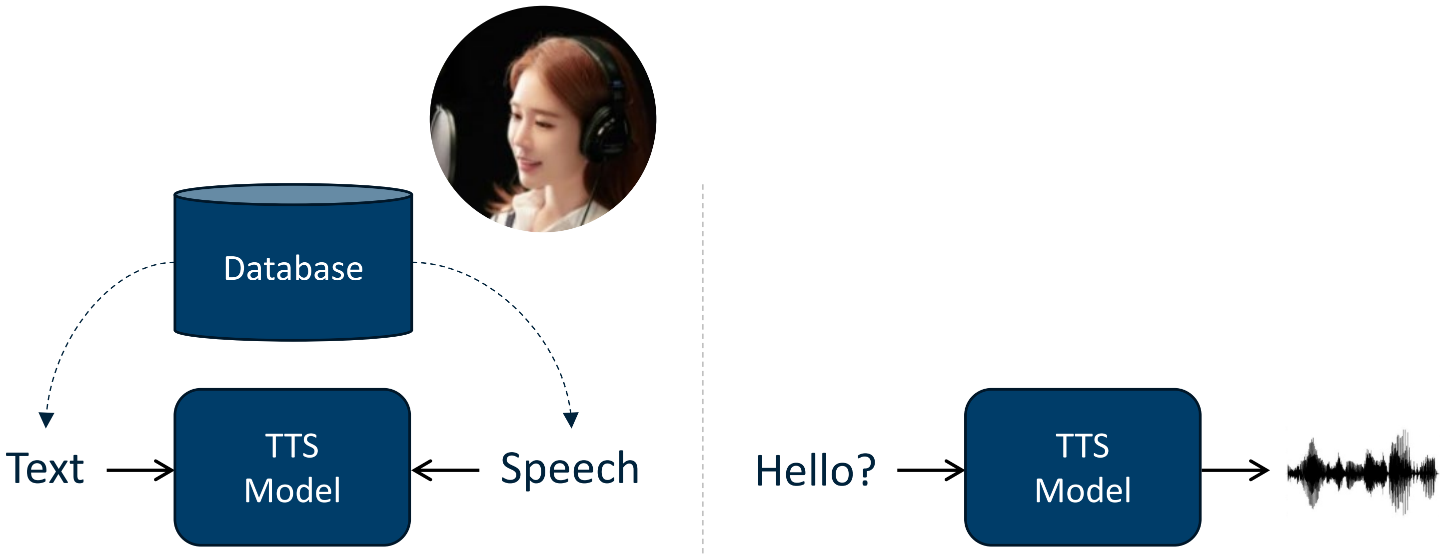
Zero-shot Voice Cloning

Eunwoo Song / NAVER Cloud

Contents

Introduction

Deep learning-based TTS system

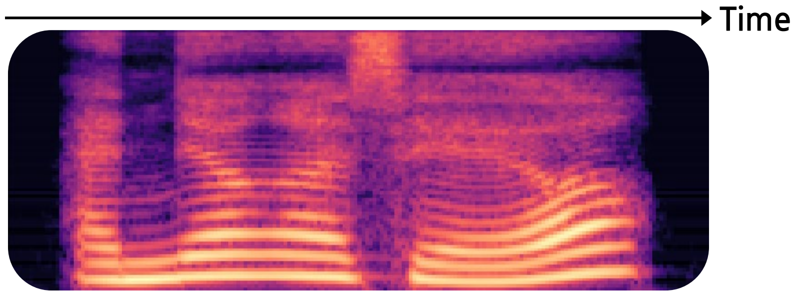


Training **Inference**

Human-like voice quality ☺

Introduction

Deep learning-based TTS system

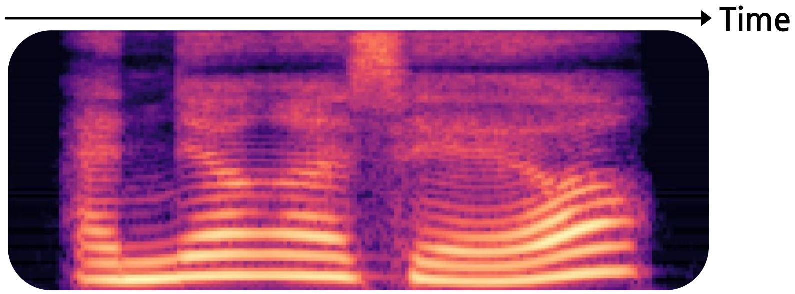


Text → Acoustic Model → Acoustic Parameters → Vocoding Model → 

Acoustic model + Vocoding model

Introduction

Deep learning-based TTS system



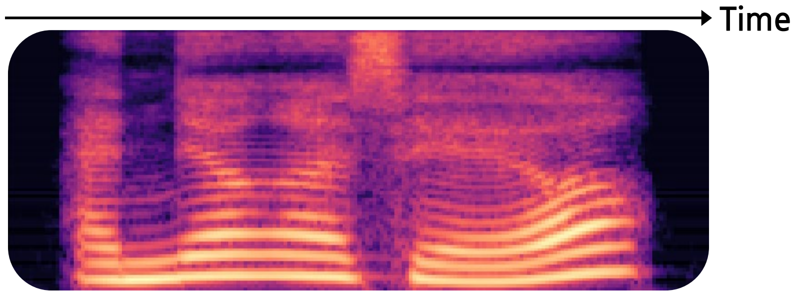
Estimating acoustic parameters from text inputs

Speaker-specific attributes
(tone, volume, timbre, speaking rate, …)

Acoustic model + Vocoding model

Introduction

Deep learning-based TTS system

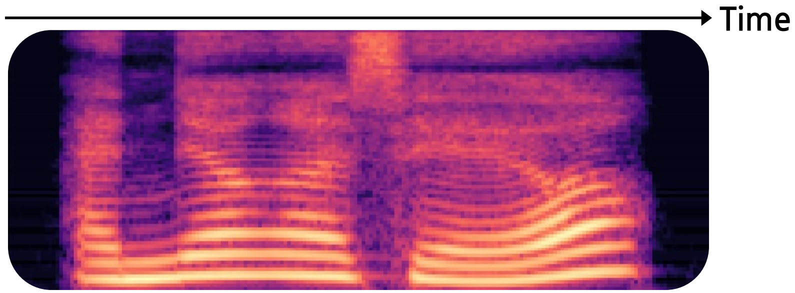


Estimating speech signals from acoustic parameters

Acoustic model + Vocoding model

Introduction

Deep learning-based TTS system



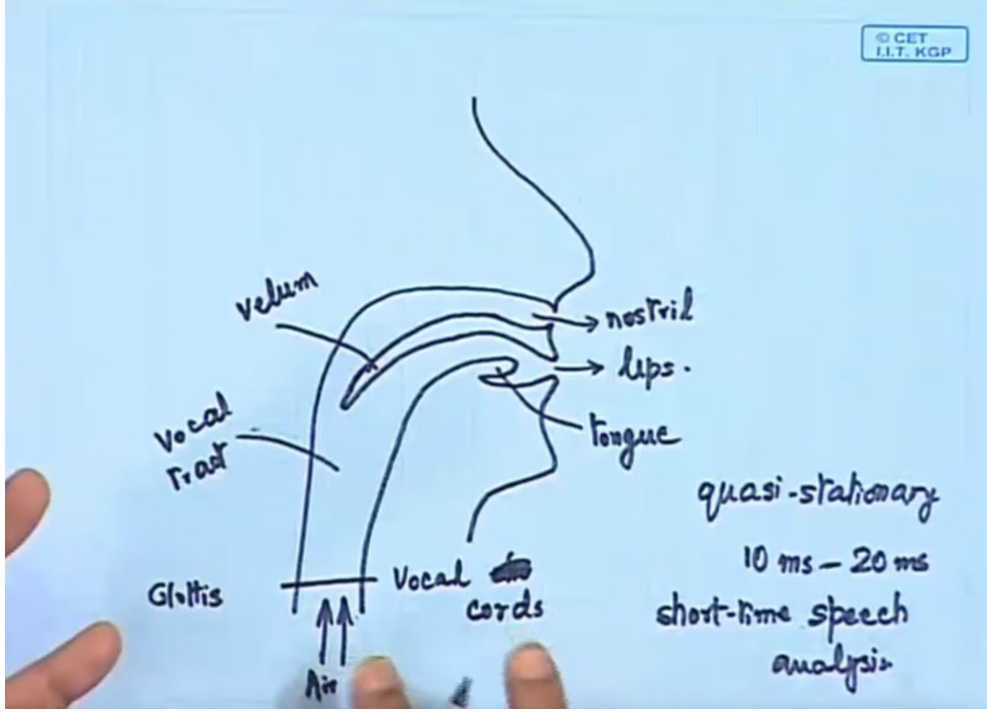
Acoustic parameters..?

Speaker-specific attributes
(tone, volume, timbre, speaking rate, …)

Speech analysis

Speech analysis

Speech production model



https://www.youtube.com/watch?v=X_JvfZiGEek

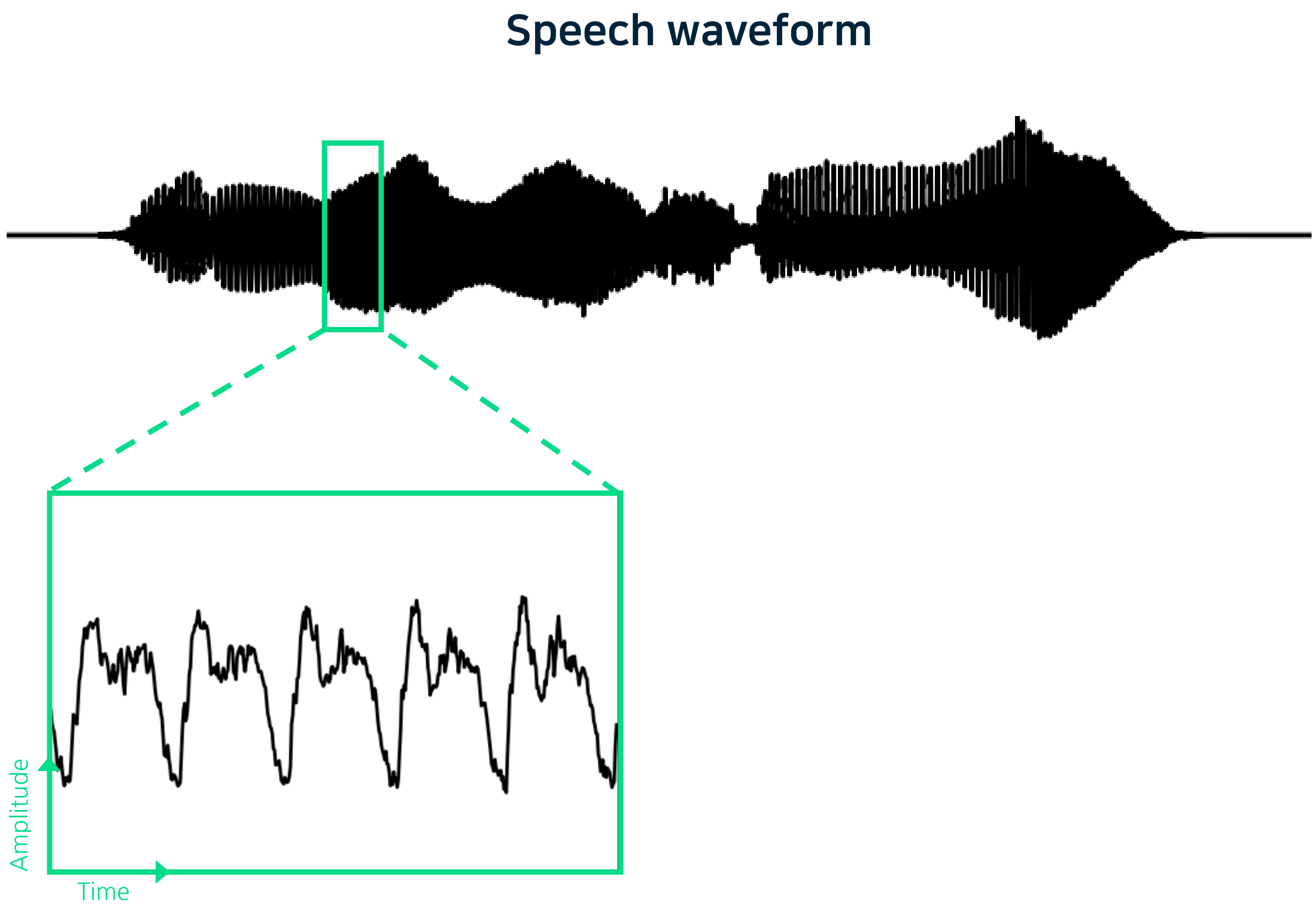
- Vocal cords
  - Voiced sound : quasi-periodic
  - Unvoiced sound : noisy

→ **목소리의 톤을 결정 (아↘아↗)**

- Vocal tract
  - Shaping voice color

→ **발음을 결정 (아/에/이/오/우)**

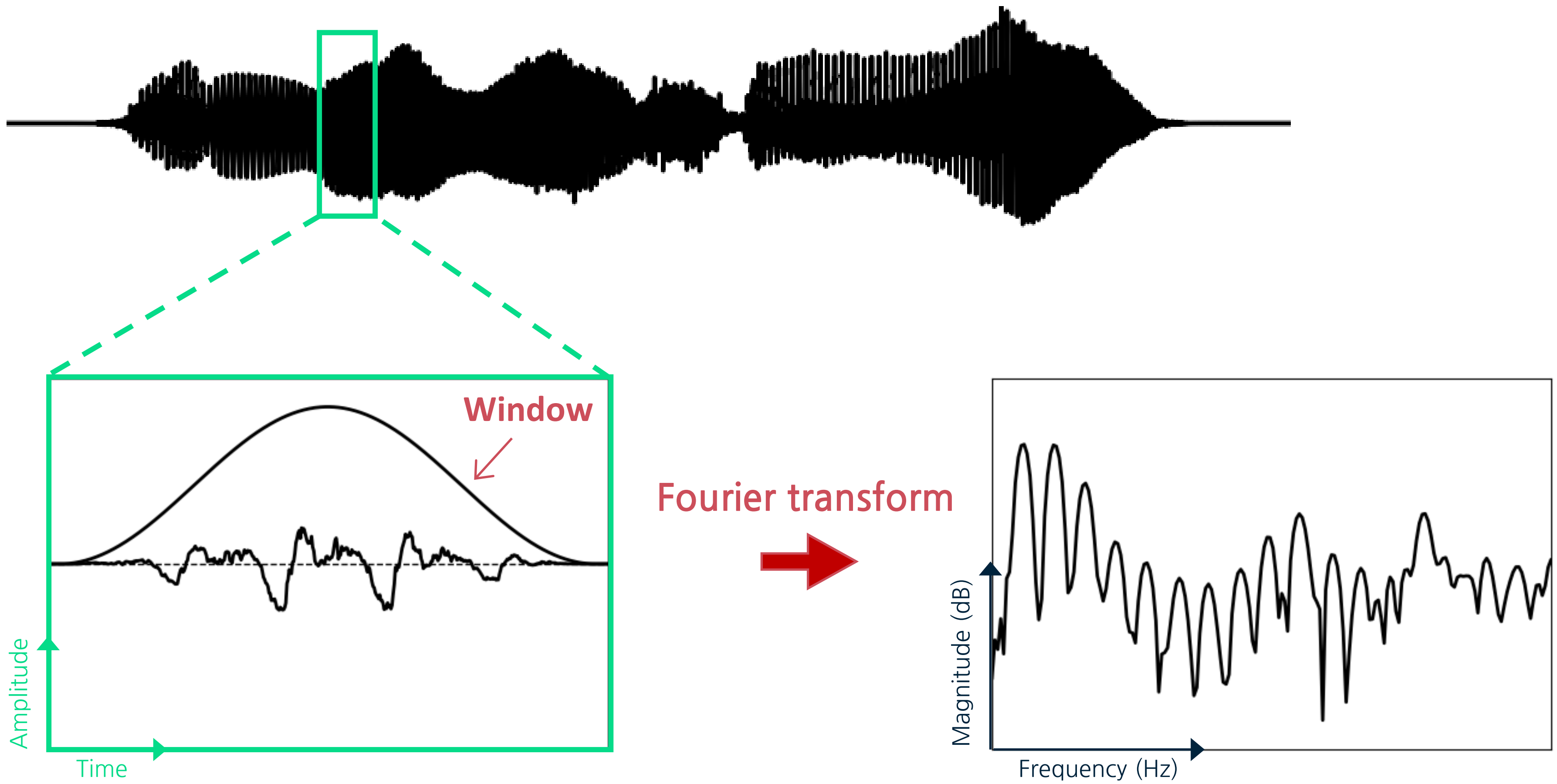
Speech analysis



음성 신호는 시간 축에서 특정한 에너지를 갖는 파형의 형태로 존재합니다

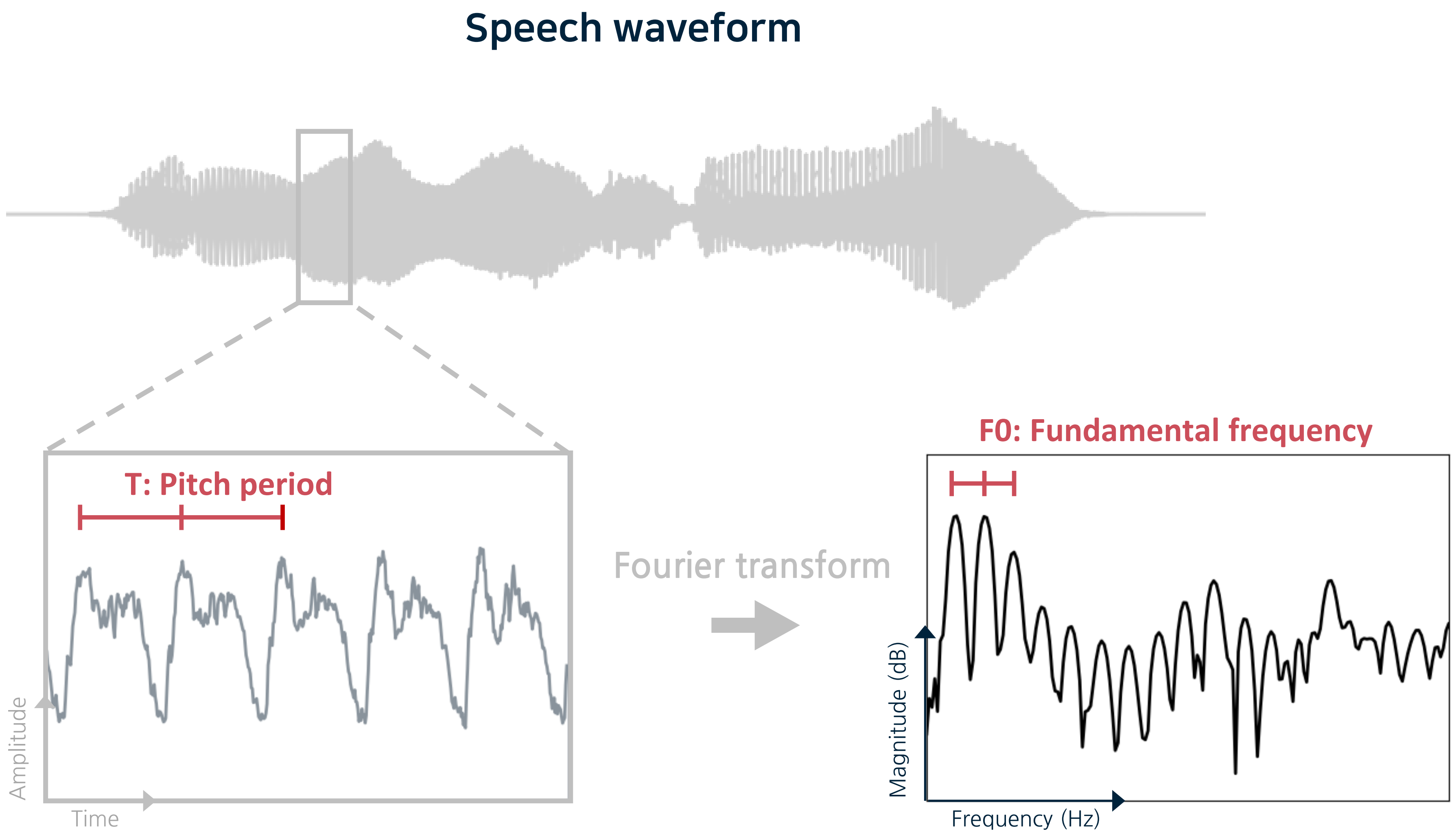
Speech analysis

Speech waveform



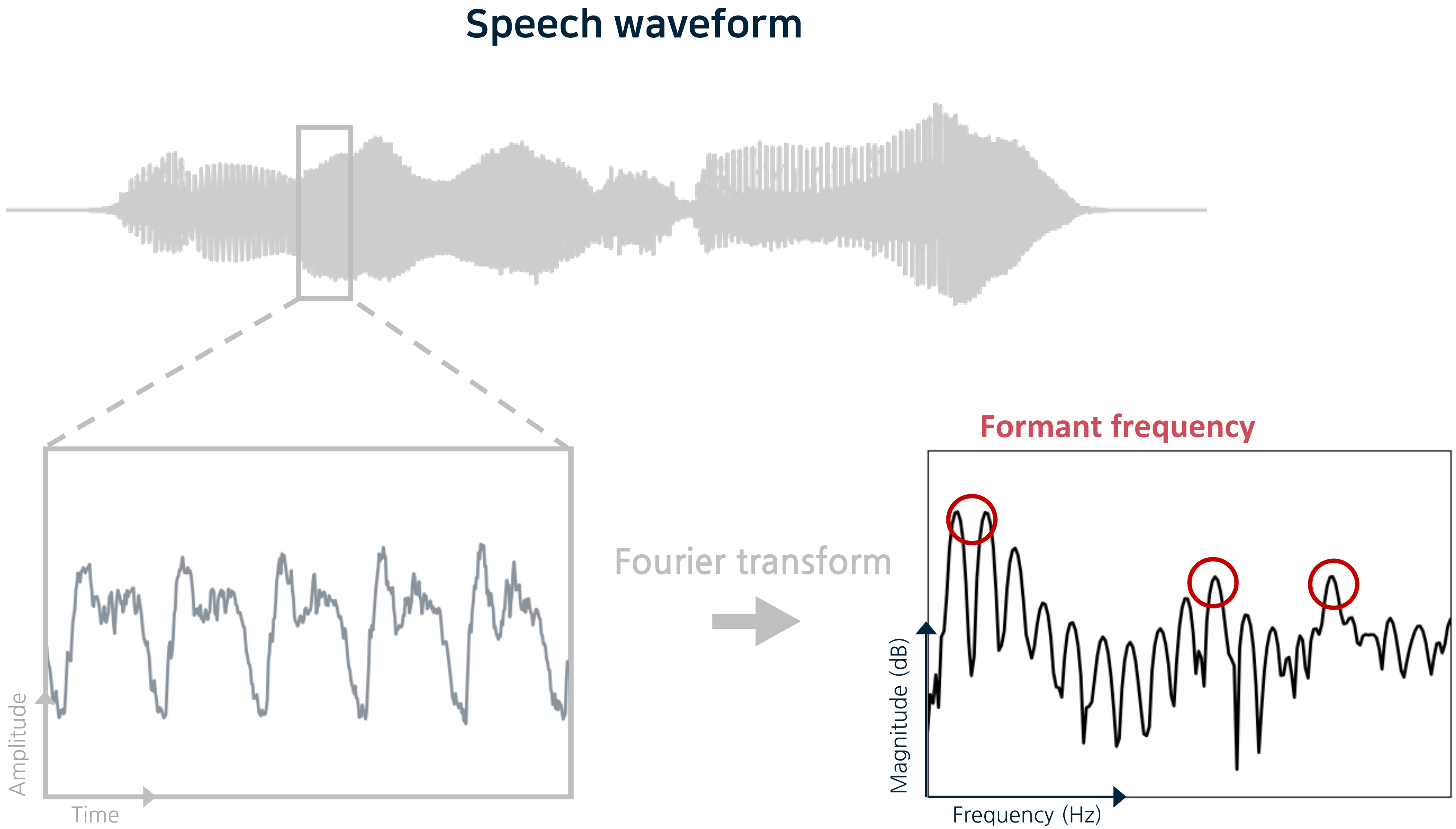
Fourier 변환을 통해 주파수 축에서 음성을 관찰할 수 있습니다

Speech analysis



F0: 목소리의 톤을 표현하는 파라미터 (아↘아↗)

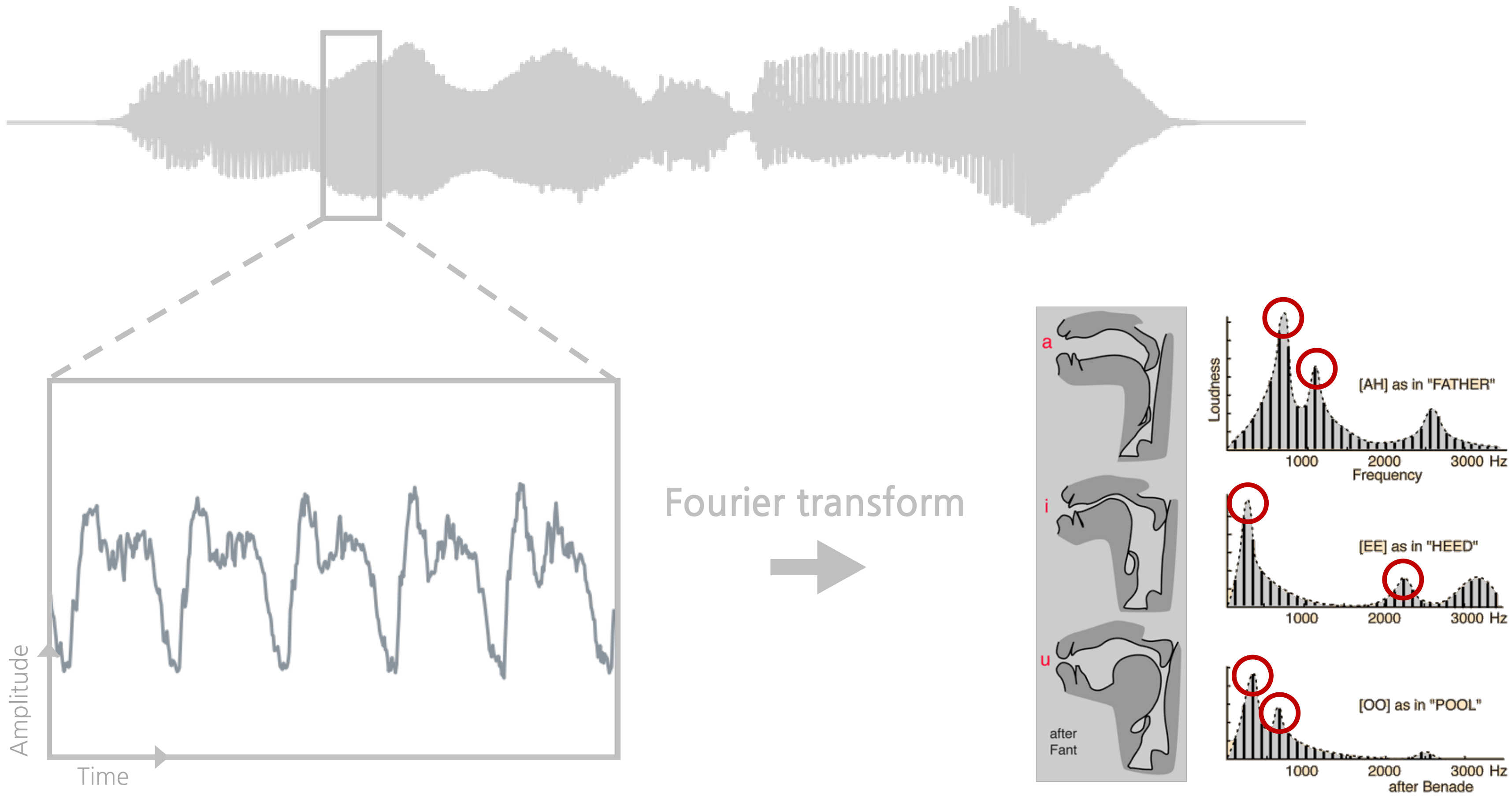
Speech analysis



Formant: 발음을 표현하는 파라미터 (아/에/이/오/우)

Speech analysis

Speech waveform

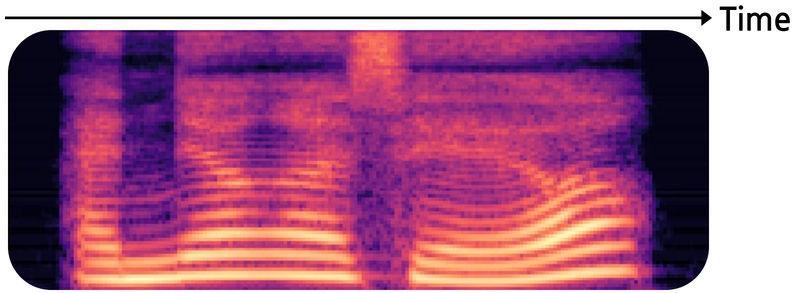


http://hyperphysics.phy-astr.gsu.edu/hbase/Music/vowel.html

Formant: 발음을 표현하는 파라미터 (아/에/이/오/우)

Speech analysis

Speech waveform

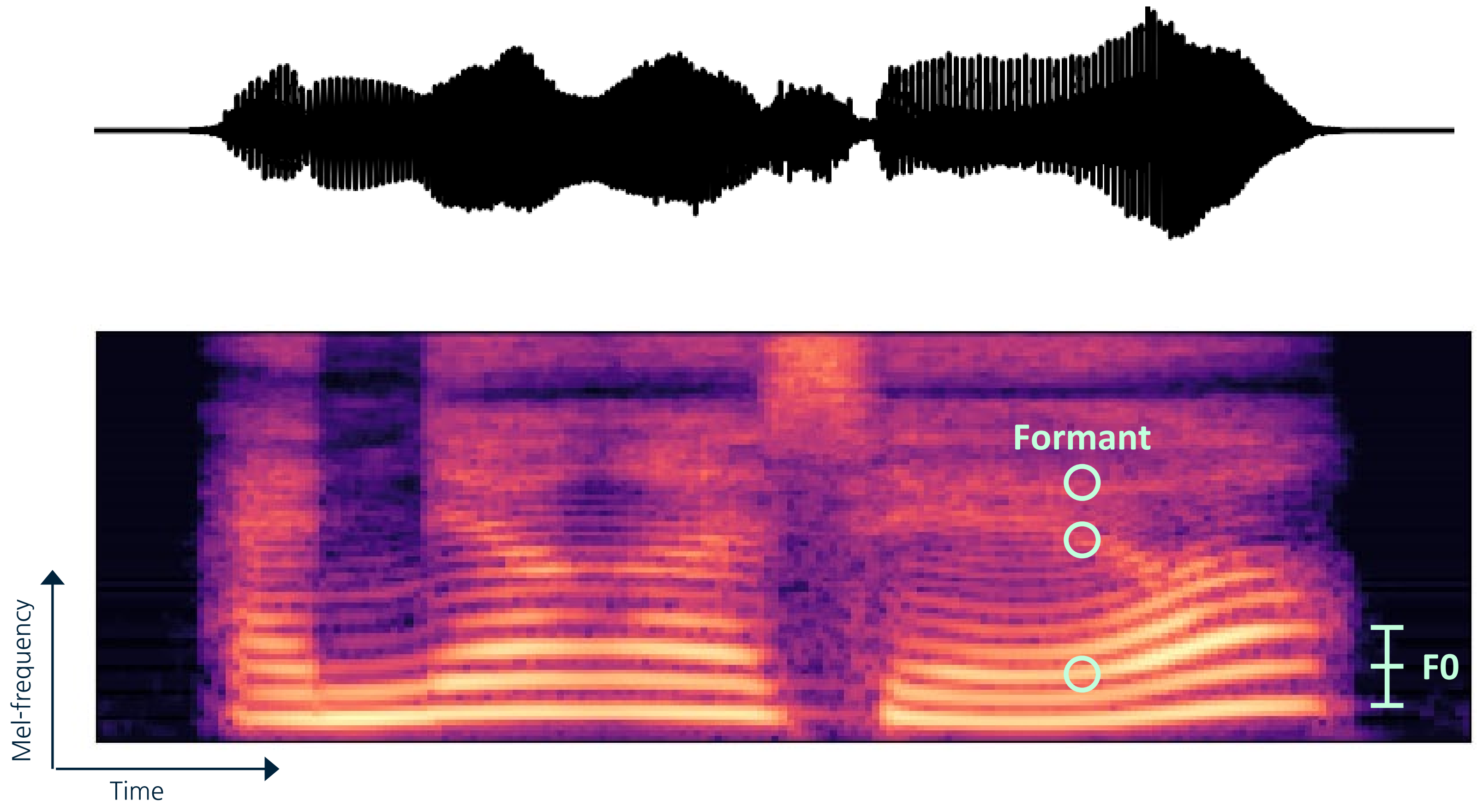


Acoustic parameters..?

Speaker-specific attributes
(tone, volume, timbre, speaking rate, …)

Speech analysis

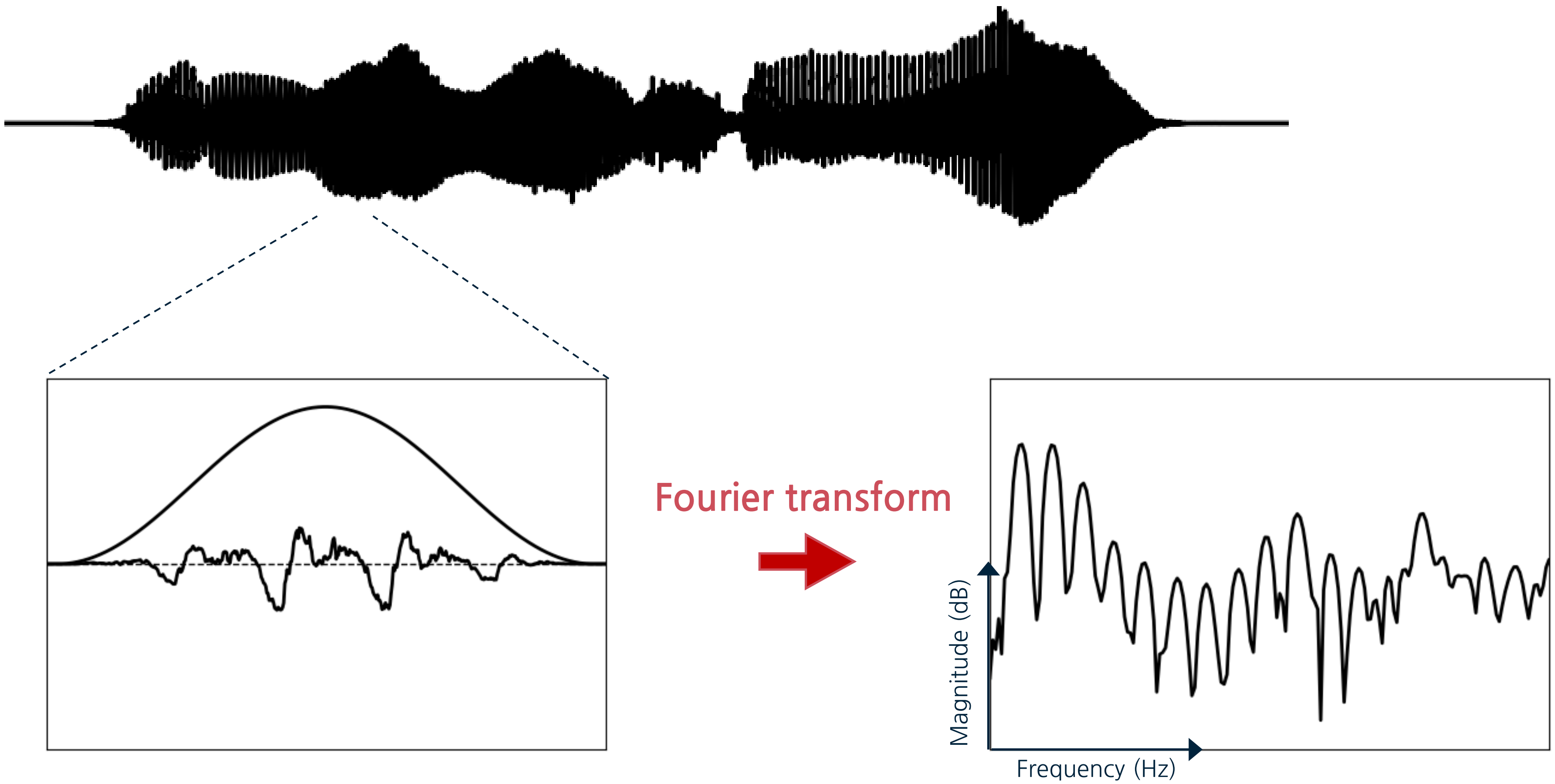
Mel-spectrogram



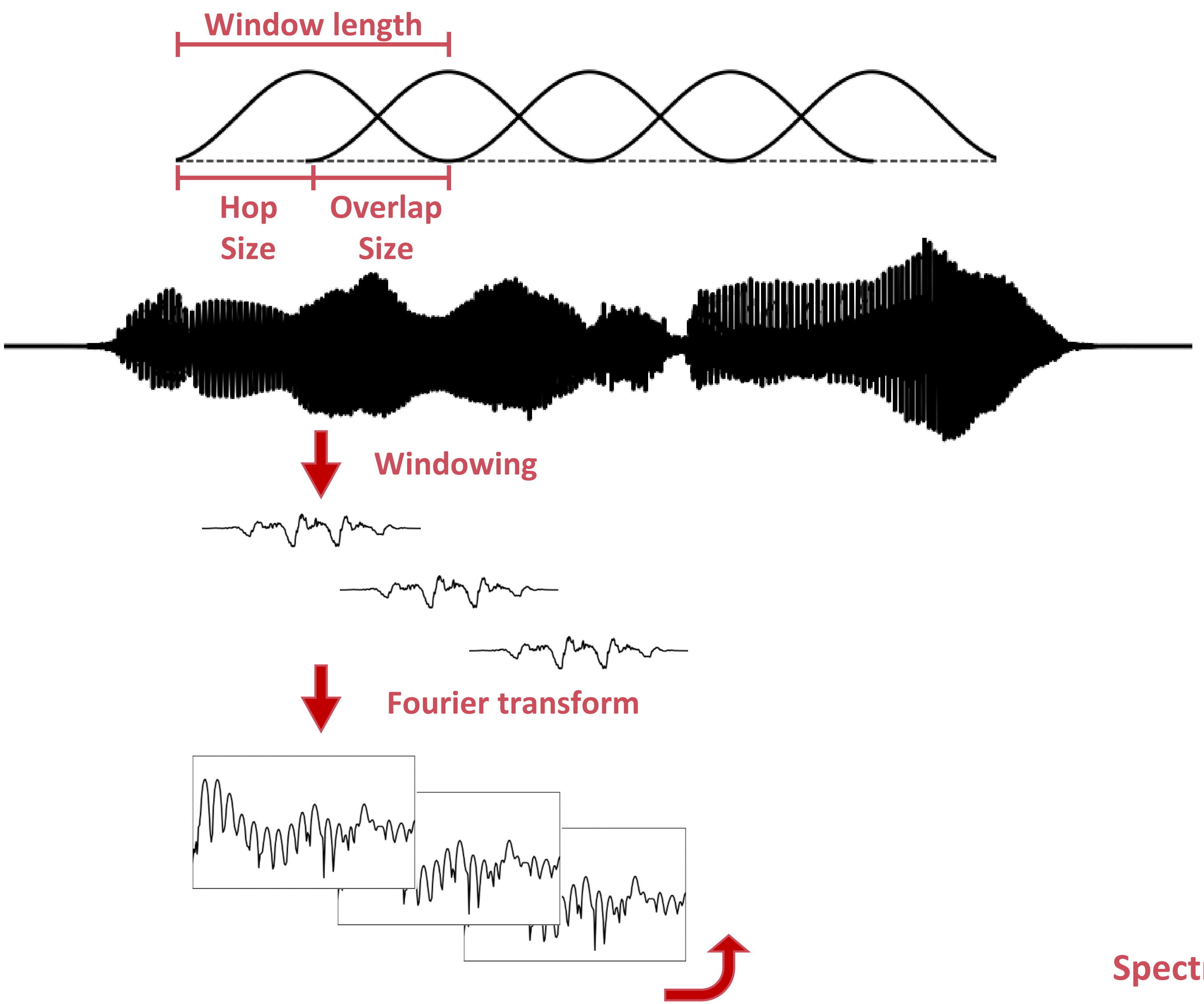
Mel-spectrogram: 음성의 다양한 특성들을 시간-주파수 축으로 표현

Speech analysis

Mel-spectrogram



복잡해 보이는 시간 축 신호를 주파수 축에서 보면 음성을 분석하기 쉬워집니다

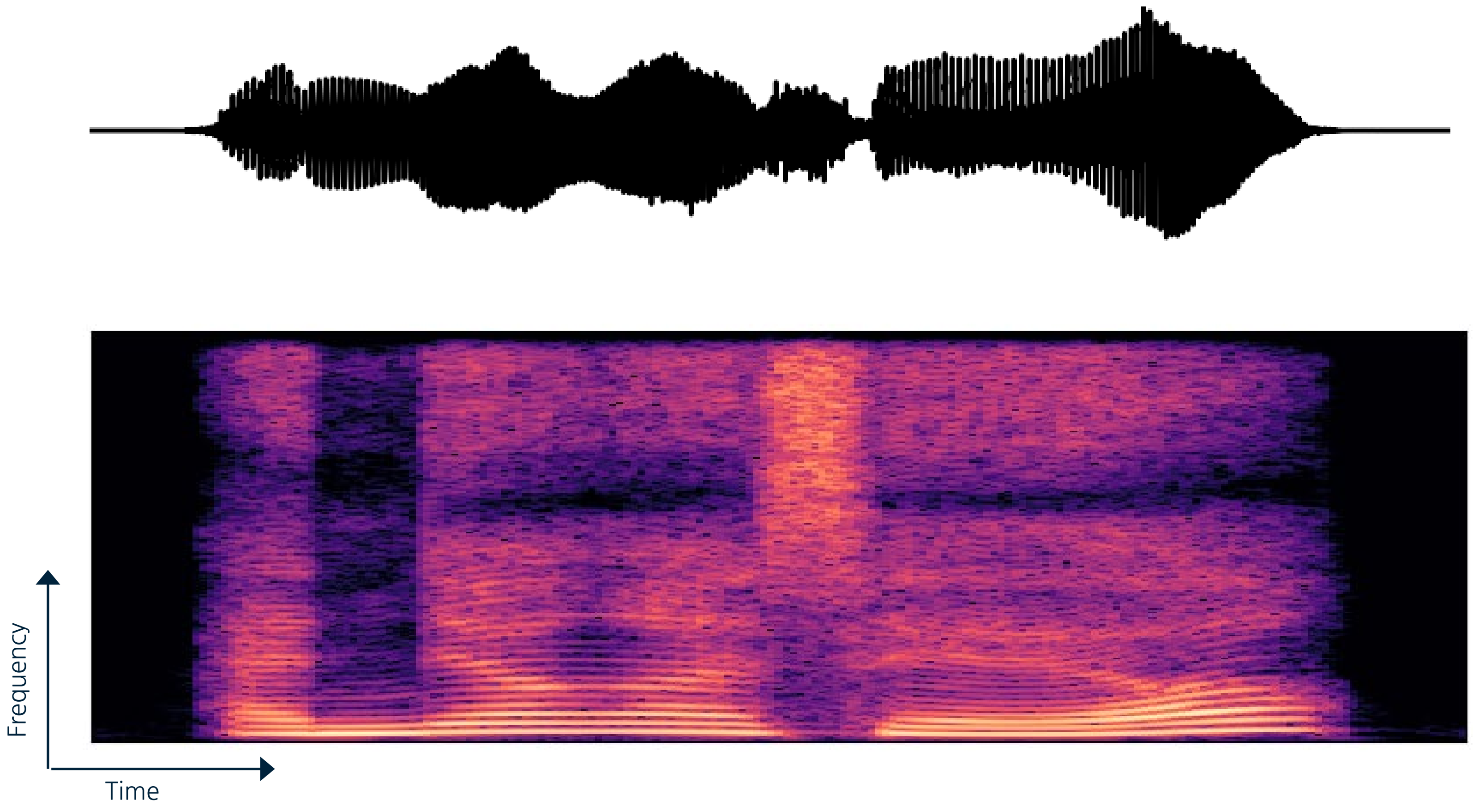


Spectrogram

STFT 신호를 시간 축으로 붙인 2D 이미지

Speech analysis

Mel-spectrogram

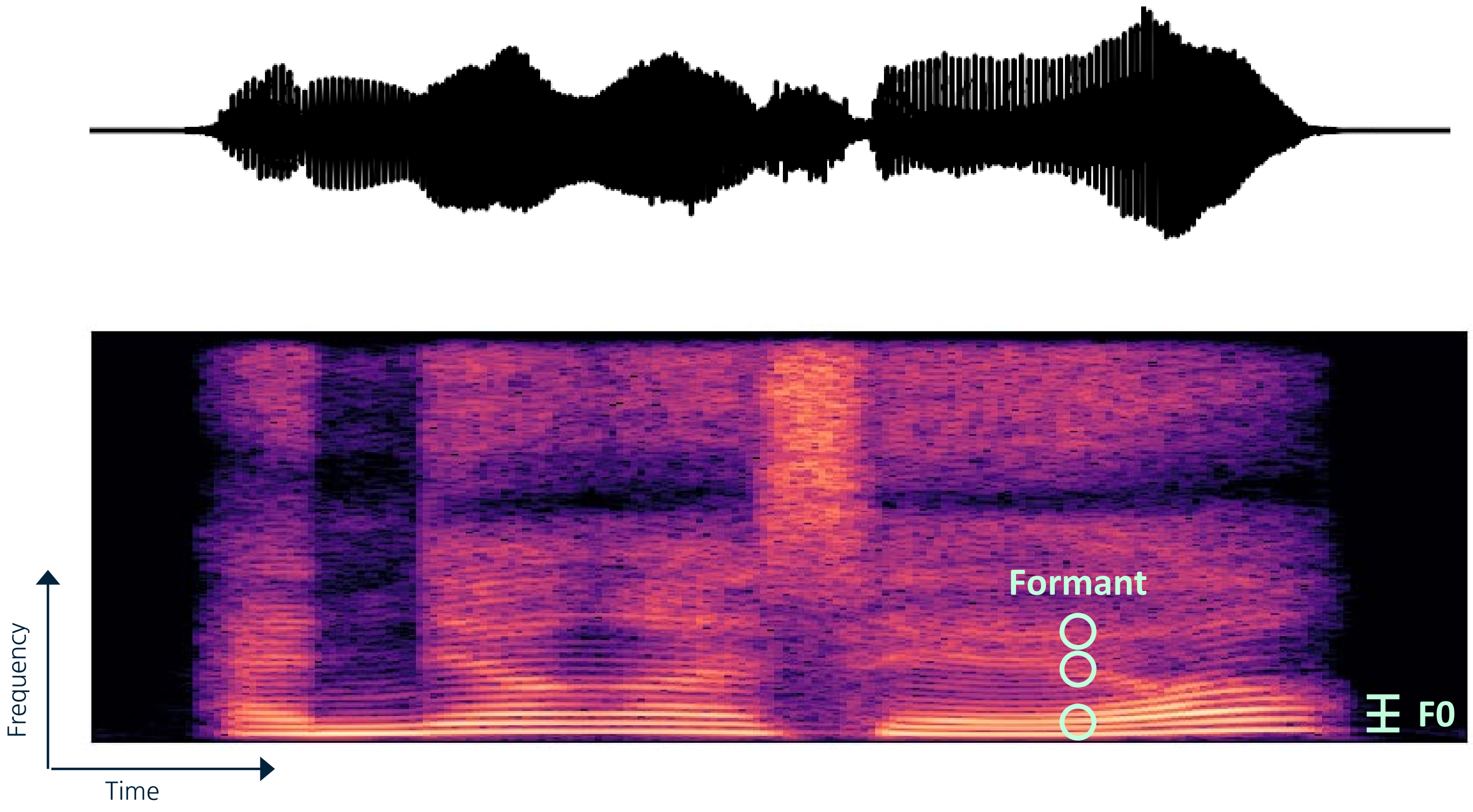


Spectrogram

STFT 신호를 시간 축으로 붙인 2D 이미지

Speech analysis

Mel-spectrogram

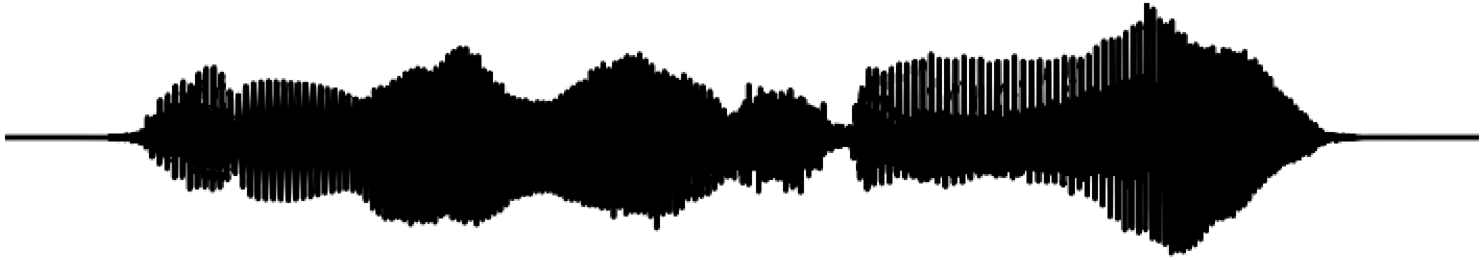


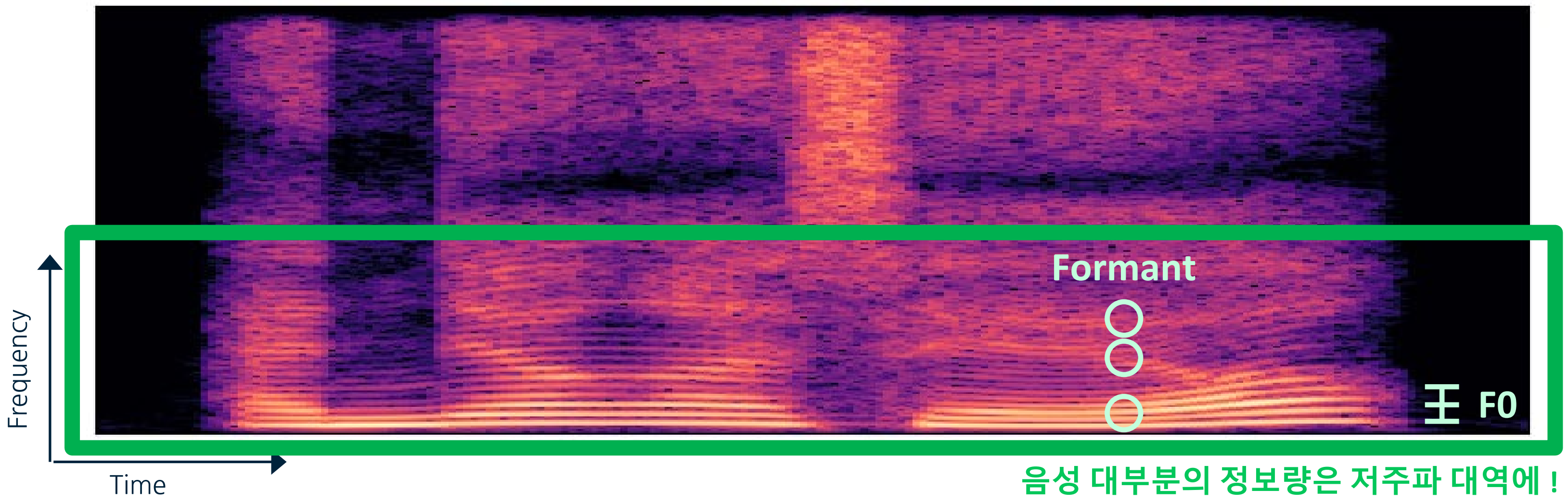
Spectrogram

음성을 시간-주파수 축에서 분석할 수 있게 되었습니다

Speech analysis

Mel-spectrogram





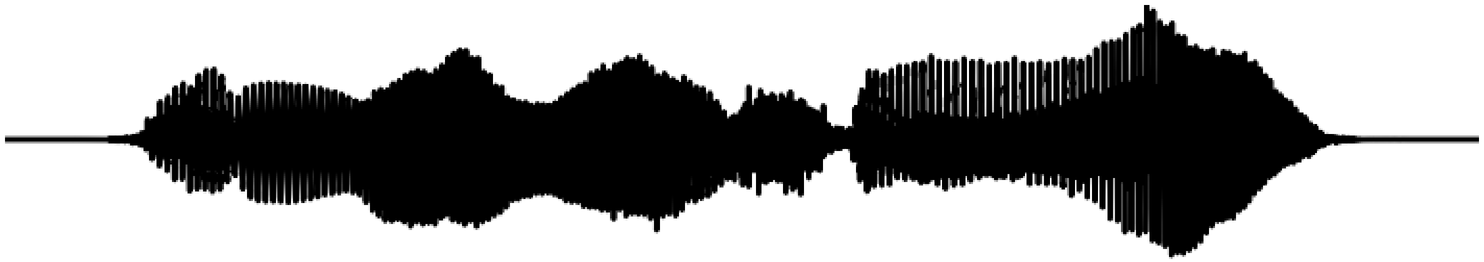
음성 대부분의 정보량은 저주파 대역에 !

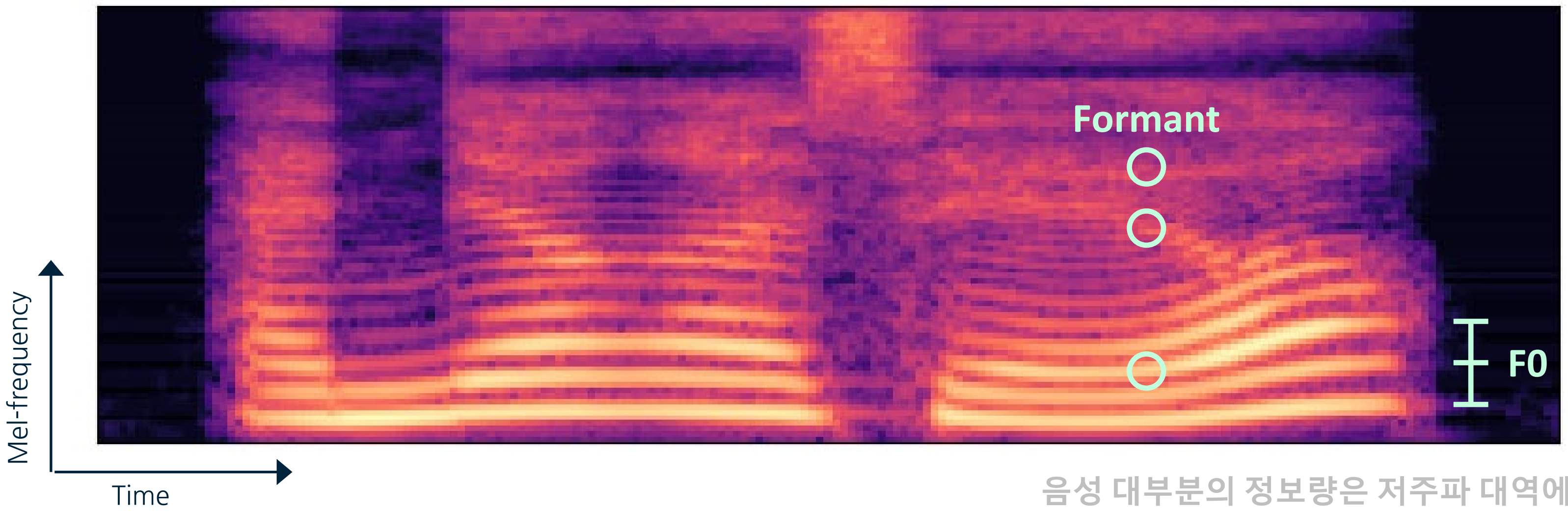
저주파 대역의 정보량에 집중할 수 있다면?

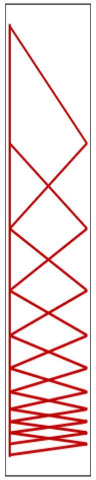
음성을 시간-주파수 축에서 분석할 수 있게 되었습니다

Speech analysis

Mel-spectrogram







주파수 축으로
Mel-filterbank 적용

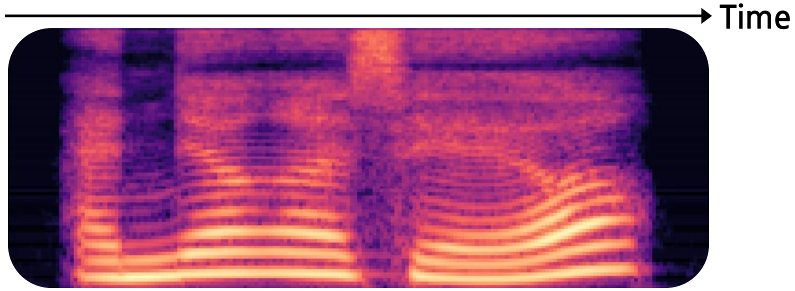
음성 대부분의 정보량은 저주파 대역에 !

저주파 대역의 정보량에 집중할 수 있다면?

모델이 음성 신호를 이해하기 쉬워집니다 ← 음성을 시간-주파수 축에서 분석을 더 잘 할 수 있습니다

Speech analysis

Deep learning-based TTS system



Text → Acoustic Model → Acoustic Parameters → Vocoding Model →

Estimating acoustic parameters from text inputs

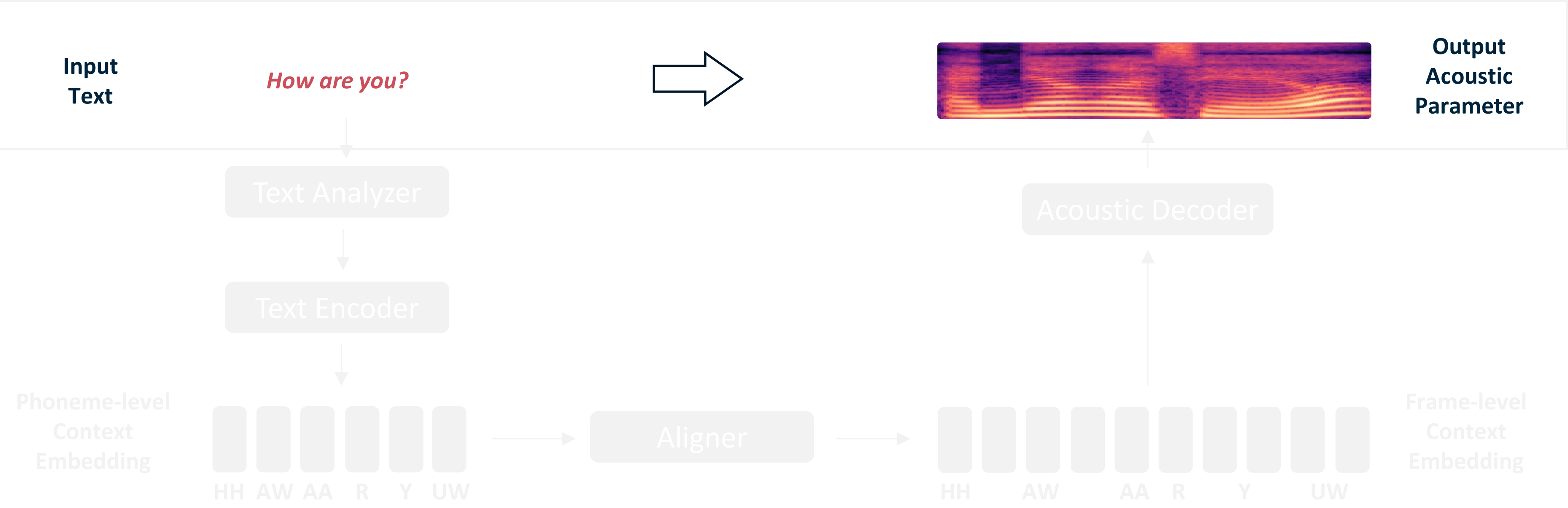
Speaker-specific attributes
(tone, volume, timbre, speaking rate, …)

주어진 입력 텍스트로 부터 사람의 음성 특성을 모델링 하는 태스크

TTS acoustic model

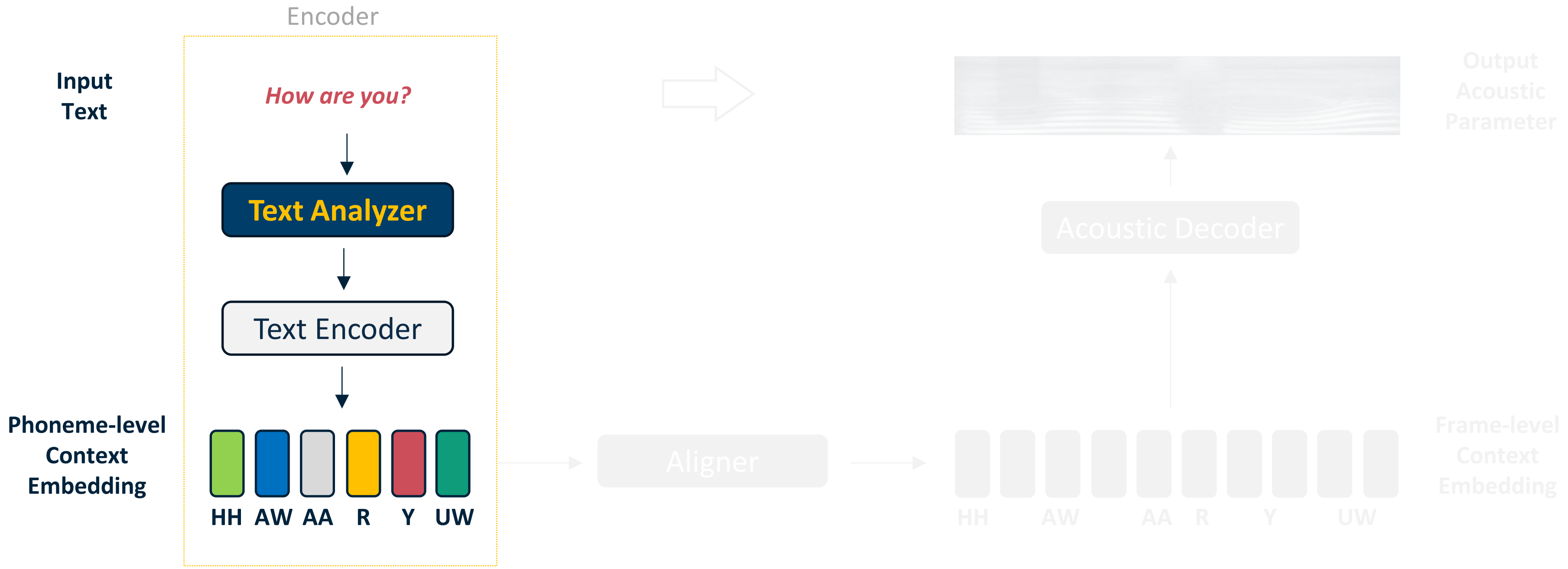
TTS acoustic model

How to generate acoustic parameters?



TTS acoustic model

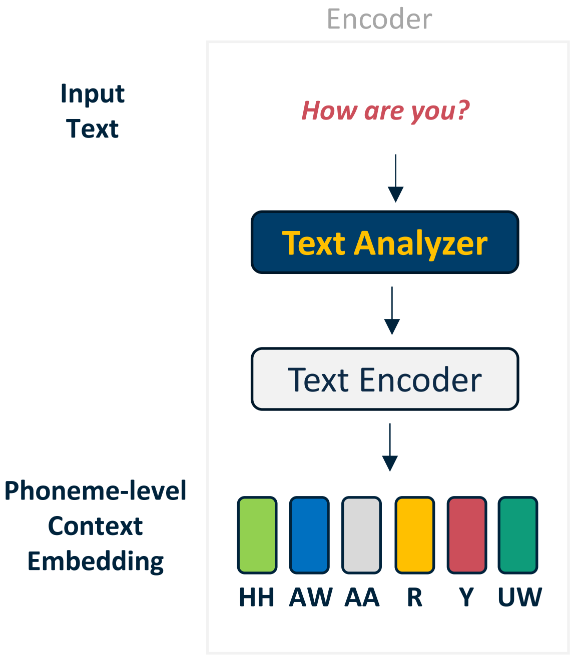
How to generate acoustic parameters?



Text analyzer extracts phoneme sequence from the given text

TTS acoustic model

How to generate acoustic parameters?



Text normalization

3.2km → 삼쩌미키로미터
naver.com → 네이버닷컴
1588-7942 → 이로팔팔칠구사이

Break prediction

삼쩌미키로미터 → 삼쩌미V키로미터
네이버닷컴 → 네이버닷컴
이로팔팔칠구사이 → 이로팔팔V칠구사이

Grapheme to phoneme conversion

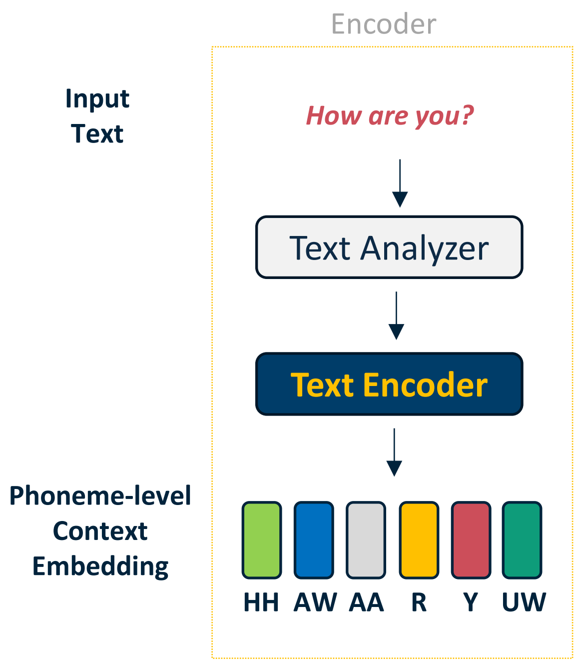
삼쩌미V키로미터 → ㅅ/ㅏ/ㅁ/ㅉ/ㅓ/ㅁ/ㅣ//ㅋ/ㅣ/ㄹ/ㅗ/ㅁ/ㅣ/ㅌ/ㅓ
네이버닷컴 → ㄴ/ㅔ/ㅣ/ㅂ/ㅓ/ㄷ/ㅏ/ㄷ/ㅋ/ㅓ/ㅁ
이로팔팔V칠구사이 → ㅣ/ㄹ/ㅗ/ㅍ/ㅏ/ㄹ/ㅍ/ㅏㄹ // ㅊ/ㅣ/ㄹ/ㄱ/ㅜ/ㅅ/ㅏ/ㅣ

Text analyzer extracts phoneme sequence from the given text

음소: 음운론상의 최소 단위

TTS acoustic model

How to generate acoustic parameters?



Text normalization

3.2km → 삼쩜미키로미터
naver.com → 네이버닫컴
1588-7942 → 이로팔팔칠구사이

Break prediction

삼쩜미키로미터 → 삼쩜미V키로미터
네이버닫컴 → 네이버닫컴
이로팔팔칠구사이 → 이로팔팔V칠구사이

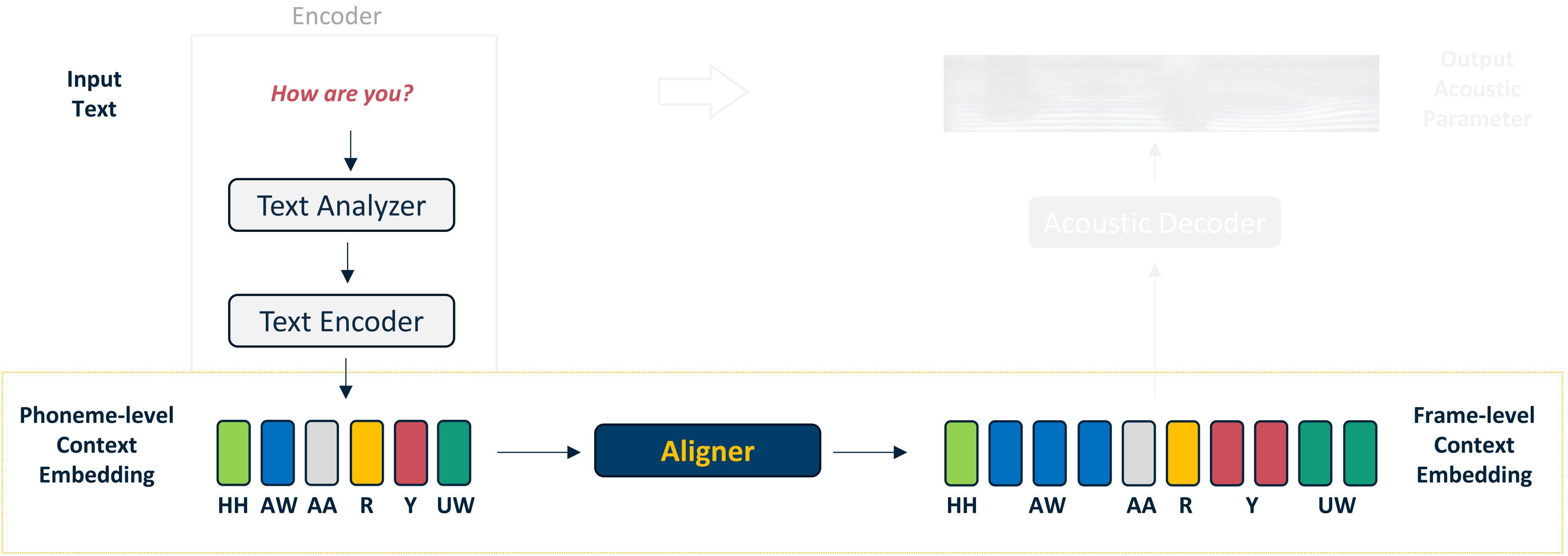
Grapheme to phoneme conversion

삼쩜미V키로미터 → ㅅ/ㅏ/ㅁ/ㅉ/ㅓ/ㅁ/ㅣ//ㅋ/ㅣ/ㄹ/ㅗ/ㅁ/ㅣ/ㅌ/ㅓ
네이버닫컴 → ㄴ/ㅔ/ㅣ/ㅂ/ㅓ/ㄷ/ㅏ/ㄷ/ㅋ/ㅓ/ㅁ
이로팔팔V칠구사이 → ㅣ/ㄹ/ㅗ/ㅍ/ㅏ/ㄹ/ㅍ/ㅏㄹ // ㅊ/ㅣ/ㄹ/ㄱ/ㅜ/ㅅ/ㅏ/ㅣ

Text encoder extracts high-level **context features** from the given phoneme sequence

TTS acoustic model

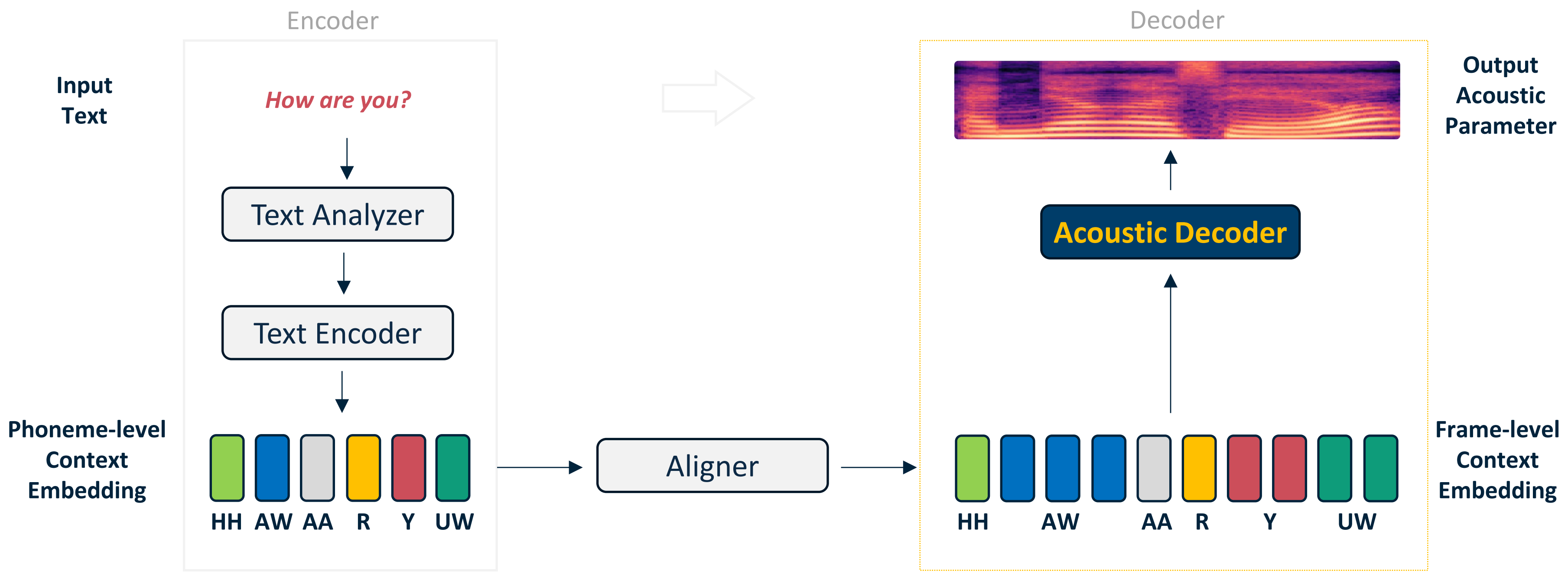
How to generate acoustic parameters?



Aligner upamples context embeddings from phoneme-level to frame-level

TTS acoustic model

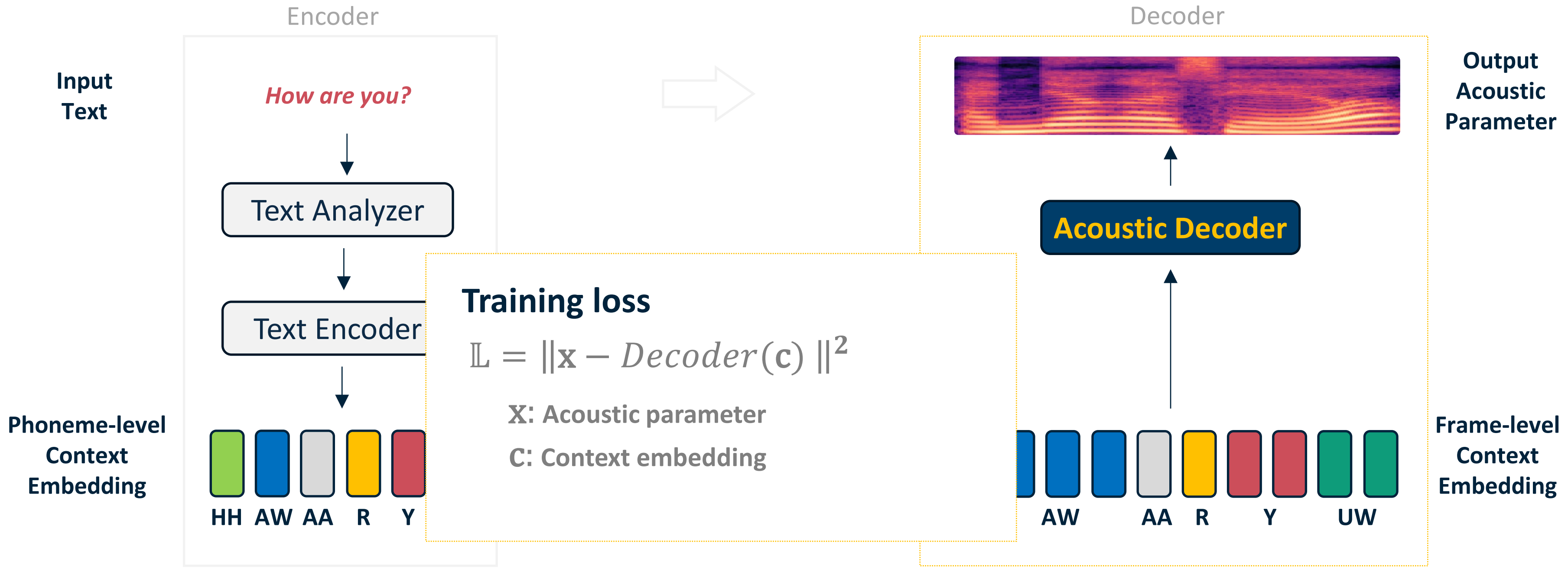
How to generate acoustic parameters?



Acoustic decoder predicts acoustic parameters from the given context embeddings

TTS acoustic model

How to generate acoustic parameters?



Training loss

$$\mathbb{L} = \|\mathbf{x} - Decoder(\mathbf{c})\|^2$$

$\mathbf{X}$: Acoustic parameter

$\mathbf{C}$: Context embedding

Acoustic decoder predicts acoustic parameters from the given context embeddings

TTS acoustic model

Statistical parametric speech synthesis (2023)

STATISTICAL PARAMETRIC SPEECH SYNTHESIS USING DEEP NEURAL NETWORKS

Heiga Zen, Andrew Senior, Mike Schuster

Google

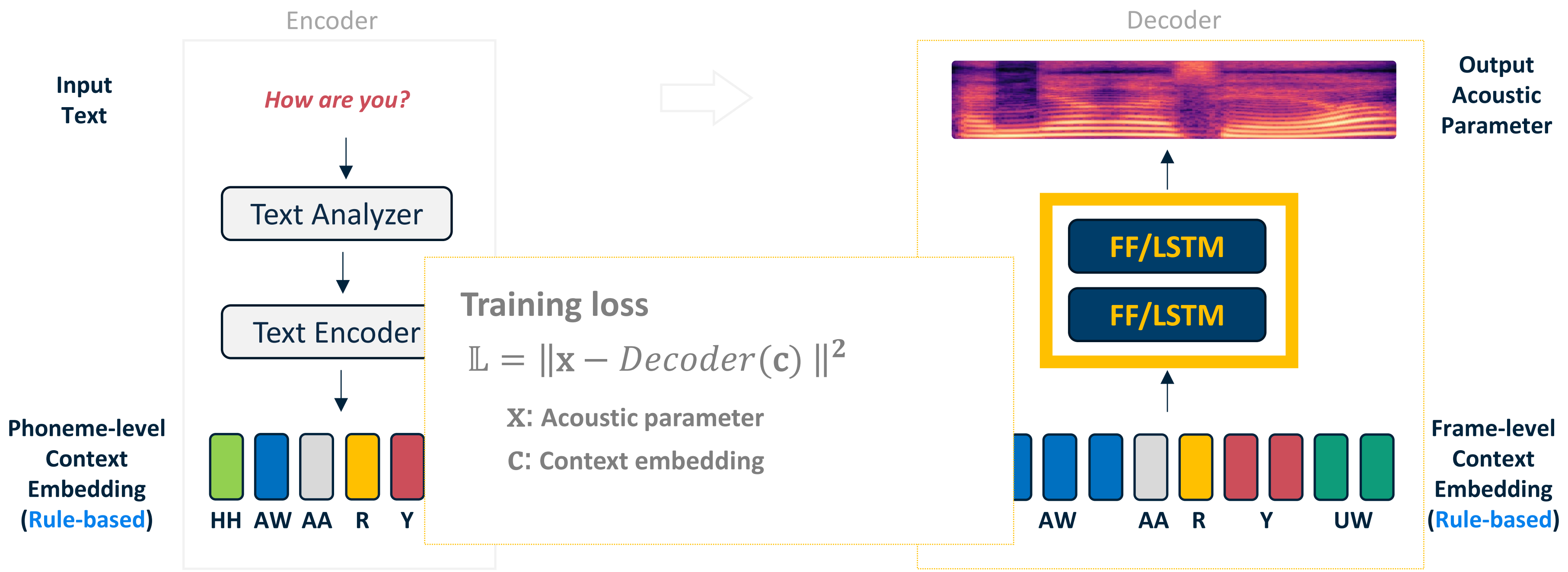
{heigazen,andrewsenior,schuster}@google.com

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.

TTS acoustic model

Statistical parametric speech synthesis (2023)



The first DNN model for the TTS acoustic model

TTS acoustic model

Tacotron 2 (2018)

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]

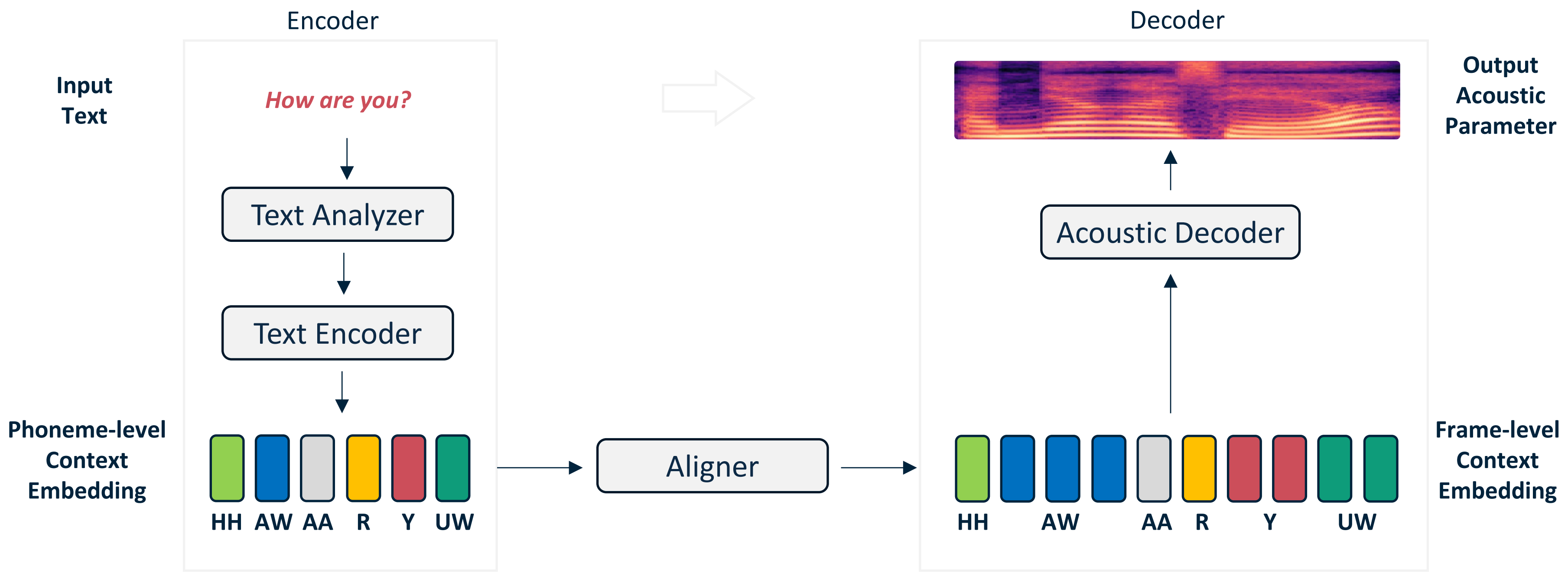
[1]Google, Inc., [2]University of California, Berkeley,
{jonathanasdf,rpang,yonghui}@google.com

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.

TTS acoustic model

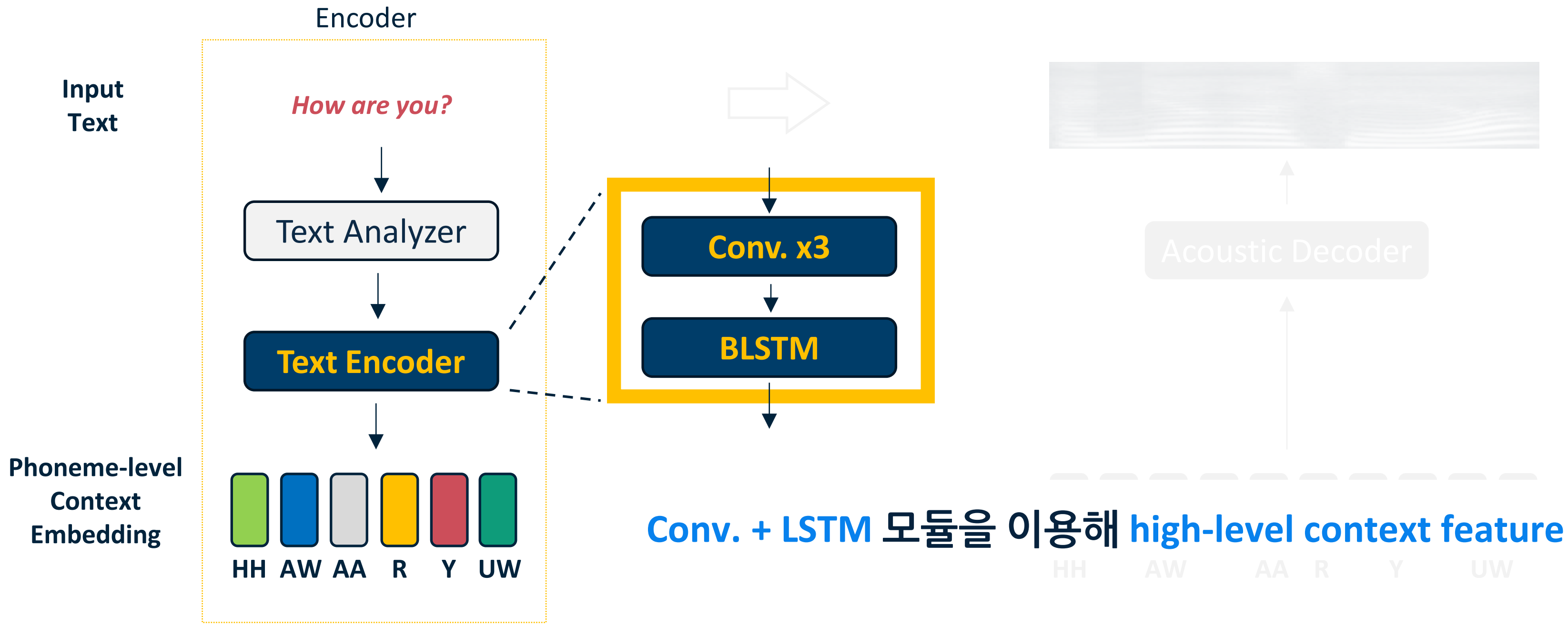
Tacotron 2 (2018)



The first seq2seq model for TTS acoustic model

TTS acoustic model

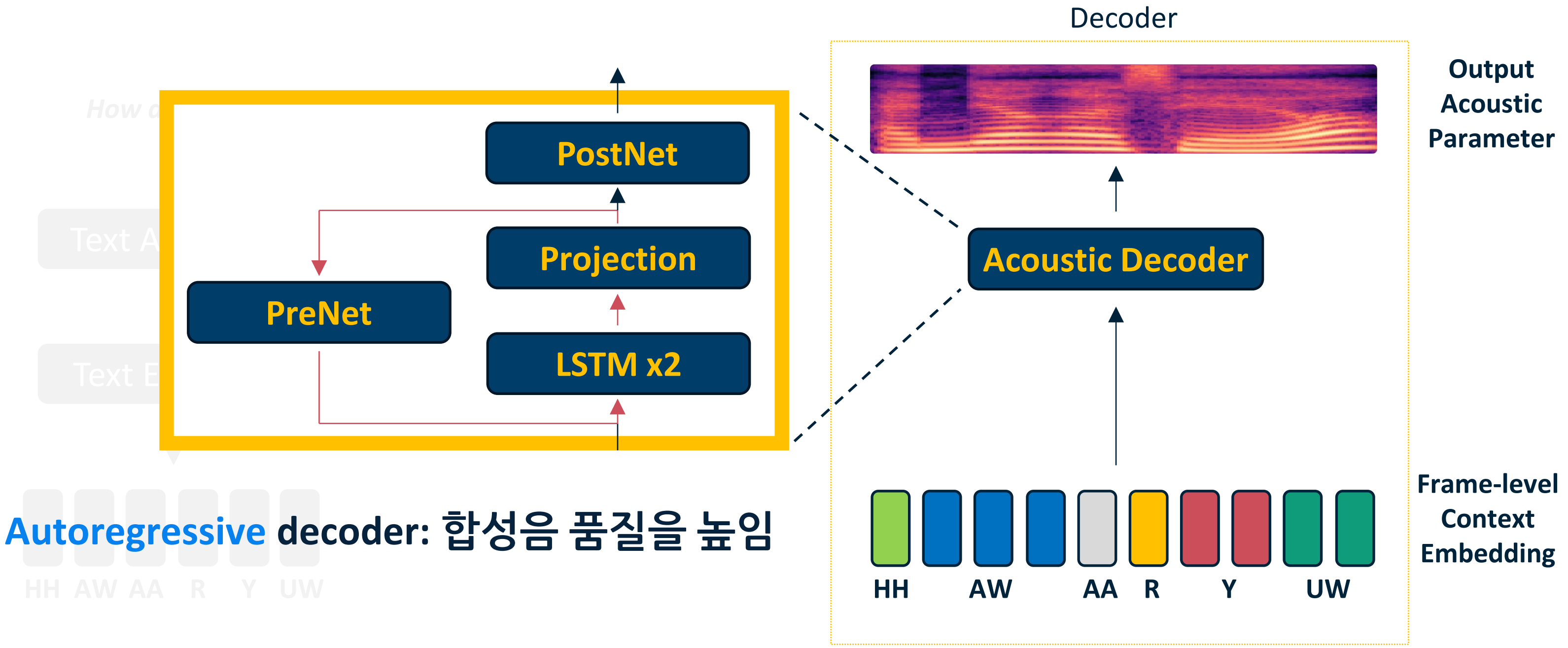
Tacotron 2 (2018)



Conv. + LSTM 모듈을 이용해 high-level context feature 를 얻어낼 수 있음

TTS acoustic model

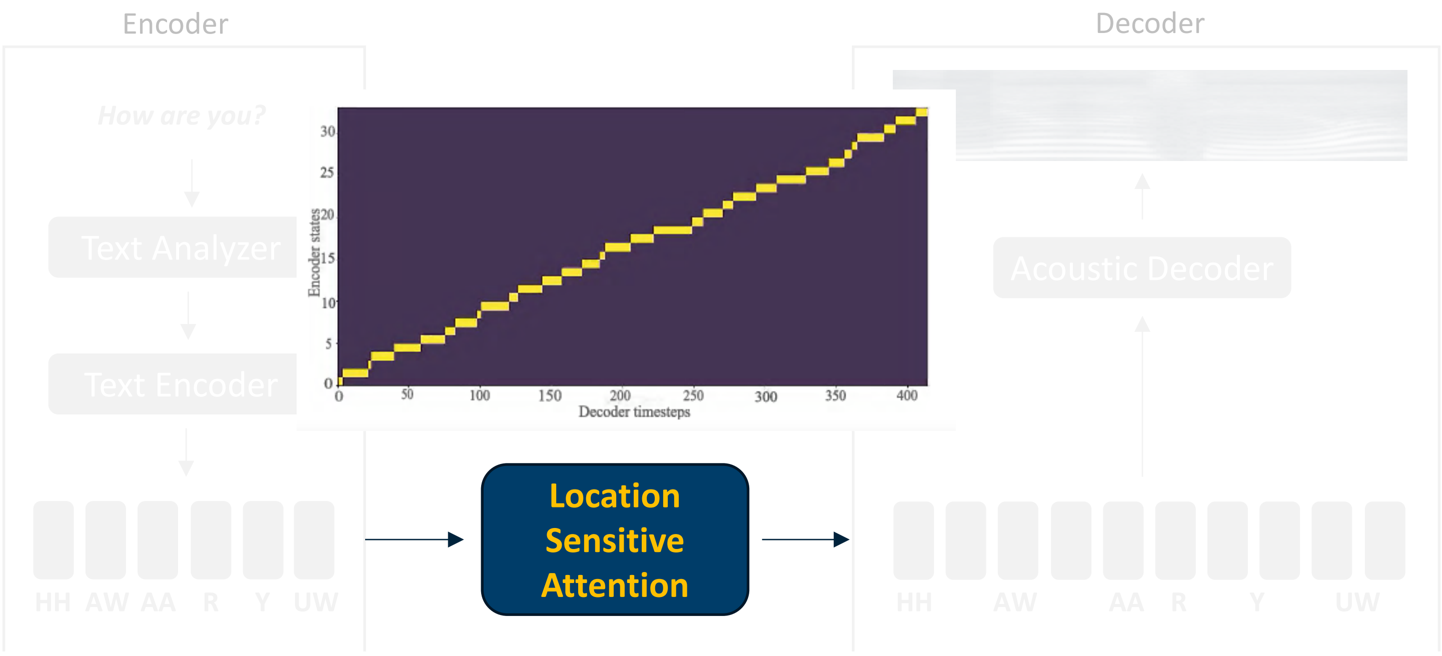
Tacotron 2 (2018)



Autoregressive decoder: 합성음 품질을 높임

TTS acoustic model

Tacotron 2 (2018)



Attention 메카니즘을 이용해 인코더-디코더 사이의 alignment 를 얻어낼 수 있음

TTS acoustic model

Tacotron 2 (2018)

System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 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

TTS acoustic model

FastSpeech 2 (2020)

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

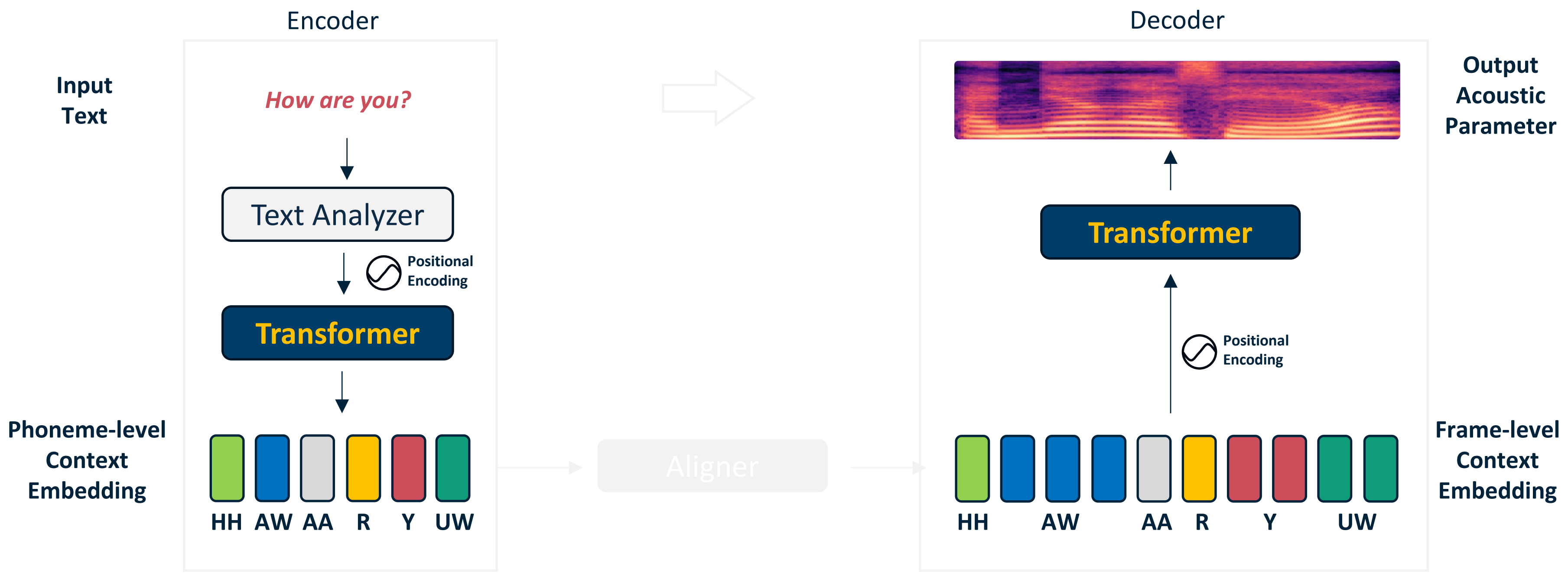
[3]Microsoft Azure Speech
Sheng.Zhao@microsoft.com

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/`.

TTS acoustic model

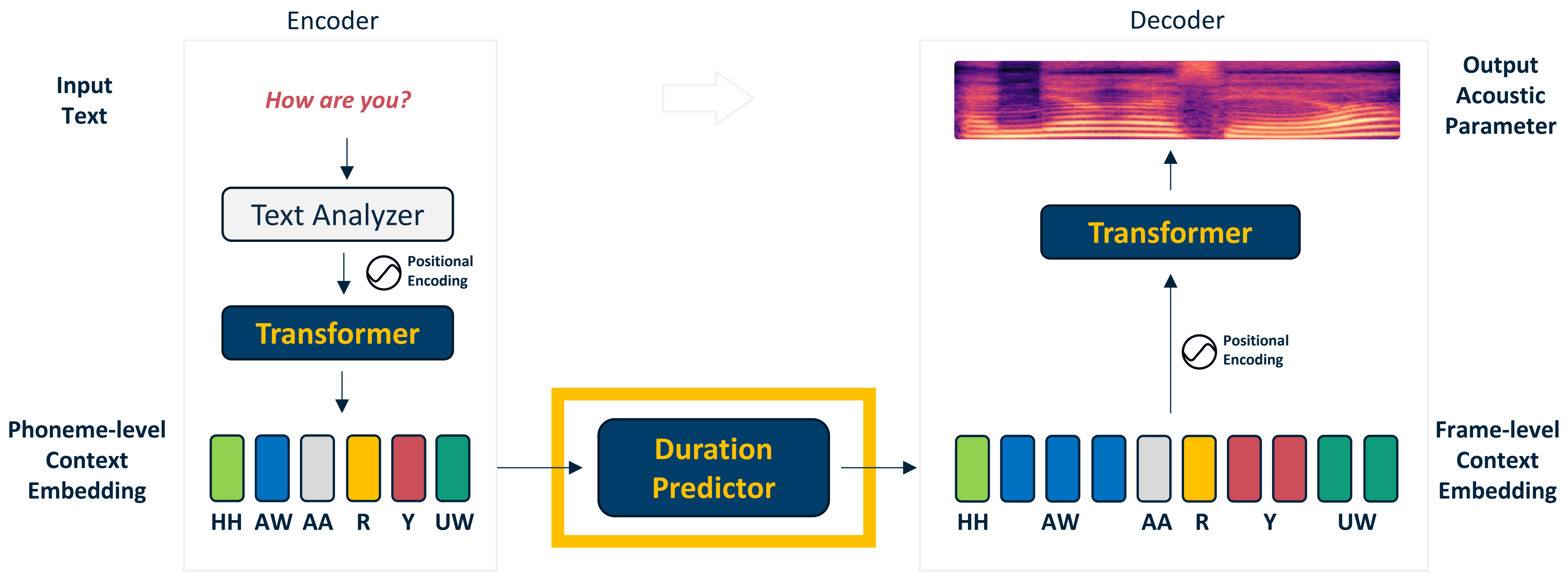
FastSpeech 2 (2020)



트랜스포머 기반의 인코더-디코더 사용

TTS acoustic model

FastSpeech 2 (2020)



Duration predictor-based alignment

파라미터 복원을 병렬로(non-autoregressive) 처리함으로써 생성 속도를 개선

Zero-shot Voice Cloning

Zero-shot voice cloning

Recording constraint

	Conventional TTS	Voice cloning
Recording amount	> 30~60 min	< Few seconds
Speaking type	Script reading	Spontaneous speaking
Speaker	Professional voice actor	Non-professional
Recording amount	Clean studio	Anywhere
TTS quality	Natural	Unnatural

Zero-shot voice cloning

Recording constraint

	Conventional TTS	Voice cloning
Recording amount	> 30~60 min	< Few seconds
Speaking type	Script reading	Spontaneous speaking
Speaker	Professional voice actor	Non-professional
Recording amount	Clean studio	Anywhere
TTS quality	Natural	Unnatural

Recording quality matters: Poor recording → TTS degradation

Zero-shot voice cloning

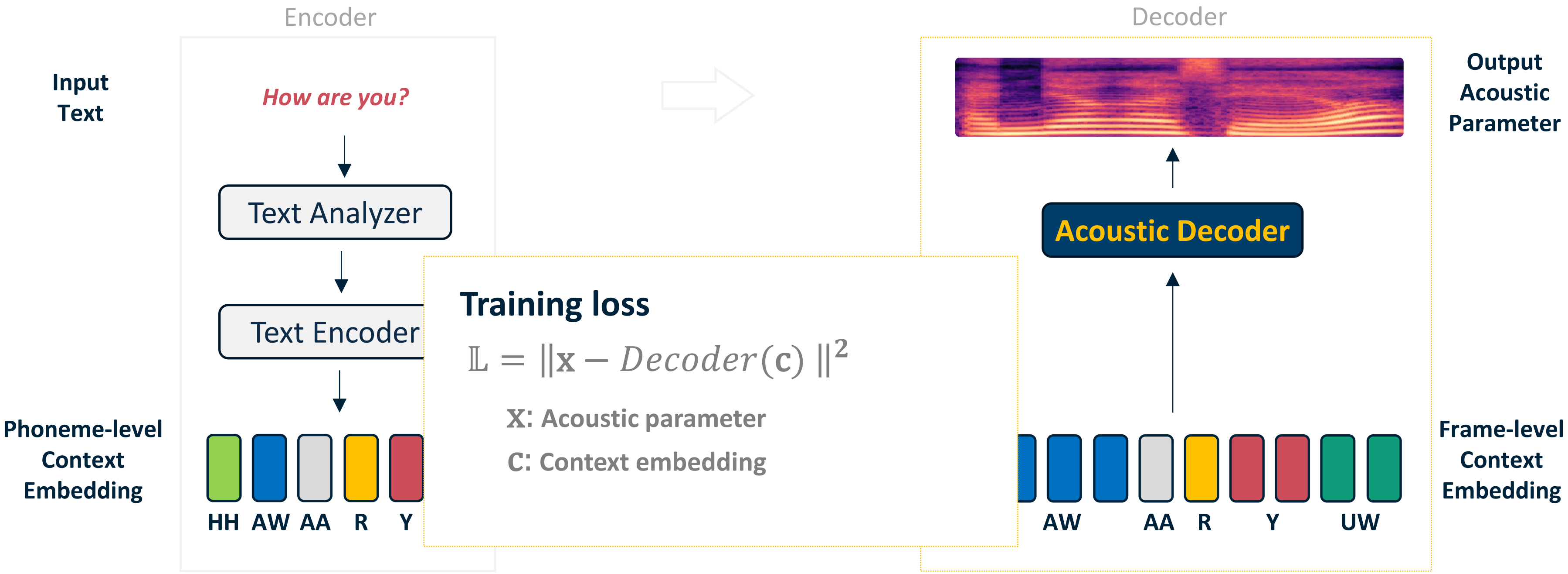
Recording constraint

	Conventional TTS	Voice cloning
Recording amount	> 30~60 min	< Few seconds
Speaking type	Script reading	Spontaneous speaking
Speaker	Professional voice actor	Non-professional
Recording amount	Clean studio	Anywhere
TTS quality	Natural	**Very natural**

Recording quality ~~matters~~: Poor recording → TTS ~~degradation~~

Zero-shot voice cloning

Recall – Conventional TTS



Training loss

$$\mathbb{L} = ||\mathbf{x} - Decoder(\mathbf{c})\ ||^2$$

$\mathbf{X}$: Acoustic parameter

$\mathbf{C}$: Context embedding

The model directly learns characteristic of the target voice..

→ **Output quality** is heavily **dependent** on **target** data

Zero-shot voice cloning

Key solution: Applying audio infilling task

Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale

Matthew Le* **Apoorv Vyas*** **Bowen Shi*** **Brian Karrer*** **Leda Sari** **Rashel Moritz**
Mary Williamson **Vimal Manohar** **Yossi Adi†** **Jay Mahadeokar** **Wei-Ning Hsu***

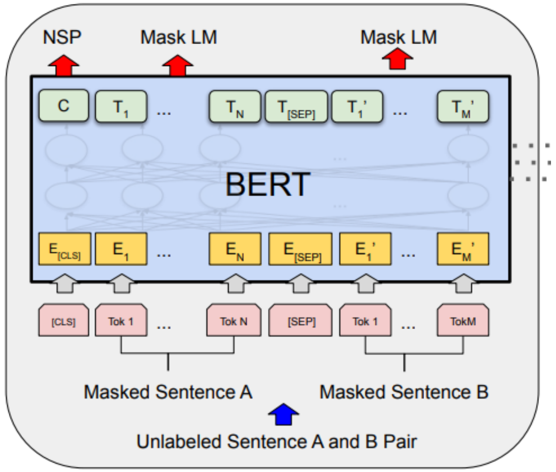
Fundamental AI Research (FAIR), Meta

Abstract

Large-scale generative models such as GPT and DALL-E have revolutionized natural language processing and computer vision research. These models not only generate high fidelity text or image outputs, but are also generalists which can solve tasks not explicitly taught. In contrast, speech generative models are still primitive in terms of scale and task generalization. In this paper, we present Voicebox, the most versatile text-guided generative model for speech at scale. Voicebox is a non-autoregressive flow-matching model trained to infill speech, given audio context and text, trained on over 50K hours of speech that are neither filtered nor enhanced. Similar to GPT, Voicebox can perform many different tasks through in-context learning, but is more flexible as it can also condition on future context. Voicebox can be used for mono or cross-lingual zero-shot text-to-speech synthesis, noise removal, content editing, style conversion, and diverse sample generation. In particular, Voicebox outperforms the state-of-the-art zero-shot TTS model VALL-E on both intelligibility (5.9% vs 1.9% word error rates) and audio similarity (0.580 vs 0.681) while being up to 20 times faster. Audio samples can be found in `https://voicebox.metademolab.com`.

Zero-shot voice cloning

Key solution: Applying audio infilling task



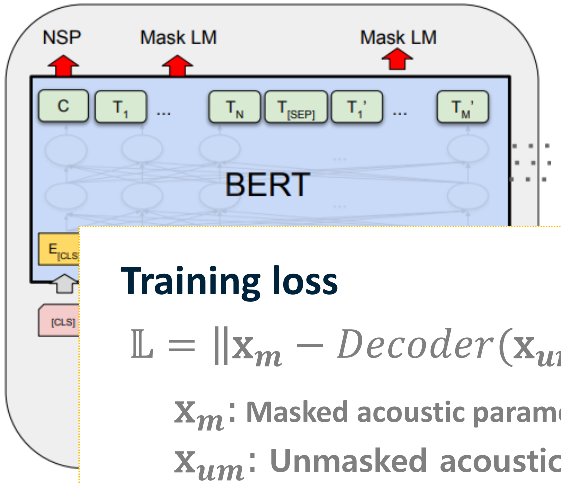
J. Devlin, et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

Inspired by BERT's masked language modeling,

the model is trained to predict masked acoustic parameters using neighboring acoustic information

Zero-shot voice cloning

Key solution: Applying audio infilling task



Training loss

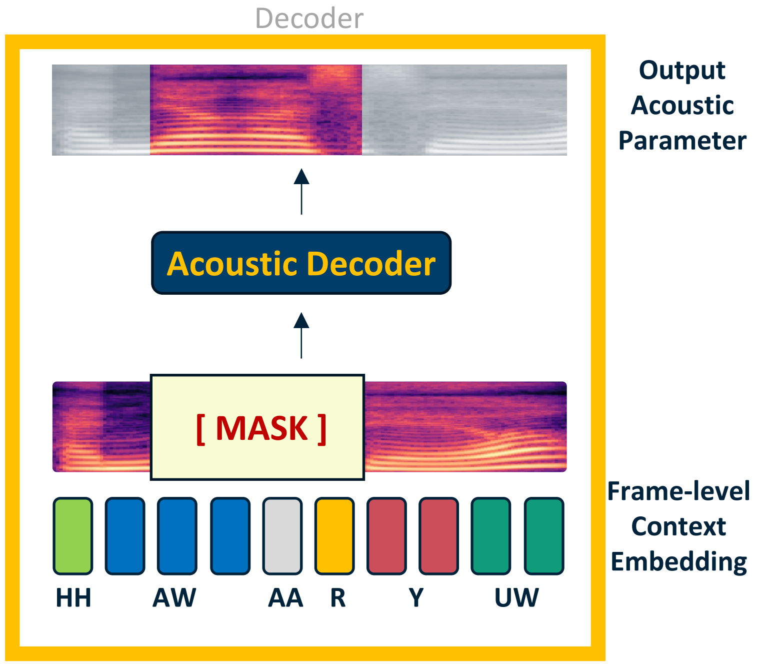
$$\mathbb{L} = \|\mathbf{x}_m - Decoder(\mathbf{x}_{um}, \mathbf{c})\|^2$$

$\mathbf{x}_m$: Masked acoustic parameter

$\mathbf{x}_{um}$: Unmasked acoustic parameter

C: Context embedding

J. Devlin, et al., "BERT: F
understanding," *arXiv pre*



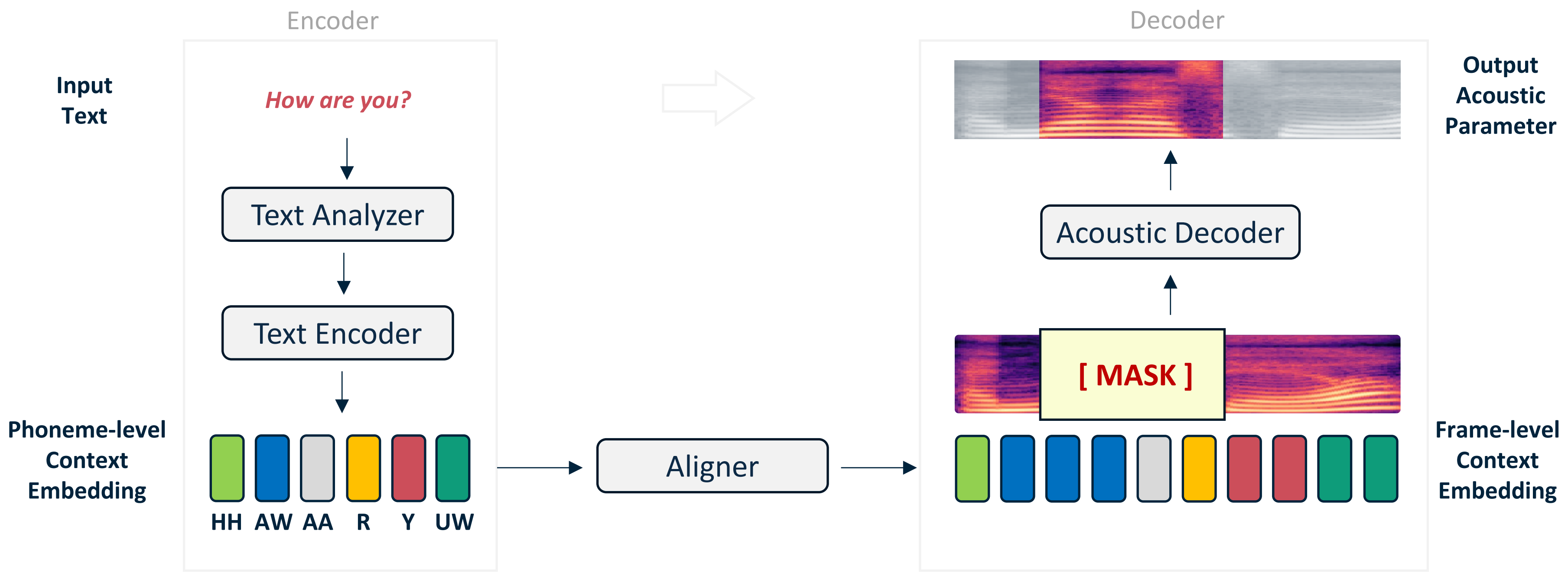
Inspired by BERT's masked language modeling,

the model is trained to predict masked acoustic parameters using neighboring acoustic information

Zero-shot voice cloning

Key solution: Applying audio infilling task

Training with large-scale speech corpora



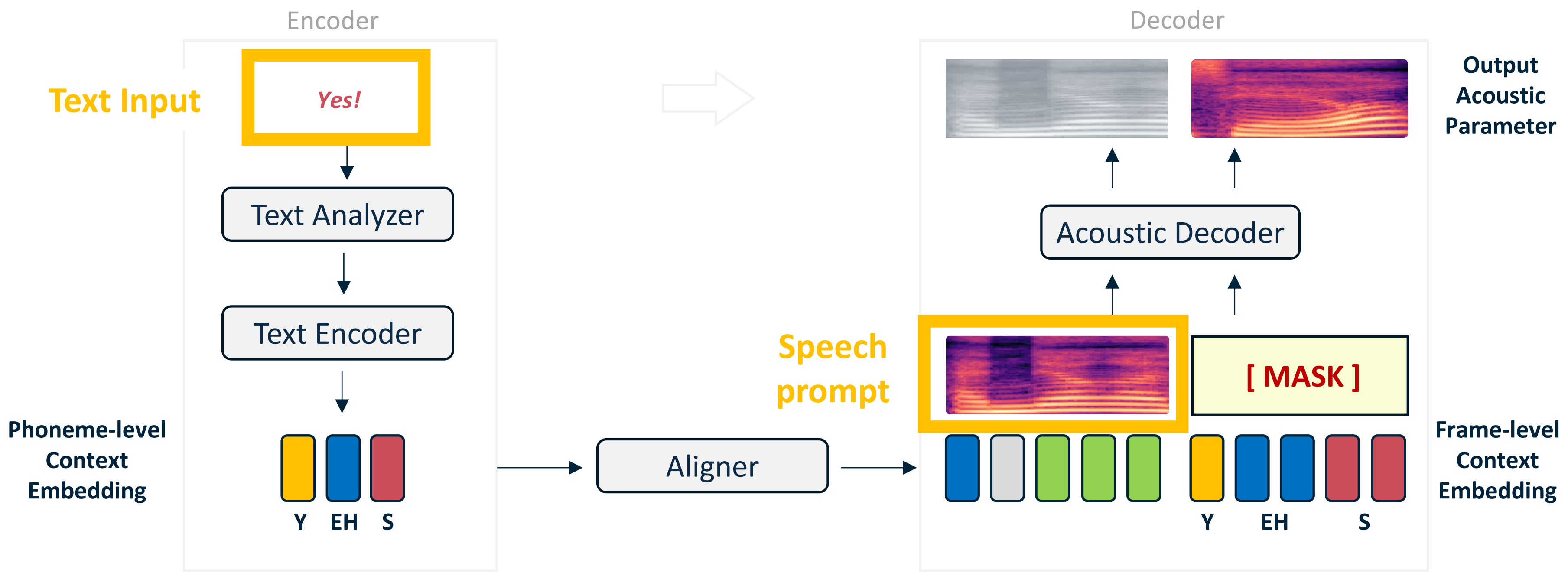
The model focuses on relationship between adjacent acoustic parameters,

rather than reconstructing the target data

Zero-shot voice cloning

Key solution: Applying audio infilling task

Inference with 5~10 seconds speech prompt



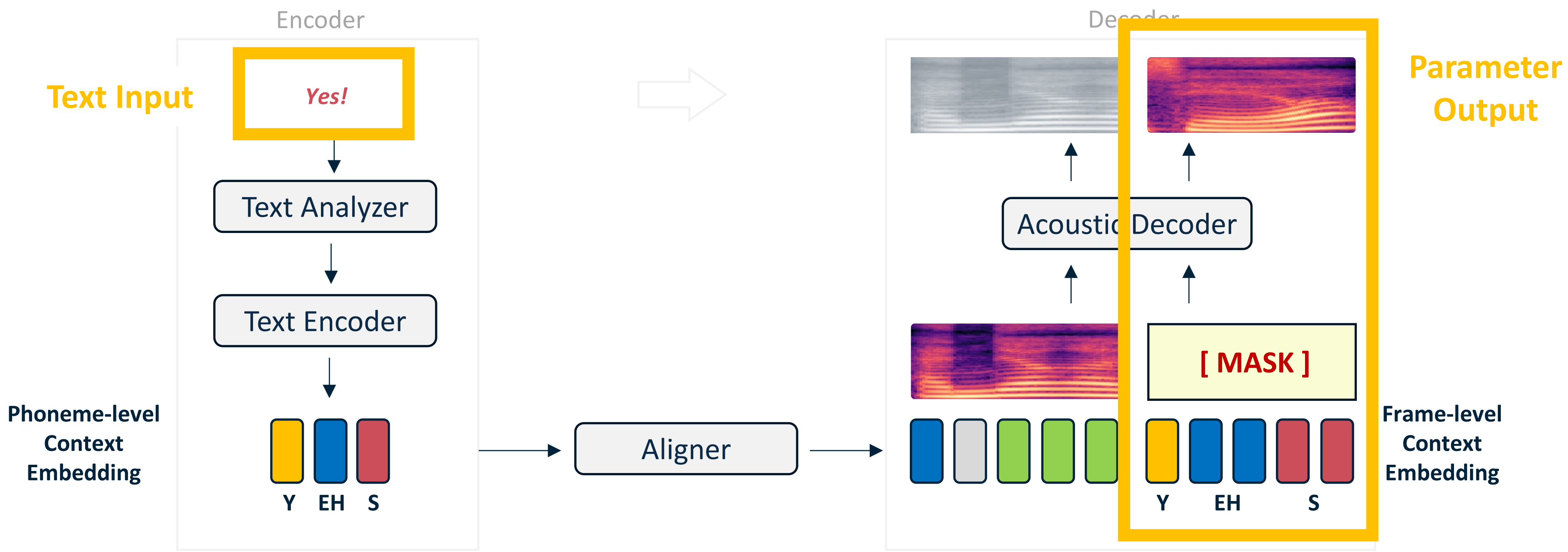
From the given text and speech prompt,

the model generates corresponding acoustic parameters

Zero-shot voice cloning

Key solution: Applying audio infilling task

Inference with 5~10 seconds speech prompt



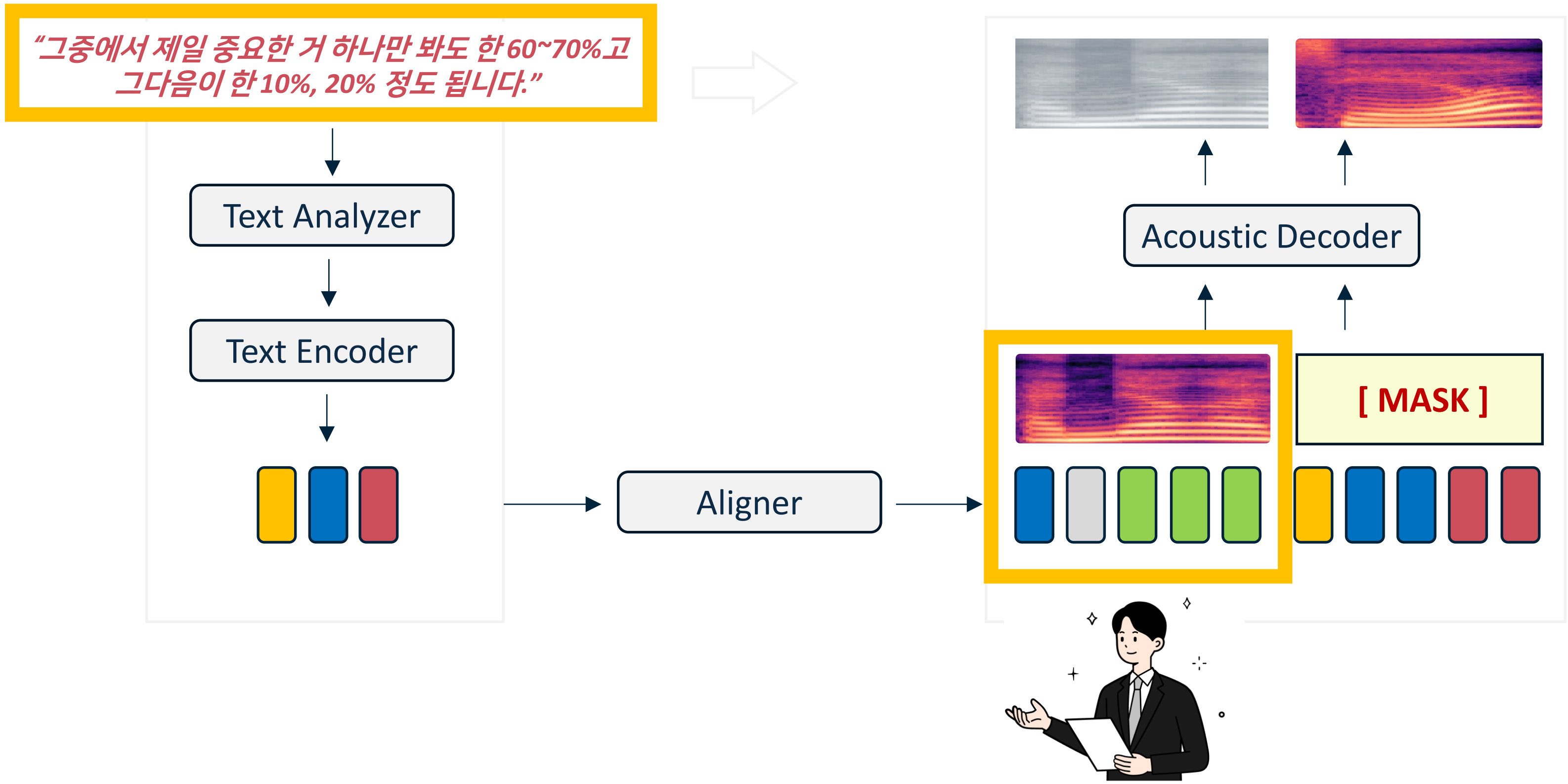
From the given text and speech prompt,

the model generates corresponding acoustic parameters

Zero-shot voice cloning

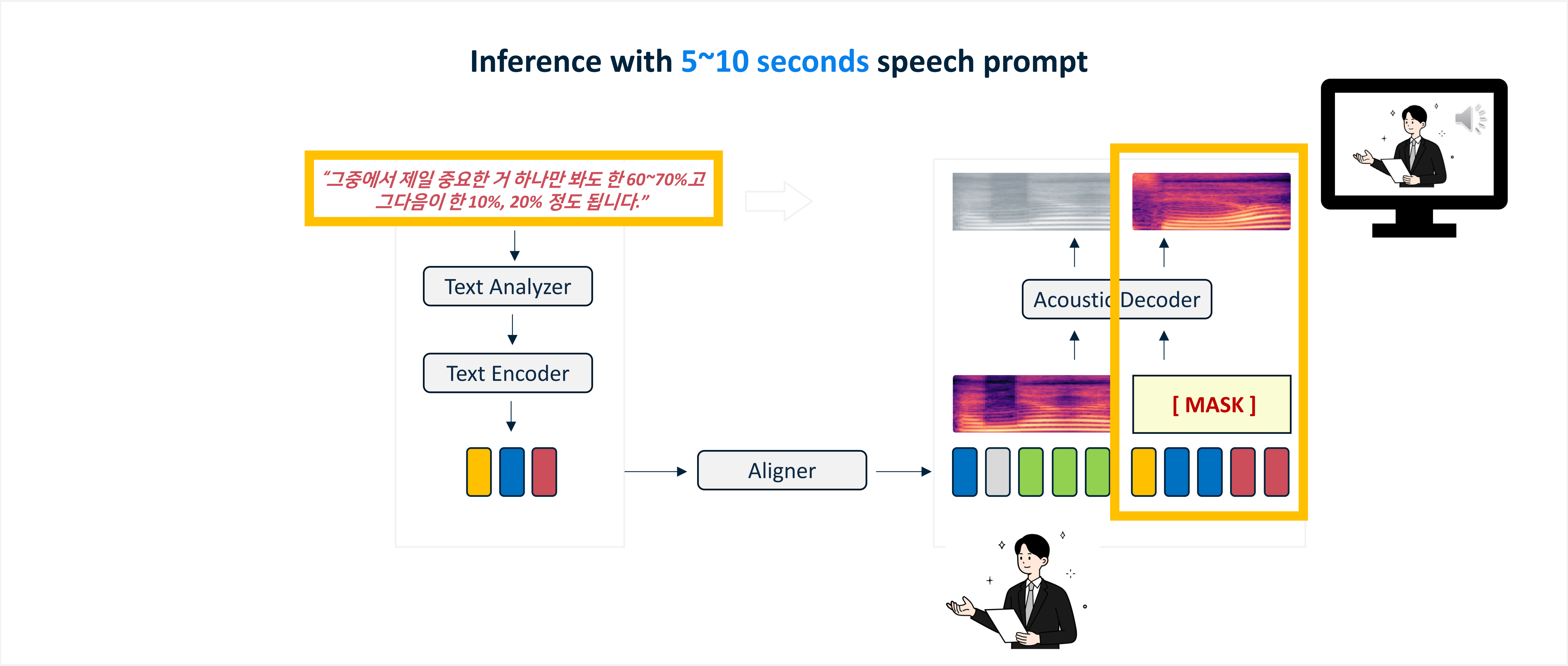
Key solution: Applying audio infilling task

Inference with 5~10 seconds speech prompt



Zero-shot voice cloning

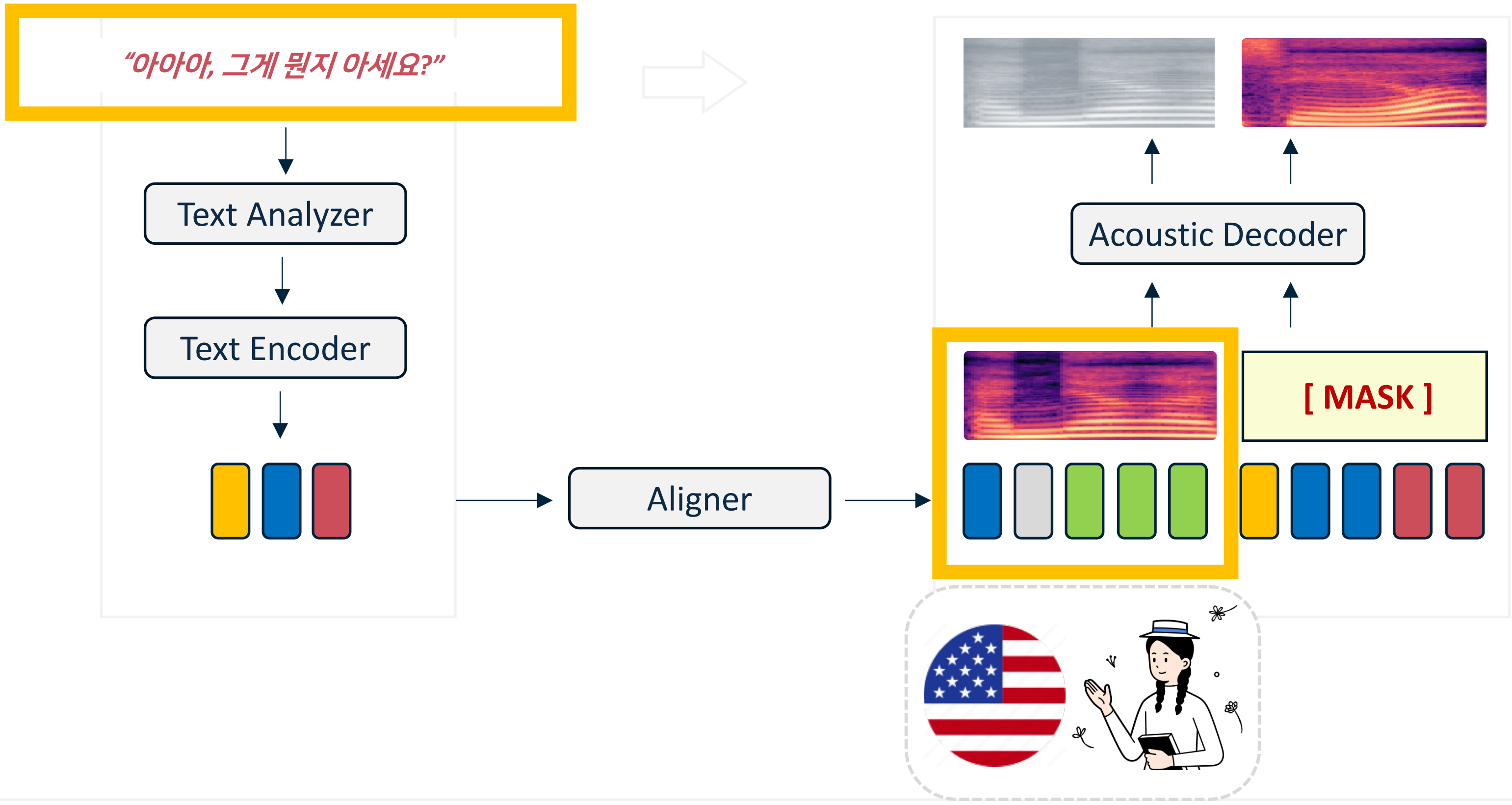
Key solution: Applying audio infilling task



Zero-shot voice cloning

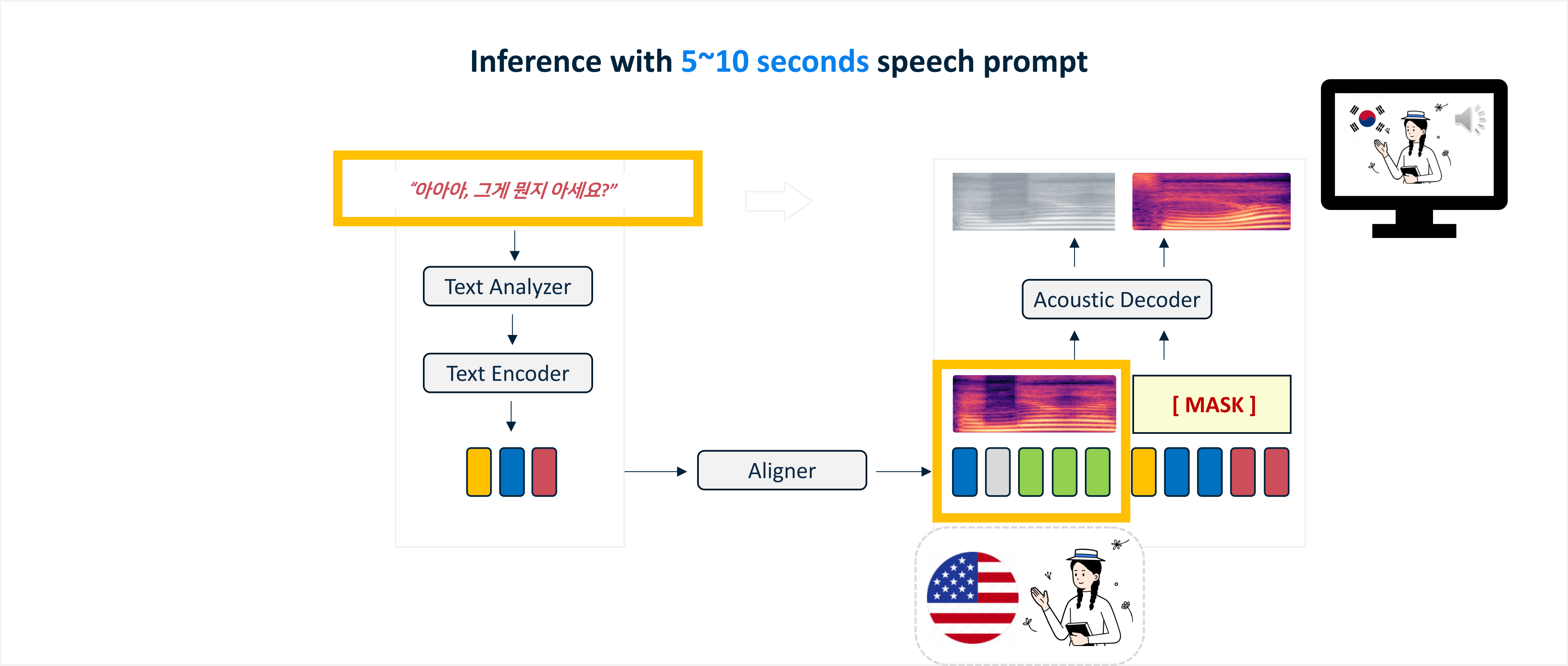
Key solution: Applying audio infilling task

Inference with 5~10 seconds speech prompt



Zero-shot voice cloning

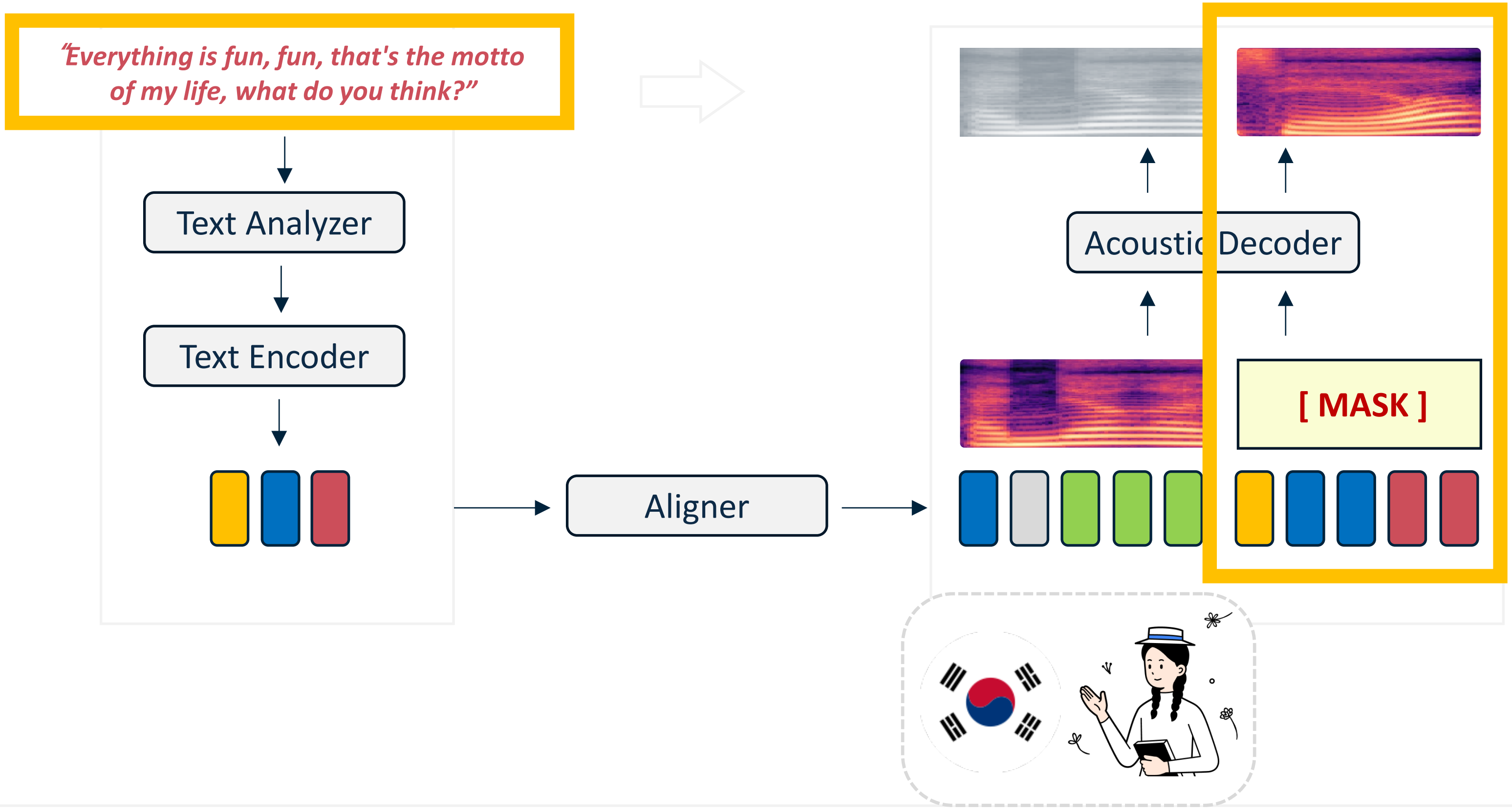
Key solution: Applying audio infilling task



Zero-shot voice cloning

Key solution: Applying audio infilling task

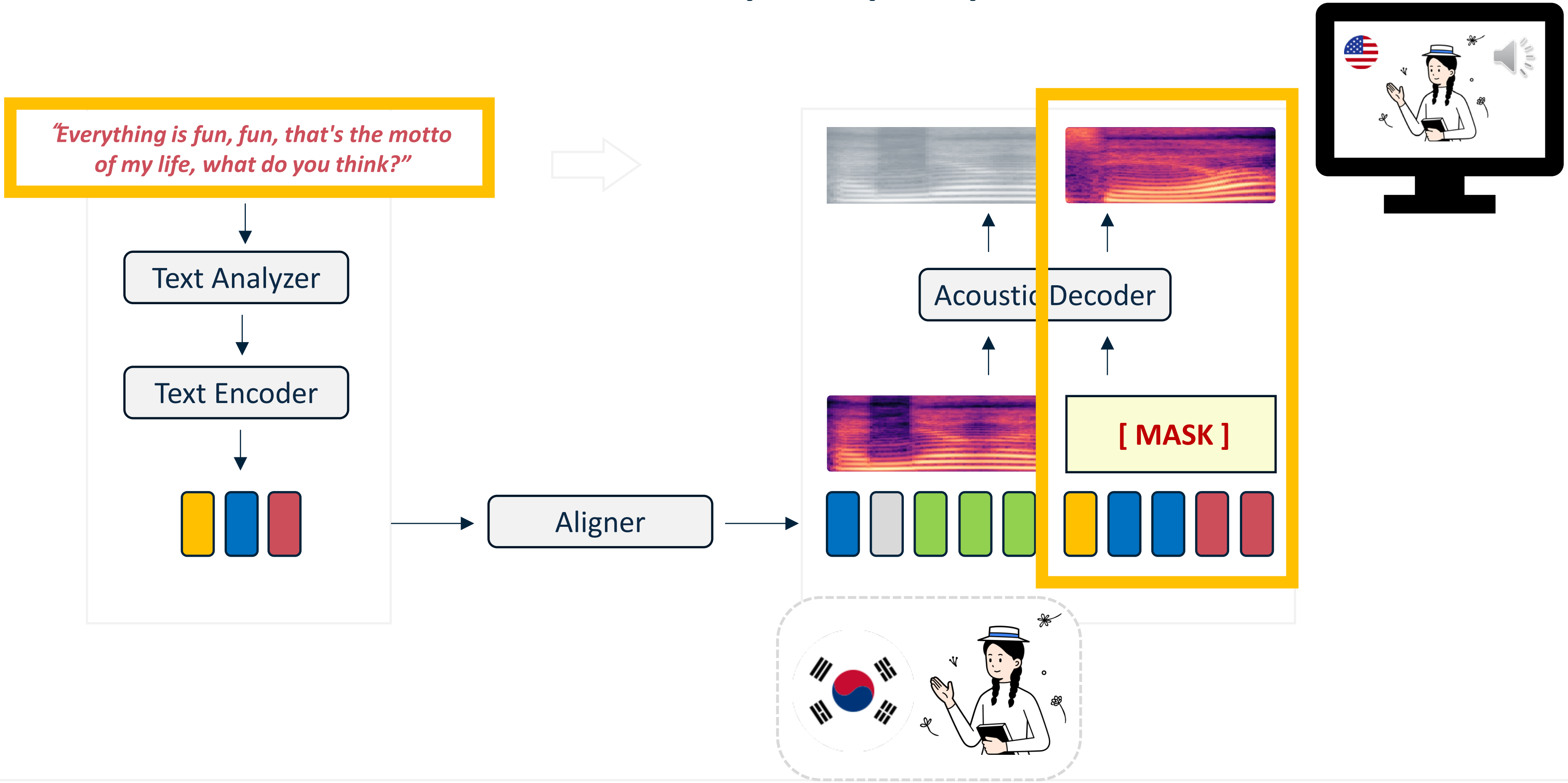
Inference with 5~10 seconds speech prompt



Zero-shot voice cloning

Key solution: Applying audio infilling task

Inference with 5~10 seconds speech prompt



Zero-shot voice cloning

Overcoming the recording constraint

	Conventional TTS	Voice cloning
Speaker	Professional voice actor	Non-professional
Recording environment	Clean studio	Anywhere
Recording amount	> 30~60 min	< Few seconds
Speaking type	Clean studio	Anywhere
Model size	0.03B	0.41B
Inference speed	Real time x5 (CPU)	Real time x5 (GPU)
TTS quality	Natural	**Very natural**

Zero-shot voice cloning

Evaluations

	Conventional TTS	Voice cloning
Dataset	4 Korean speakers (2 male + 2 female)	
	~1h / speaker	4~8s / speaker
Speaker similarity (SECS)↑	68.0 %	78.3%
Intelligibility (CER)↓	1.8%	1.1%
Naturalness (MOS)↑	4.2	4.4

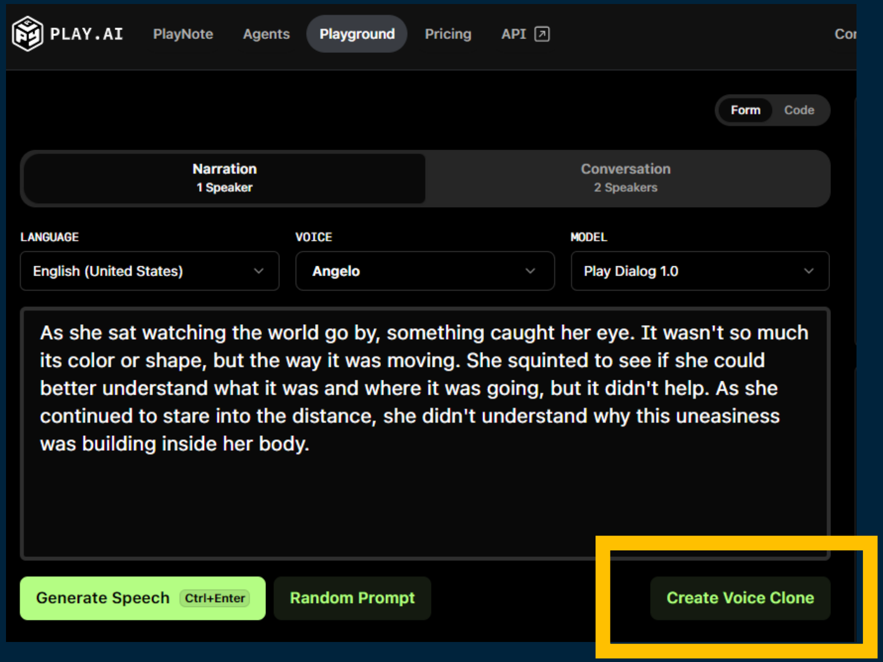
SECS; speaker embedding cosine similarity: 스피커 임베딩 벡터간의 유사도
CER; character error rate: 입력 텍스트 ↔ 출력 음성의 ASR 결과(텍스트)간의 오류율
MOS; mean opinion score: 전문가 청취평가(1~5 scale)

Zero-shot voice cloning

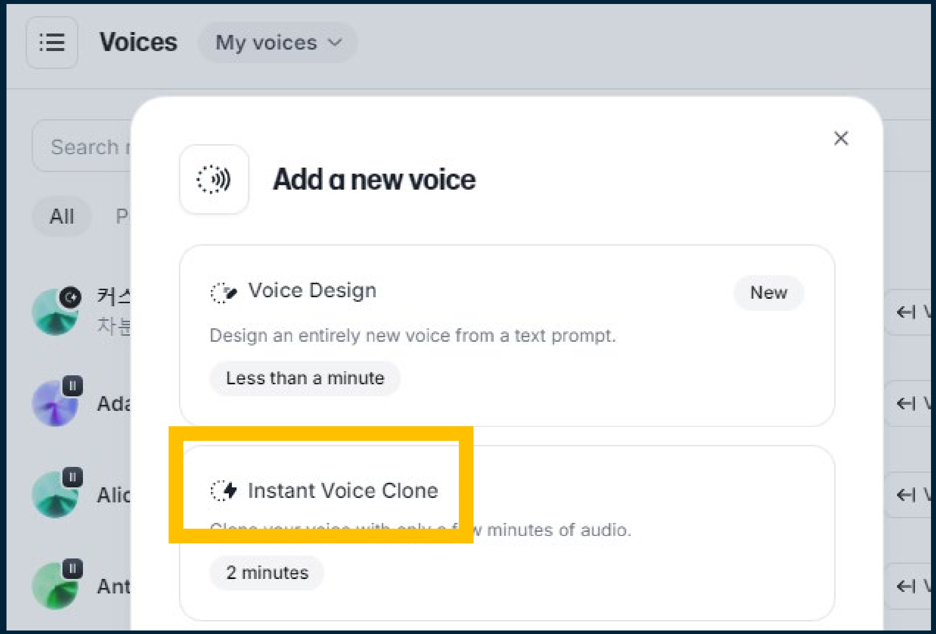
Examples

Recording	Conventional TTS	Voice cloning
		
	슈팅 사정거리 안에서 기회가 왔을 때는 좀 더 과감하게 시도를 해 줘야죠.	
		
	사건을 배당받은 서울중앙지검 공공형사수사부는 기초 자료 검토를 시작했습니다.	

Zero-shot voice cloning

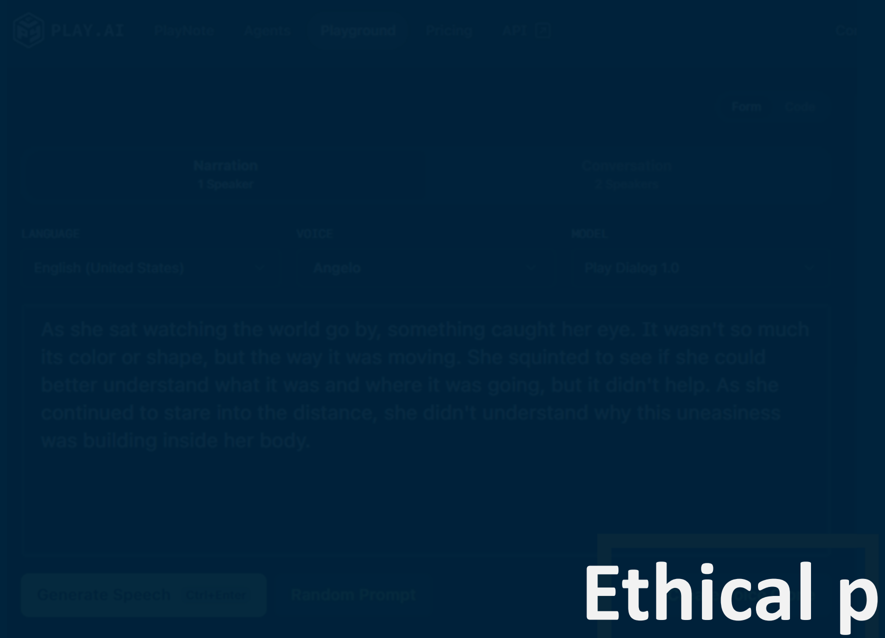


https://play.ht/



https://elevenlabs.io/

Zero-shot voice cloning



Ethical problem ?

https://play.ht/

https://elevenlabs.io/

Q/A

gregorio.song@gmail.com