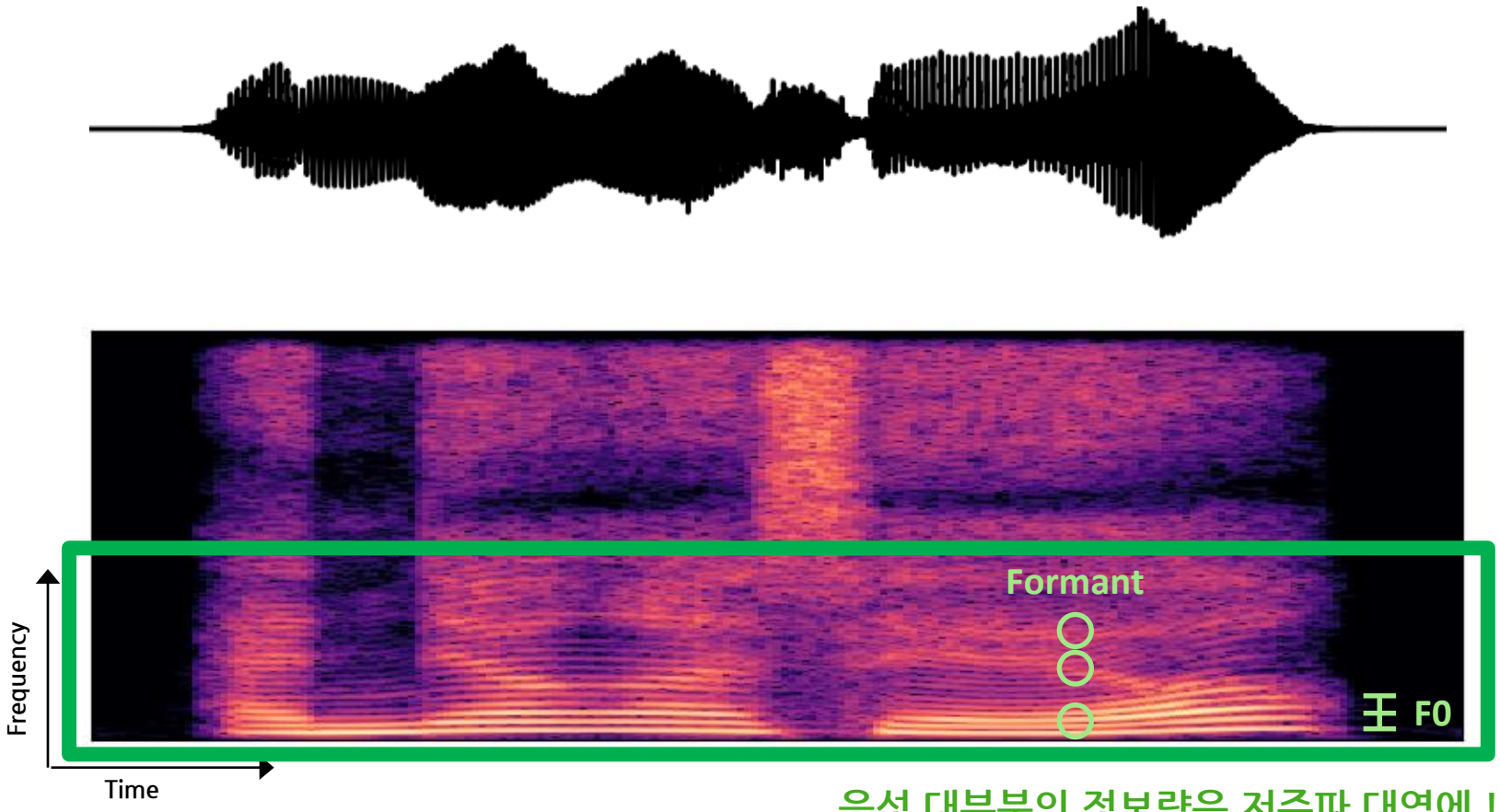

음성 대부분의 정보량은 저주파 대역에 !

저주파 대역의 정보량에 집중할 수 있다면?

음성 신호를 시간-주파수 축에서 분석할 수 있게 되었습니다

CLOVA

# Spectrogram

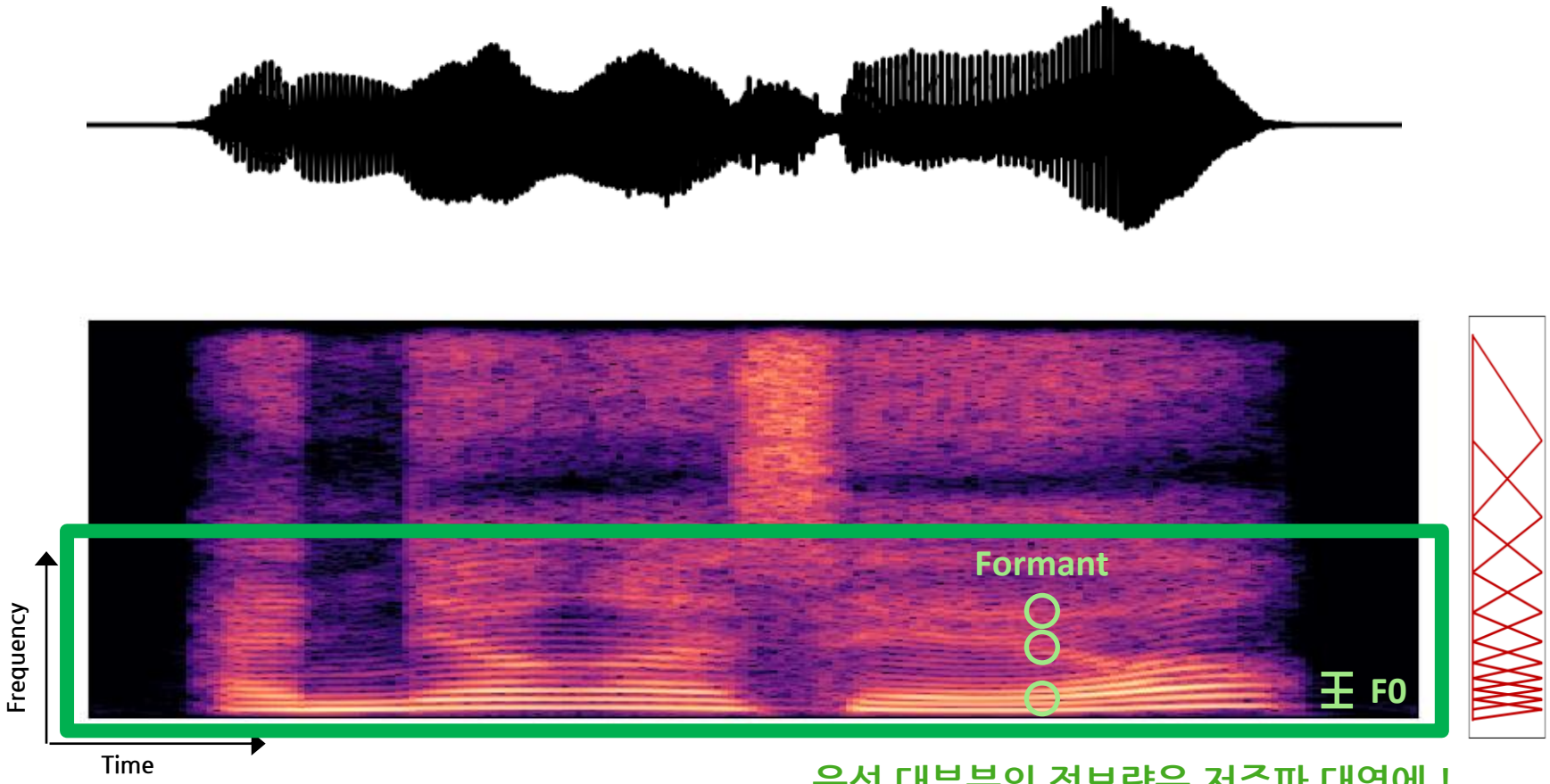



음성 대부분의 정보량은 저주파 대역에 !

저주파 대역의 정보량에 집중할 수 있다면?

음성 신호를 시간-주파수 축에서 **분석을 더 잘 ...**

# Spectrogram




주파수 축으로
Mel-filterbank 적용

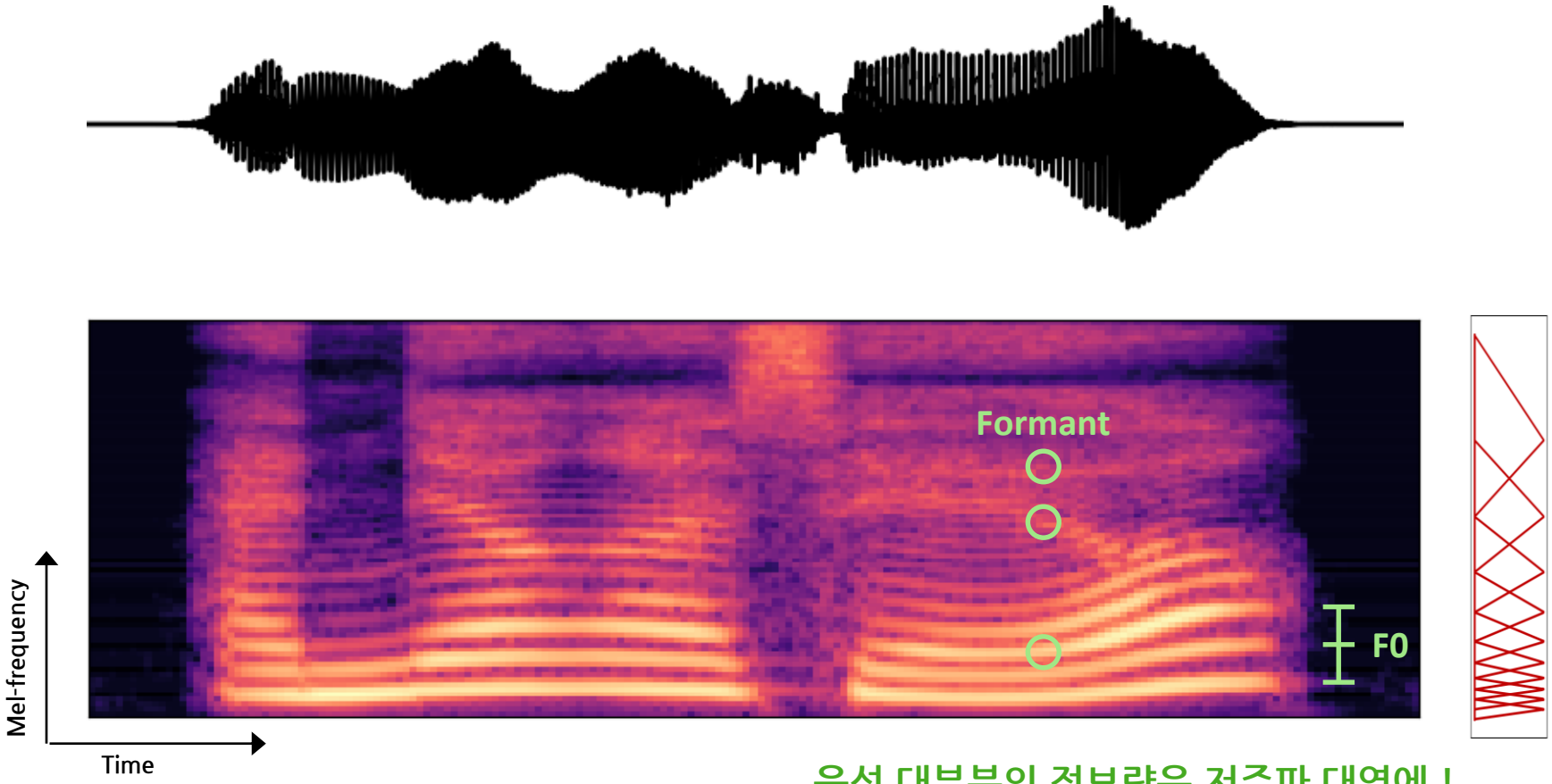
음성 대부분의 정보량은 저주파 대역에 !

저주파 대역의 정보량에 집중할 수 있다면?

음성 신호를 시간-주파수 축에서 분석을 더 잘 ...

CLOVA

# Mel-spectrogram




주파수 축으로

Mel-filterbank 적용

음성 대부분의 정보량은 저주파 대역에 !

저주파 대역의 정보량에 집중할 수 있다면?

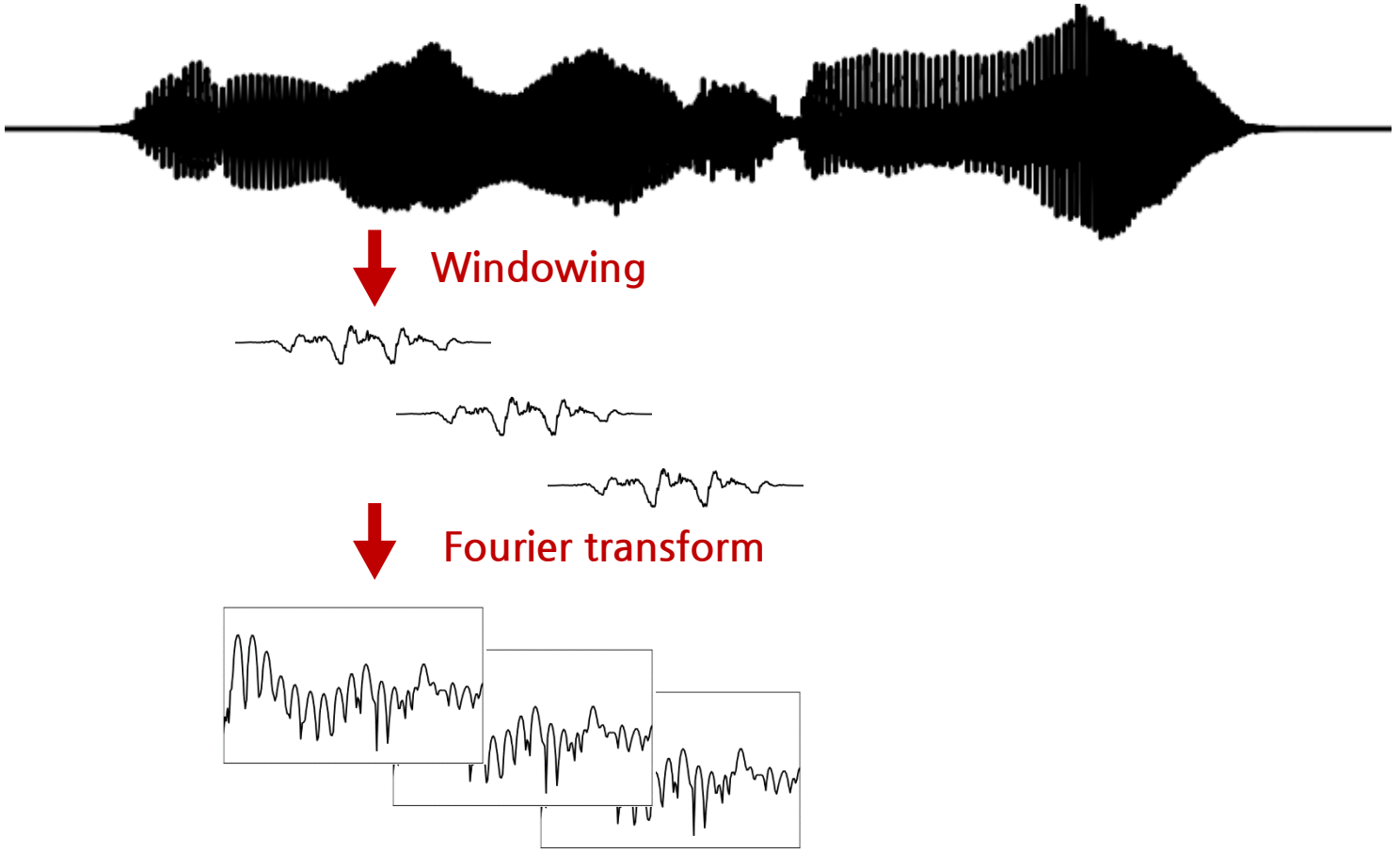
음성 신호를 시간-주파수 축에서 분석을 더 잘 할 수 있습니다!

CLOVA

# 정리하자면..

## 1. Short-time Fourier transform (STFT)

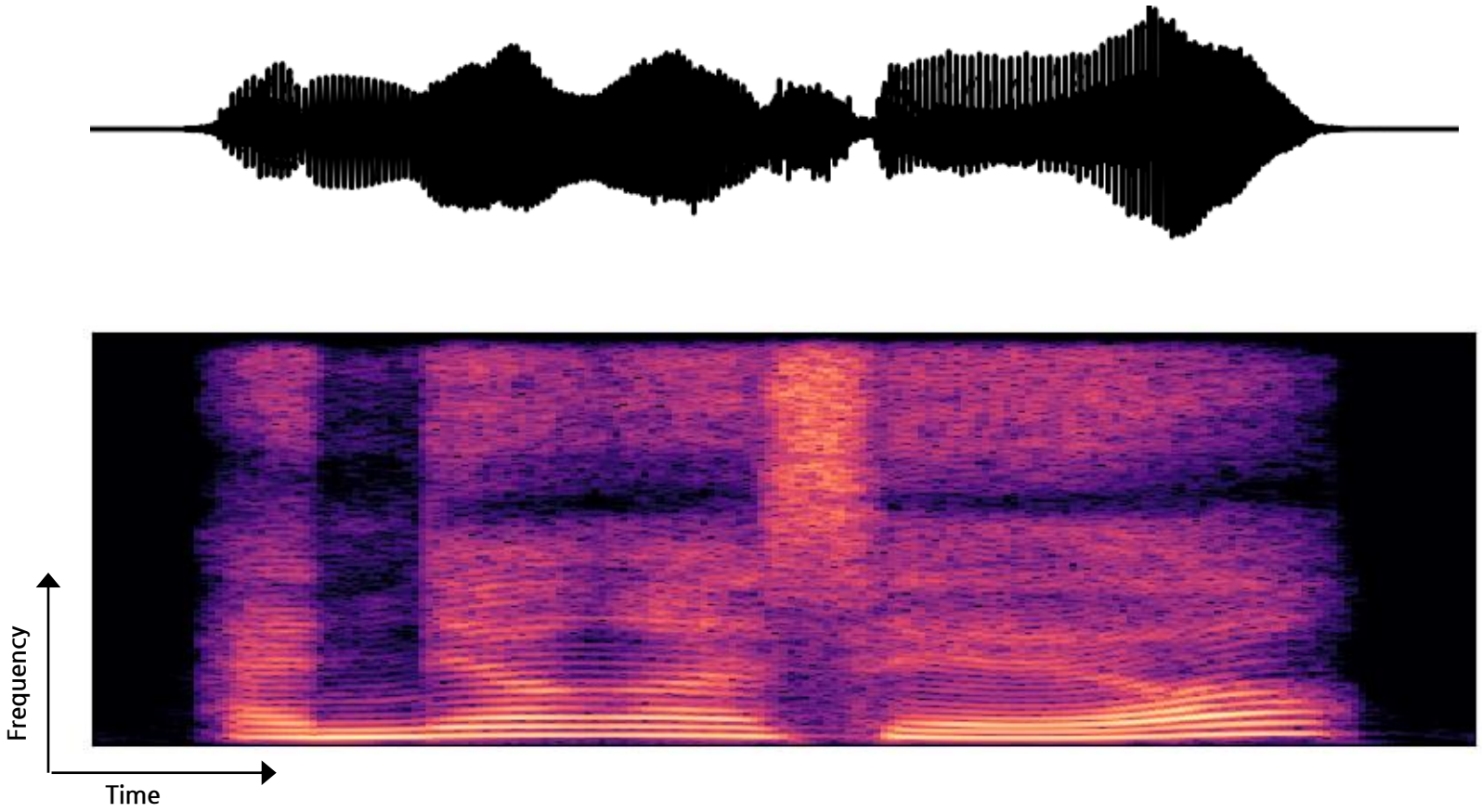
일정한 Hop 사이즈로 Fourier transform 적용합니다



# 정리하자면..

## 2. Spectrogram

STFT magnitude 신호를 붙여 시간-주파수 축의 2D 이미지를 만듭니다

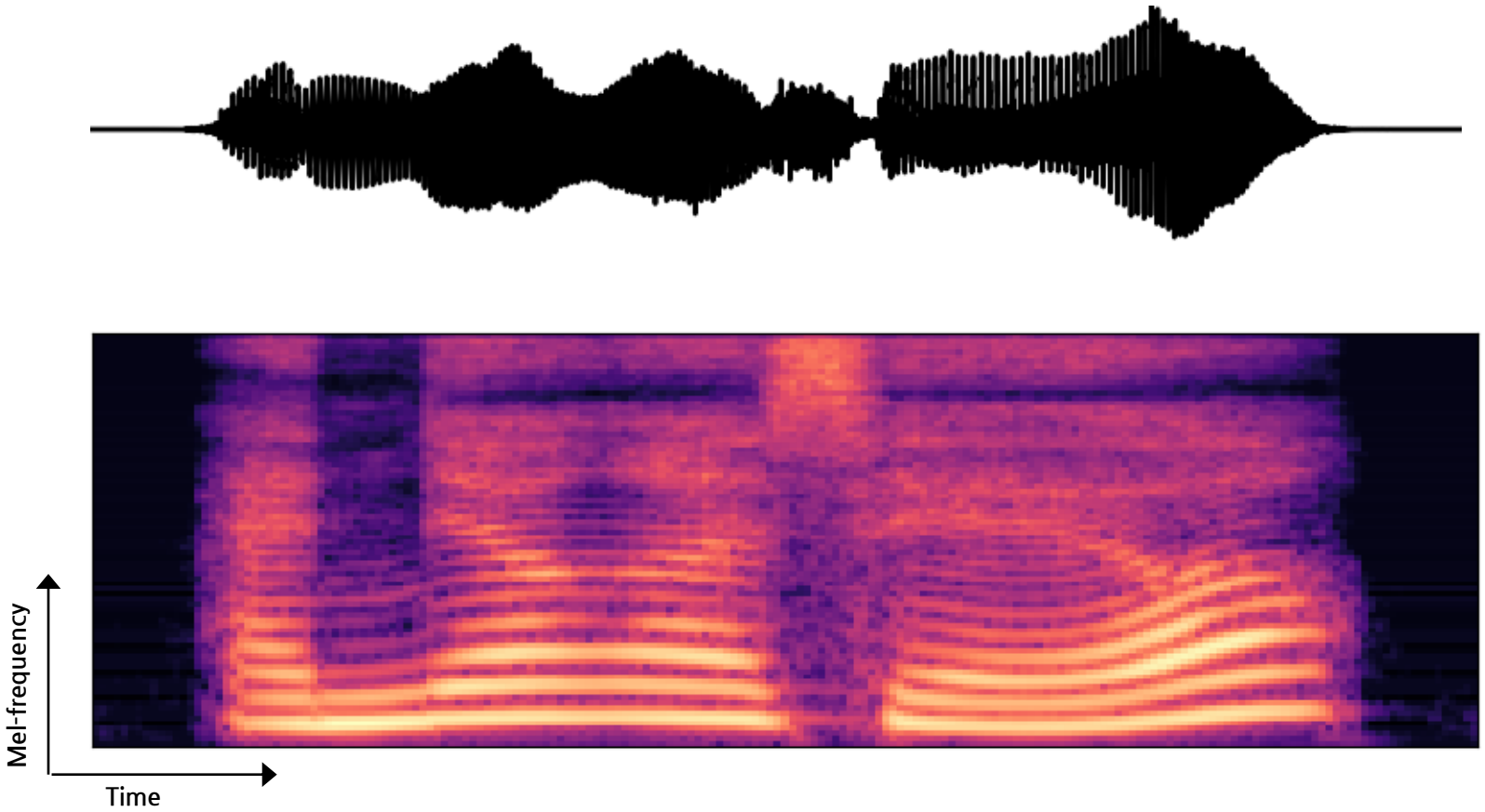


CLOVA

# 정리하자면..

## 3. Mel-Spectrogram

주파수 축으로 Mel-filterbank 를 적용합니다

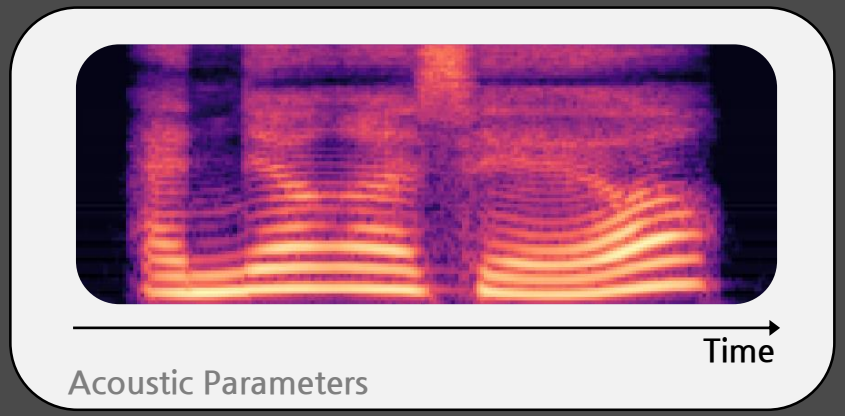



음성 신호를 시간-주파수 축에서 분석할 수 있습니다 = **딥러닝** 모델이 음성 신호를 **이해**하기 쉬워집니다

CLOVA

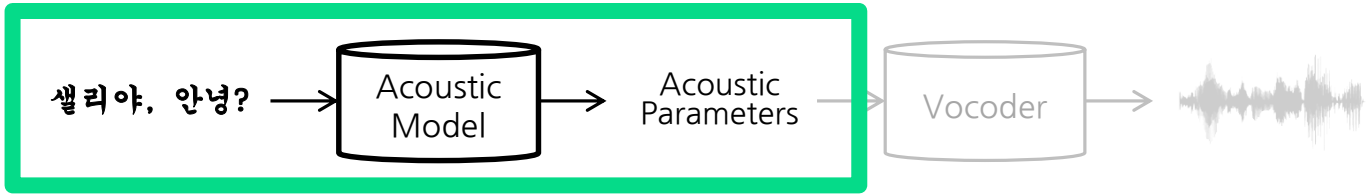
# 정리하자면..

Acoustic model 과 vocoder 를 연결하는 **매개체** 역할을 하는 것이 Mel-spectrogram

CLOVA

# Acoustic model



# Overview

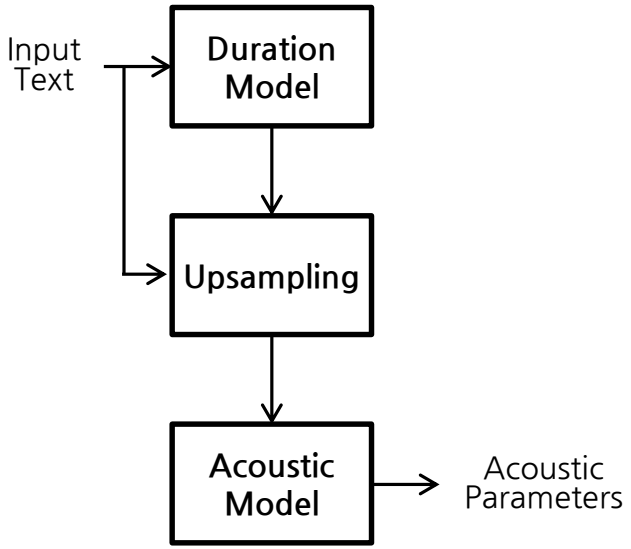
Acoustic model 은 Text 로부터 Acoustic Parameter 를 추정하는 역할을 합니다

# Overview

Acoustic model 은 Text 로부터 Acoustic Parameter 를 추정하는 역할을 합니다

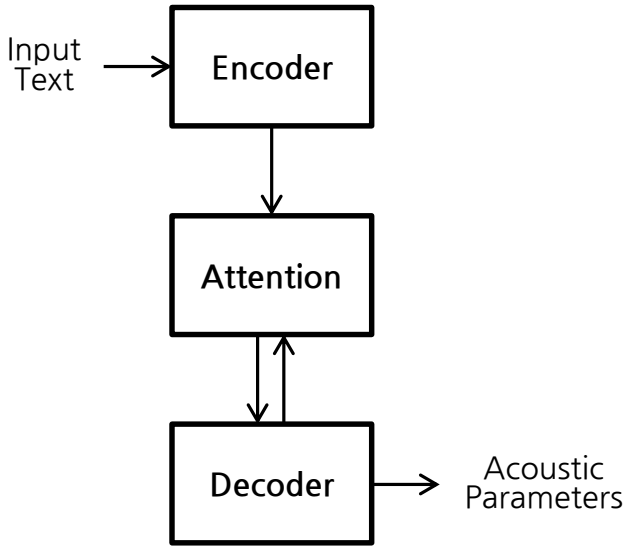
## Statistical parametric speech synthesis

- Simple deep learning model (FF+LSTM)




## End-to-end speech synthesis

- Seq2seq model

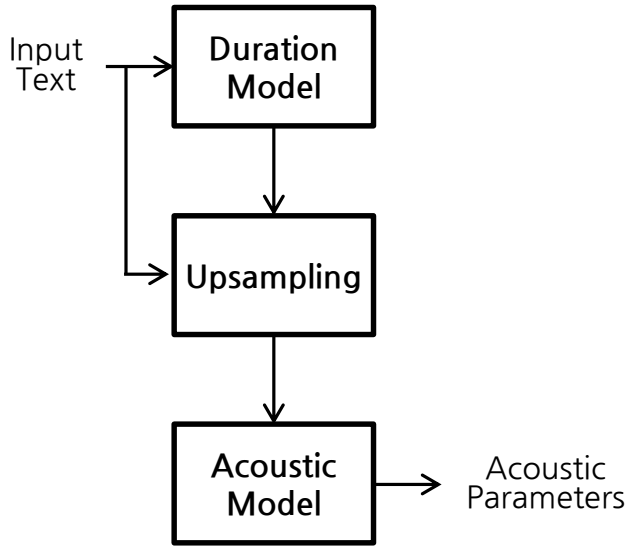


# Overview

Acoustic model 은 Text 로부터 Acoustic Parameter 를 추정하는 역할을 합니다

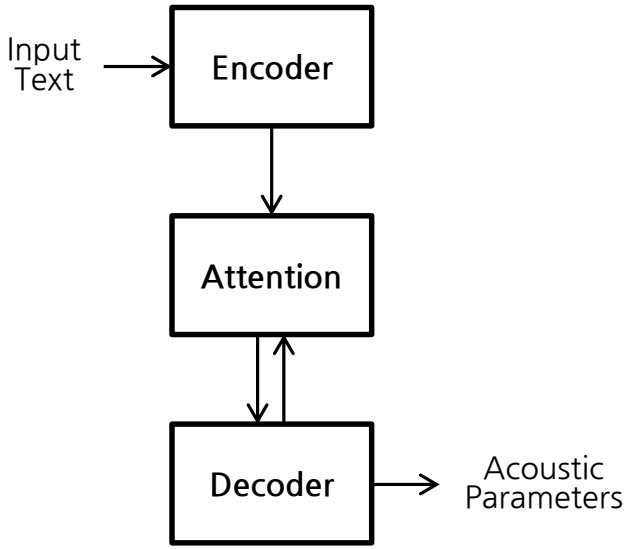
## Statistical parametric speech synthesis

- Simple deep learning model (FF+LSTM)




## End-to-end speech synthesis

- Seq2seq model



# Statistical parametric speech synthesis (SPSS)

SPSS 교과서

## STATISTICAL PARAMETRIC SPEECH SYNTHESIS USING DEEP NEURAL NETWORKS

*Heiga Zen, Andrew Senior, Mike Schuster*

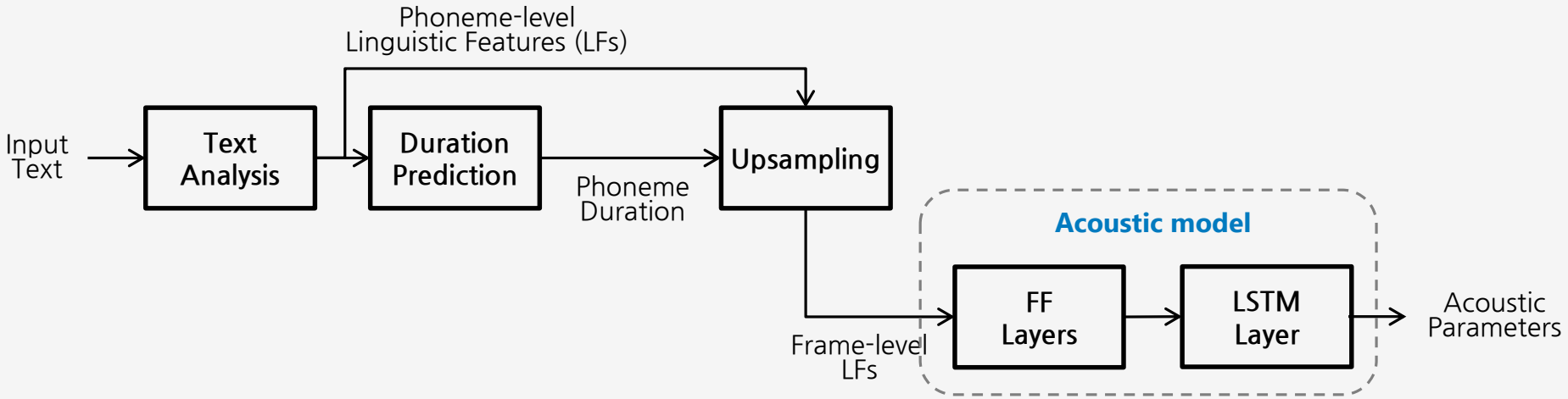
Google

{heigazen,andrewsenior,schuster}@google.com

**ABSTRACT**

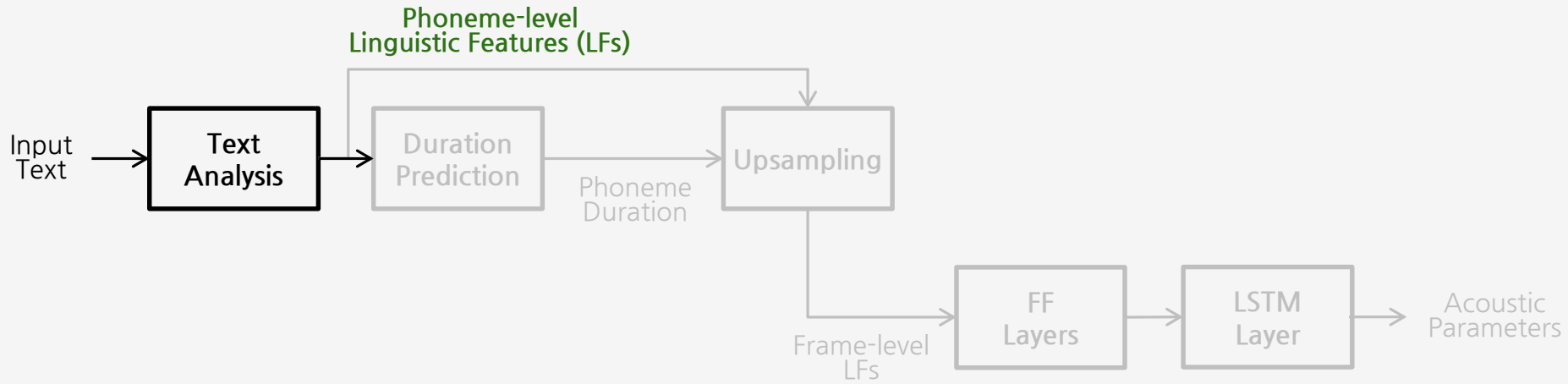
Conventional approaches to statistical parametric speech synthesis typically use decision tree-clustered context-dependent hidden Markov models (HMMs) to represent probability densities of speech parameters given texts. Speech parameters are generated from the probability densities to maximize their output probabilities, then a speech waveform is reconstructed from the generated parameters. This approach is reasonably effective but has a couple of limitations, *e.g.* decision trees are inefficient to model complex context dependencies. This paper examines an alternative scheme that is based on a deep neural network (DNN). The relationship between input texts and their acoustic realizations is modeled by a DNN. The use of the DNN can address some limitations of the conventional approach. Experimental results show that the DNN-based systems outperformed the HMM-based systems with similar numbers of parameters.

# Statistical parametric speech synthesis (SPSS)



# Statistical parametric speech synthesis (SPSS)

Text analyzer: Generates phoneme-level linguistic features (Phoneme: 음운론상의 최소 단위)




"안녕하세요"

Phoneme ID
Break
Punctuation
Position
.
.

a-N| n-jv-O| h-a| s-e| jo

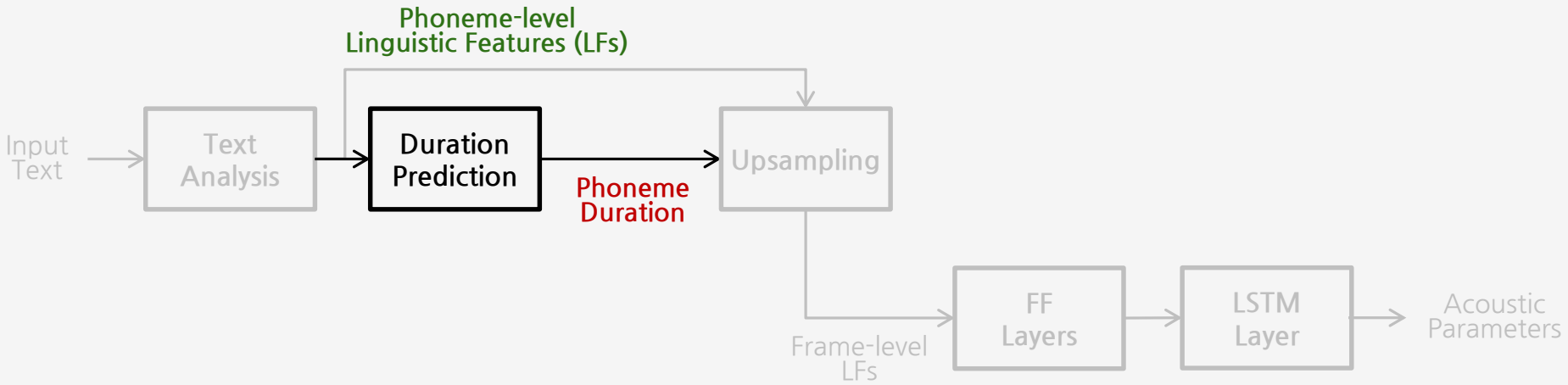
Linguistic features

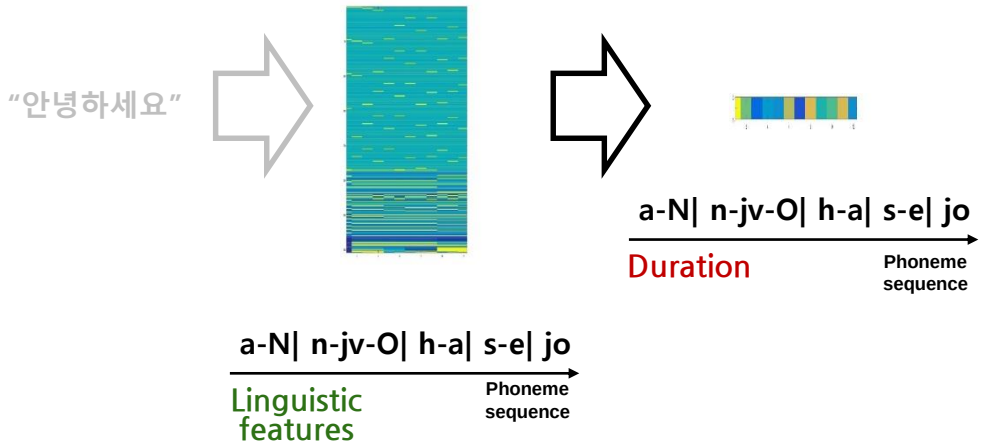
Phoneme sequence

```
WD=[안녕하세요] PR=[a00 NX13 n00 jv00 OX13 h00 a03 s00 e03 jo04] BR=[6] OWD
WD=[눈이] PR=[n00 u03 n00 i04] OWD=[눈이] OPR=[누니] ONPR=[누니] DOM=[0] E
WD=[마주치자] PR=[m00 a03 z00 u03 c00 i03 z00 a04] BR=[6] OWD=[마주치자] OPR
WD=[가쁜] PR=[g00 a03 B00 U00 NX14] OWD=[가쁜] OPR=[가쁜] ONPR=[가쁜] DOM
WD=[숨] PR=[s00 u00 MX14] BR=[3] OWD=[숨] OPR=[숨] ONPR=[숨] DOM=[0] EMO
WD=[사이로] PR=[s00 a03 i03 r00 o04] OWD=[사이로] OPR=[사이로] ONPR=[사이로]
WD=[미소] PR=[m00 i03 s00 o04] OWD=[미소] OPR=[미소] ONPR=[미소] DOM=[0] E
WD=[섞인] PR=[s00 v03 G00 i04] BR=[3] OWD=[섞인] OPR=[서끼] ONPR=[서끼] DOM
WD=[인사가] PR=[n00 i00 NX13 s00 a03 g00 a04] OWD=[인사가] OPR=[닌사가] ONP
WD=[배어] PR=[b00 e03 v04] OWD=[배어] OPR=[배어] ONPR=[베어] DOM=[0] EMO=
WD=[나온다] PR=[n00 a03 o00 NX13 d00 a04] PUNCT=[.] BR=[7] OWD=[나온다.] OP
```

# Statistical parametric speech synthesis (SPSS)

Duration model: Predicts phoneme duration

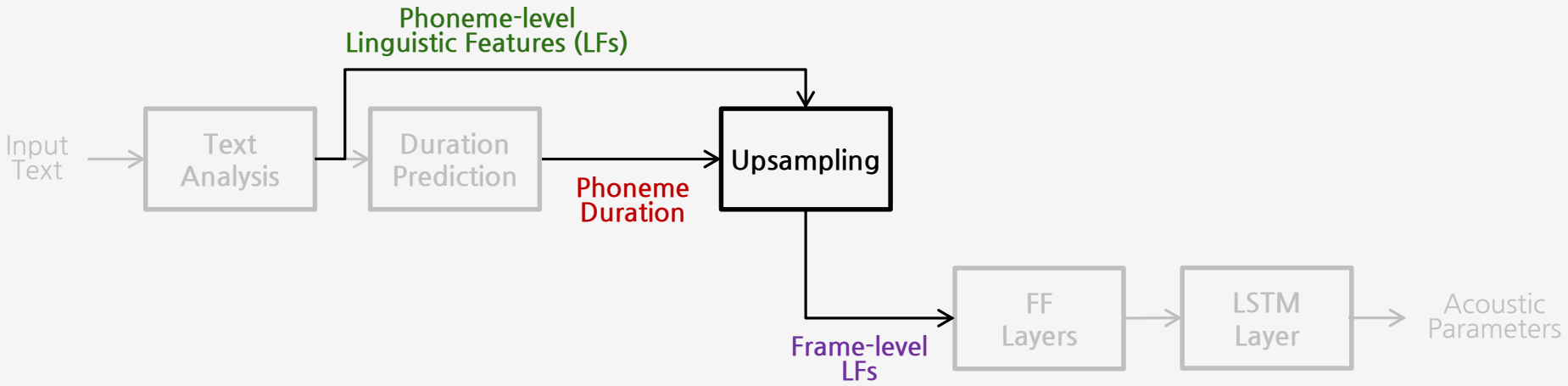


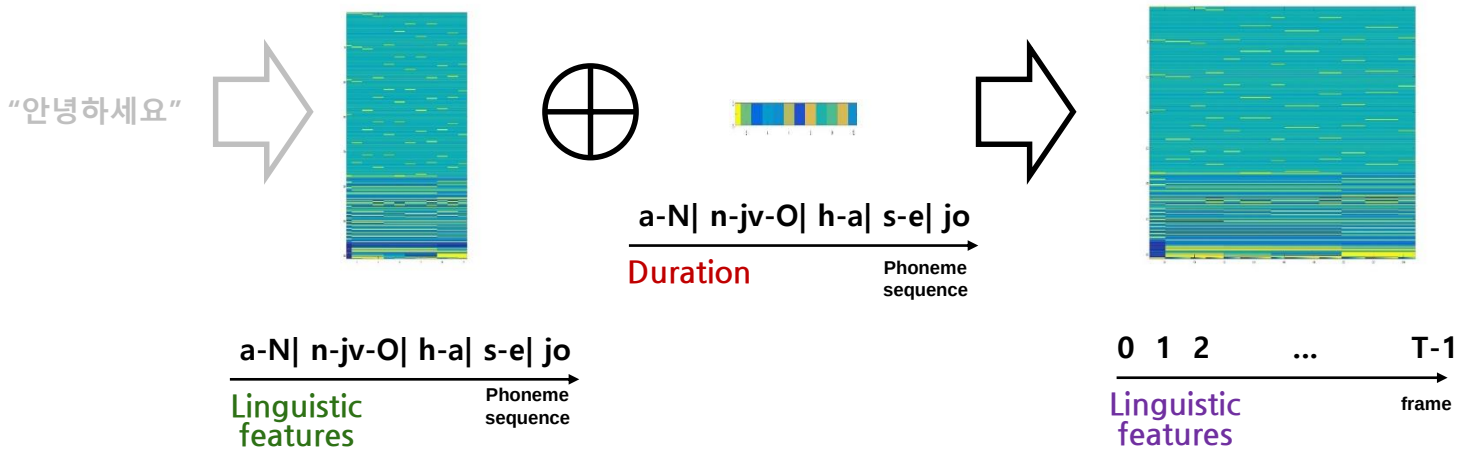



CLOVA

# Statistical parametric speech synthesis (SPSS)

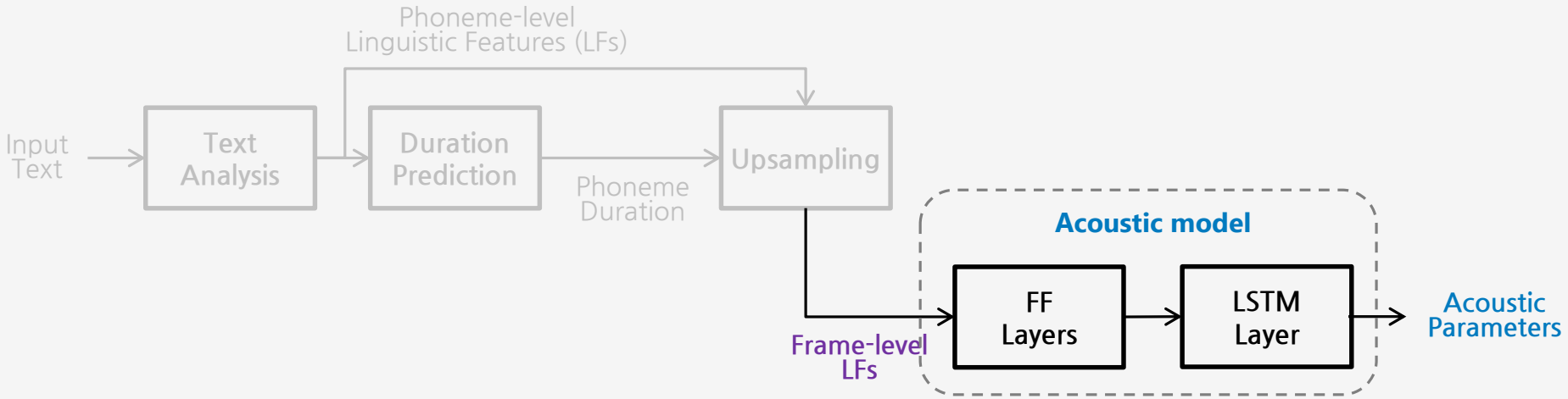
Linguistic upsampler: Generates frame-level linguistic features

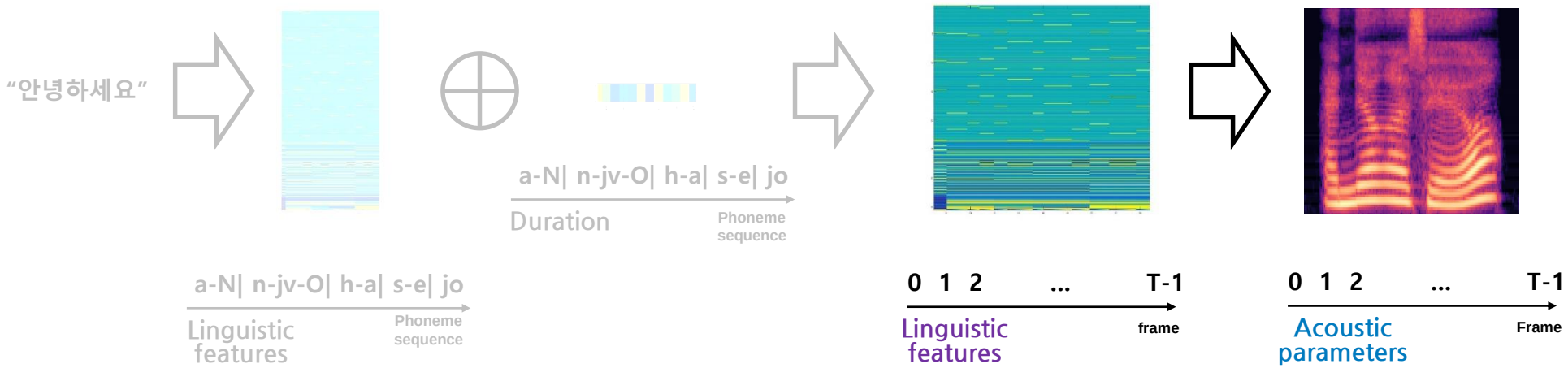


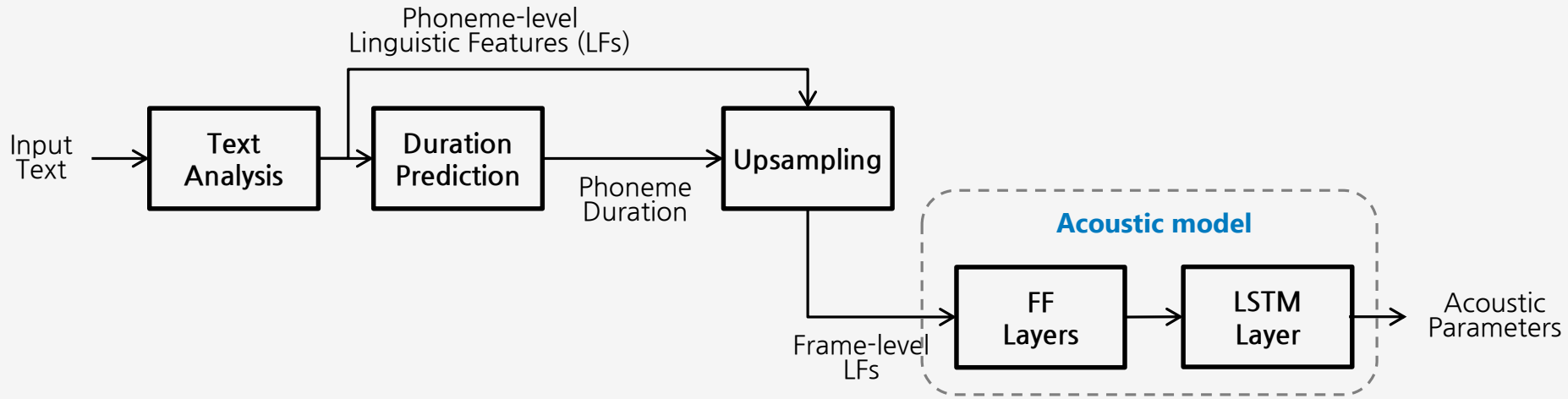
# Statistical parametric speech synthesis (SPSS)

Acoustic model: Predicts frame-level acoustic parameters



# Statistical parametric speech synthesis (SPSS)



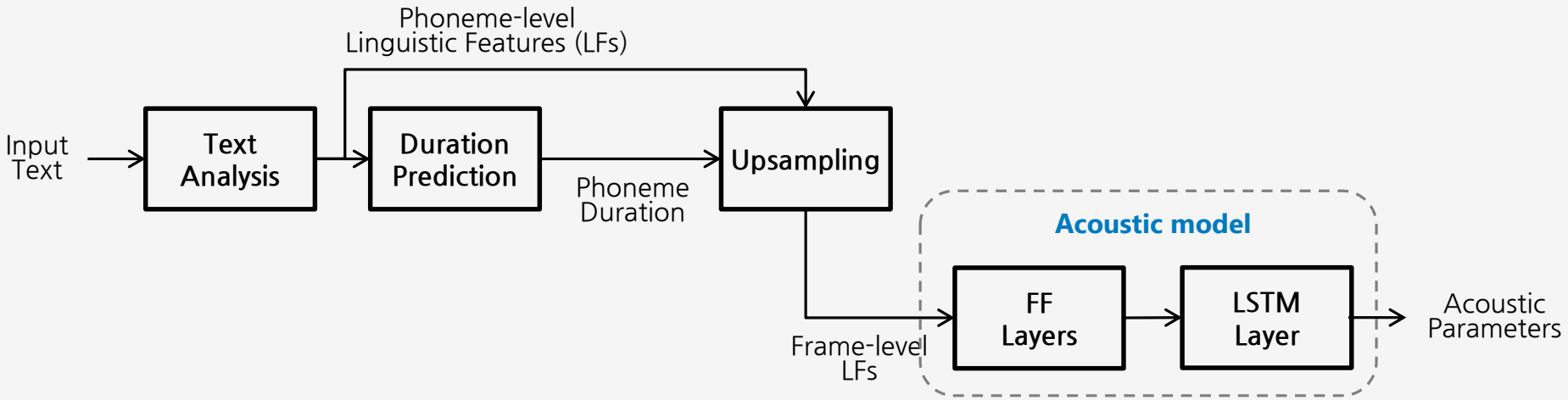

Simple and compact

1:1 mapping between linguistic and acoustic features

가볍다 + 빠르다

안정적이다

# Statistical parametric speech synthesis (SPSS)




Simple and compact

1:1 mapping between linguistic and acoustic features

합성음 **품질**이 좋지 않다

Phoneme segmentation 을 위한 **비용**이 많이 든다

CLOVA

# Text-to-speech

## End-to-end speech synthesis

Tacotron

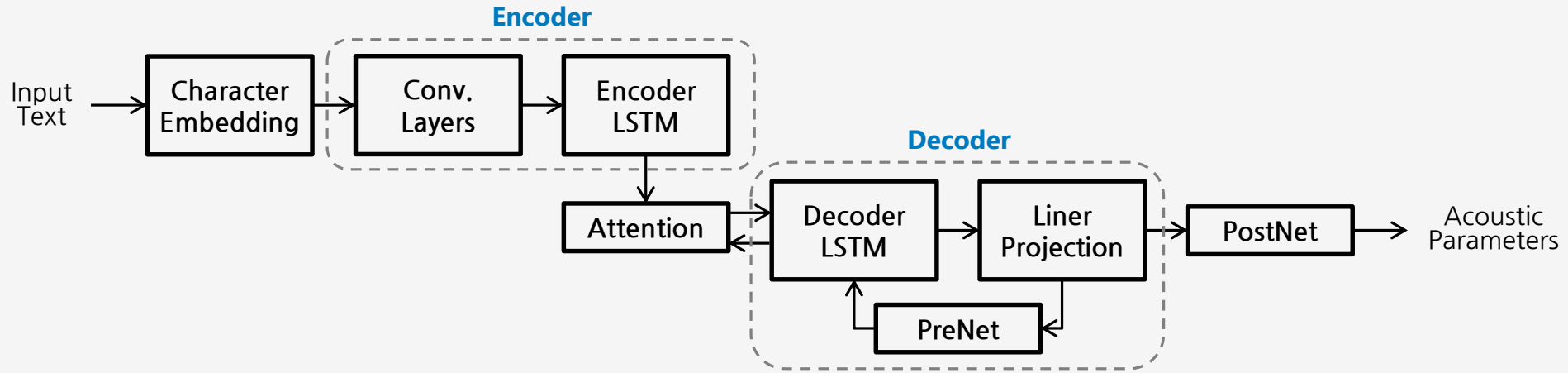
Tacotron 2

FastSpeech

FastSpeech 2

# Overview

(Text) **Encoder** 와 (Acoustic Parameter) **Decoder** 를 만들고, **Attention** 으로 Alignment 를 잡아주면 됩니다.




## Tacotron 2

# Tacotron

Google 에서 발표한 대표적인 End-to-end 음성 합성 모델

## TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS

**Yuxuan Wang*, RJ Skerry-Ryan*, Daisy Stanton, Yonghui Wu, Ron J. Weiss†, Navdeep Jaitly,**

**Zongheng Yang, Ying Xiao*, Zhifeng Chen, Samy Bengio†, Quoc Le, Yannis Agiomyrgiannakis,**

**Rob Clark, Rif A. Saurous***

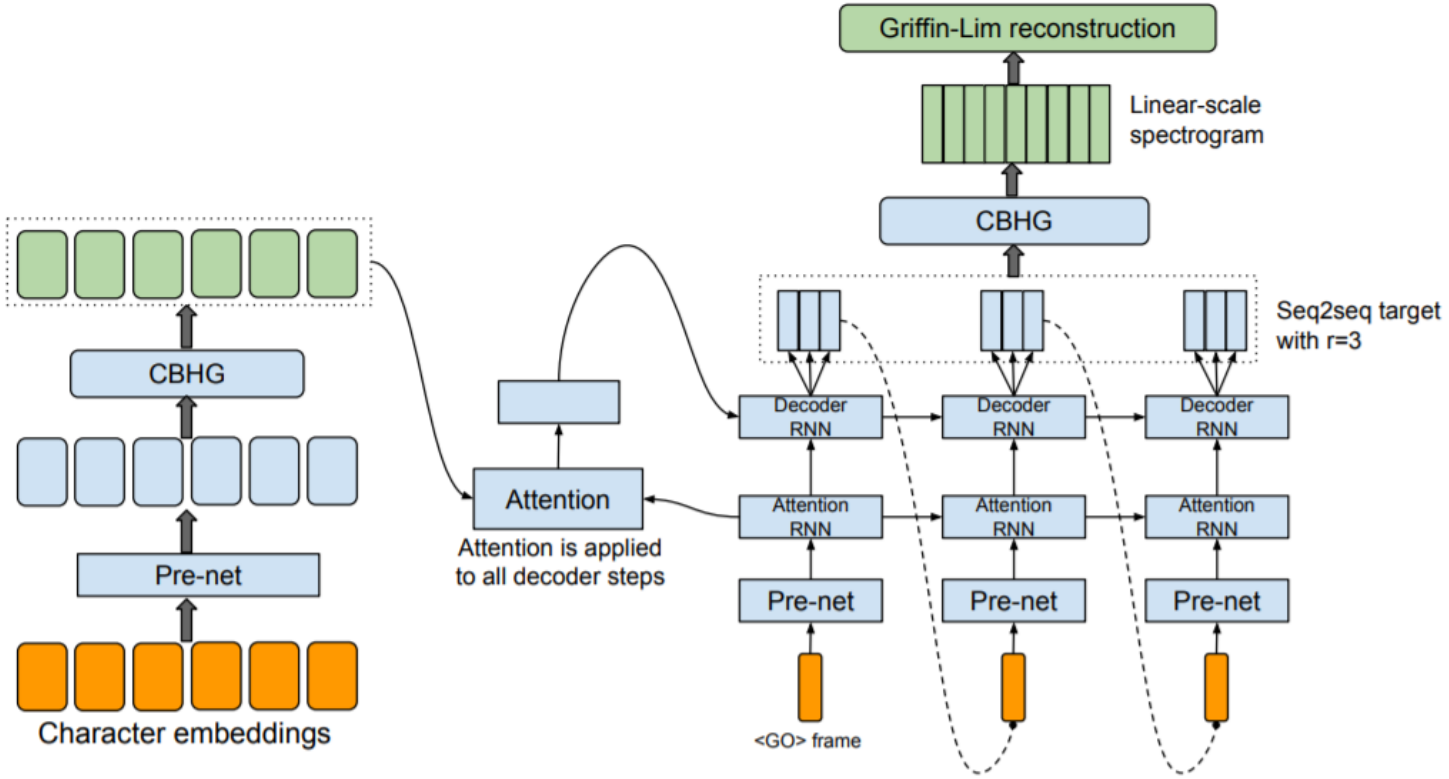
Google, Inc.
{yxwang,rjryan,rif}@google.com

### ABSTRACT

A text-to-speech synthesis system typically consists of multiple stages, such as a text analysis frontend, an acoustic model and an audio synthesis module. Building these components often requires extensive domain expertise and may contain brittle design choices. In this paper, we present Tacotron, an end-to-end generative text-to-speech model that synthesizes speech directly from characters. Given <text, audio> pairs, the model can be trained completely from scratch with random initialization. We present several key techniques to make the sequence-to-sequence framework perform well for this challenging task. Tacotron achieves a 3.82 subjective 5-scale mean opinion score on US English, outperforming a production parametric system in terms of naturalness. In addition, since Tacotron generates speech at the frame level, it's substantially faster than sample-level autoregressive methods.

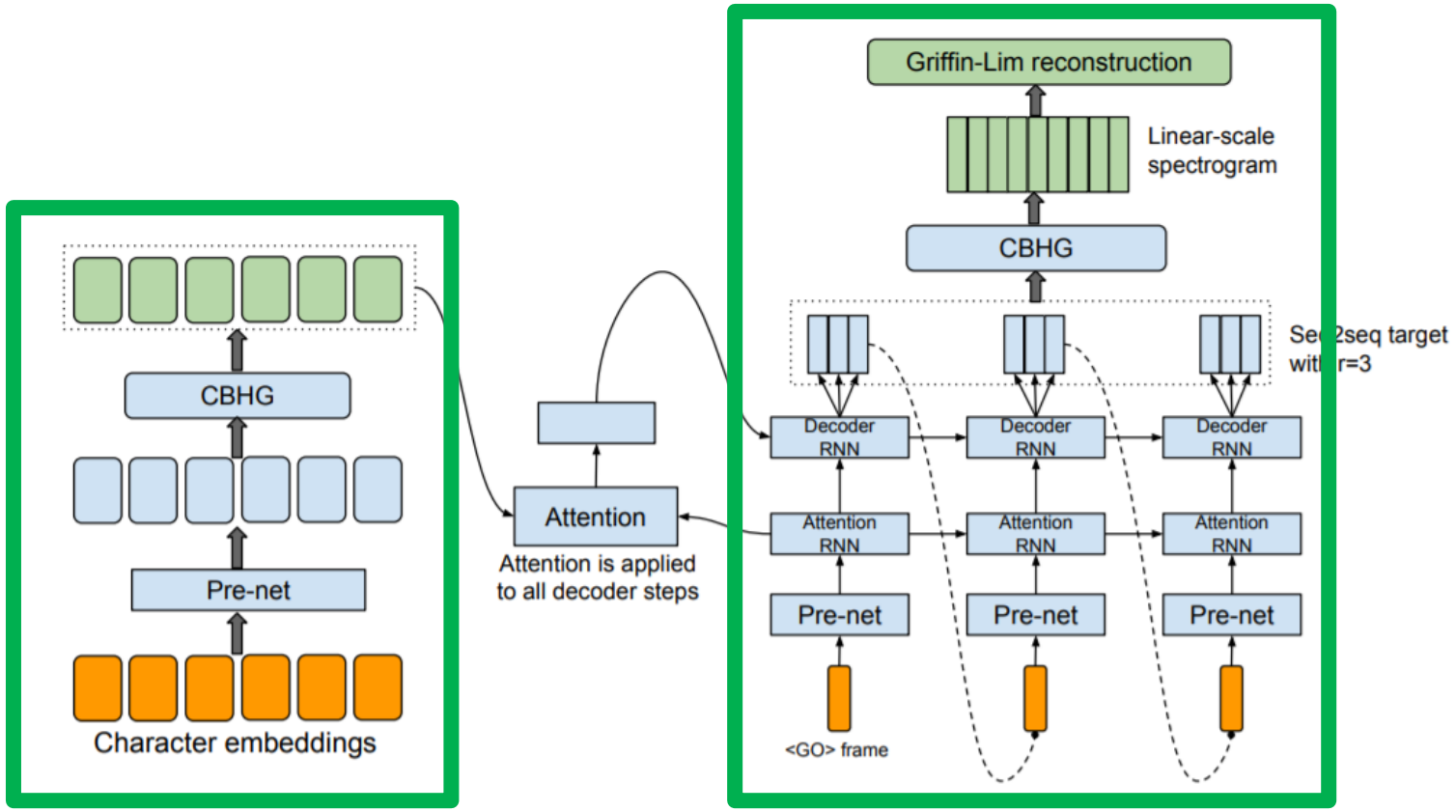
# Tacotron

Google 에서 발표한 대표적인 End-to-end 음성 합성 모델



# Tacotron

Google 에서 발표한 대표적인 End-to-end 음성 합성 모델

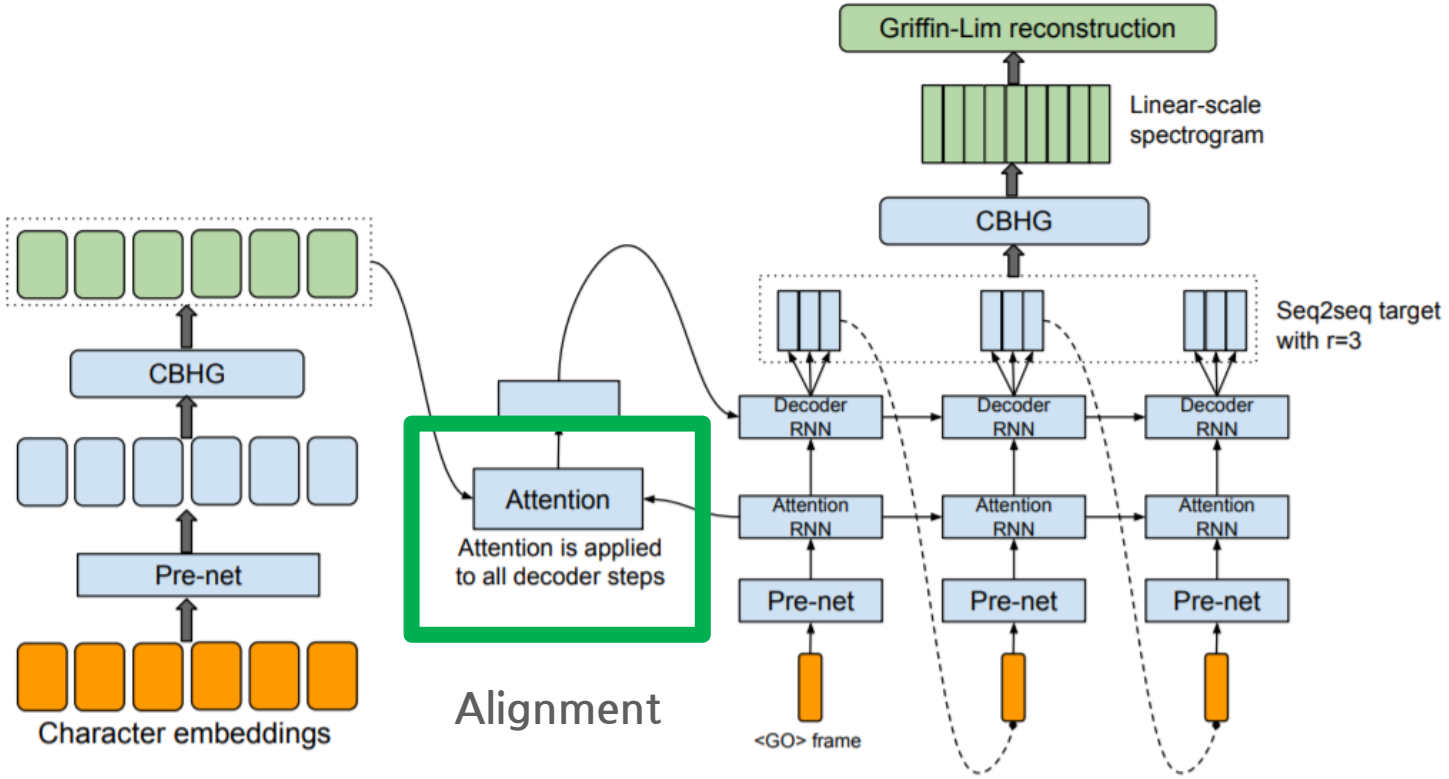



Encoder

Decoder

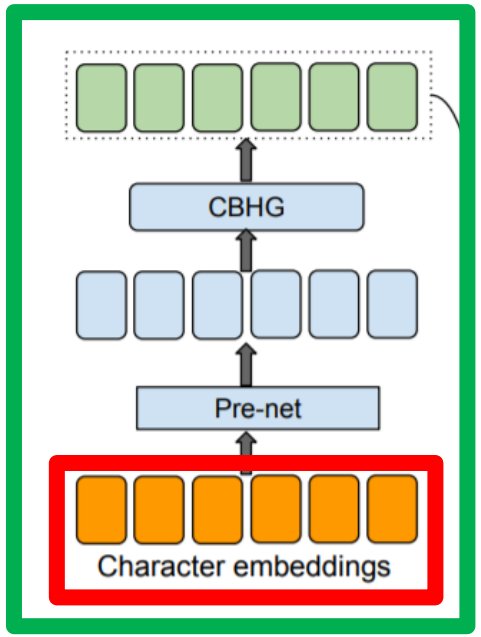
# Tacotron

Google 에서 발표한 대표적인 End-to-end 음성 합성 모델



# Tacotron

Google 에서 발표한 대표적인 End-to-end 음성 합성 모델

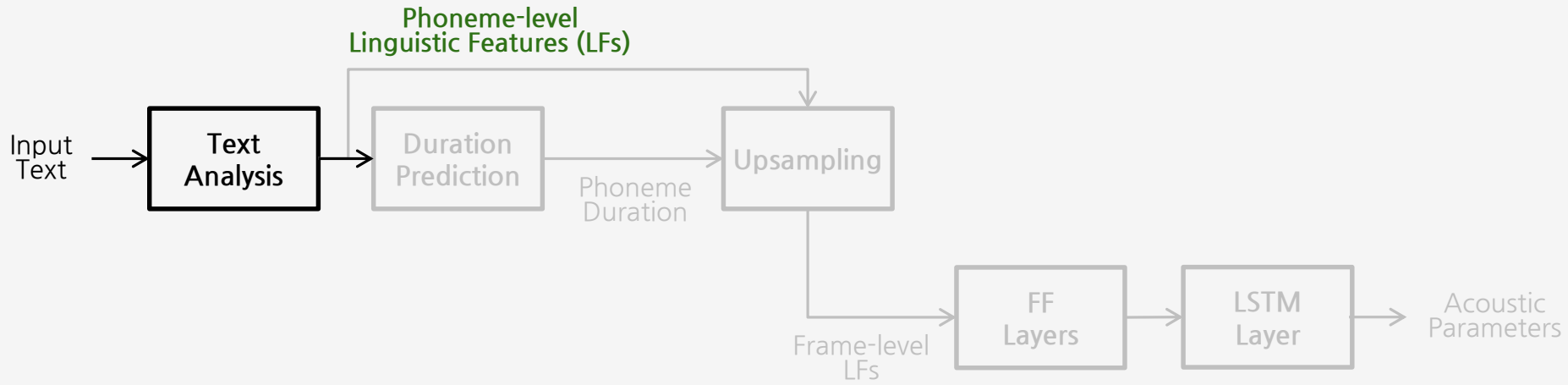



Encoder

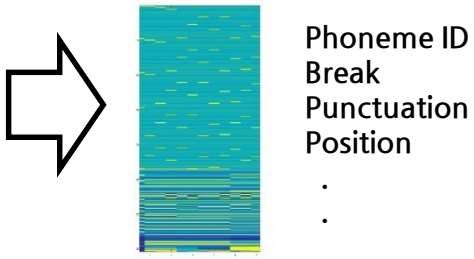
Input: Linguistic feature 가 아닌 character embedding
또는 phoneme

# Recall: Linguistic feature

Text analyzer: Generates phoneme-level linguistic features (Phoneme: 음운론상의 최소 단위)




"안녕하세요"



Phoneme ID
Break
Punctuation
Position
.
.

a-N| n-jv-O| h-a| s-e| jo

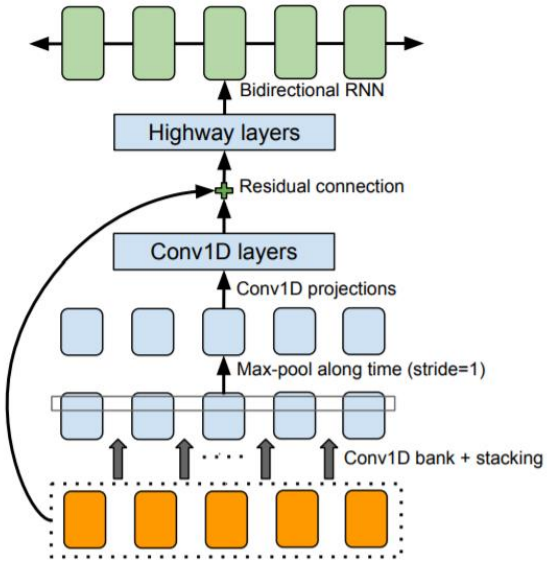
Linguistic features

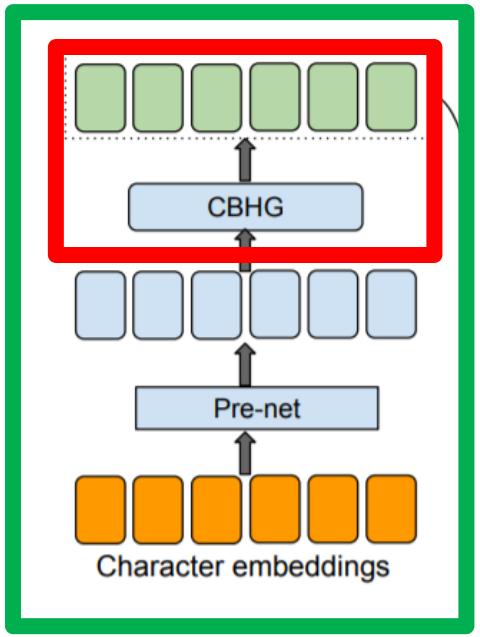
Phoneme sequence

```
WD=[안녕하세요] PR=[a00 NX13 n00 jv00 OX13 h00 a03 s00 e03 jo04] BR=[6] OWD
WD=[눈이] PR=[n00 u03 n00 i04] OWD=[눈이] OPR=[누니] ONPR=[누니] DOM=[0] E
WD=[마주치자] PR=[m00 a03 z00 u03 c00 i03 z00 a04] BR=[6] OWD=[마주치자] OPR
WD=[가쁜] PR=[g00 a03 B00 U00 NX14] OWD=[가쁜] OPR=[가쁜] ONPR=[가쁜] DOM
WD=[숨] PR=[s00 u00 MX14] BR=[3] OWD=[숨] OPR=[숨] ONPR=[숨] DOM=[0] EMO
WD=[사이로] PR=[s00 a03 i03 r00 o04] OWD=[사이로] OPR=[사이로] ONPR=[사이로]
WD=[미소] PR=[m00 i03 s00 o04] OWD=[미소] OPR=[미소] ONPR=[미소] DOM=[0] E
WD=[섞인] PR=[s00 v03 G00 i04] BR=[3] OWD=[섞인] OPR=[서끼] ONPR=[서끼] DOM
WD=[인사가] PR=[n00 i00 NX13 s00 a03 g00 a04] OWD=[인사가] OPR=[닌사가] ONP
WD=[배어] PR=[b00 e03 v04] OWD=[배어] OPR=[배어] ONPR=[베어] DOM=[0] EMO=
WD=[나온다] PR=[n00 a03 o00 NX13 d00 a04] PUNCT=[.] BR=[7] OWD=[나온다.] OP
```

# Tacotron

Google 에서 발표한 대표적인 End-to-end 음성 합성 모델




Encoder

Input: Linguistic feature 가 아닌 character embedding

또는 phoneme

대신 CBHG 모듈을 통해 high-level context feature 를 얻어낼 수 있음

Conv 1D bank + Highway network + GRU

Conv 1D bank: 다양한 커널 사이즈를 갖는 Conv 1D 집합

Highway network: Residual connection 의 비율 조절

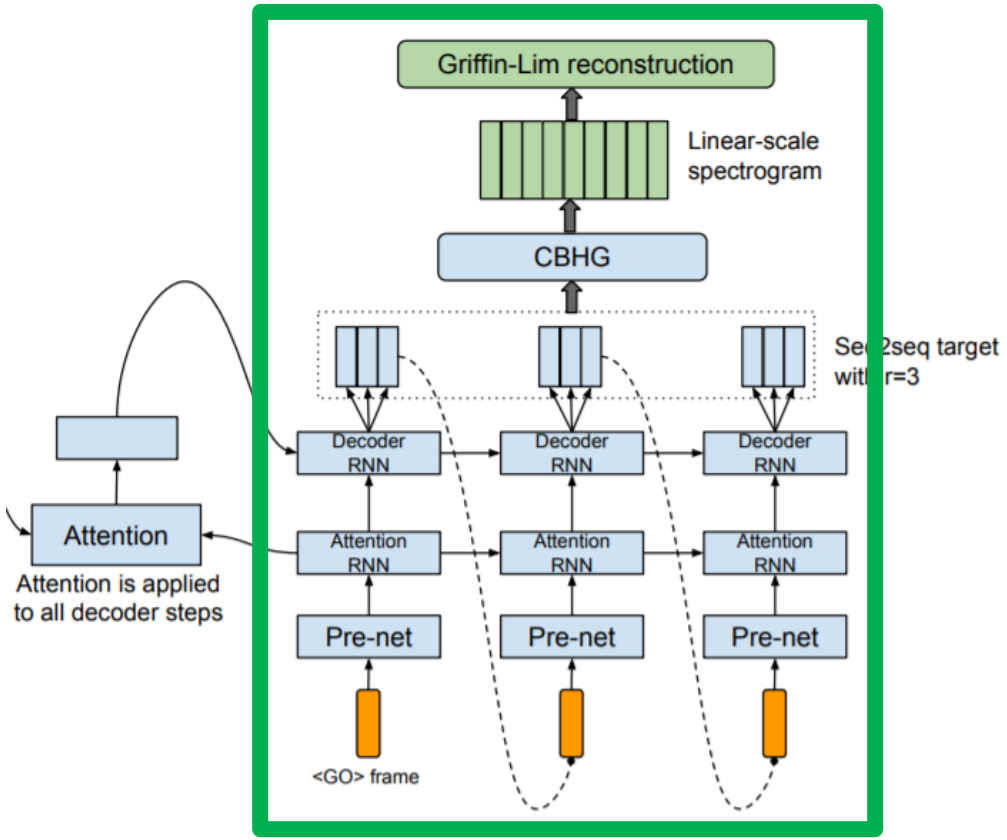
$$\mathbf{y} = H(\mathbf{x}, \mathbf{W_H}) \cdot T(\mathbf{x}, \mathbf{W_T}) + \mathbf{x} \cdot (1 - T(\mathbf{x}, \mathbf{W_T})).$$

# Tacotron

Google 에서 발표한 대표적인 End-to-end 음성 합성 모델

## Autoregressive decoder: 합성음 품질을 높임

현재 frame 을 디코딩 할 때, 이전 frame 의 정보를 사용




Decoder

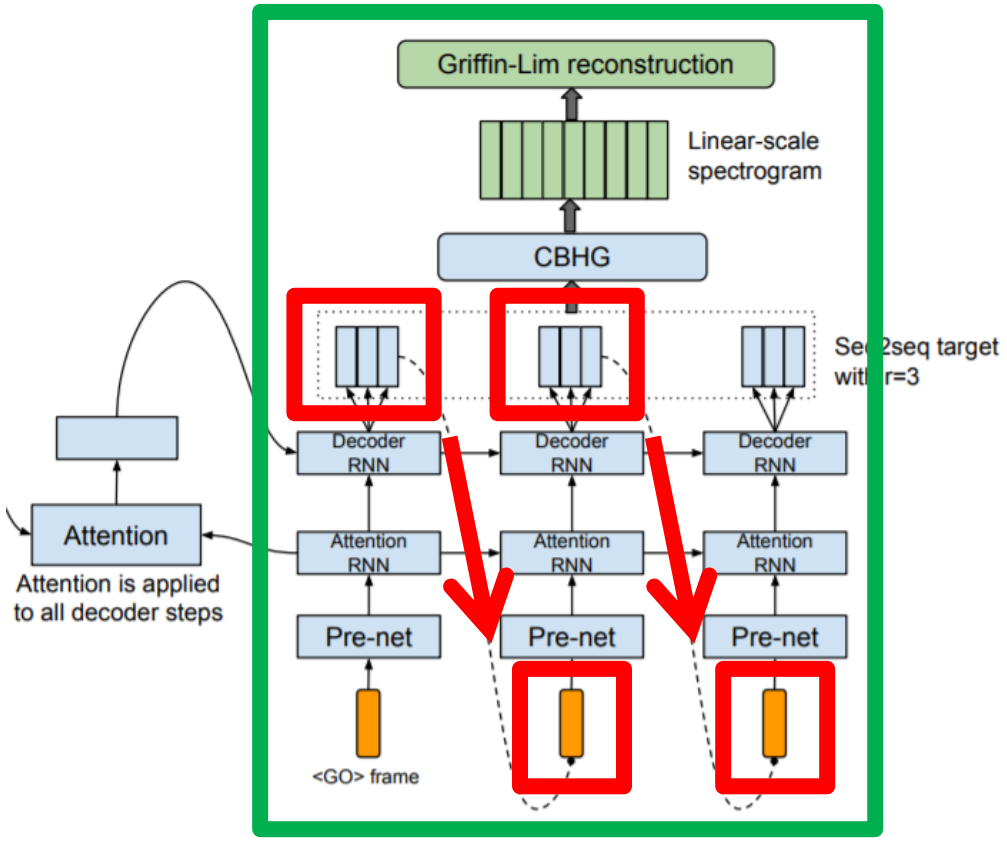
CLOVA

# Tacotron

Google 에서 발표한 대표적인 End-to-end 음성 합성 모델

## Autoregressive decoder: 합성음 품질을 높임

현재 frame 을 디코딩 할 때, 이전 frame 의 정보를 사용




Decoder

# Tacotron

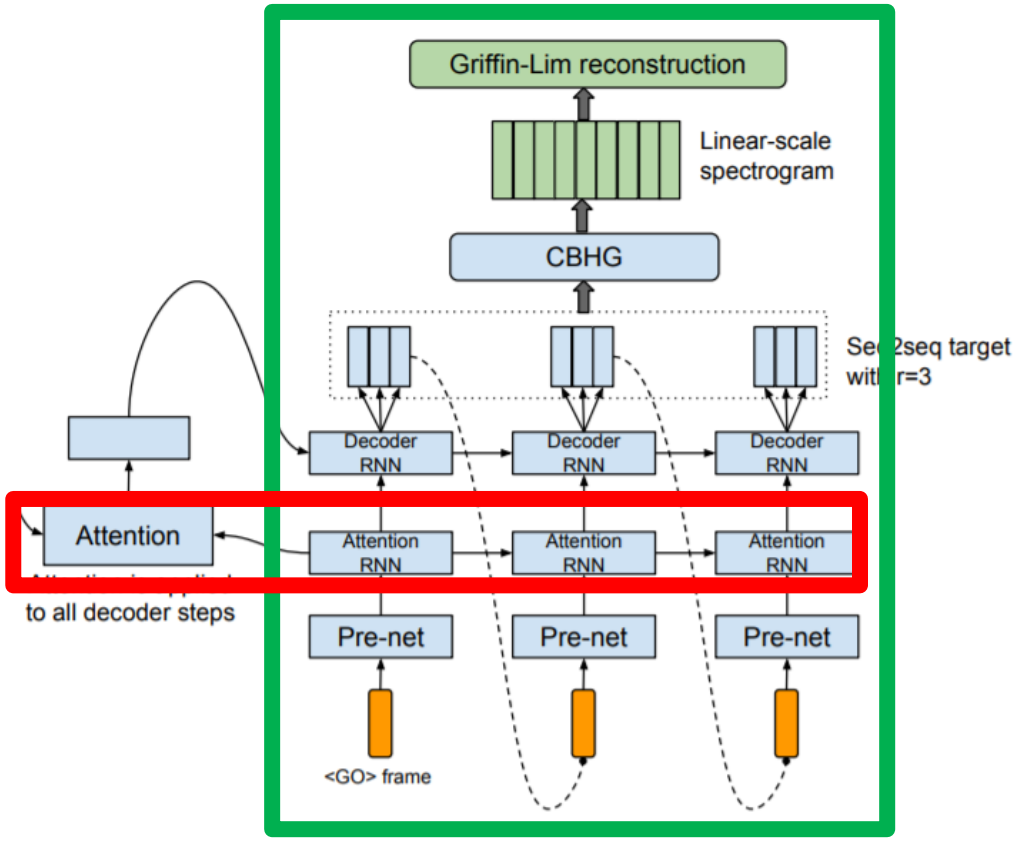
Google 에서 발표한 대표적인 End-to-end 음성 합성 모델

## Autoregressive decoder: 합성음 품질을 높임

현재 frame 을 디코딩 할 때, 이전 frame 의 정보를 사용

## Attention: Text-Mel alignment

Duration model 역할을 attention 으로 대체




Decoder

# Tacotron

Google 에서 발표한 대표적인 End-to-end 음성 합성 모델

## Autoregressive decoder: 합성음 품질을 높임

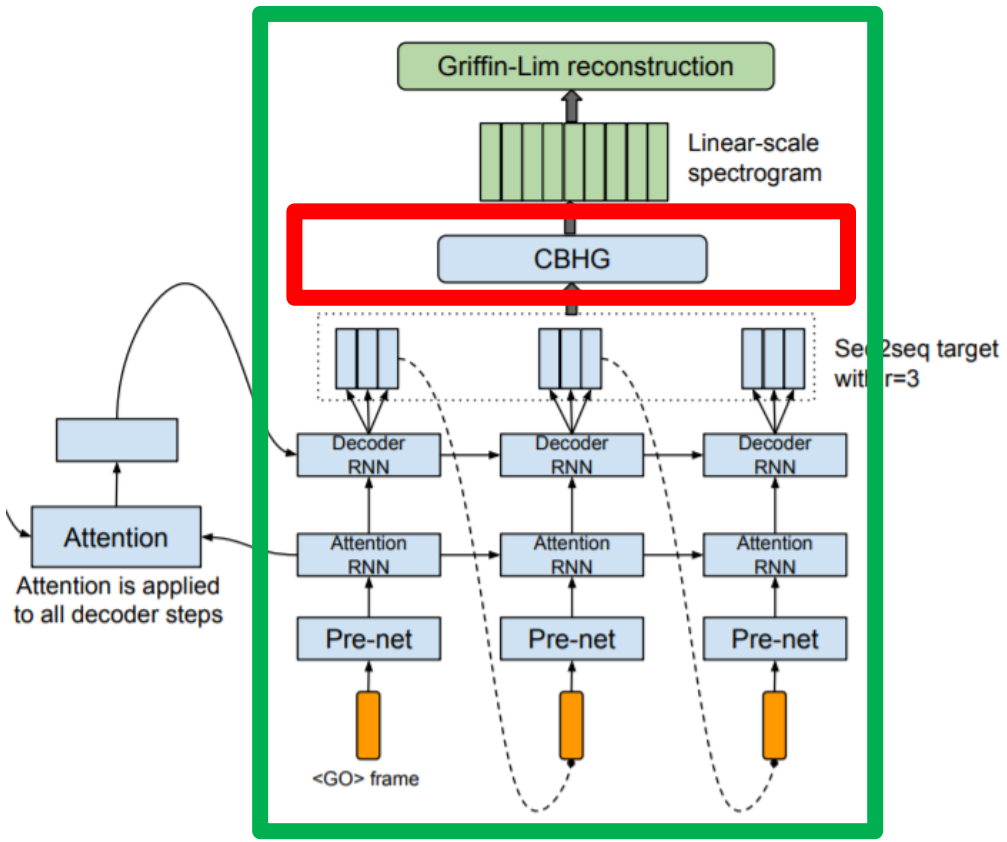
현재 frame 을 디코딩 할 때, 이전 frame 의 정보를 사용

## Attention: Text-Mel alignment

Duration model 역할을 attention 으로 대체

## CBHG: Mel-to-linear conversion

Neural 보코더가 없던 시절 → Griffin-Lim 적용을 위한 모듈




Decoder

# Tacotron

Google 에서 발표한 대표적인 End-to-end 음성 합성 모델

Table 2: 5-scale mean opinion score evaluation.

| | mean opinion score |
|---|---|
| Tacotron | $3.82 \pm 0.085$ |
| Parametric | $3.69 \pm 0.109$ |
| Concatenative | $4.09 \pm 0.119$ |

MOS test: 합성음을 듣고 1점(듣기싫은) ~ 5점(사람같은) 사이의 점수를 주는 방법

Concatenative: Unit 접합 방식의 TTS, 품질이 좋은 방면, 많은 양의 녹음 필요

Phoneme segmentation, etc

복잡한 feature engineering 을 최소화 하면서도,

기존 parametric 방식 보다 좋은 품질의 합성음을 만들 수 있게 되었습니다.

https://google.github.io/tacotron/publications/tacotron/index.html

CLOVA

# Text-to-speech

## End-to-end speech synthesis

Tacotron

Tacotron 2

FastSpeech

FastSpeech 2

# Tacotron 2

Google 에서 발표한 두 번째 End-to-end 음성 합성 모델

## NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS

*Jonathan Shen*[1], *Ruoming Pang*[1], *Ron J. Weiss*[1], *Mike Schuster*[1], *Navdeep Jaitly*[1], *Zongheng Yang*[*2], *Zhifeng Chen*[1], *Yu Zhang*[1], *Yuxuan Wang*[1], *RJ Skerry-Ryan*[1], *Rif A. Saurous*[1], *Yannis Agiomyrgiannakis*[1], *and Yonghui Wu*[1]

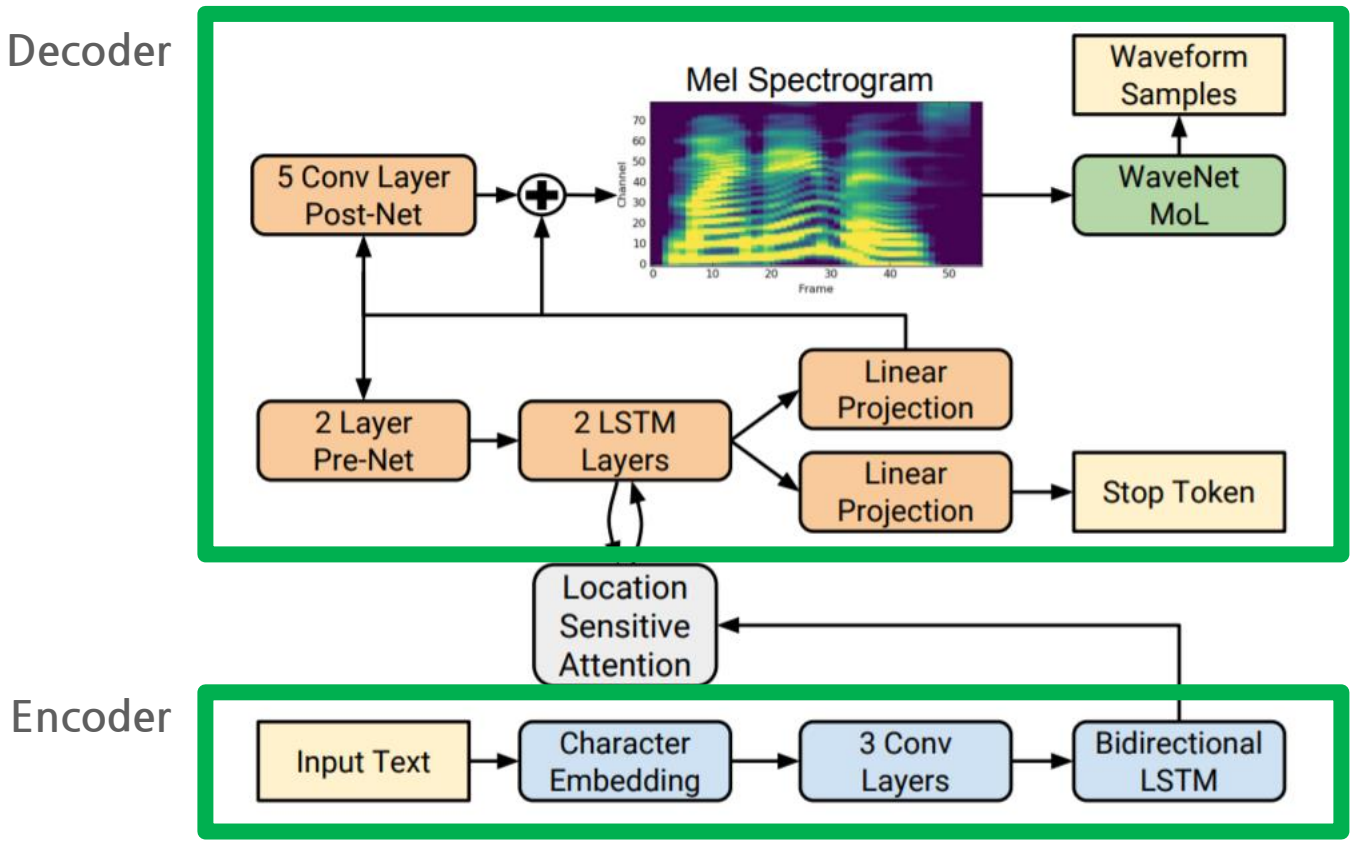
[1]Google, Inc., [2]University of California, Berkeley,
{jonathanasdf,rpang,yonghui}@google.com

**ABSTRACT**

This paper describes Tacotron 2, a neural network architecture for speech synthesis directly from text. The system is composed of a recurrent sequence-to-sequence feature prediction network that maps character embeddings to mel-scale spectrograms, followed by a modified WaveNet model acting as a vocoder to synthesize time-domain waveforms from those spectrograms. Our model achieves a mean opinion score (MOS) of 4.53 comparable to a MOS of 4.58 for professionally recorded speech. To validate our design choices, we present ablation studies of key components of our system and evaluate the impact of using mel spectrograms as the conditioning input to WaveNet instead of linguistic, duration, and $F_0$ features. We further show that using this compact acoustic intermediate representation allows for a significant reduction in the size of the WaveNet architecture.

# Tacotron 2

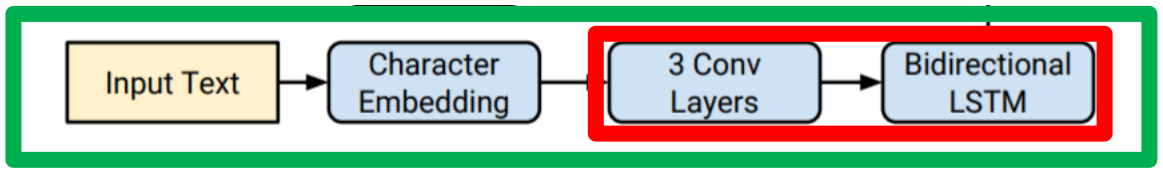
Google 에서 발표한 두 번째 End-to-end 음성 합성 모델



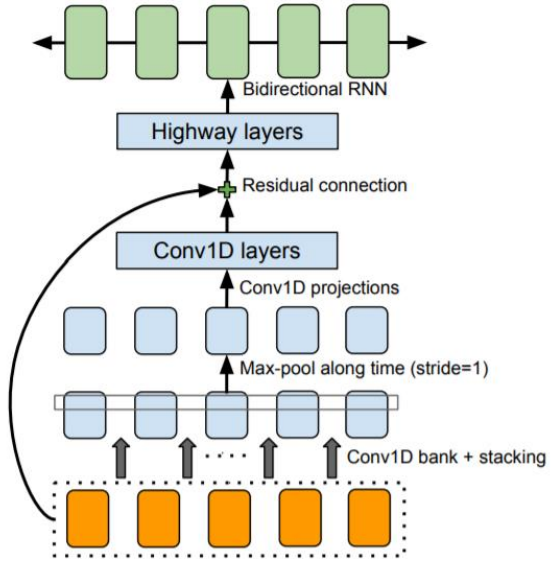
# Tacotron 2

Google 에서 발표한 두 번째 End-to-end 음성 합성 모델

Encoder




Conv + BLSTM: Tacotron 1 보다 간단한 구조

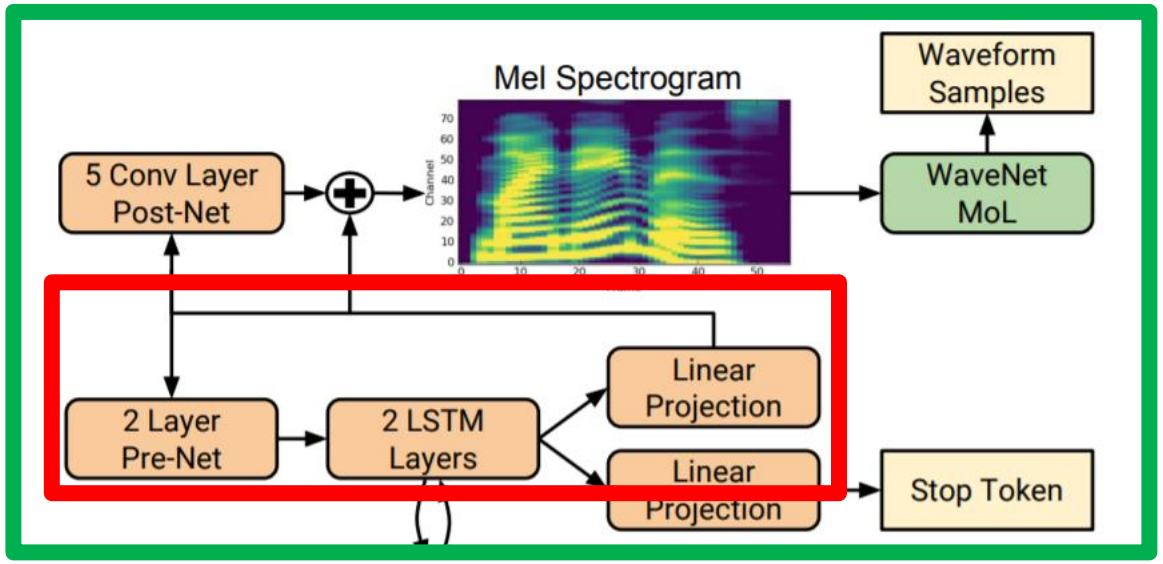



CBHG

# Tacotron 2

Google 에서 발표한 두 번째 End-to-end 음성 합성 모델

Decoder

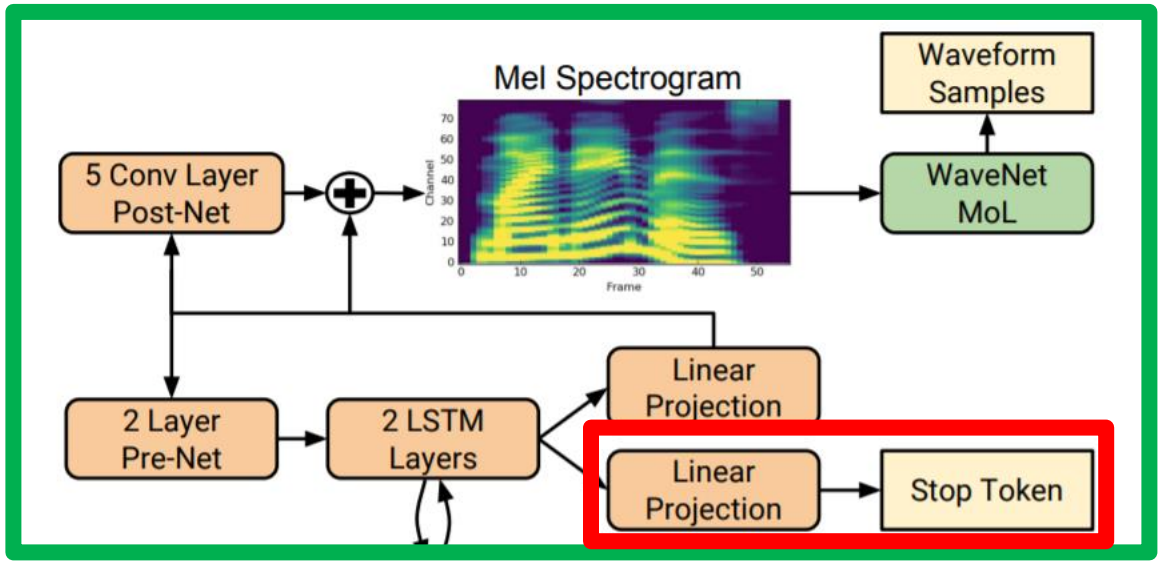



Autoregressive decoder: 합성음 품질을 높임

# Tacotron 2

Google 에서 발표한 두 번째 End-to-end 음성 합성 모델

Decoder




Autoregressive decoder: 합성음 품질을 높임

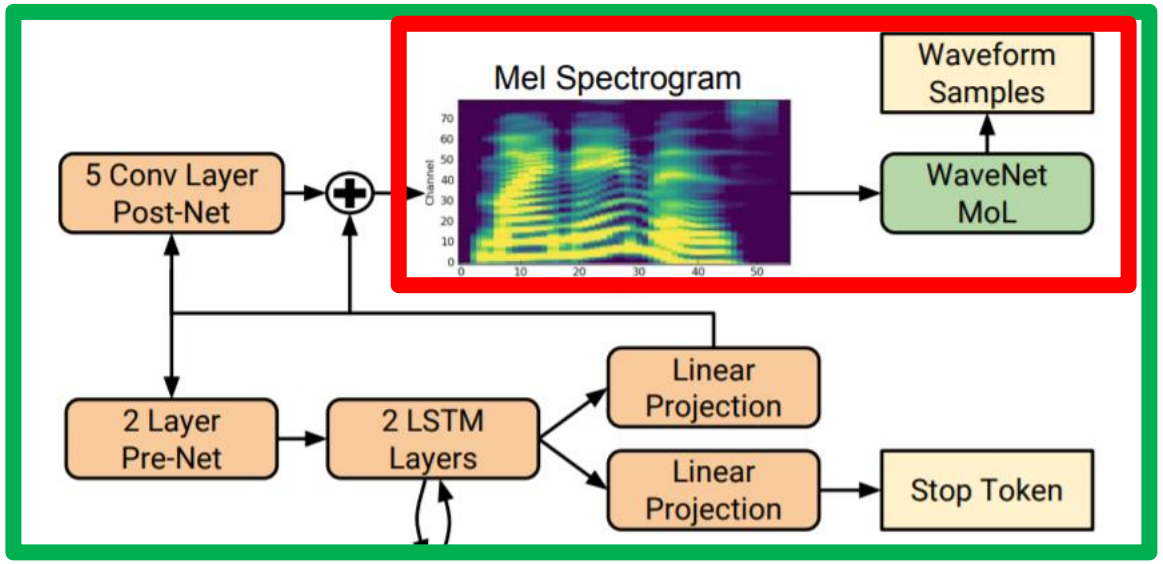
Stop token: 발화의 종료 시점을 추정할 수 있음

cf. Tacotron 1: 발화 종료와 상관 없이 일정 길이 만큼 음성을 생성해야 했음

# Tacotron 2

Google 에서 발표한 두 번째 End-to-end 음성 합성 모델

Decoder

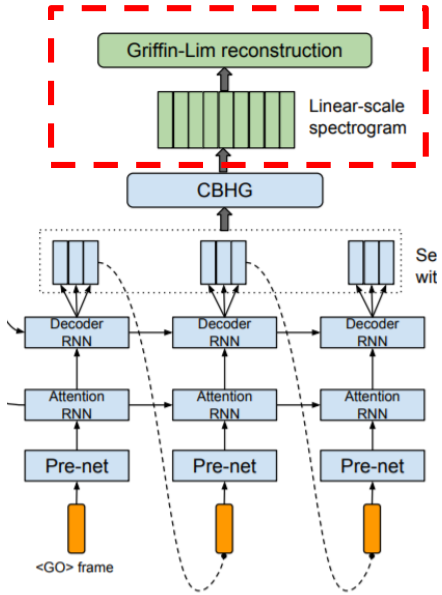



Autoregressive decoder: 합성음 품질을 높임

Stop token: 발화의 종료 시점을 추정할 수 있음

WaveNet 보코더: 합성음 품질을 더 더 더욱 높임

Neural TTS 패러다임을 이끌어낸 주인공 (?)




Tacotron 1 decoder

# Tacotron 2

Google 에서 발표한 두 번째 End-to-end 음성 합성 모델

| System | MOS |
|---|---|
| Parametric | $3.492 \pm 0.096$ |
| Tacotron (Griffin-Lim) | $4.001 \pm 0.087$ |
| Concatenative | $4.166 \pm 0.091$ |
| WaveNet (Linguistic) | $4.341 \pm 0.051$ |
| Ground truth | $4.582 \pm 0.053$ |
| Tacotron 2 (this paper) | $\mathbf{4.526 \pm 0.066}$ |

End-to-end acoustic model + WaveNet vocoder

당시 최고 합성 모델인 Concatenative 보다 우수한, 녹음에 가까운 수준의 음성 합성 모델

https://ai.googleblog.com/2017/12/tacotron-2-generating-human-like-speech.html

# Summary

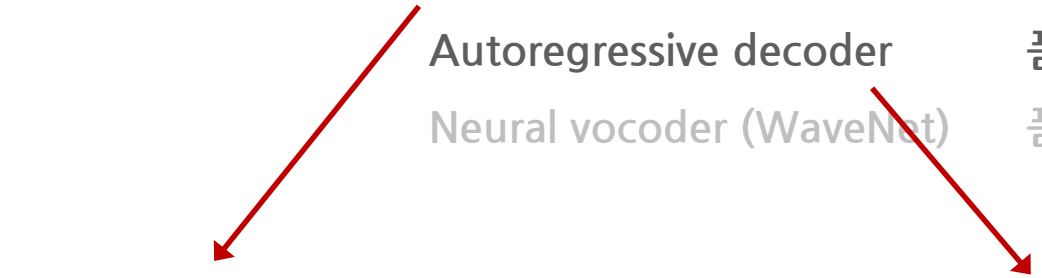
Tacotron 1, 2

| | |
|---|---|
| Seq2seq + attention | 복잡한 feature engineering 을 최소화 |
| Autoregressive decoder | 품질 향상 |
| Neural vocoder (WaveNet) | 품질 향상 |

# Summary

Tacotron 1, 2

| | |
|---|---|
| Seq2seq + attention | 복잡한 feature engineering 을 최소화 |
| Autoregressive decoder | 품질 향상 |
| Neural vocoder (WaveNet) | 품질 향상 |



Alignment failure 문제에 취약함

합성 속도 느림

CLOVA

# Text-to-speech

## End-to-end speech synthesis

Tacotron

Tacotron 2

FastSpeech

FastSpeech 2

# FastSpeech

MS 에서 발표한 대표적인 End-to-end 음성 합성 모델

## FastSpeech: Fast, Robust and Controllable Text to Speech

**Yi Ren**[*]
Zhejiang University
rayeren@zju.edu.cn

**Yangjun Ruan**[*]
Zhejiang University
ruanyj3107@zju.edu.cn

**Xu Tan**
Microsoft Research
xuta@microsoft.com

**Tao Qin**
Microsoft Research
taoqin@microsoft.com

**Sheng Zhao**
Microsoft STC Asia
Sheng.Zhao@microsoft.com

**Zhou Zhao**[†]
Zhejiang University
zhaozhou@zju.edu.cn

**Tie-Yan Liu**
Microsoft Research
tyliu@microsoft.com

**Abstract**

Neural network based end-to-end text to speech (TTS) has significantly improved the quality of synthesized speech. Prominent methods (e.g., Tacotron 2) usually first generate mel-spectrogram from text, and then synthesize speech from the mel-spectrogram using vocoder such as WaveNet. Compared with traditional concatenative and statistical parametric approaches, neural network based end-to-end models suffer from slow inference speed, and the synthesized speech is usually not robust (i.e., some words are skipped or repeated) and lack of controllability (voice speed or prosody control). In this work, we propose a novel feed-forward network based on Transformer to generate mel-spectrogram in parallel for TTS. Specifically, we extract attention alignments from an encoder-decoder based teacher model for phoneme duration prediction, which is used by a length regulator to expand the source phoneme sequence to match the length of the target mel-spectrogram sequence for parallel mel-spectrogram generation. Experiments on the LJSpeech dataset show that our parallel model matches autoregressive models in terms of speech quality, nearly eliminates the problem of word skipping and repeating in particularly hard cases, and can adjust voice speed smoothly. Most importantly, compared with autoregressive Transformer TTS, our model speeds up mel-spectrogram generation by 270x and the end-to-end speech synthesis by 38x. Therefore, we call our model FastSpeech. [3]

# FastSpeech

MS 에서 발표한 대표적인 End-to-end 음성 합성 모델

## FastSpeech: Fast, Robust and Controllable Text to Speech

**Yi Ren**[*]
Zhejiang University
rayeren@zju.edu.cn

**Yangjun Ruan**[*]
Zhejiang University
ruanyj3107@zju.edu.cn

**Xu Tan**
Microsoft Research
xuta@microsoft.com

**Tao Qin**
Microsoft Research
taoqin@microsoft.com

**Sheng Zhao**
Microsoft STC Asia
Sheng.Zhao@microsoft.com

**Zhou Zhao**[†]
Zhejiang University
zhaozhou@zju.edu.cn

**Tie-Yan Liu**
Microsoft Research
tyliu@microsoft.com

**Abstract**

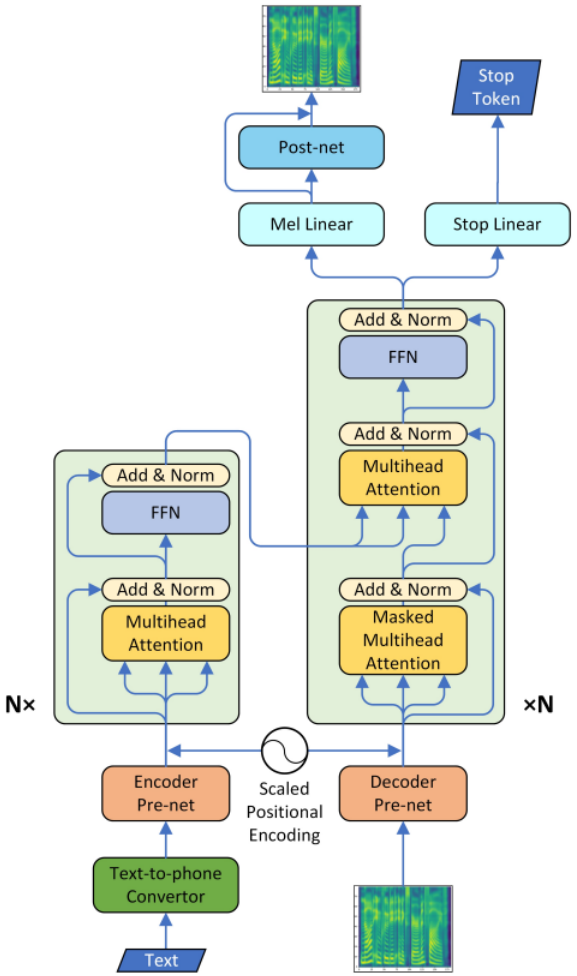
Neural network based end-to-end text to speech (TTS) has significantly improved the quality of synthesized speech. Prominent methods (e.g., Tacotron 2) usually first generate mel-spectrogram from text, and then synthesize speech from the mel-spectrogram using vocoder such as WaveNet. Compared with traditional concatenative and statistical parametric approaches, neural network based end-to-end models suffer from slow inference speed, and the synthesized speech is usually not robust (i.e., some words are skipped or repeated) and lack of controllability (voice speed or prosody control). In this work, we propose a novel feed-forward network based on Transformer to generate mel-spectrogram in parallel for TTS. Specifically, we extract attention alignments from an encoder-decoder based teacher model for phoneme duration prediction, which is used by a length regulator to expand the source phoneme sequence to match the length of the target mel-spectrogram sequence for parallel mel-spectrogram generation. Experiments on the LJSpeech dataset show that our parallel model matches autoregressive models in terms of speech quality, nearly eliminates the problem of word skipping and repeating in particularly hard cases, and can adjust voice speed smoothly. Most importantly, compared with autoregressive Transformer TTS, our model speeds up mel-spectrogram generation by 270x and the end-to-end speech synthesis by 38x. Therefore, we call our model FastSpeech. [3]

Autoregressive Tacotron 모델이 갖고 있던 문제점을 극복하기 위한

1. 빠른 inference 방법과
2. Alignment failure 줄이는 방법을 제안

# 참고: Transformer TTS

MS 에서 발표한 대표적인 End-to-end 음성 합성 모델




**Neural Speech Synthesis with Transformer Network**

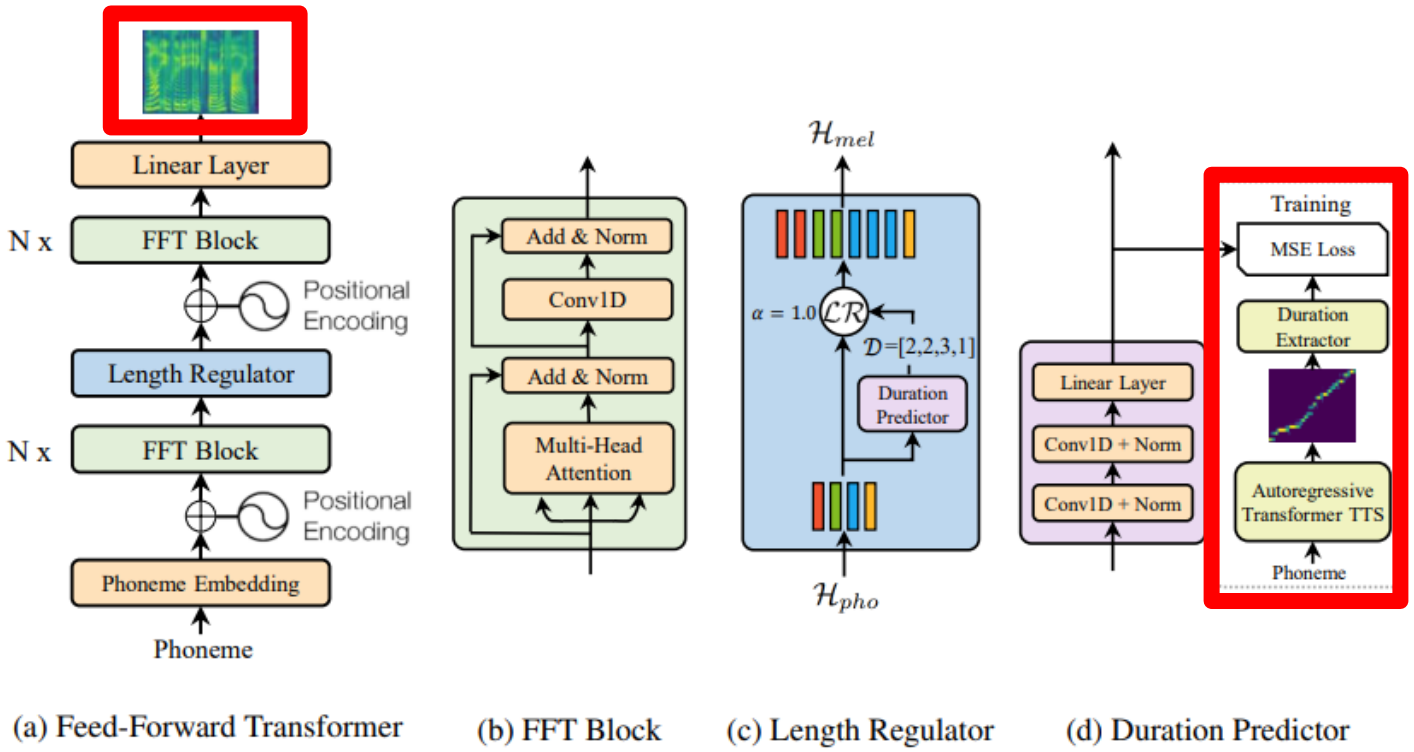
**Naihan Li*[1,4], Shujie Liu[2], Yanqing Liu[3], Sheng Zhao[3], Ming Liu[1,4], Ming Zhou[2]**
[1]University of Electronic Science and Technology of China
[2]Microsoft Research Asia
[3]Microsoft STC Asia
[4]CETC Big Data Research Institute Co.,Ltd, Guizhou, China
lnhzsbls1994@163.com
{shujliu,yanqliu,szhao,mingzhou}@microsoft.com
csmliu@uestc.edu.cn

Tacotron 의 RNN 구조를 Transformer 형태로 변경

1. Training 병렬화 (efficiency ↑)
2. Long-term dependency problem 해결

# FastSpeech

MS 에서 발표한 대표적인 End-to-end 음성 합성 모델




(a) Feed-Forward Transformer (b) FFT Block (c) Length Regulator (d) Duration Predictor

Knowledge distillation from teacher (AR Transformer) to FastSpeech (non-AR Transformer)

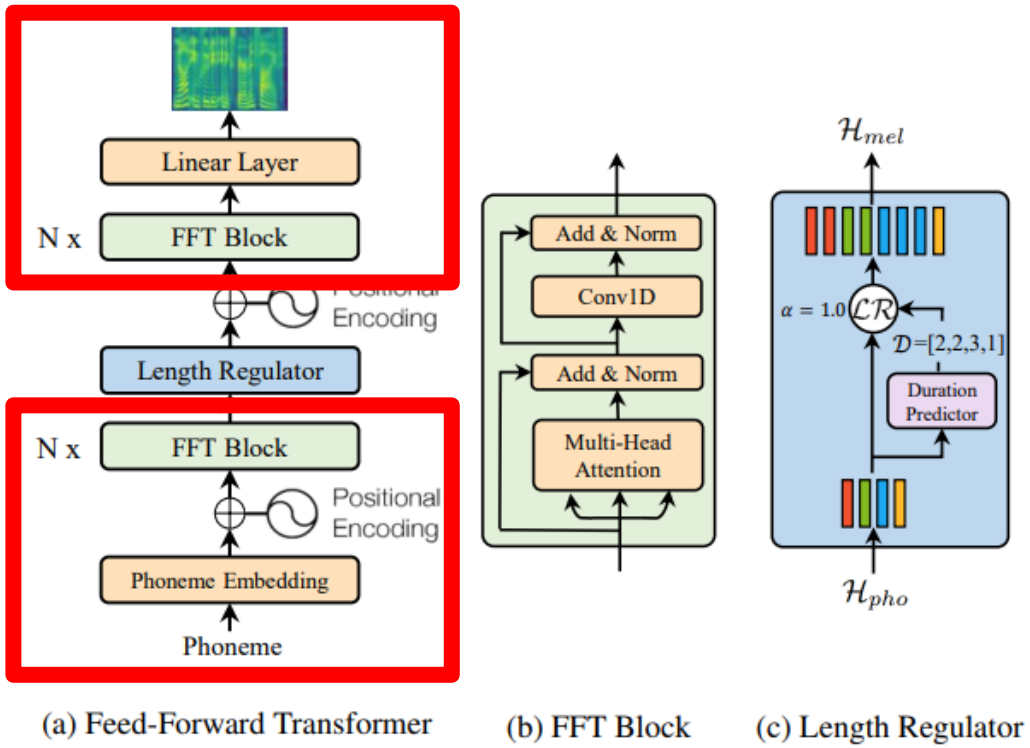
AR: Autoregressive

# FastSpeech

MS 에서 발표한 대표적인 End-to-end 음성 합성 모델

Decoder

Encoder



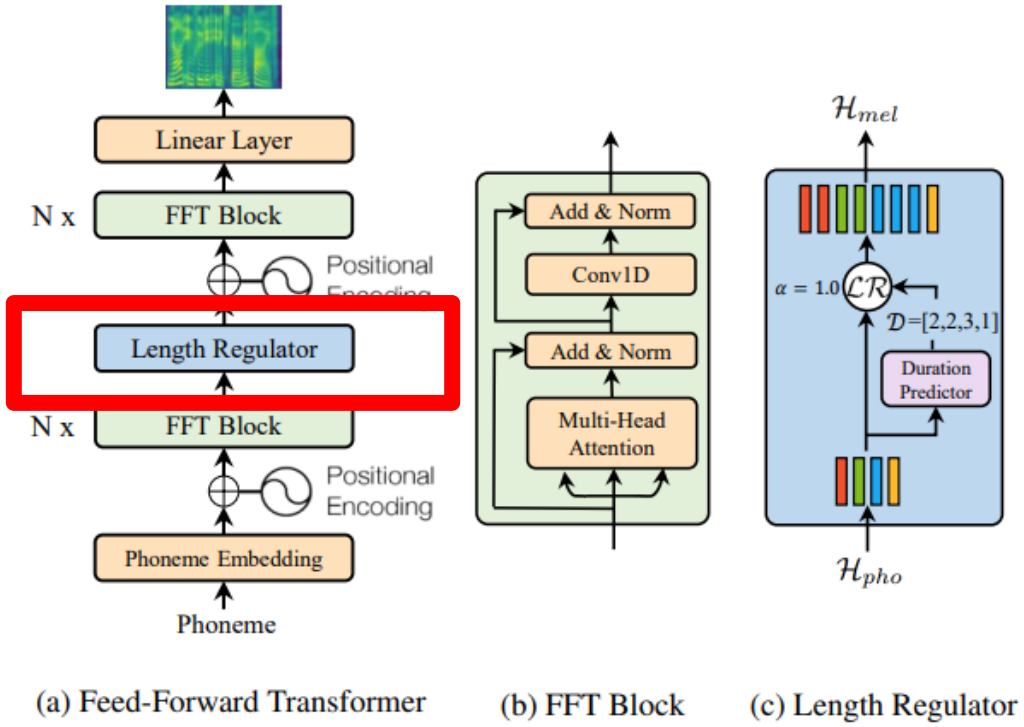

(a) Feed-Forward Transformer (b) FFT Block (c) Length Regulator

Transformer 기반의 encoder-decoder model

Input 안정성: Phoneme > character

# FastSpeech

MS 에서 발표한 대표적인 End-to-end 음성 합성 모델




(a) Feed-Forward Transformer (b) FFT Block (c) Length Regulator

## Transformer 기반의 encoder-decoder model

Input 안정성: Phoneme > character

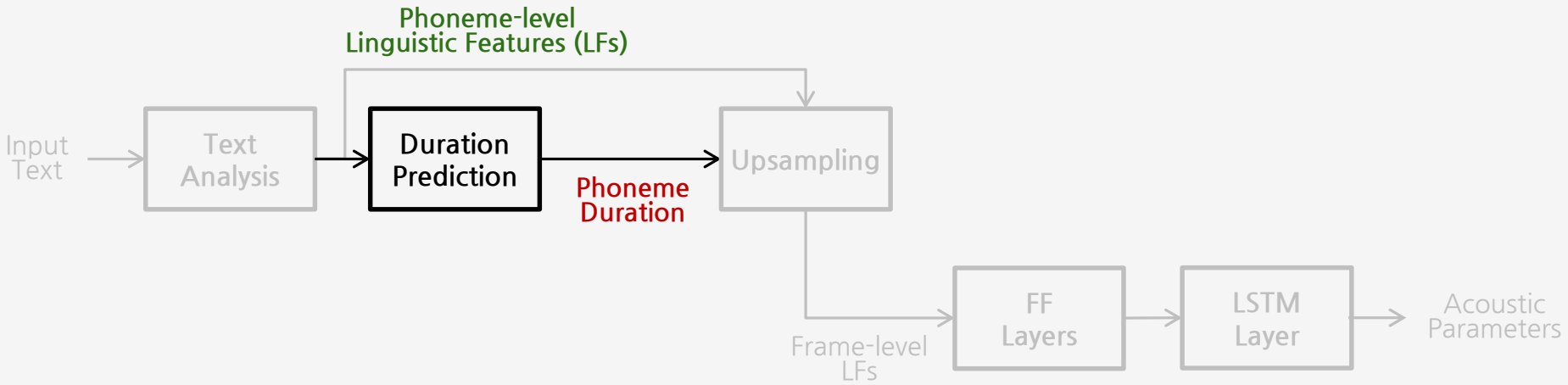
## Length regulator: Alignment 안정성 확보

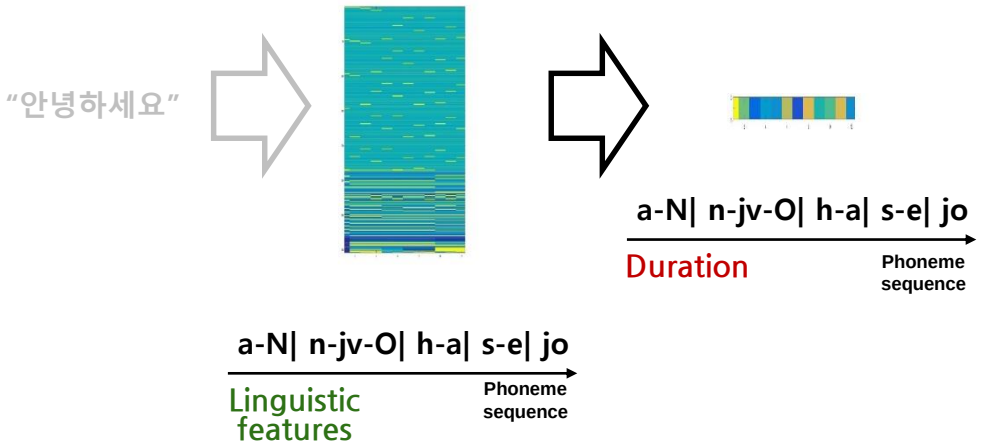
Phoneme 단위로 duration 을 구하고, decoder 맞게 upsampling

1. Decoder 에서 한번에 병렬처리 가능
2. Skipping, repeating 이슈 해결
3. Phoneme 단위로 발화 속도 조절 가능

# Statistical parametric speech synthesis (SPSS)

Duration model: Predicts phoneme duration

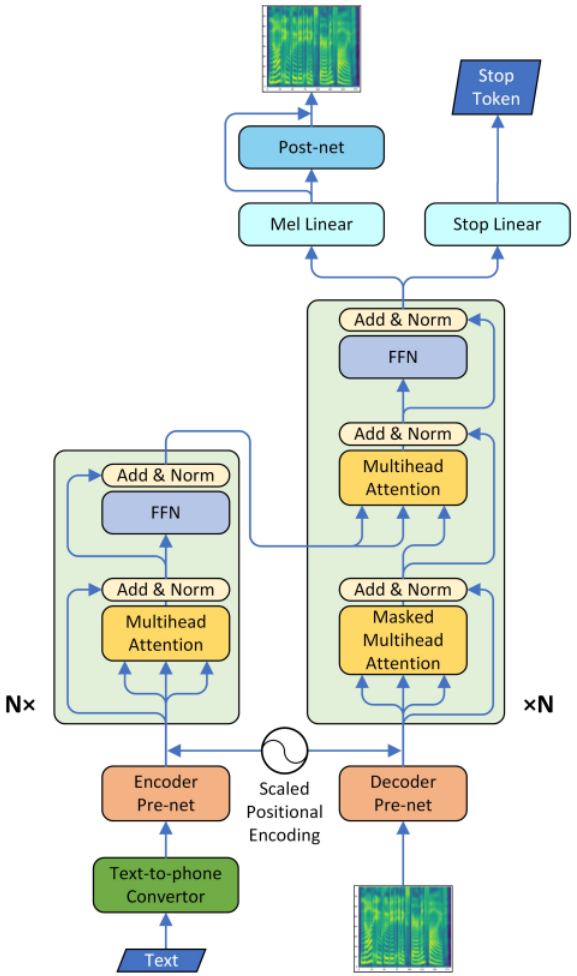



한계점: Phoneme segmentation 을 위한 **비용**이 많이 든다

# 참고: Transformer TTS

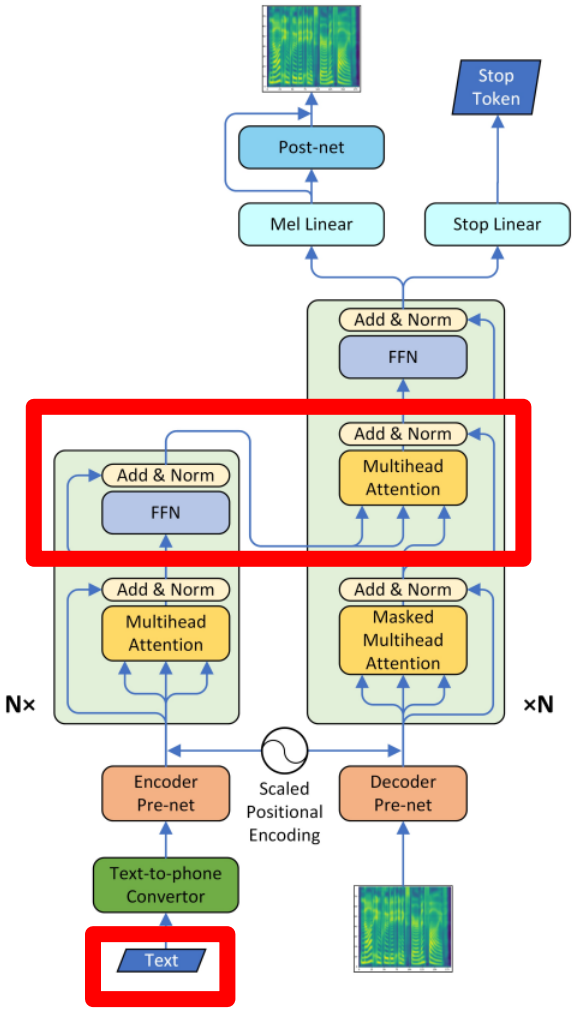
MS 에서 발표한 대표적인 End-to-end 음성 합성 모델




1. Teacher (AR Transformer) TTS 학습

# 참고: Transformer TTS

MS 에서 발표한 대표적인 End-to-end 음성 합성 모델



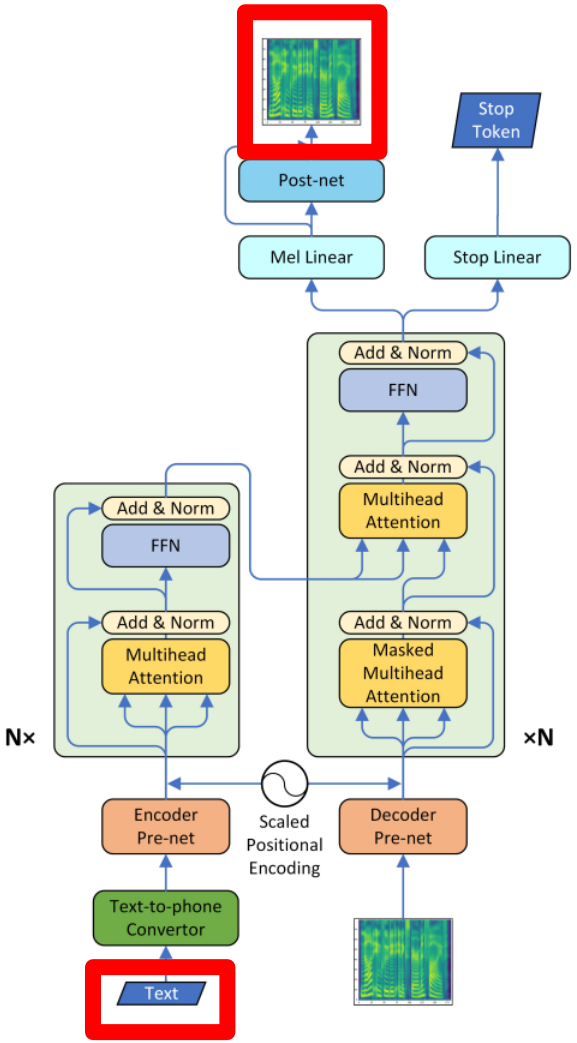

1. Teacher (AR Transformer) TTS 학습

2. Input (text) - output (mel) 간의 attention map 구하고

Phoneme duration 추출하는 역할 → FastSpeech 학습에 사용

# 참고: Transformer TTS

MS 에서 발표한 대표적인 End-to-end 음성 합성 모델




1. Teacher (AR Transformer) TTS 학습

2. Input (text) - output (mel) 간의 attention map 구하고

   Phoneme duration 추출하는 역할 → FastSpeech 학습에 사용

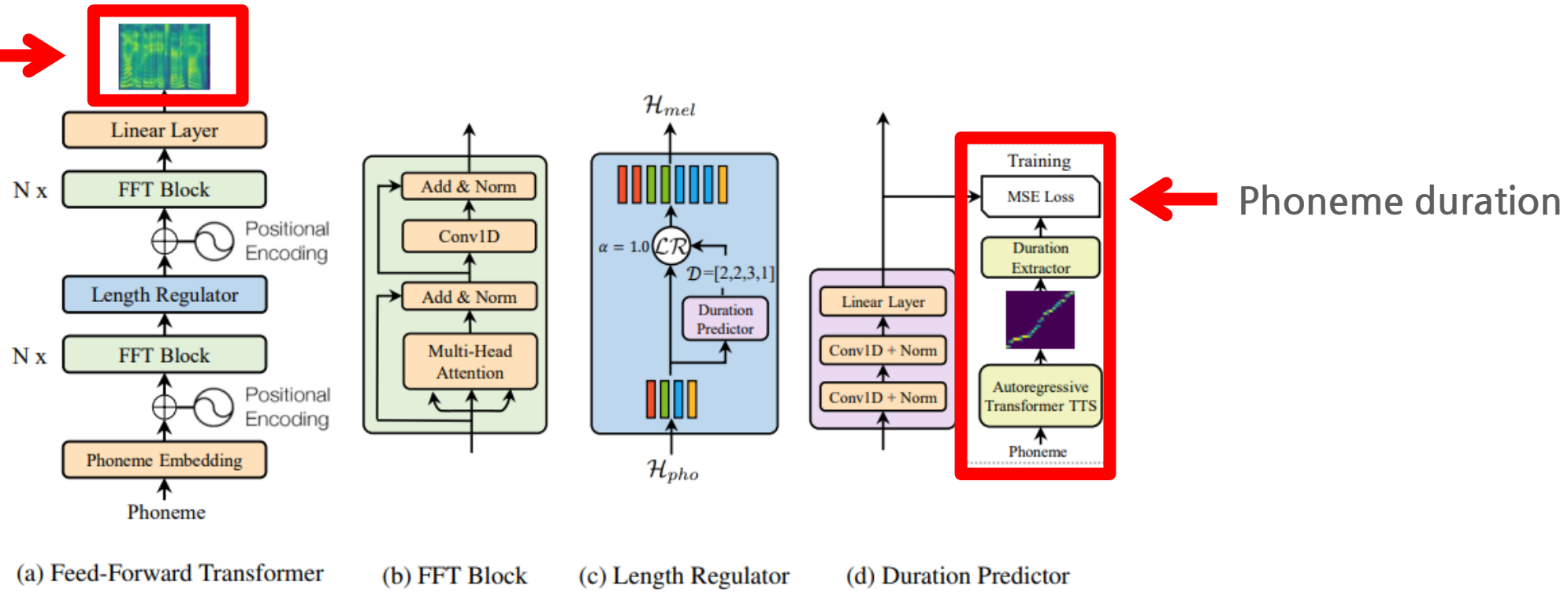
3. Output (mel) 생성한다

   Phoneme duration 과 align 되어있는 output → FastSpeech 학습에 사용

# FastSpeech

MS 에서 발표한 대표적인 End-to-end 음성 합성 모델

Output (mel)




(a) Feed-Forward Transformer (b) FFT Block (c) Length Regulator (d) Duration Predictor

Phoneme duration

Knowledge distillation from teacher (AR Transformer) to FastSpeech (non-AR Transformer)

AR: Autoregressive

CLOVA

# FastSpeech

MS 에서 발표한 대표적인 End-to-end 음성 합성 모델

| Method | MOS |
|---|---|
| *GT* | $4.41 \pm 0.08$ |
| *GT (Mel + WaveGlow)* | $4.00 \pm 0.09$ |
| *Tacotron 2 [22] (Mel + WaveGlow)* | $3.86 \pm 0.09$ |
| *Merlin [28] (WORLD)* | $2.40 \pm 0.13$ |
| *Transformer TTS [14] (Mel + WaveGlow)* | $3.88 \pm 0.09$ |
| *FastSpeech (Mel + WaveGlow)* | $3.84 \pm 0.08$ |

AR model (Tacotron, Transformer) 비슷한 품질

| Method | Latency (s) | Speedup |
|---|---|---|
| *Transformer TTS [14] (Mel)* | $6.735 \pm 3.969$ | / |
| *FastSpeech (Mel)* | $0.025 \pm 0.005$ | 269.40× |
| *Transformer TTS [14] (Mel + WaveGlow)* | $6.895 \pm 3.969$ | / |
| *FastSpeech (Mel + WaveGlow)* | $0.180 \pm 0.078$ | 38.30× |

V100 GPU 1장 기준

합성 속도는 약 270 배 (0.025 RT)

https://speechresearch.github.io/fastspeech/

# FastSpeech

MS 에서 발표한 대표적인 End-to-end 음성 합성 모델

| Method | MOS |
|---|---|
| *GT* | $4.41 \pm 0.08$ |
| *GT (Mel + WaveGlow)* | $4.00 \pm 0.09$ |
| *Tacotron 2 [22] (Mel + WaveGlow)* | $3.86 \pm 0.09$ |
| *Merlin [28] (WORLD)* | $2.40 \pm 0.13$ |
| *Transformer TTS [14] (Mel + WaveGlow)* | $3.88 \pm 0.09$ |
| *FastSpeech (Mel + WaveGlow)* | $3.84 \pm 0.08$ |

AR model (Tacotron, Transformer) 비슷한 품질

| Method | Repeats | Skips | Error Sentences | Error Rate |
|---|---|---|---|---|
| *Tacotron 2* | 4 | 11 | 12 | 24% |
| *Transformer TTS* | 7 | 15 | 17 | 34% |
| *FastSpeech* | 0 | 0 | 0 | 0% |

V100 GPU 1장 기준

합성 **속도**는 약 270 배 (0.025 RT)

Alignment **안정성** 확보

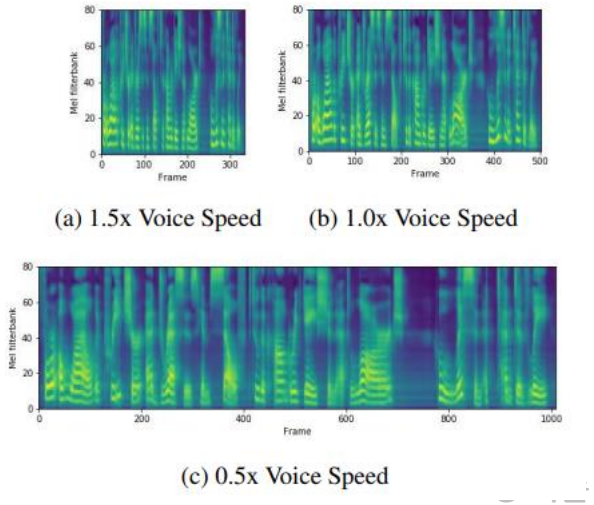
https://speechresearch.github.io/fastspeech/

# FastSpeech

MS 에서 발표한 대표적인 End-to-end 음성 합성 모델

| Method | MOS |
|---|---|
| *GT* | $4.41 \pm 0.08$ |
| *GT (Mel + WaveGlow)* | $4.00 \pm 0.09$ |
| *Tacotron 2 [22] (Mel + WaveGlow)* | $3.86 \pm 0.09$ |
| *Merlin [28] (WORLD)* | $2.40 \pm 0.13$ |
| *Transformer TTS [14] (Mel + WaveGlow)* | $3.88 \pm 0.09$ |
| *FastSpeech (Mel + WaveGlow)* | $3.84 \pm 0.08$ |

AR model (Tacotron, Transformer) 비슷한 품질




(a) 1.5x Voice Speed (b) 1.0x Voice Speed

(c) 0.5x Voice Speed

합성 **속도**는 약 270 배 (0.025 RT)

Alignment **안정성** 확보

**발화 속도** 조절 가능

https://speechresearch.github.io/fastspeech/

CLOVA

# Text-to-speech

## End-to-end speech synthesis

Tacotron

Tacotron 2

FastSpeech

FastSpeech 2

# FastSpeech 2

MS 에서 발표한 두번째 Non-AR End-to-end 음성 합성 모델

## FastSpeech 2: Fast and High-Quality End-to-End Text to Speech

**Yi Ren**[1]*, **Chenxu Hu**[1]*, **Xu Tan**[2], **Tao Qin**[2], **Sheng Zhao**[3], **Zhou Zhao**[1]†, **Tie-Yan Liu**[2]

[1]Zhejiang University
{rayeren,chenxuhu,zhaozhou}@zju.edu.cn

[2]Microsoft Research Asia
{xuta,taoqin,tyliu}@microsoft.com

[3]Microsoft Azure Speech
Sheng.Zhao@microsoft.com

### Abstract

Non-autoregressive text to speech (TTS) models such as FastSpeech (Ren et al., 2019) can synthesize speech significantly faster than previous autoregressive models with comparable quality. The training of FastSpeech model relies on an autoregressive teacher model for duration prediction (to provide more information as input) and knowledge distillation (to simplify the data distribution in output), which can ease the one-to-many mapping problem (i.e., multiple speech variations correspond to the same text) in TTS. However, FastSpeech has several disadvantages: 1) the teacher-student distillation pipeline is complicated and time-consuming, 2) the duration extracted from the teacher model is not accurate enough, and the target mel-spectrograms distilled from teacher model suffer from information loss due to data simplification, both of which limit the voice quality. In this paper, we propose FastSpeech 2, which addresses the issues in FastSpeech and better solves the one-to-many mapping problem in TTS by 1) directly training the model with ground-truth target instead of the simplified output from teacher, and 2) introducing more variation information of speech (e.g., pitch, energy and more accurate duration) as conditional inputs. Specifically, we extract duration, pitch and energy from speech waveform and directly take them as conditional inputs in training and use predicted values in inference. We further design FastSpeech 2s, which is the first attempt to directly generate speech waveform from text in parallel, enjoying the benefit of fully end-to-end inference. Experimental results show that 1) FastSpeech 2 achieves a 3x training speed-up over FastSpeech, and FastSpeech 2s enjoys even faster inference speed; 2) FastSpeech 2 and 2s outperform FastSpeech in voice quality, and FastSpeech 2 can even surpass autoregressive models. Audio samples are available at https://speechresearch.github.io/fastspeech2/.

# FastSpeech 2

MS 에서 발표한 두번째 Non-AR End-to-end 음성 합성 모델

## FastSpeech 2: Fast and High-Quality End-to-End Text to Speech

**Yi Ren**[1]*, **Chenxu Hu**[1]*, **Xu Tan**[2], **Tao Qin**[2], **Sheng Zhao**[3], **Zhou Zhao**[1]†, **Tie-Yan Liu**[2]

[1]Zhejiang University
{rayeren,chenxuhu,zhaozhou}@zju.edu.cn

[2]Microsoft Research Asia
{xuta,taoqin,tyliu}@microsoft.com

[3]Microsoft Azure Speech
Sheng.Zhao@microsoft.com

### Abstract

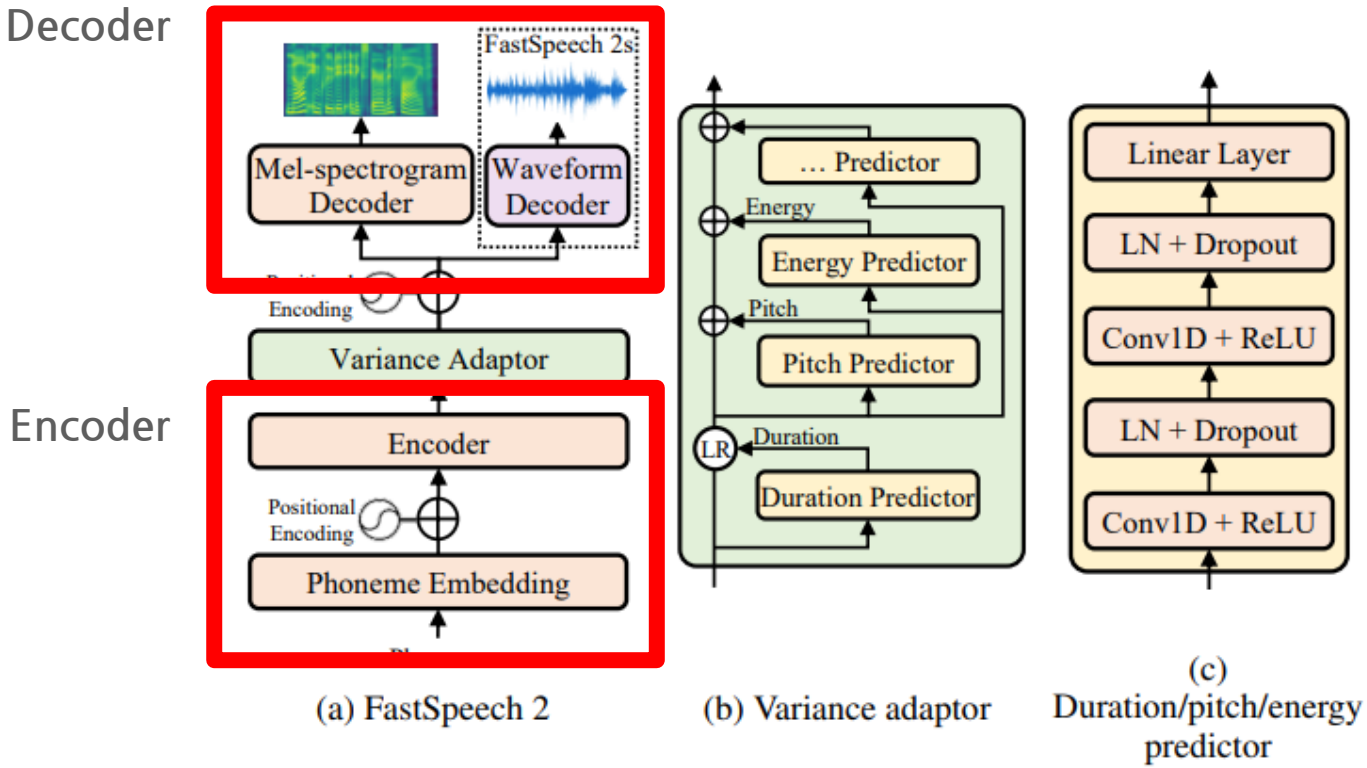
Non-autoregressive text to speech (TTS) models such as FastSpeech (Ren et al., 2019) can synthesize speech significantly faster than previous autoregressive models with comparable quality. The training of FastSpeech model relies on an autoregressive teacher model for duration prediction (to provide more information as input) and knowledge distillation (to simplify the data distribution in output), which can ease the one-to-many mapping problem (i.e., multiple speech

eral disadvantages: 1) the teacher-student distillation pipeline is complicated and time-consuming, 2) the duration extracted from the teacher model is not accurate enough, and the target mel-spectrograms distilled from teacher model suffer from information loss due to data simplification, both of which limit the voice quality. In this paper, we propose FastSpeech 2, which addresses the is-

Knowledge distillation 없이 학습하자!

# FastSpeech 2

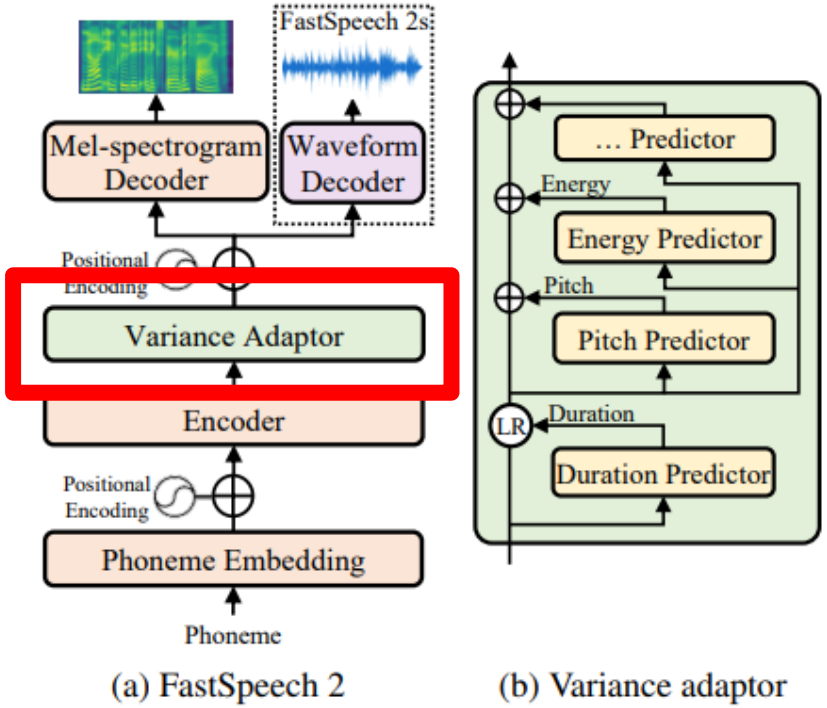
MS 에서 발표한 두번째 Non-AR End-to-end 음성 합성 모델




(a) FastSpeech 2 (b) Variance adaptor (c) Duration/pitch/energy predictor

# FastSpeech 2

MS 에서 발표한 두번째 Non-AR End-to-end 음성 합성 모델



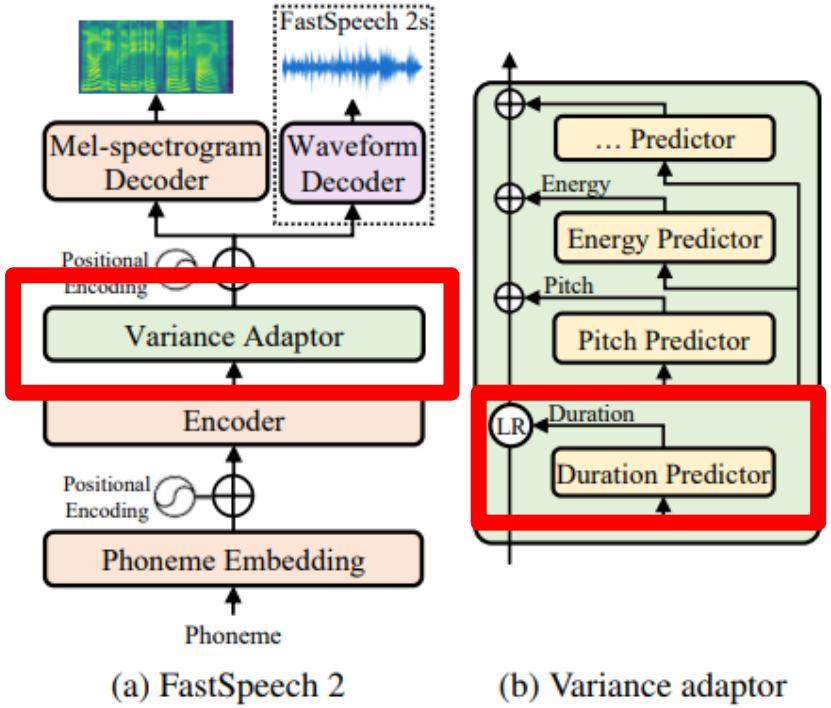

Transformer 기반의 encoder-decoder model

Variance adaptor: Controllability 제공

- Duration predictor & Length regulator
- Pitch regulator & Energy regulator

# FastSpeech 2

MS 에서 발표한 두번째 Non-AR End-to-end 음성 합성 모델




Transformer 기반의 encoder-decoder model

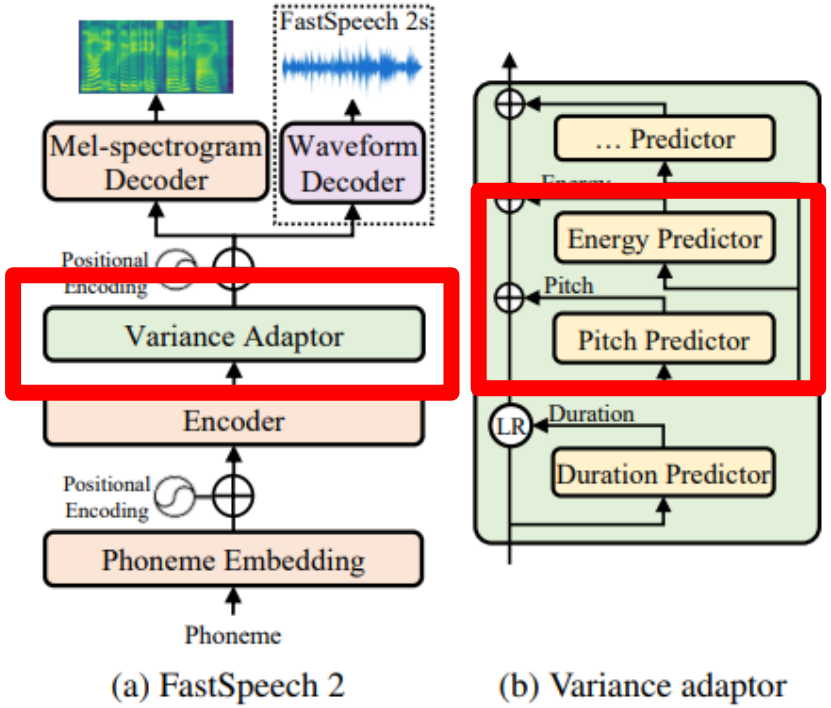
Variance adaptor: Controllability 제공

- Duration predictor & Length regulator
- Pitch regulator & Energy regulator

좀 비싸도 phoneme segmentation 하고
Knowledge distillation 없이 학습하자!

# FastSpeech 2

MS 에서 발표한 두번째 Non-AR End-to-end 음성 합성 모델




Transformer 기반의 encoder-decoder model

Variance adaptor: Controllability 제공

- Duration predictor & Length regulator
- Pitch regulator & Energy regulator

좀 비싸도 phoneme segmentation 하고
Knowledge distillation 없이 학습하자!

대신 서비스에 맞게 다양한 기능 넣자!

CLOVA

# FastSpeech 2

MS 에서 발표한 두번째 Non-AR End-to-end 음성 합성 모델

| Method | MOS |
|---|---|
| *GT* | $4.30 \pm 0.07$ |
| *GT (Mel + PWG)* | $3.92 \pm 0.08$ |
| *Tacotron 2 (Shen et al., 2018) (Mel + PWG)* | $3.70 \pm 0.08$ |
| *Transformer TTS (Li et al., 2019) (Mel + PWG)* | $3.72 \pm 0.07$ |
| *FastSpeech (Ren et al., 2019) (Mel + PWG)* | $3.68 \pm 0.09$ |
| *FastSpeech 2 (Mel + PWG)* | $3.83 \pm 0.08$ |
| *FastSpeech 2s* | $3.71 \pm 0.09$ |

| Method | Training Time (h) | Inference Speed (RTF) | Inference Speedup |
|---|---|---|---|
| *Transformer TTS (Li et al., 2019)* | 38.64 | $9.32 \times 10^{-1}$ | / |
| *FastSpeech (Ren et al., 2019)* | 53.12 | $1.92 \times 10^{-2}$ | 48.5× |
| *FastSpeech 2* | **17.02** | $1.95 \times 10^{-2}$ | 47.8× |
| *FastSpeech 2s* | 92.18 | $\mathbf{1.80 \times 10^{-2}}$ | **51.8**× |

V100 GPU 1장 기준

AR model (Tacotron, Transformer) 보다 품질도 좋고

FastSpeech 하고 합성 속도도 비슷하면서

합성음 컨트롤이 가능함

https://speechresearch.github.io/fastspeech2/

CLOVA

# Summary

Text-to-speech (TTS) 란 기계가 사람 처럼 **텍스트를 읽어주는** 기술입니다


딥러닝 TTS 기술은 크게 acoustic model 과 vocoding model 두가지 모듈로 구성됩니다

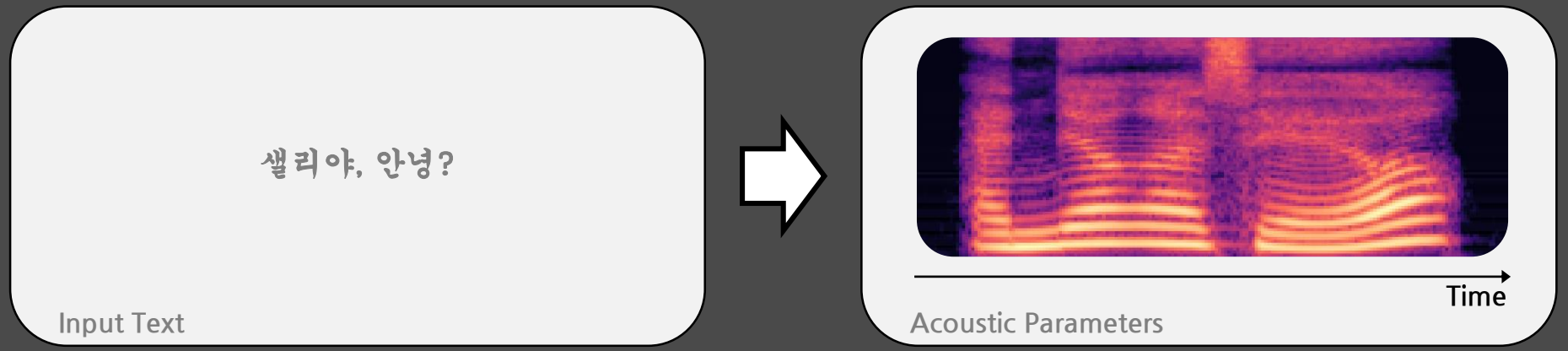
CLOVA

# Summary

Text-to-speech (TTS) 란 기계가 사람 처럼 **텍스트를 읽어주는** 기술입니다

톤의 높낮이, 음색, 어조, 강세 등

**텍스트에서 Acoustic Parameter 를 추정**

CLOVA

# Summary

Text-to-speech (TTS) 란 기계가 사람 처럼 **텍스트를 읽어주는** 기술입니다

Acoustic Parameter 에서 음성 신호를 생성


CLOVA

# Summary

Text-to-speech (TTS) 란 기계가 사람 처럼 **텍스트를 읽어주는** 기술입니다

톤의 높낮이, 음색, 어조, 강세 등

**텍스트에서 Acoustic Parameter 를 추정**


## Statistical Parametric Speech Synthesis

가볍고, 빠르고, 안정적 but 품질이 아쉬움

CLOVA

# Summary

Text-to-speech (TTS) 란 기계가 사람 처럼 **텍스트를 읽어주는** 기술입니다

톤의 높낮이, 음색, 어조, 강세 등

**텍스트에서 Acoustic Parameter 를 추정**


## End-to-end Speech Synthesis

AR models (Tacotron 1,2) with Attention Alignment

복잡한 Feature Engineering 최소화 하면서도 고품질의 음성을 만들 수 있음

but 느리고 안전성 떨어짐

CLOVA

# Summary

Text-to-speech (TTS) 란 기계가 사람 처럼 **텍스트를 읽어주는** 기술입니다

톤의 높낮이, 음색, 어조, 강세 등

**텍스트에서 Acoustic Parameter 를 추정**


## End-to-end Speech Synthesis

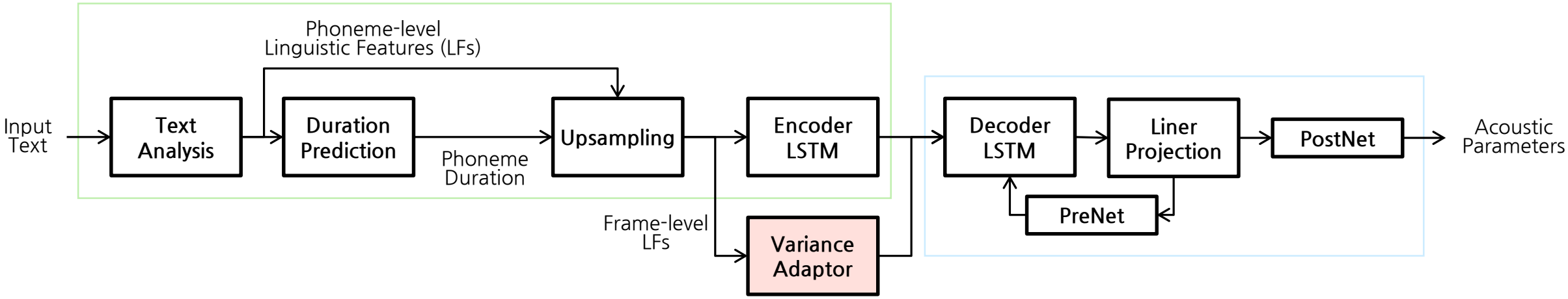
Non-AR models (FastSpeech 1,2) with External Duration Model

빠르고 안정적인 합성음을 만들 수 있음

음질은 Best-quality 일까?

CLOVA

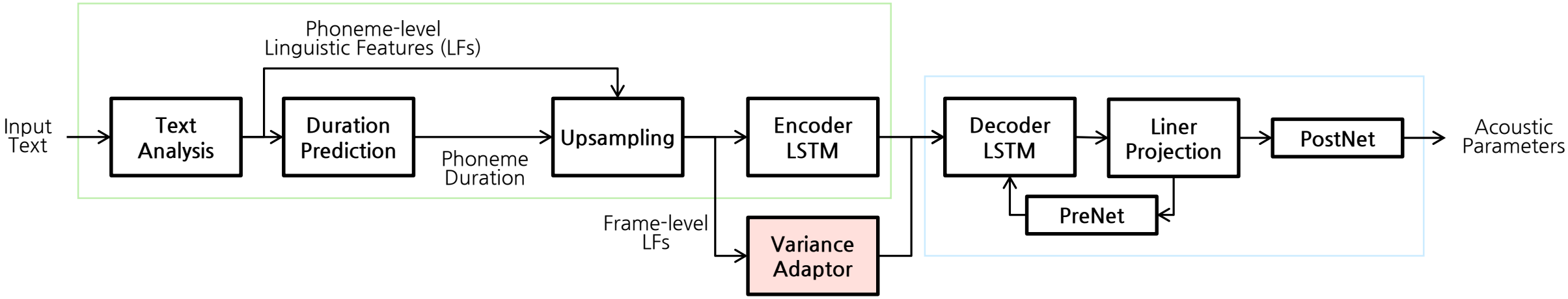
# Summary




다양한 Application 환경에 맞는 모델을 조합해서 사용합니다.

CLOVA

# Summary




정상 합성음

낮은 톤

작은 목소리

느린 목소리

높은 톤

큰 목소리

빠른 목소리

# CLOVA Voice Synthesis

## Service

CLOVA

## Whale Browser

papago

## Naver Dictionary


## Navigation


## Audio Book


## Clova Speaker

## Device


## Ai Call


## Care Call

CLOVA


+

# ‘유인나’ Voice
# 클로바 스피커 기본 목소리

CLOVA

‘오상진’ Voice
네이버 뉴스 본문 듣기 목소리

CLOVA

# 내비게이션 × Clova

내 차안의 인공지능 비서

CLOVA


변화의 시작! 꿈이 현실이 되는 해운대
클로바 케어콜이란?
인공지능 AI가 친구처럼 돌봄대상자와 안부를 묻고 답하는 서비스
AI와 어떤 대화를 할까요?
"간밤에 잠은 잘 주무셨어요?"
"아니오, 요즘 통 잠을 제대로 못자고 있네요"
"왜 그러세요? 무슨 걱정거리라도 있으세요?"
"그건 아닌데, 무릎이 아파서…"
"관절이 많이 안 좋으신가요?"
"부쩍 밤마다 많이 아프네요…"
"아이고 그러셨군요, 파스나 찜질팩이라도 해보세요.
병원에도 한번 가보시고요."
(어르신과 AI의 실제 대화 내용 중 발췌)

CLOVA

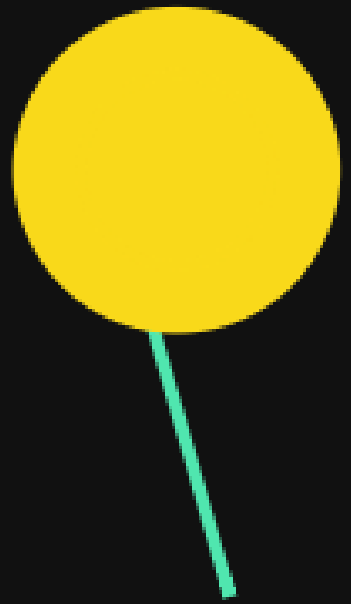
# 인공지능부터 로봇까지...네이버 실험실 거듭난 '1784'

이영아 기자 | 승인 2022.04.22 17:26

가 가

동영상에 보이스를 더하다

# CLOVA Dubbing$^{\beta}$



자연스러운 클로바보이스로 동영상에
특별한 생동감을 더해주세요

CLOVA

# Q / A

eunwoo.song@navercorp.com