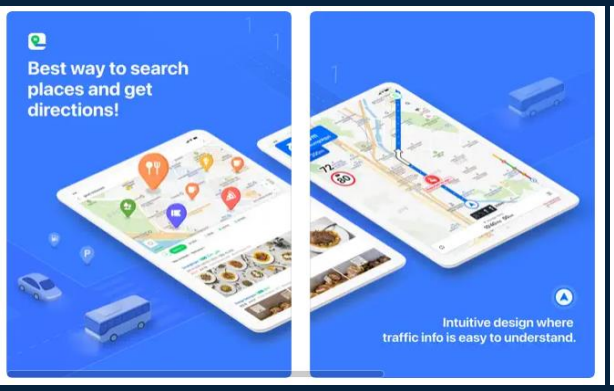
AI Human: Large-Scale Text-to-Speech Applications

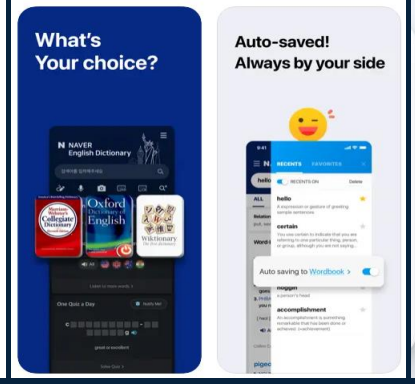
Eunwoo Song / NAVER Cloud

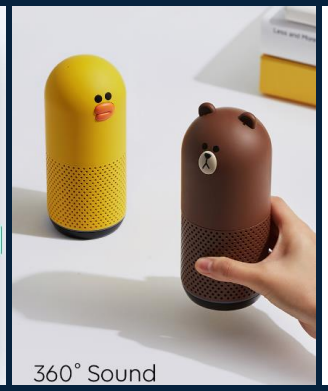
Who am I?

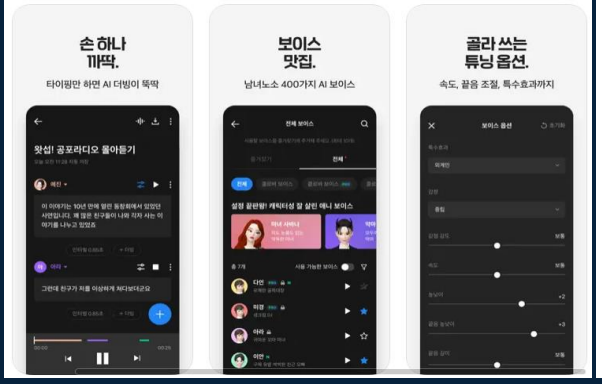
Developing Text-to-Speech engines for

- NAVER: Maps, Dictionary, News
- NAVER CLOVA: AI Speaker, Dubbing Editor
- NAVER Cloud: Papago Translator, Voice API
- …



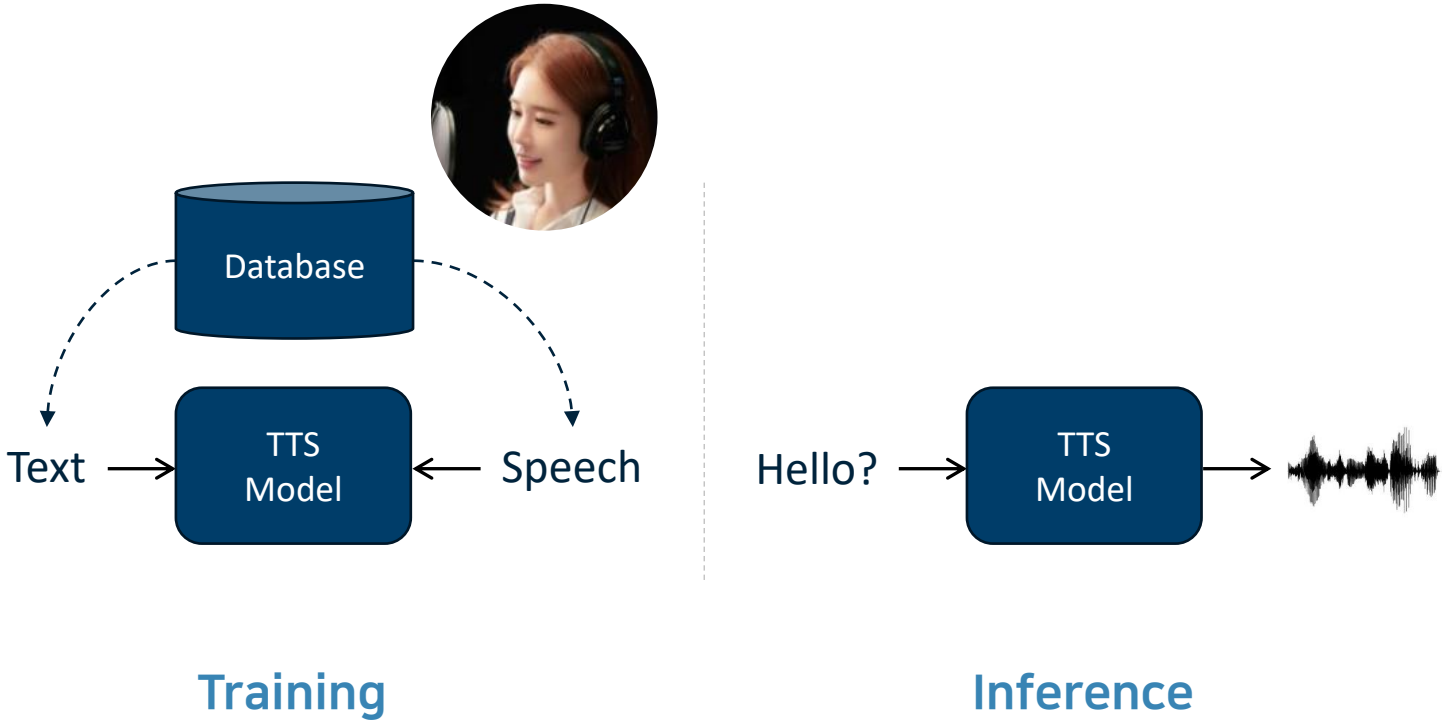




Introduction

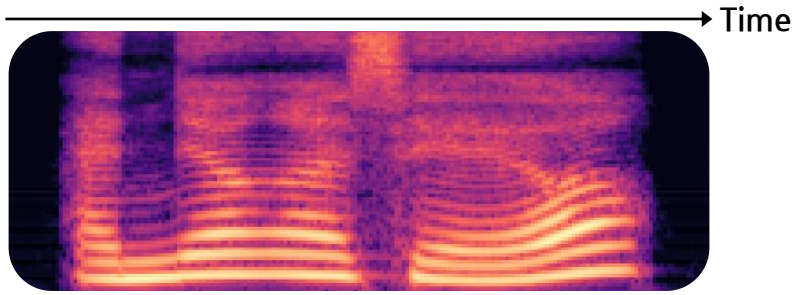
Deep learning-based TTS system



Human-like voice quality ☺

Introduction

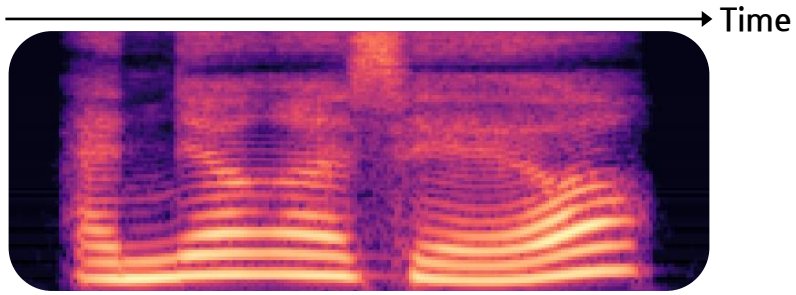
Deep learning-based TTS system



Acoustic model + Vocoding model

Introduction

Deep learning-based TTS system



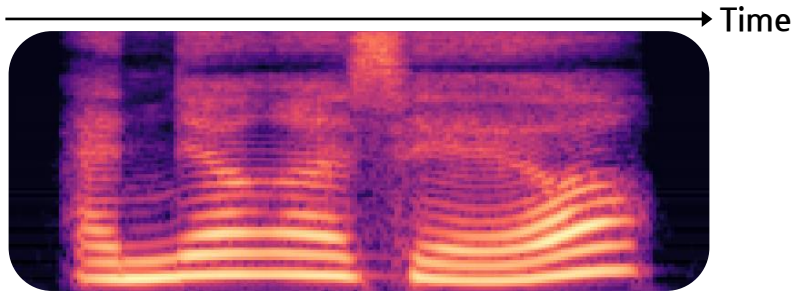
Estimating acoustic parameters from text inputs

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

Acoustic model + Vocoding model

Introduction

Deep learning-based TTS system

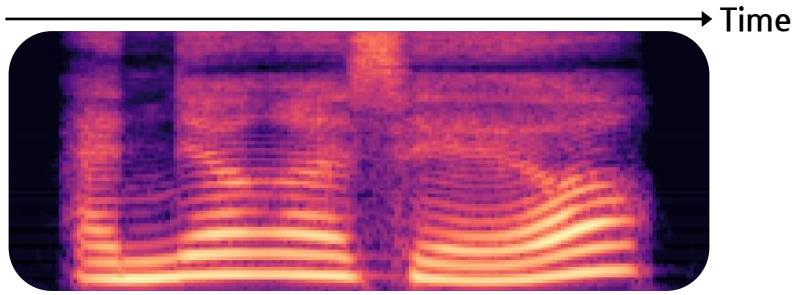


Estimating speech signals from acoustic parameters

Acoustic model + Vocoding model

Introduction

Deep learning-based TTS system

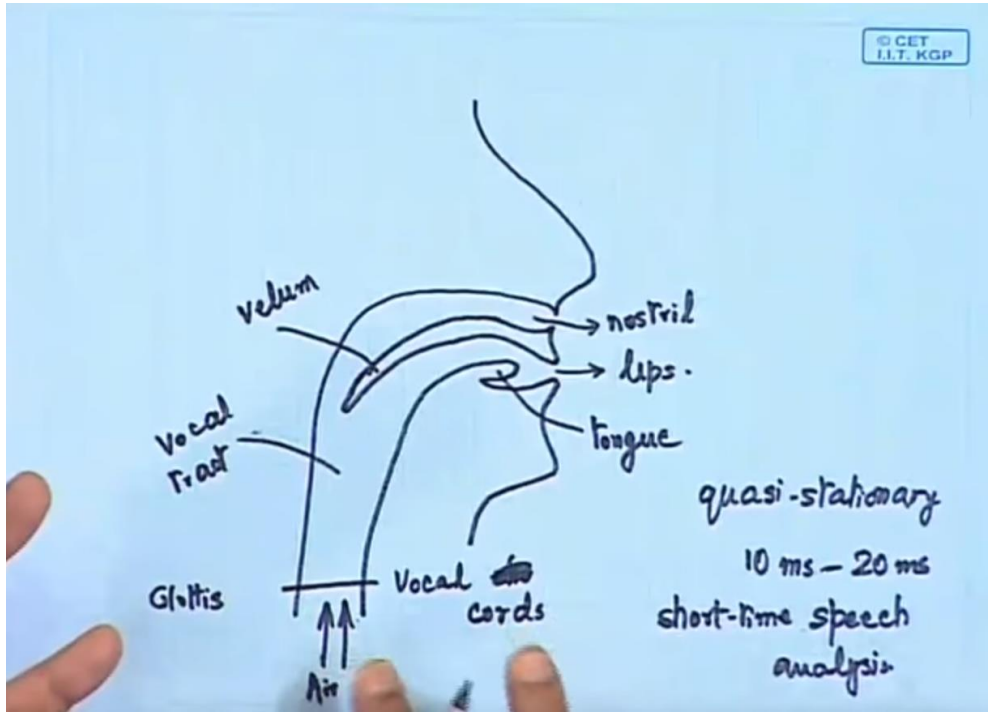


Acoustic parameters..?

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

Introduction

Speech production model



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

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

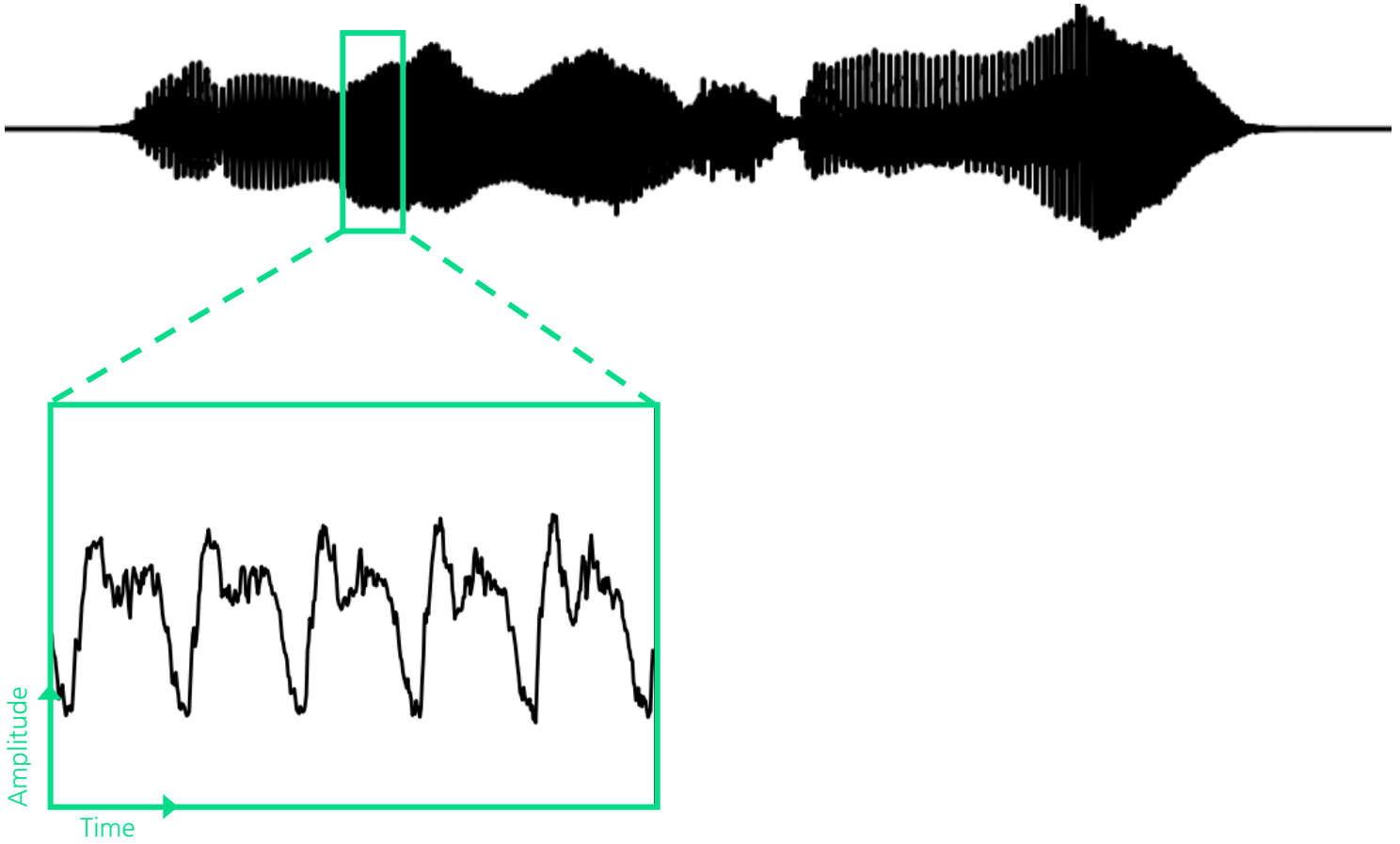
  → **목소리의 톤을 결정 (아↘아↗)**

- Vocal tract
  - Shaping voice color

  → **발음을 결정 (아/에/이/오/우)**

Introduction

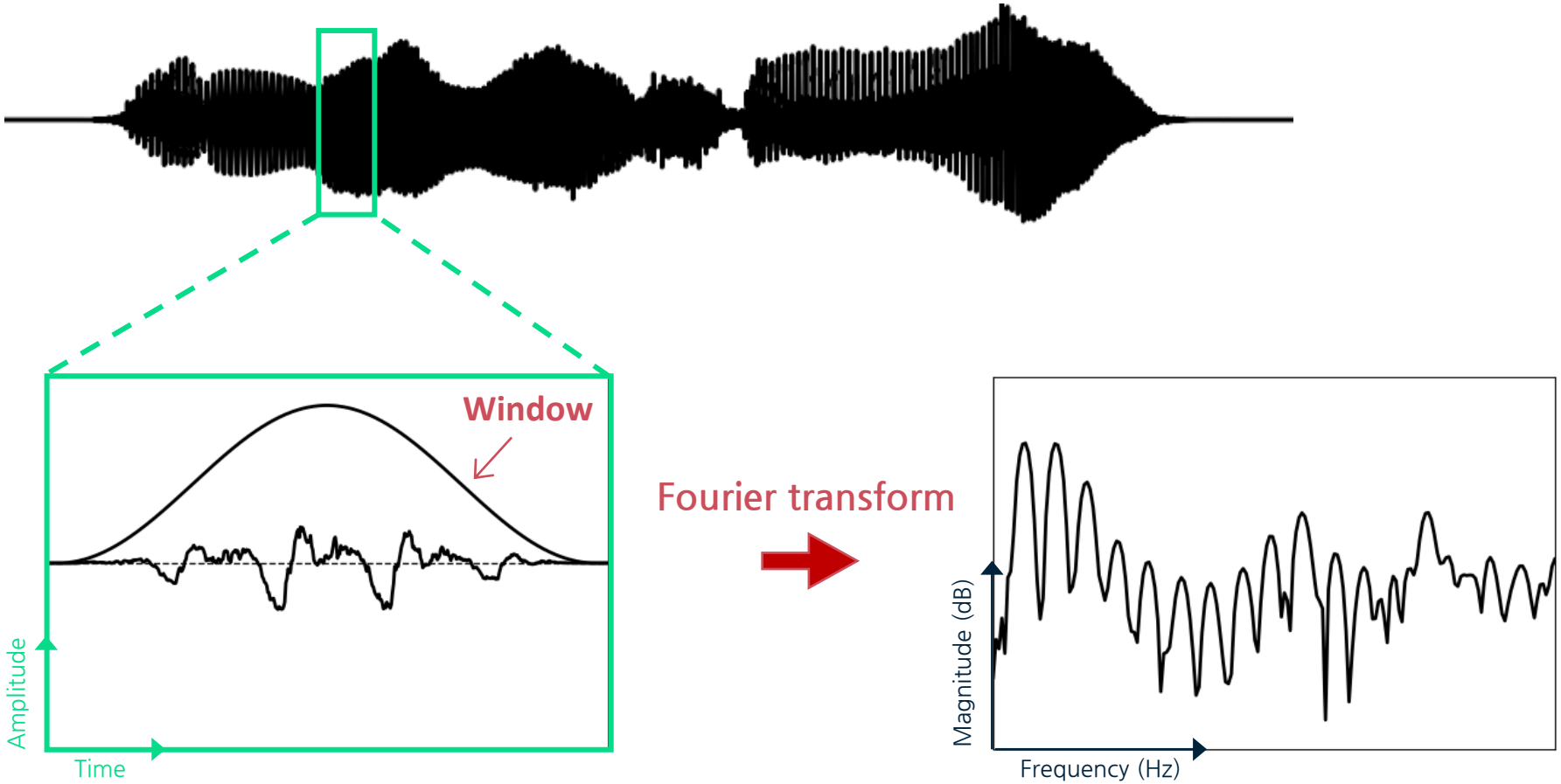
Speech analysis



음성 신호는 시간 축에서 특정한 에너지를 갖는 파형의 형태로 존재합니다

Introduction

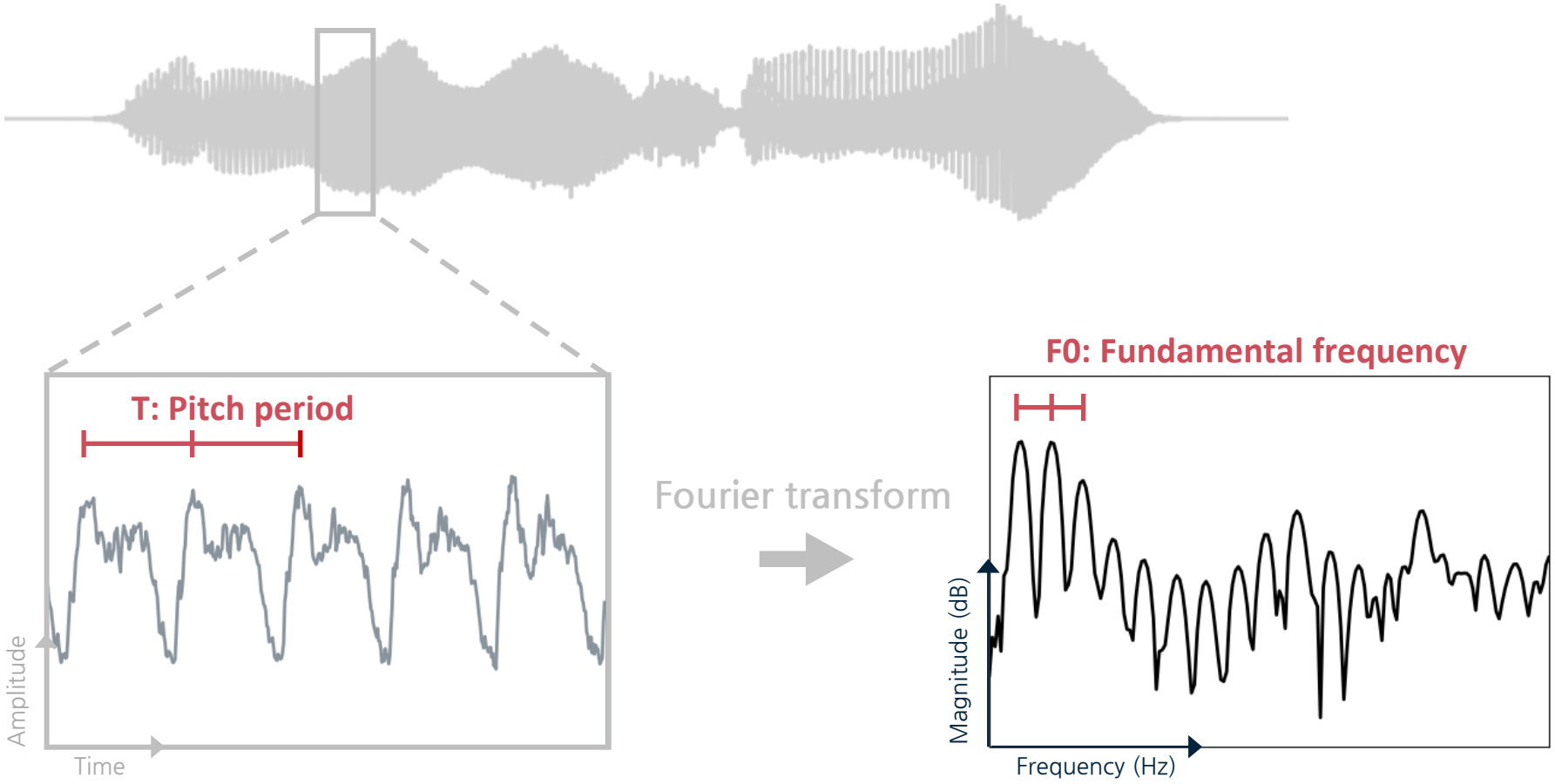
Speech analysis



Fourier 변환을 통해 주파수 축에서 음성을 관찰할 수 있습니다

Introduction

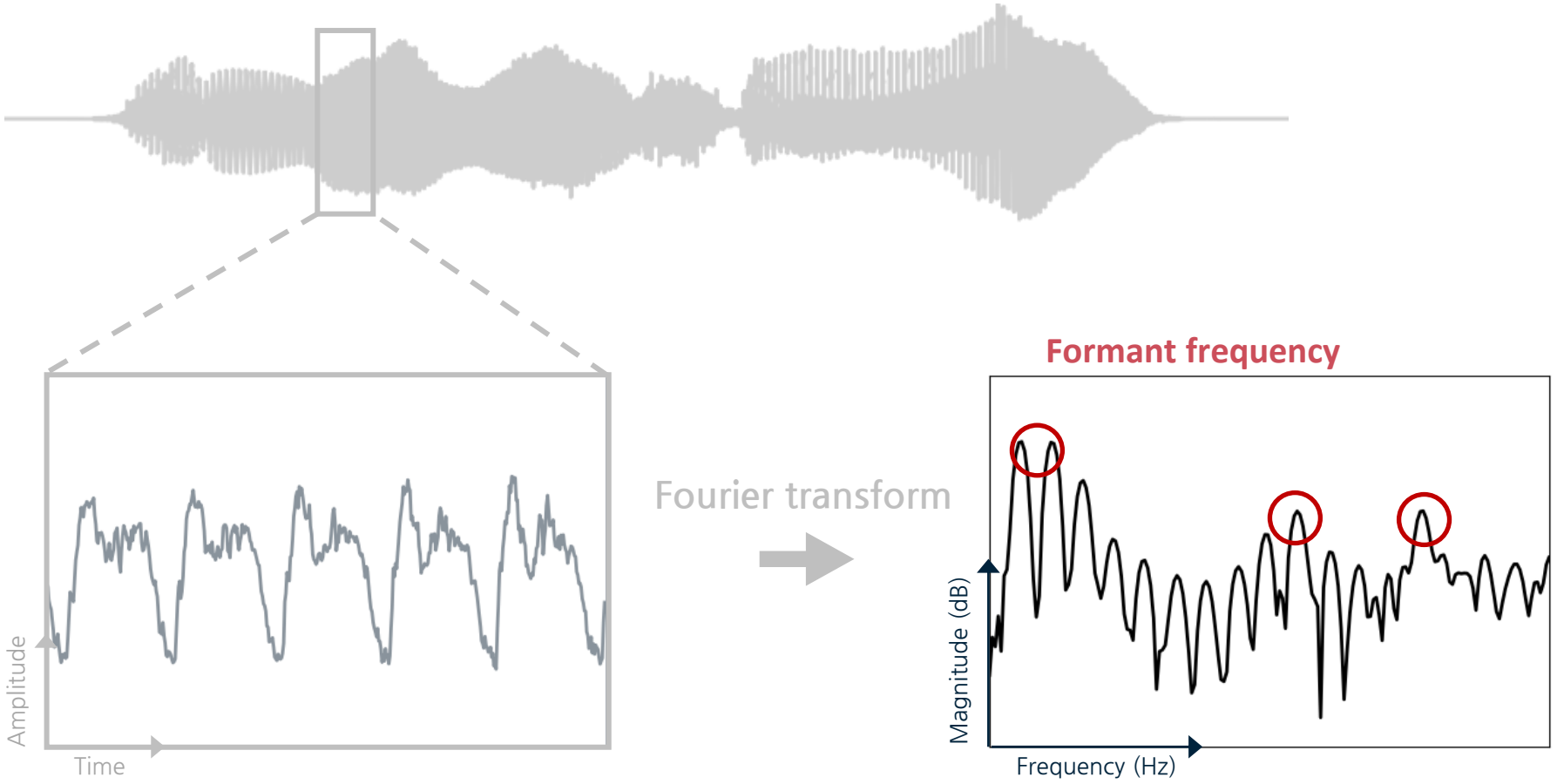
Speech analysis



F0: 목소리의 톤을 표현하는 파라미터 (아↘아↗)

Introduction

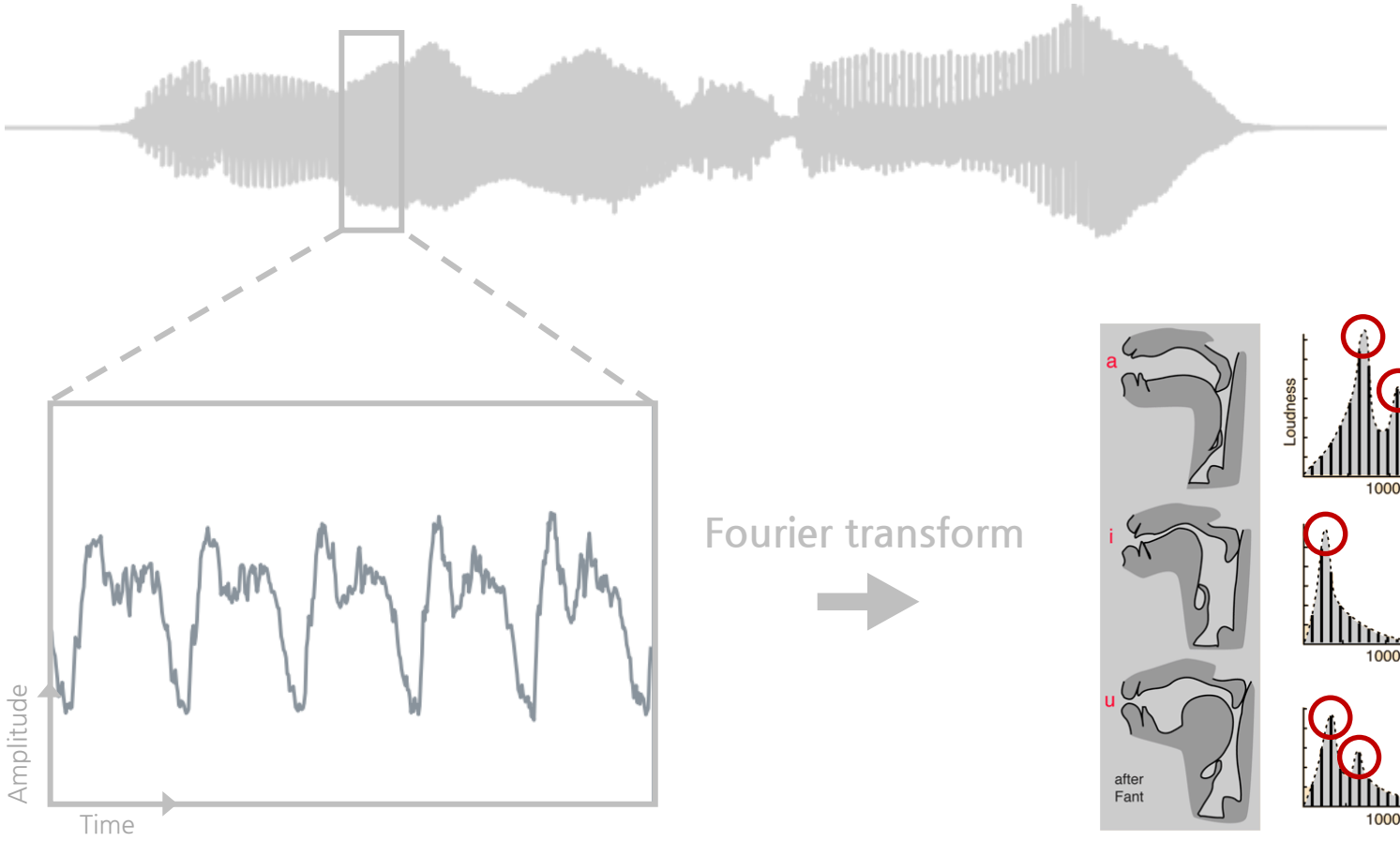
Speech analysis



Formant: 발음을 표현하는 파라미터 (아/에/이/오/우)

Introduction

Speech analysis

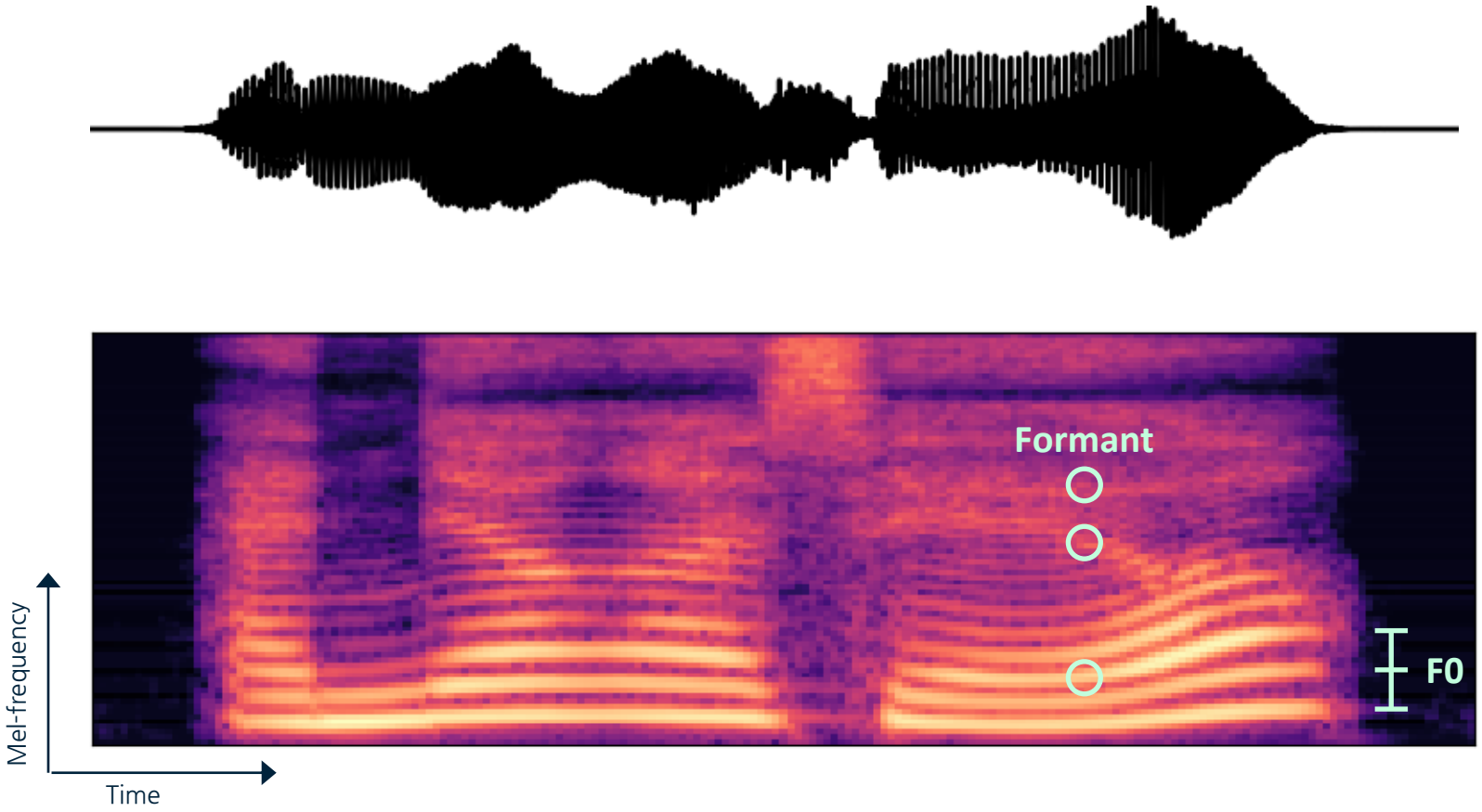


http://hyperphysics.phy-astr.gsu.edu/hbase/Music/vowel.html

Formant: 발음을 표현하는 파라미터 (아/에/이/오/우)

Introduction

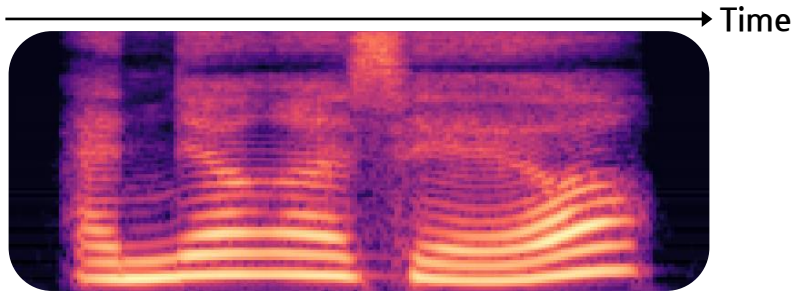
Speech analysis



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

Introduction

Deep learning-based TTS system

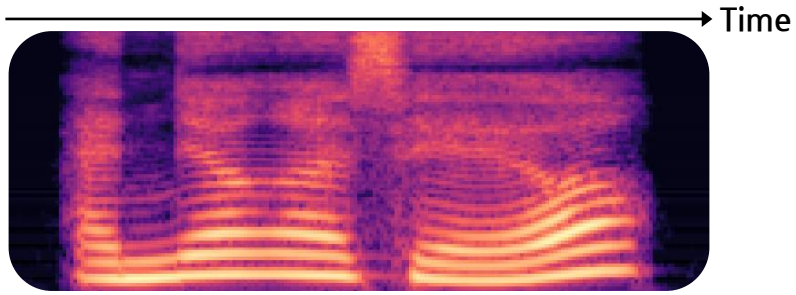


Acoustic parameters..?

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

Introduction

Deep learning-based TTS system



Estimating acoustic parameters from text inputs

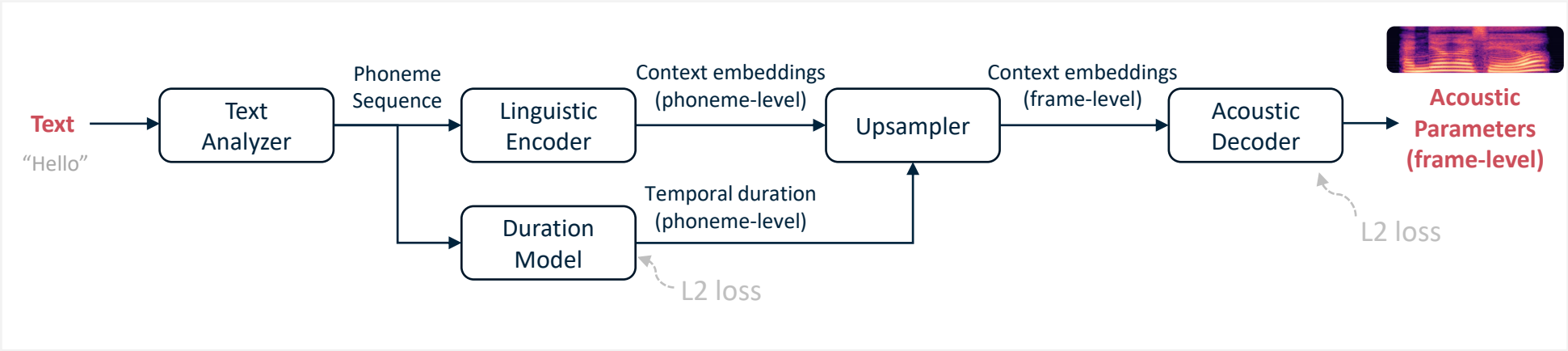
Speaker-specific attributes
(tone, volume, timbre, speaking rate, …)

주어진 입력 텍스트로 부터 사람의 음성 특성을 모델링 하는 태스크

TTS acoustic model

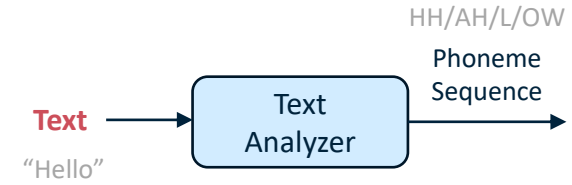
TTS acoustic model

How to generate acoustic parameters?



TTS acoustic model

How to generate acoustic parameters?



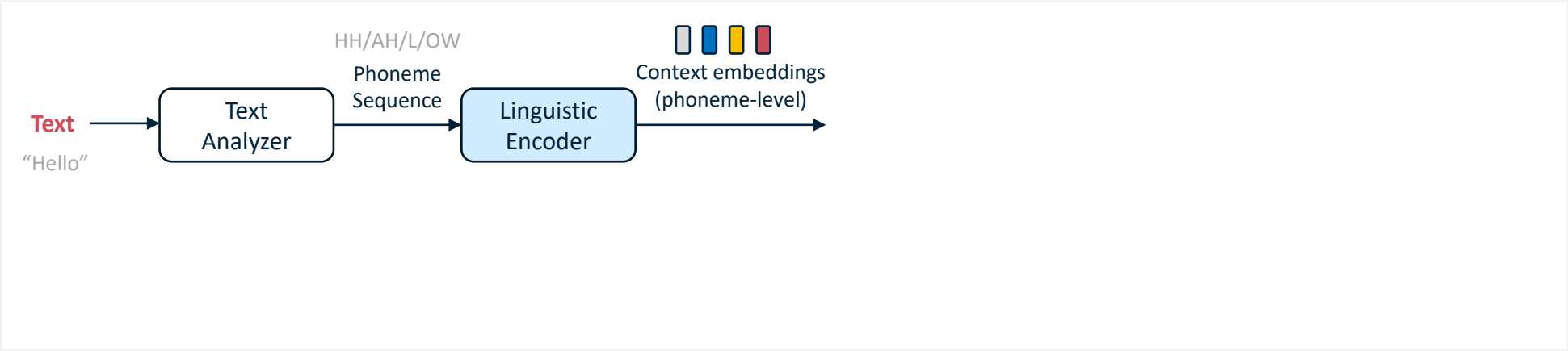
Text analyzer extracts phoneme sequence from the given text

음소: 음운론상의 최소 단위

- 3.2km → 삼쩌미키로미터 → ㅅ/ㅏ/ㅁ/ㅉ/ㅓ/ㅁ/ㅣ/ㅋ/ㅣ/ㄹ/ㅗ/ㅁ/ㅣ/ㅌ/ㅓ
- naver.com → 네이버닫컴 → ㄴ/ㅖ/ㅣ/ㅂ/ㅓ/ㄷ/ㅏ/ㄷ/ㅋ/ㅓ/ㅁ
- 1588-7942 → 이로팔팔 칠구사이 → ㅣ/ㄹ/ㅗ/ㅍ/ㅏ/ㄹ/ㅍ/ㅏㄹ/ㅊ/ㅣ/ㄹ/ㄱ/ㅜ/ㅅ/ㅏ/ㅣ

TTS acoustic model

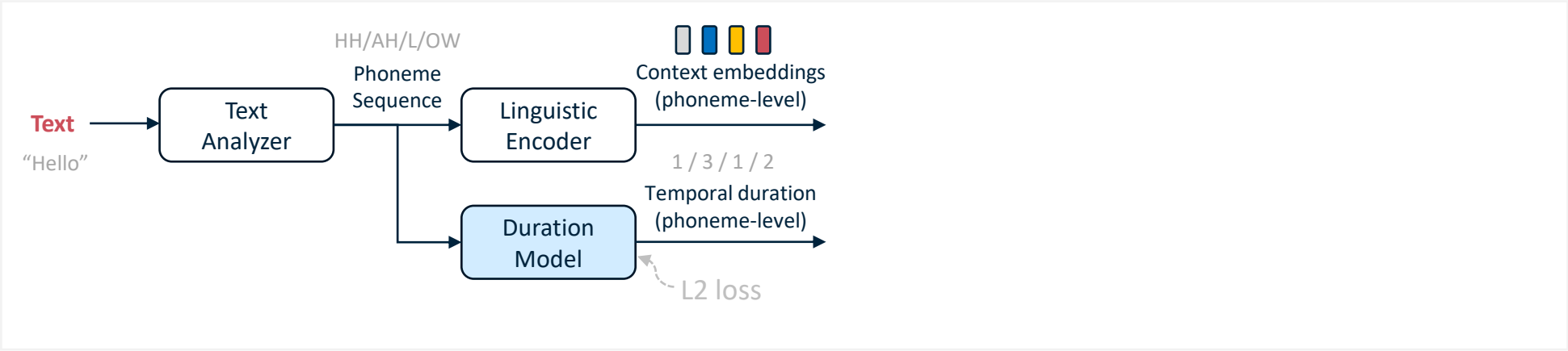
How to generate acoustic parameters?



Linguistic encoder extracts high-level context features from the given text

TTS acoustic model

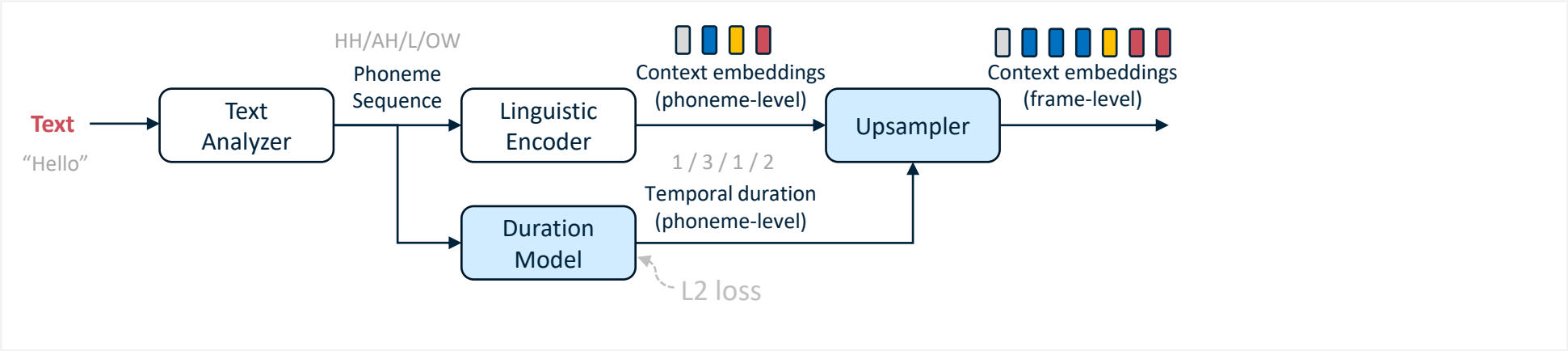
How to generate acoustic parameters?



Duration model predicts phoneme duration to upsample context embeddings into frame-level

TTS acoustic model

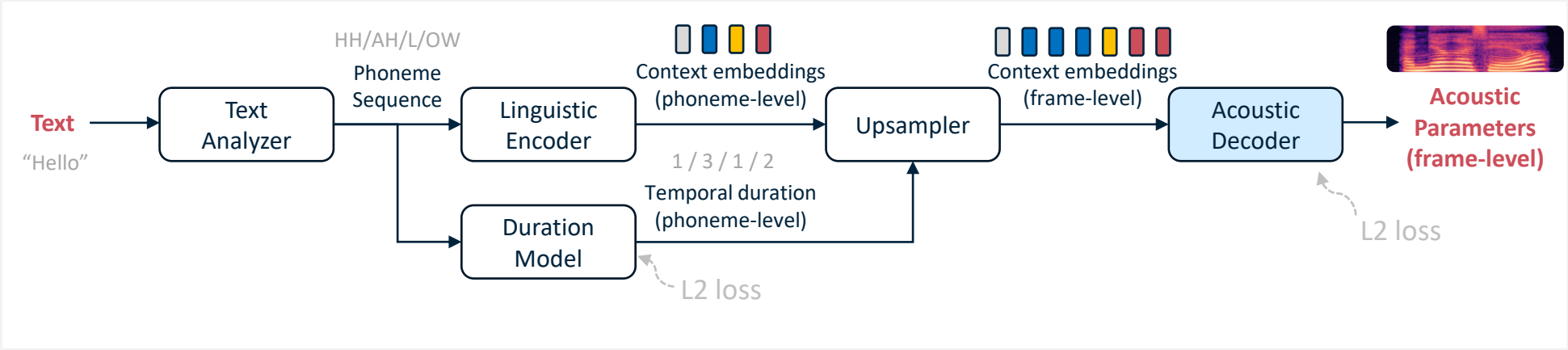
How to generate acoustic parameters?



Duration model predicts phoneme duration to upsample context embeddings into frame-level

TTS acoustic model

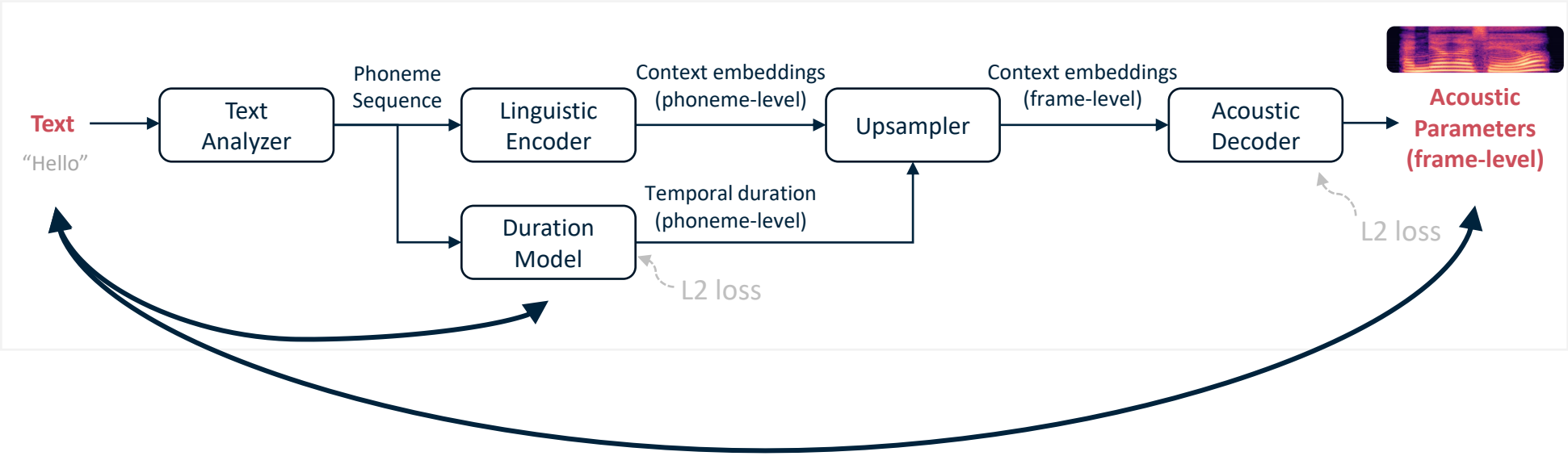
How to generate acoustic parameters?



Acoustic decoder predicts acoustic parameters from the given context embeddings

TTS acoustic model

How to generate acoustic parameters?



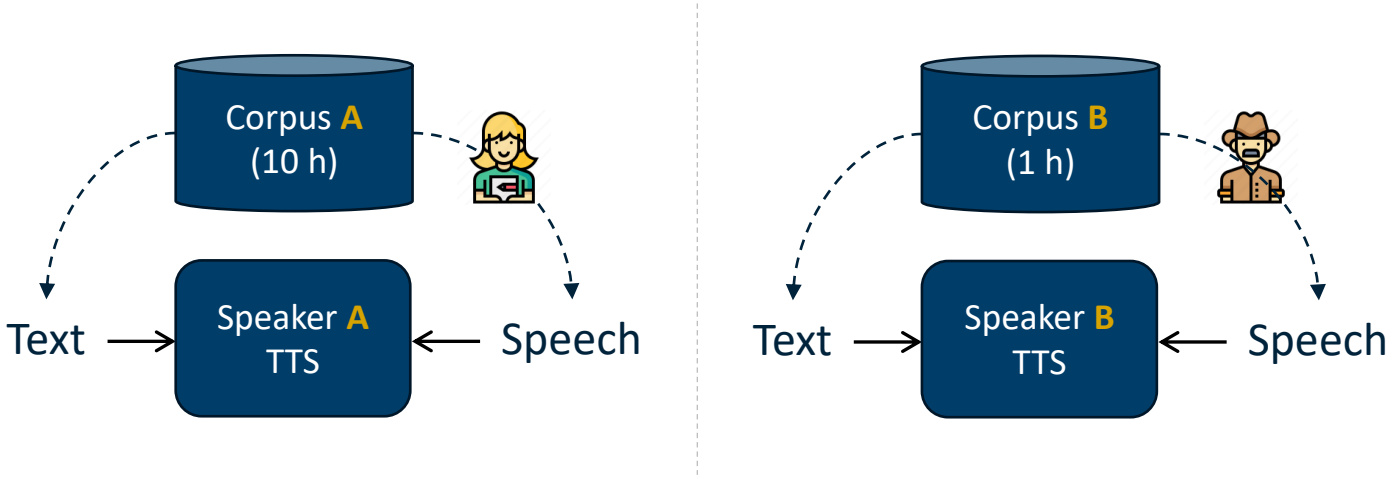
Trained by **speech-text pair**
recorded by single speaker or multiple speakers

TTS acoustic model

Multi-speaker TTS

TTS acoustic model

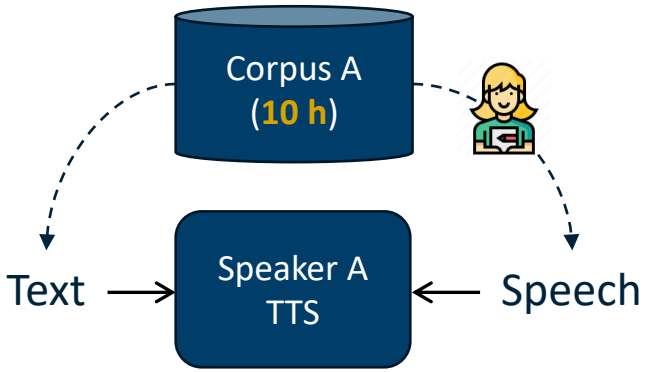
Speaker-dependent training



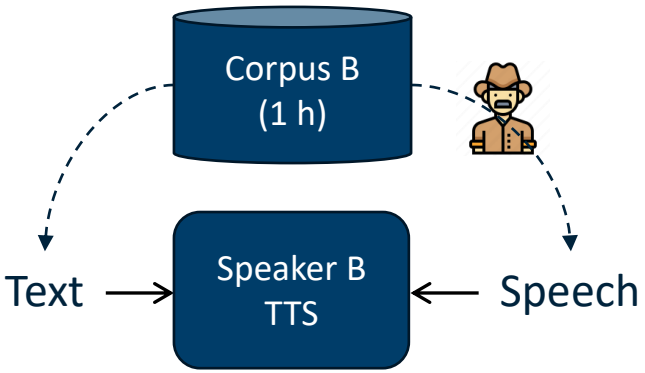
Training TTS acoustic model using target speaker’s speech corpus

TTS acoustic model

Speaker-dependent training



High-quality ☺



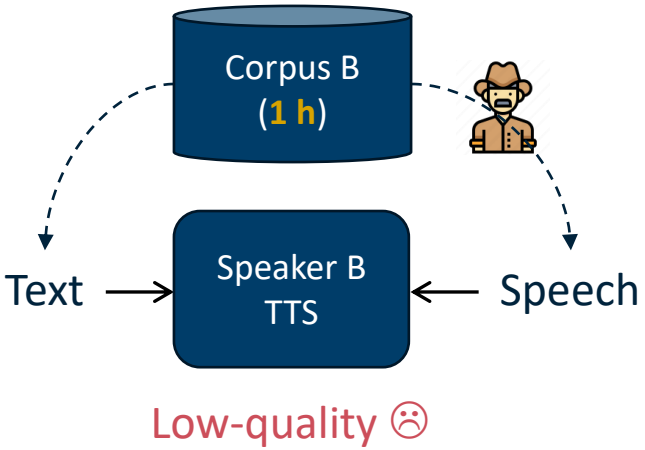
Training TTS acoustic model using target speaker’s speech corpus

This approach is highly effective as the model can learn the target speaker's speaking patterns

when plenty amount of recorded data is available (> 10 hours)

TTS acoustic model

Speaker-dependent training



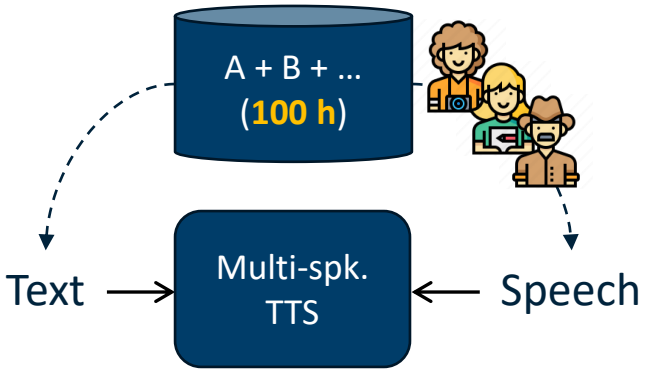
Training TTS acoustic model using target speaker's speech corpus

The synthetic quality is significantly degraded

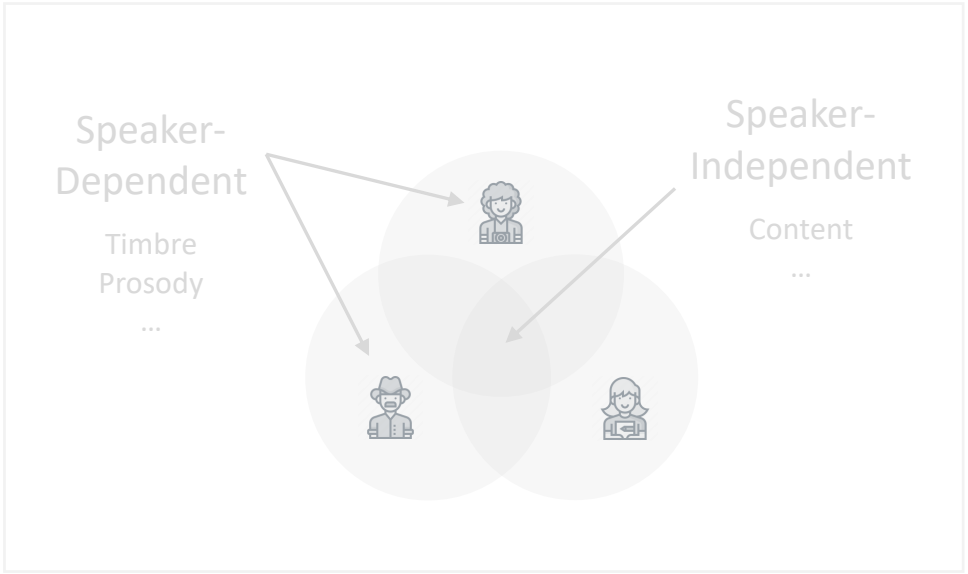
when the amount of available recorded data is limited

TTS acoustic model

Multi-speaker training



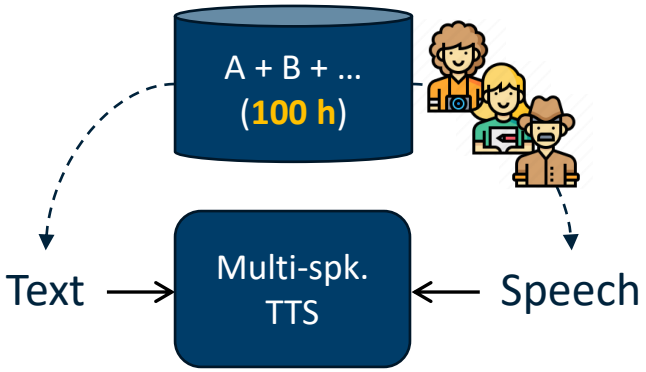
High-quality ☺



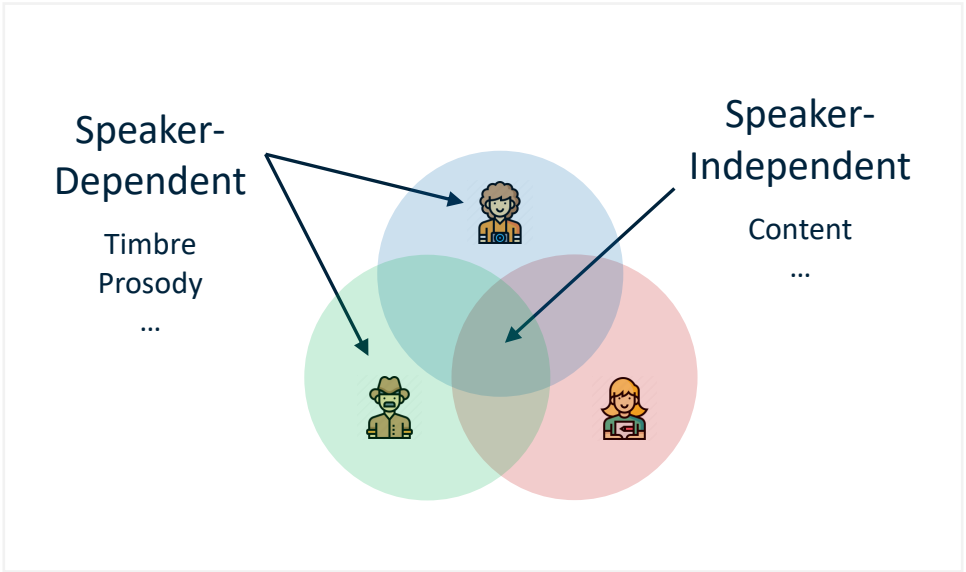
Training TTS acoustic model using speech corpora from multiple speakers

TTS acoustic model

Multi-speaker training



High-quality ☺



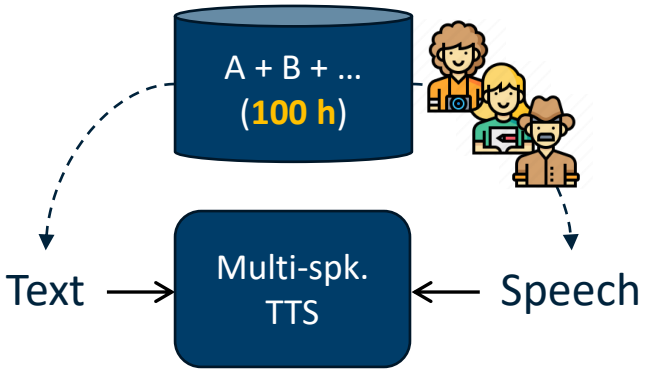
Training TTS acoustic model using speech corpora from multiple speakers

Some layers capture speaker-independent characteristics shared across different speakers

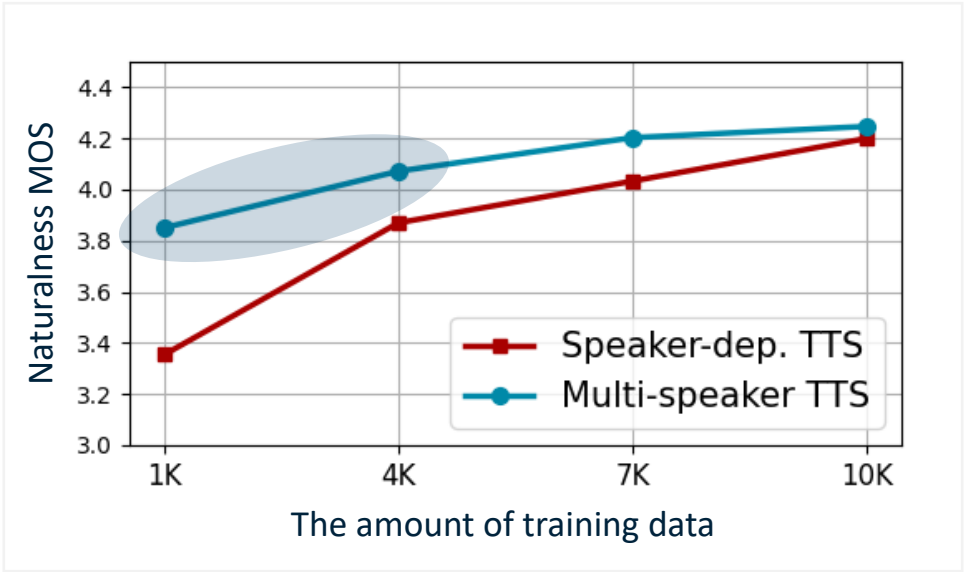
Other layers represent speaker-dependent characteristics specific to the target speaker

TTS acoustic model

Multi-speaker training

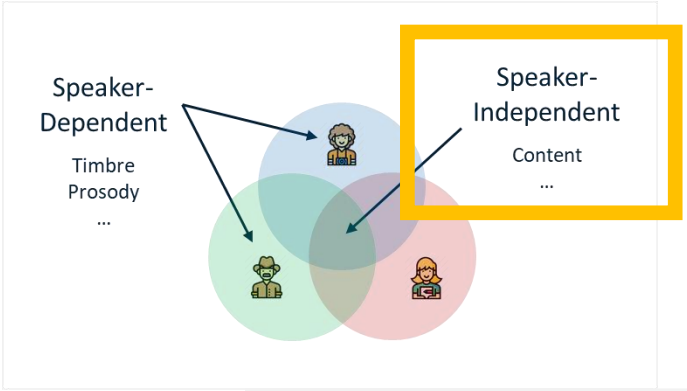


High-quality ☺



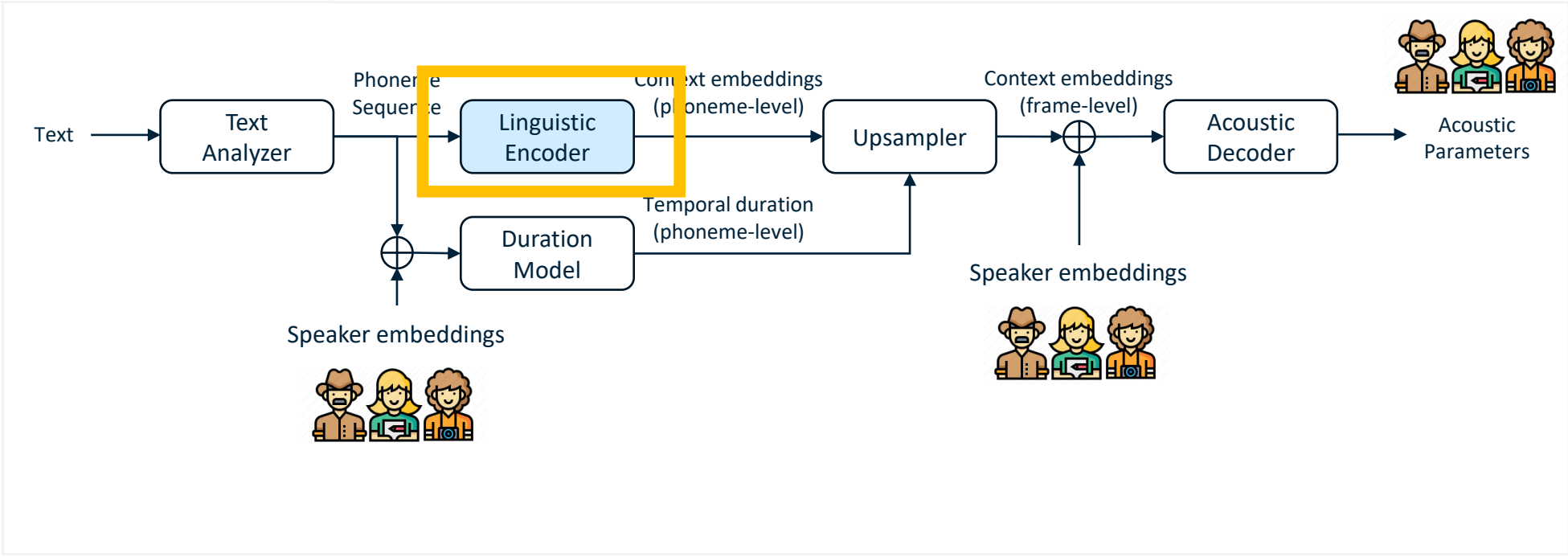
Training TTS acoustic model using speech corpora from multiple speakers

This offers a promising solution to address the data shortage problem in speaker-dependent TTS model



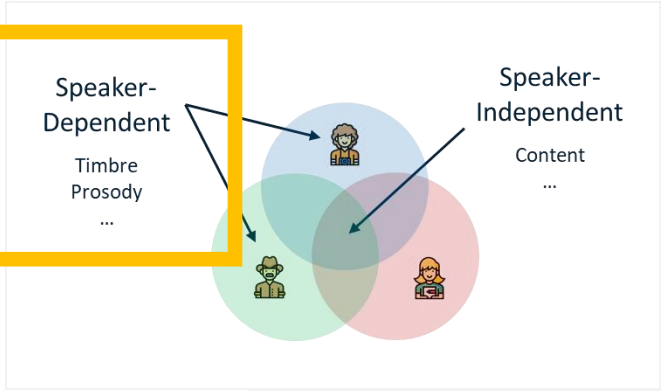
TTS acoustic model

Multi-speaker model



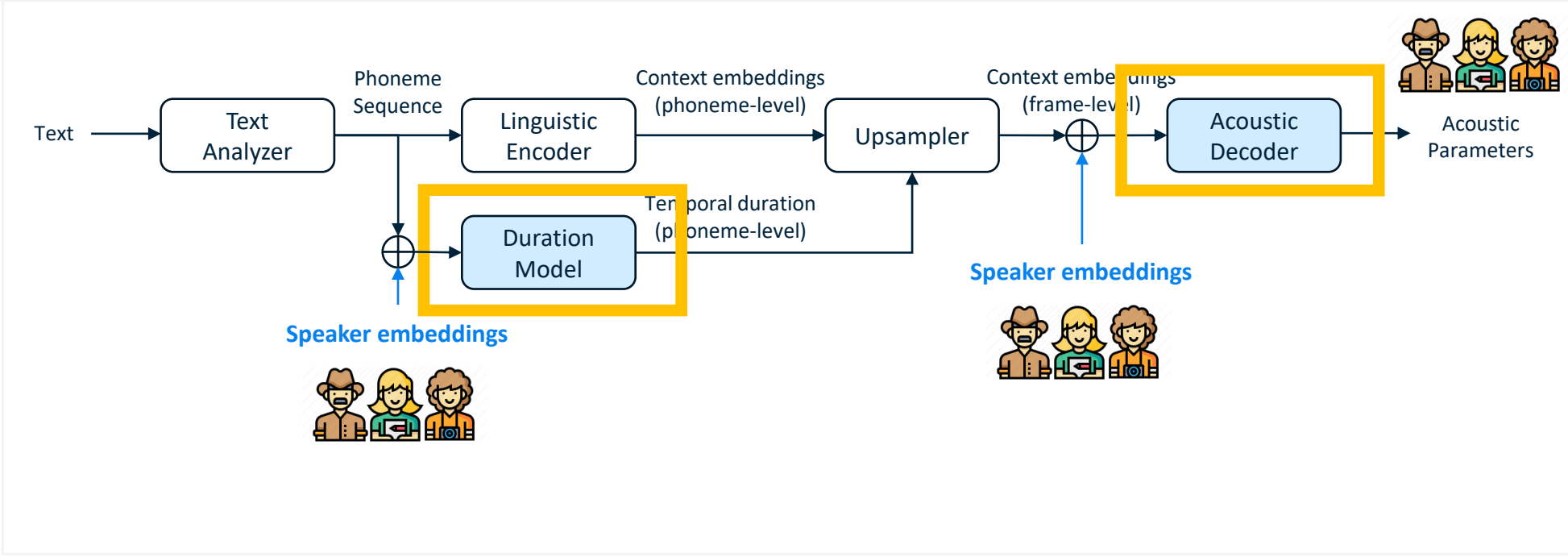
Training TTS acoustic model using speech corpora from multiple speakers with speaker embeddings

Linguistic encoder capture speaker-independent contents shared across different speakers



TTS acoustic model

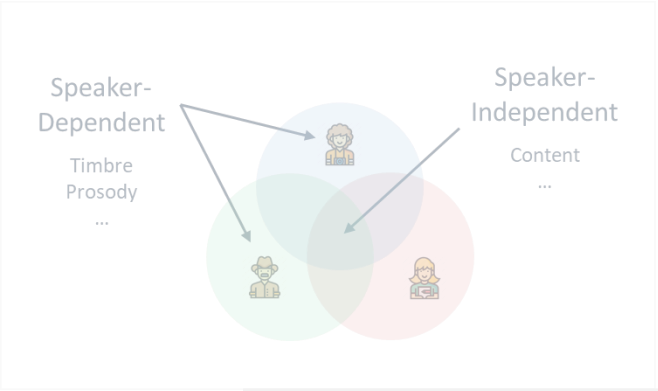
Multi-speaker model



Training TTS acoustic model using speech corpora from multiple speakers with speaker embeddings

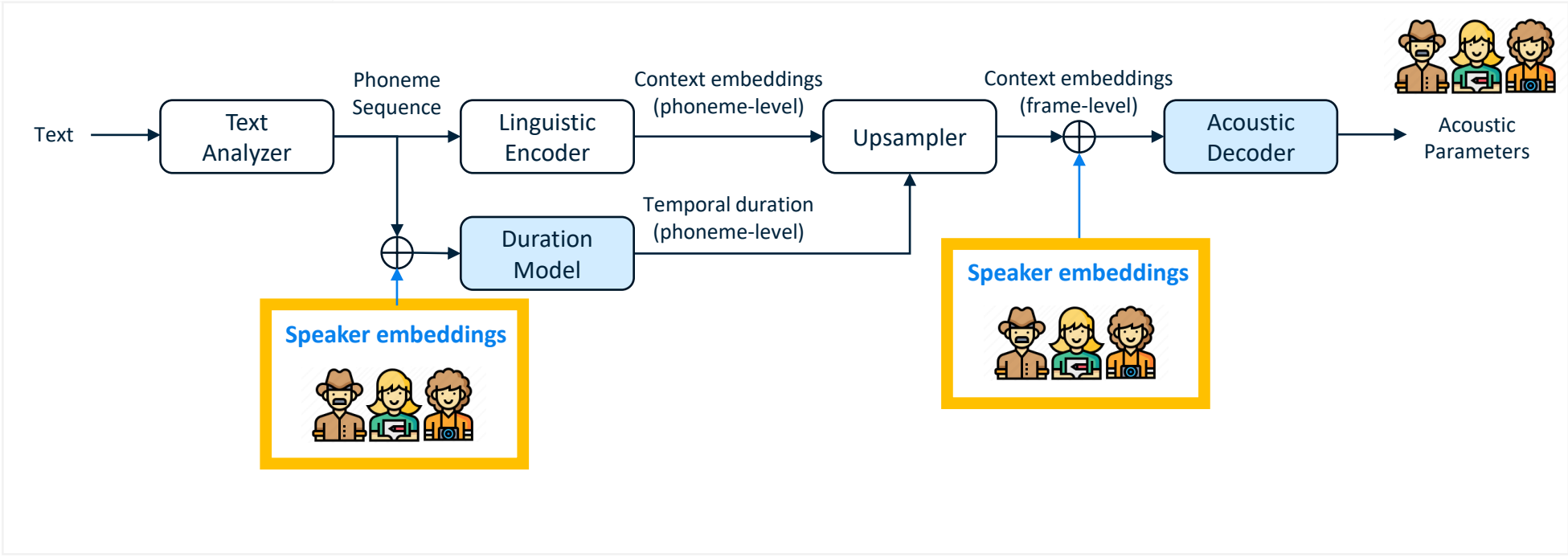
Linguistic encoder capture speaker-independent contents shared across different speakers

Acoustic decoder and duration model represent speaker-dependent timbre and prosody, respectively

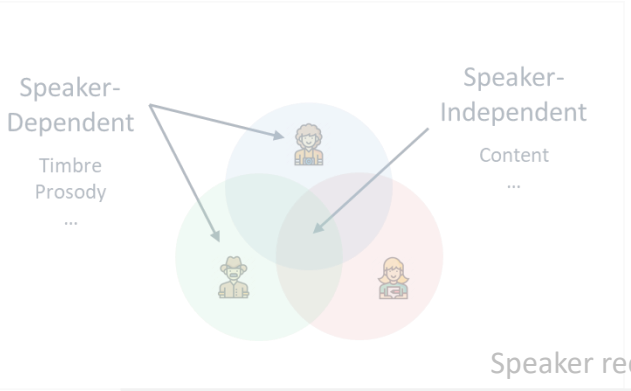


TTS acoustic model

Multi-speaker model



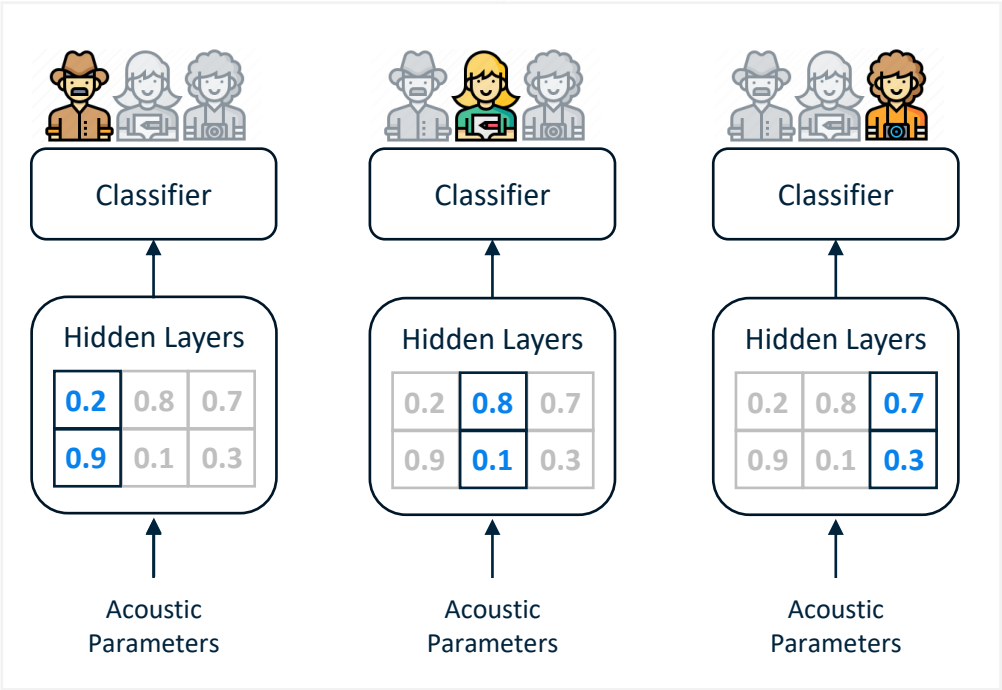
How to design the speaker embeddings ?



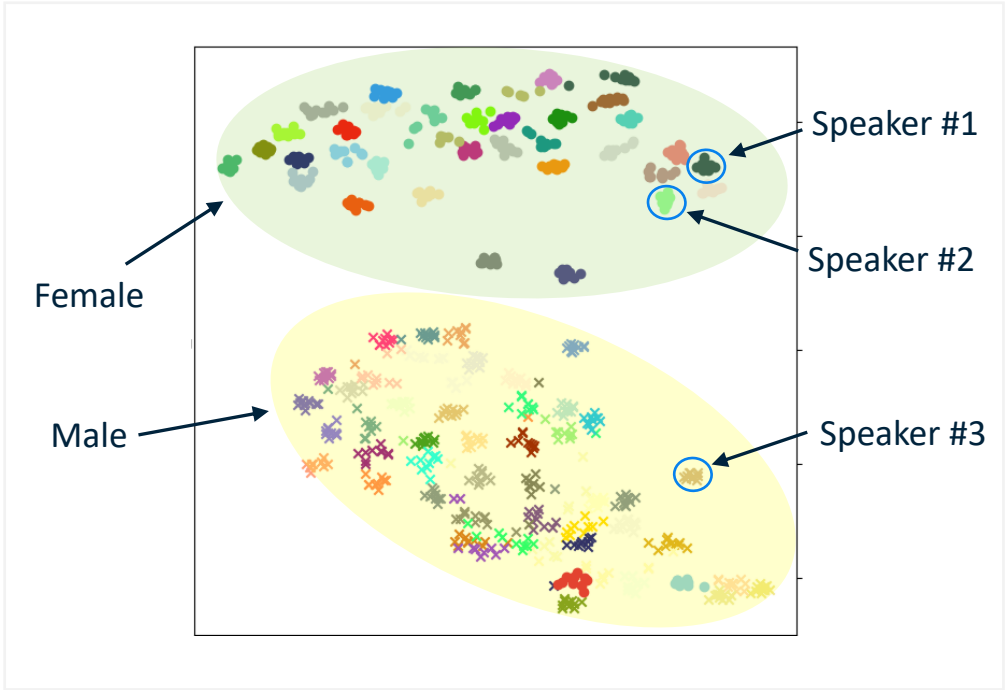
TTS acoustic model

Multi-speaker model

Speaker recognition

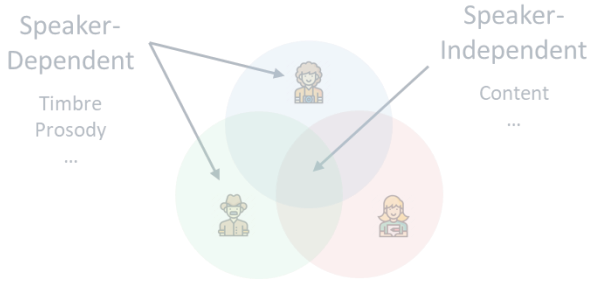


t-SNE plot of hidden representations obtained from multiple speakers



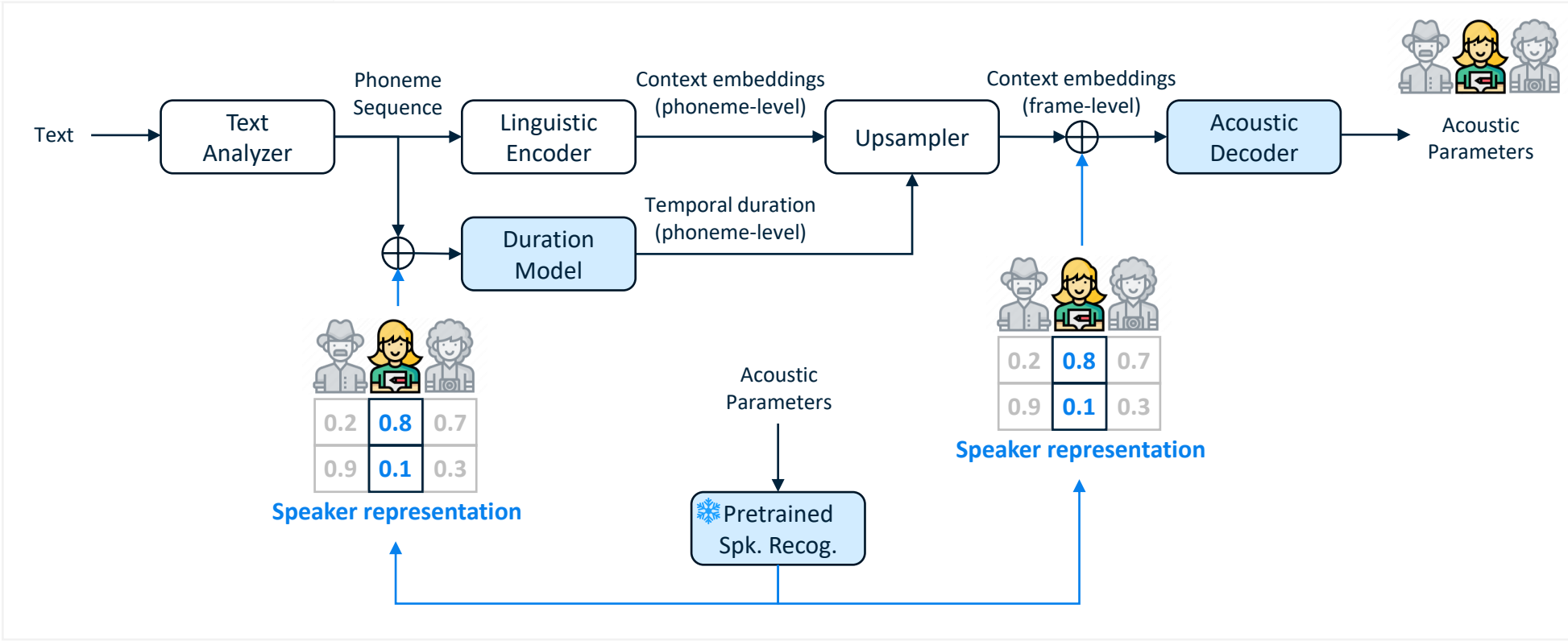
Speaker recognition model is designed to capture the speaker identity from the given acoustic parameter

Its hidden representations contain speaker-specific characteristics



TTS acoustic model

Multi-speaker model

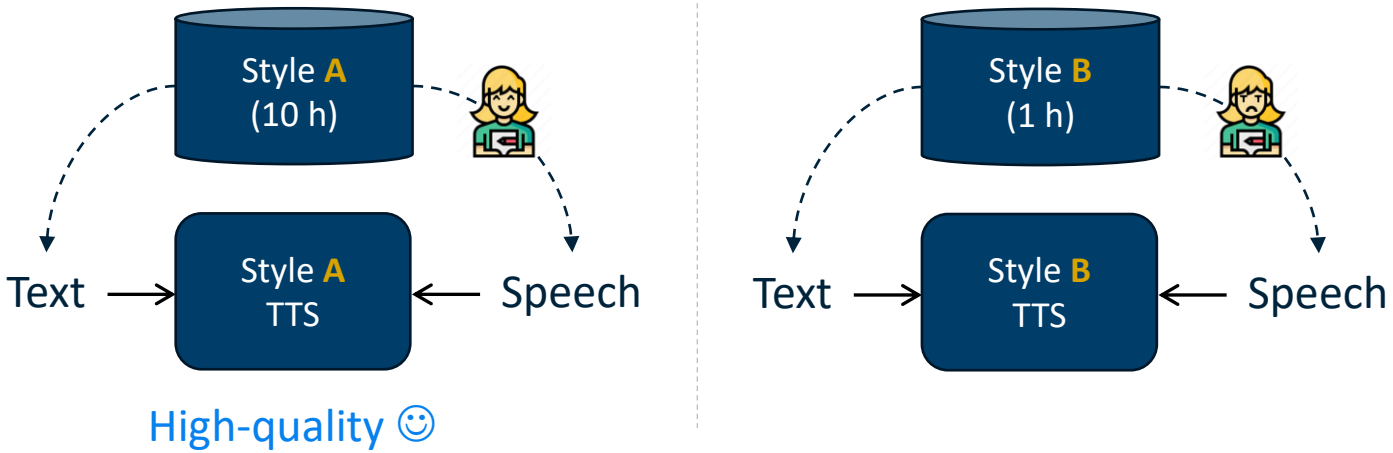


TTS acoustic model

Multi-style TTS

TTS acoustic model

Style-dependent training



High-quality ☺

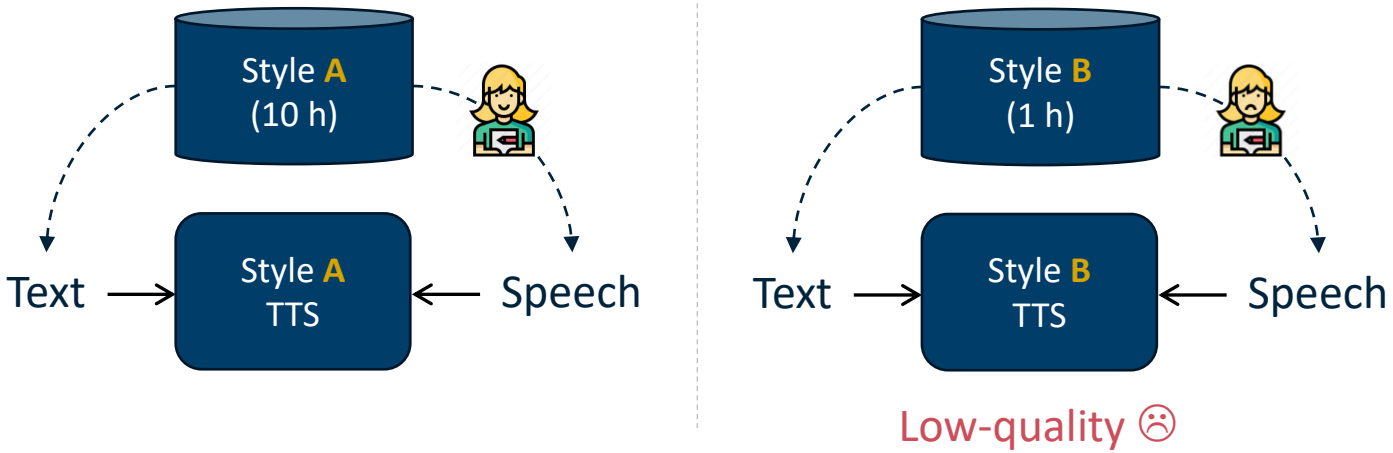
Training TTS acoustic model using speech corpus containing target speaking style

This approach is highly effective as the model can learn the target speaking style

when plenty amount of recorded data is available (> 10 hours)

TTS acoustic model

Style-dependent training



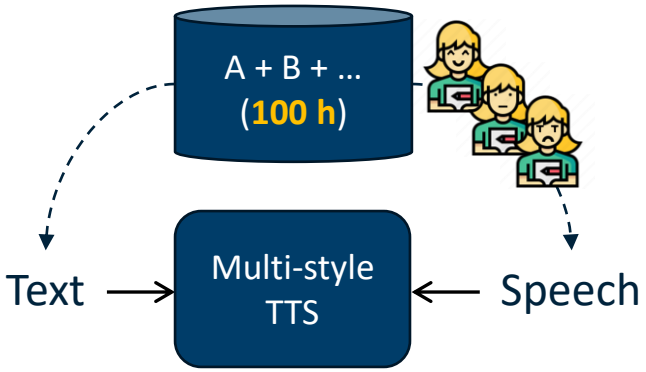
Training TTS acoustic model using speech corpus containing target speaking style

The synthetic quality is significantly degraded

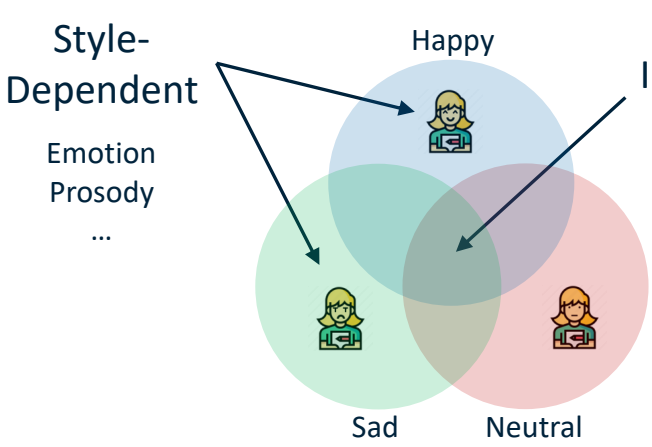
when the amount of available recorded data is limited

TTS acoustic model

Multi-style training



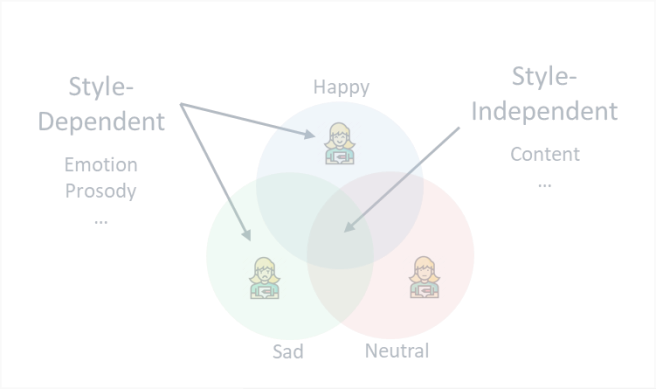
High-quality ☺



Training TTS acoustic model using speech corpora from multiple styles

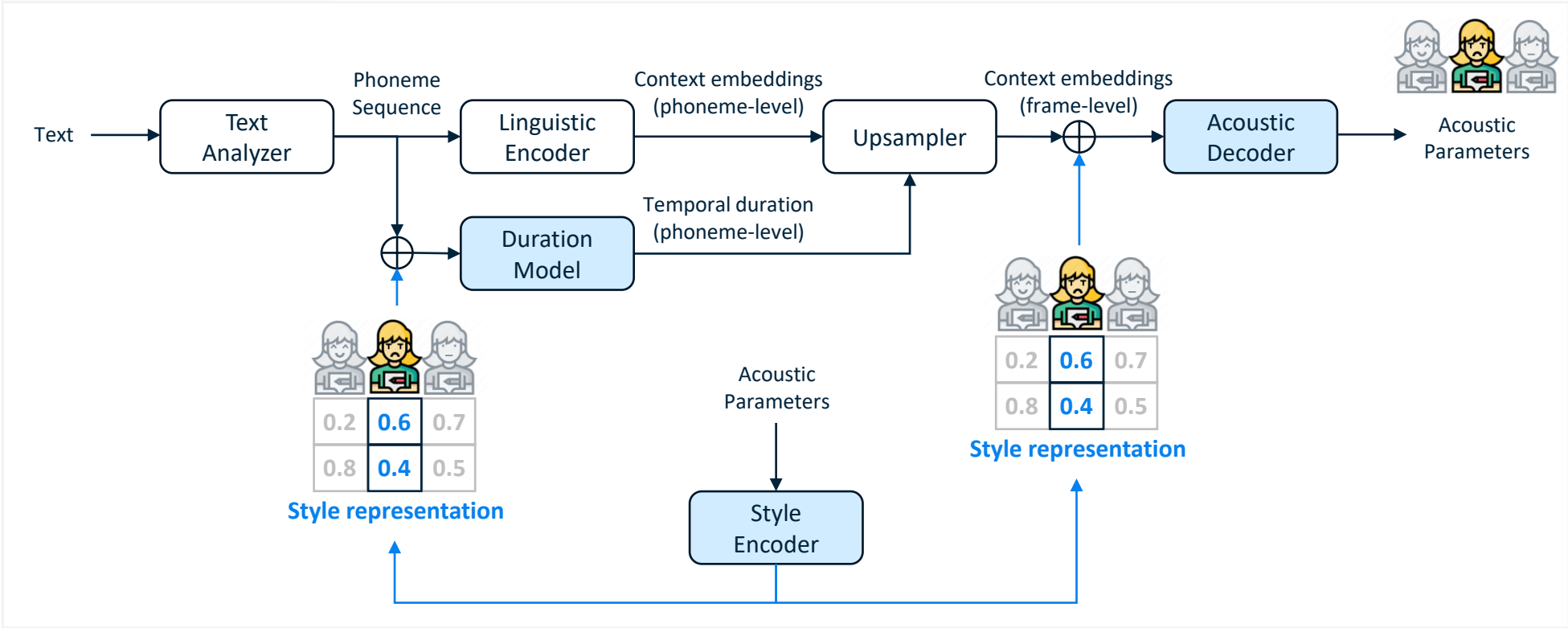
Linguistic encoders capture style-independent characteristics shared across different styles

Acoustic decoder and duration model represent style-dependent characteristics

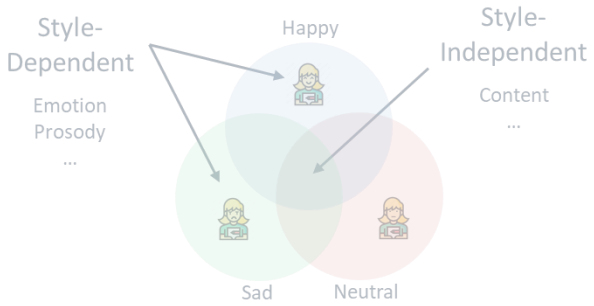


TTS acoustic model

Multi-style model



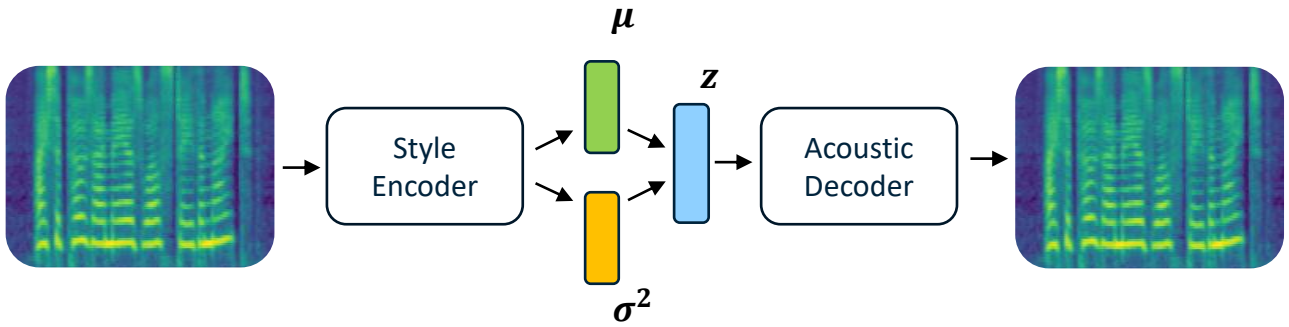
VAE-based style encoder extracts style representations from the given acoustic parameters



TTS acoustic model

Multi-style model

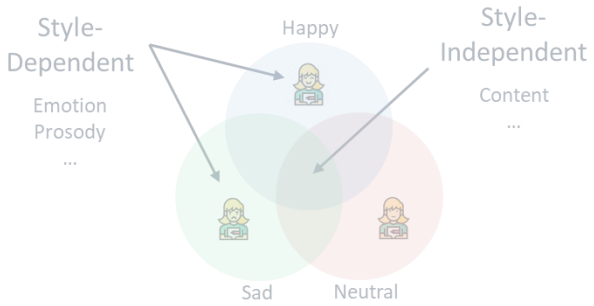
VAE-based style encoder



μ Mean vector

σ^2 Variance vector

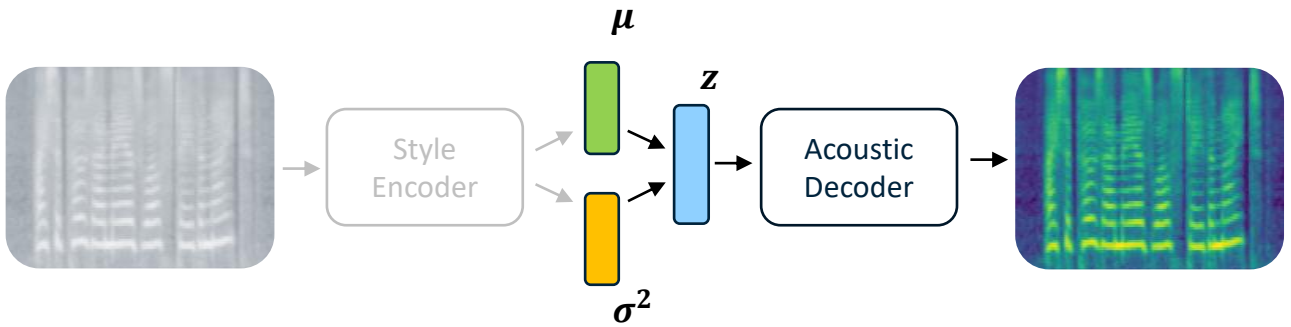
z Latent vector ~ N (μ, σ^2)

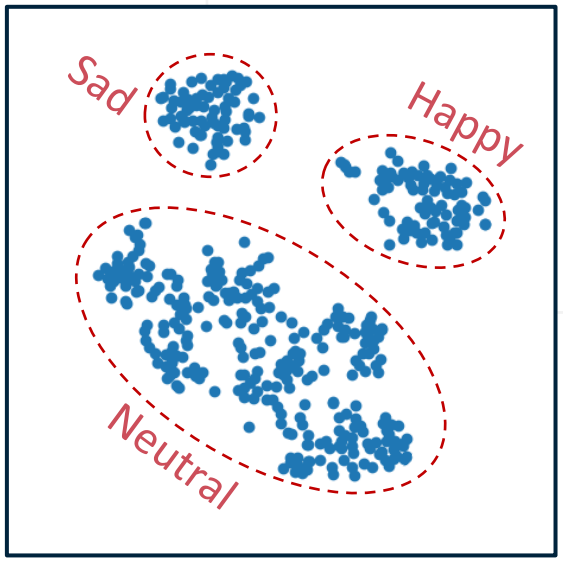


TTS acoustic model

Multi-style model

VAE-based style encoder





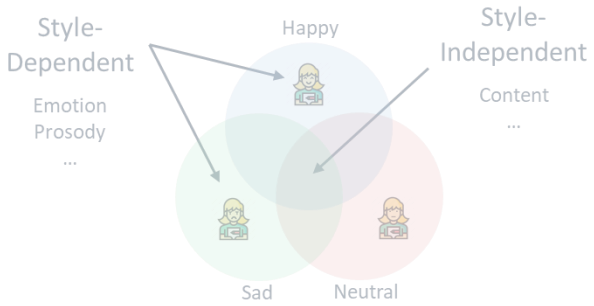
t-SNE plot of VAE representations (=μ)
obtained from multiple emotions

μ Mean vector

σ^2 Variance vector

Unsupervised style representation

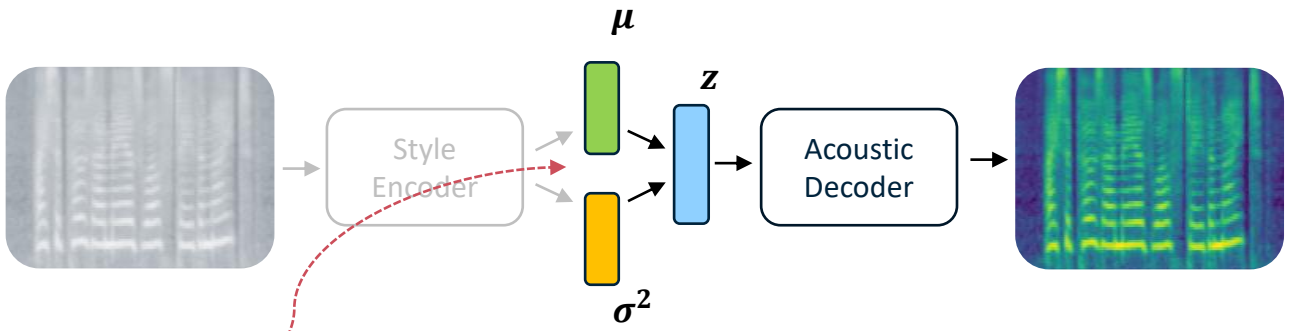
z Latent vector ~ $N(\mu, \sigma^2)$



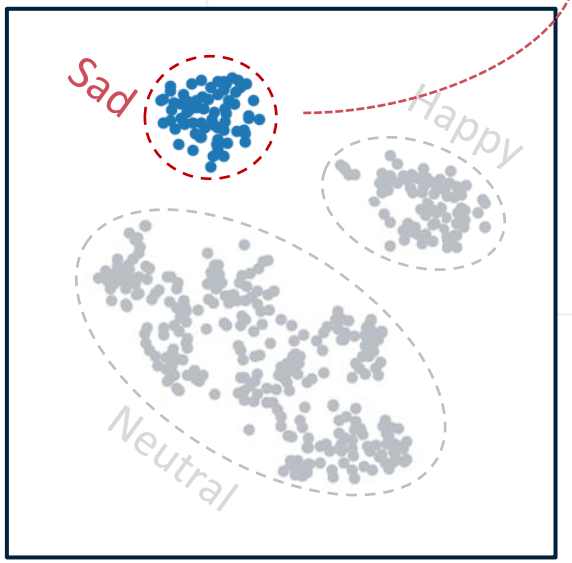
TTS acoustic model

Multi-style model

VAE-based style encoder



Sad



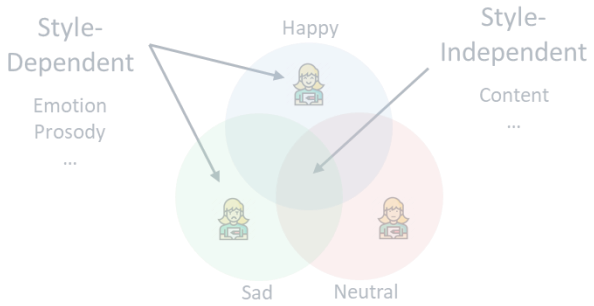
t-SNE plot of VAE representations (=μ)
obtained from multiple emotions

μ Mean vector

σ^2 Variance vector

Unsupervised style representation

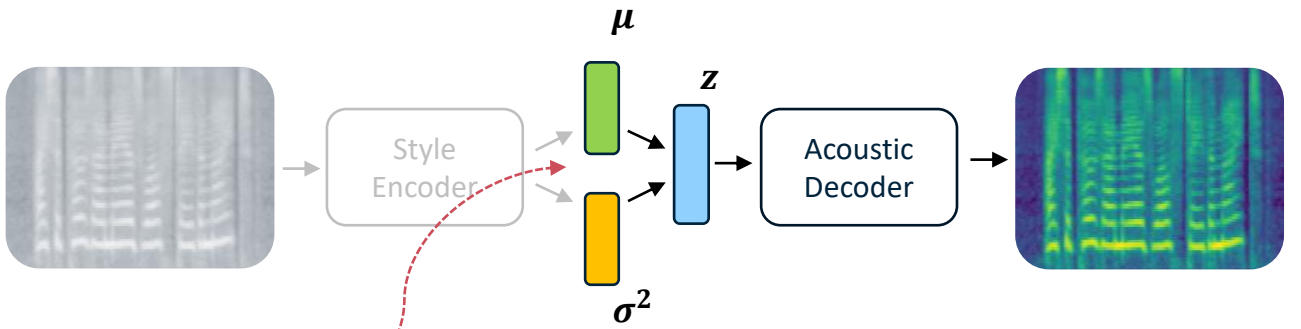
z Latent vector ~ $N(\mu, \sigma^2)$ Variations to output



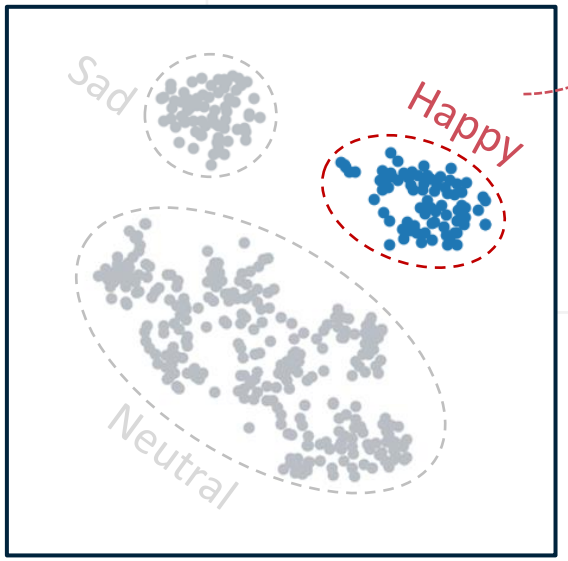
TTS acoustic model

Multi-style model

VAE-based style encoder



Happy



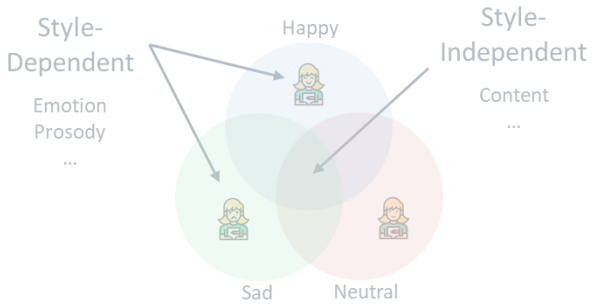
t-SNE plot of VAE representations (=μ)
obtained from multiple emotions

μ Mean vector

σ^2 Variance vector

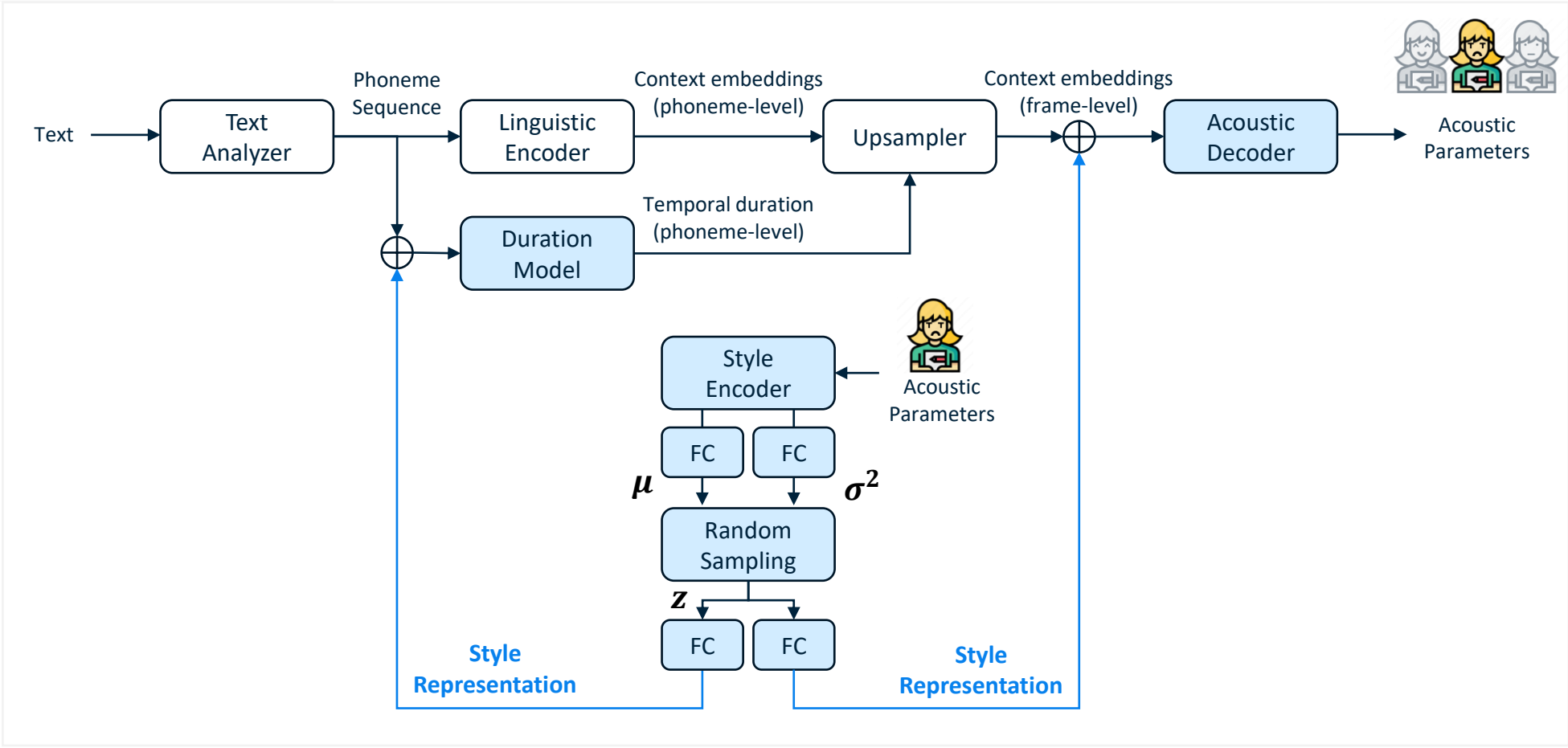
Unsupervised style representation

z Latent vector ~ $N(\mu, \sigma^2)$ Variations to output



TTS acoustic model

Multi-style model



Acoustic model

Large-scale TTS

Scaling TTS model

Recording constraint

	Conventional TTS	Custom TTS
Speaker	Professional	Non-professional
Environment	Clean studio	Anywhere
Amount	> 30~60 min	< Few seconds
Speaking type	Script reading	Spontaneous
Synthetic quality	Very natural	Unnatural

Scaling TTS model

Recording constraint

	Conventional TTS	Custom TTS
Speaker	Professional	Non-professional
Environment	Clean studio	Anywhere
Amount	> 30~60 min	< Few seconds
Speaking type	Script reading	Spontaneous
Synthetic quality	Very natural	Natural ~~Unnatural~~

The trend is shifting towards scaling up TTS model using large-scale dataset

Scaling TTS model

Voicebox model from Meta

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

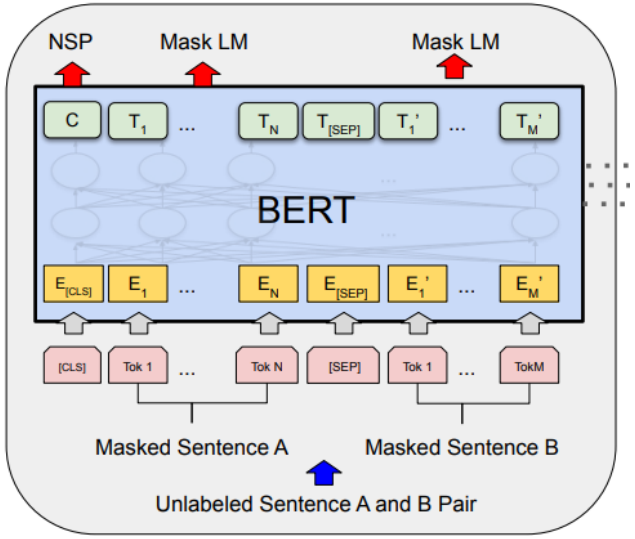
Abstract

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`.

Preliminary

Self-supervised learning representation

Model learns useful representations of data without relying on manually labeled data



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

"Unlabeled"

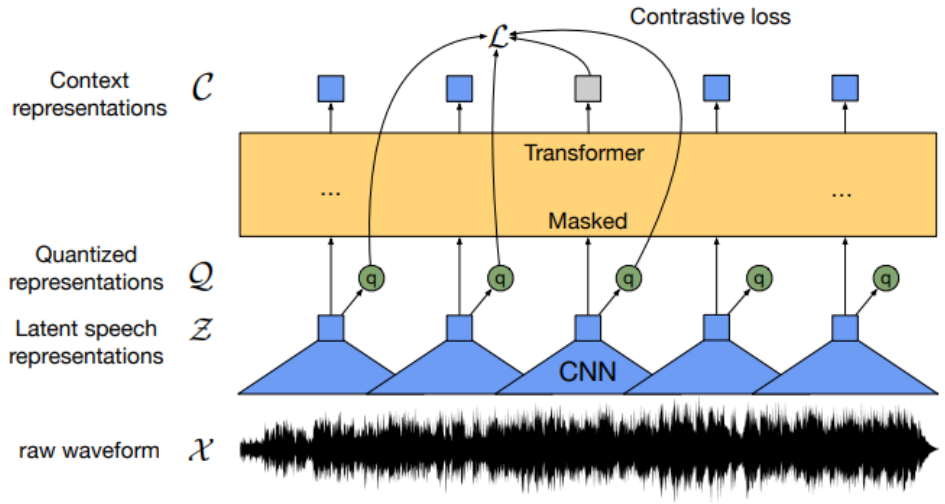
대규모 데이터를 이용하여 pretrained 모델을 구축하고

소량의 데이터로 downstream task 에 fine-tuning 할 수 있음

Preliminary

Self-supervised learning representation

Model learns useful representations of data without relying on manually labeled data



A. Baevski, et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020, pp. 12449-12460.

"Unlabeled"

대규모 데이터를 이용하여 pretrained 모델을 구축하고

소량의 데이터로 downstream task 에 fine-tuning 할 수 있음

Preliminary

Self-supervised learning representation

Model learns useful representations of data without relying on manually labeled data

In TTS generation task..?

"Unlabeled"

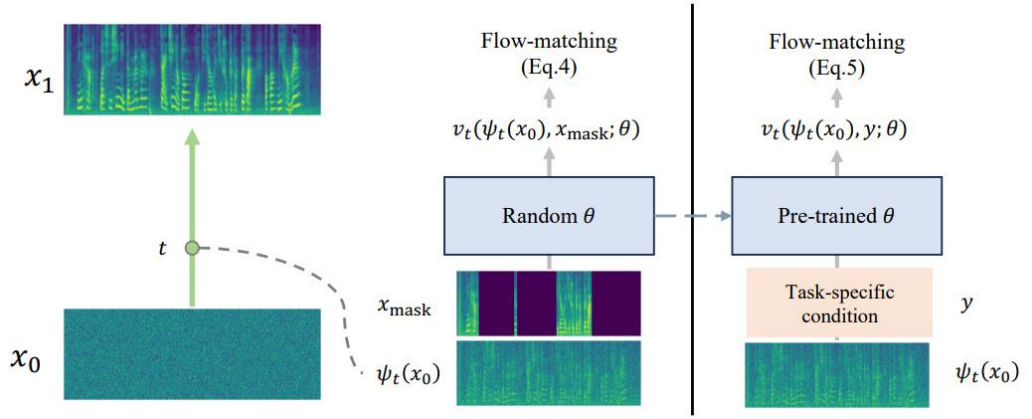
대규모 데이터를 이용하여 pretrained 모델을 구축하고

소량의 데이터로 downstream task 에 fine-tuning 할 수 있음

Preliminary

Self-supervised learning representation

Model learns useful representations of data without relying on manually labeled data



A. H. Liu, et al., "Generative pre-training for speech with flow matching," arXiv preprint arXiv:2310.16338, 2023.

"Unlabeled"

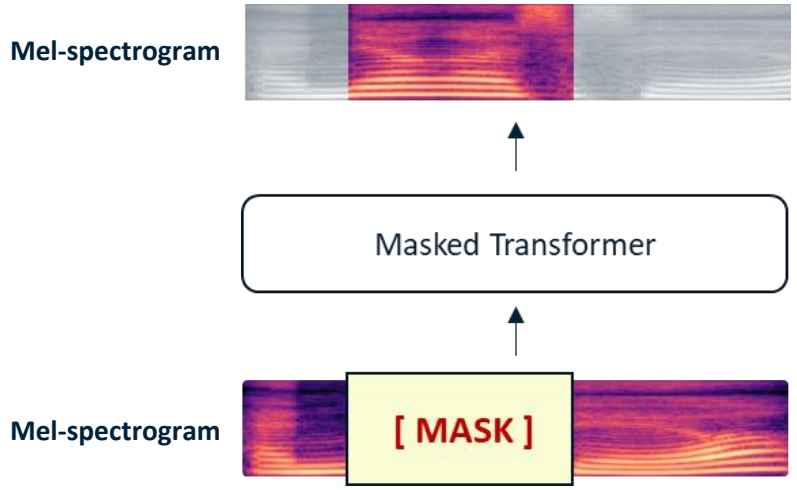
대규모 데이터를 이용하여 pretrained 모델을 구축하고

소량의 데이터로 downstream task 에 fine-tuning 할 수 있음

Preliminary

Self-supervised learning representation

Model learns useful representations of data without relying on manually labeled data



A. H. Liu, et al., “Generative pre-training for speech with flow matching,” arXiv preprint arXiv:2310.16338, 2023.

“Unlabeled”

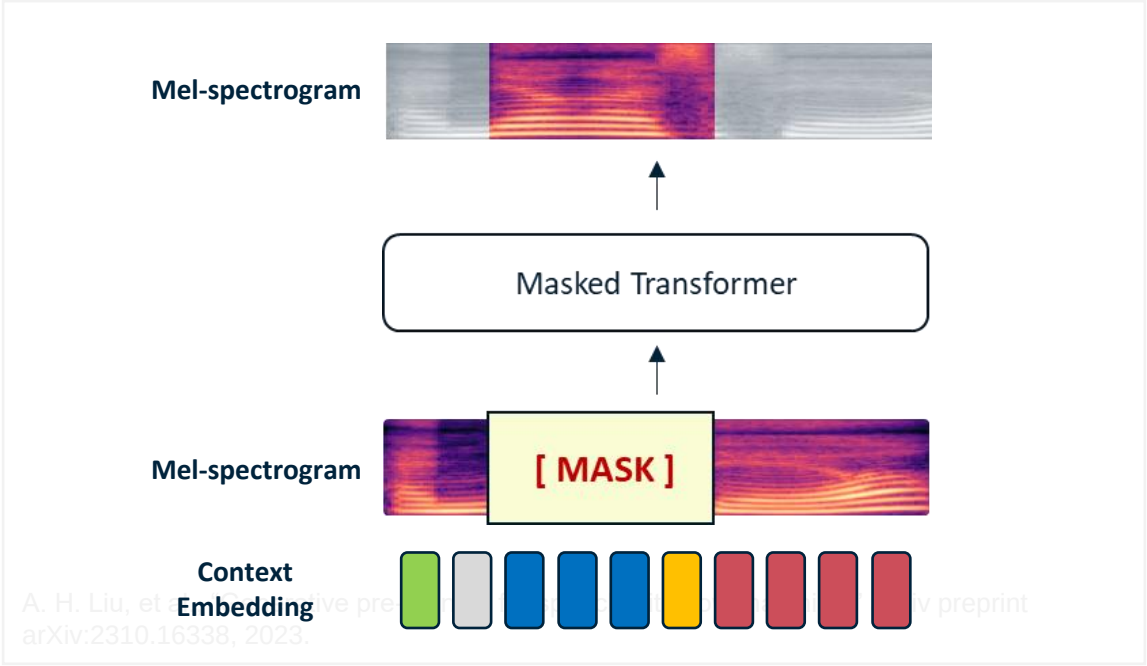
대규모 데이터를 이용하여 pretrained 모델을 구축하고

소량의 데이터로 downstream task 에 fine-tuning 할 수 있음

Preliminary

Self-supervised learning representation

Model learns useful representations of data without relying on manually labeled data



A. H. Liu, et al. "... pre-... v preprint arXiv:2310.16338, 2023.

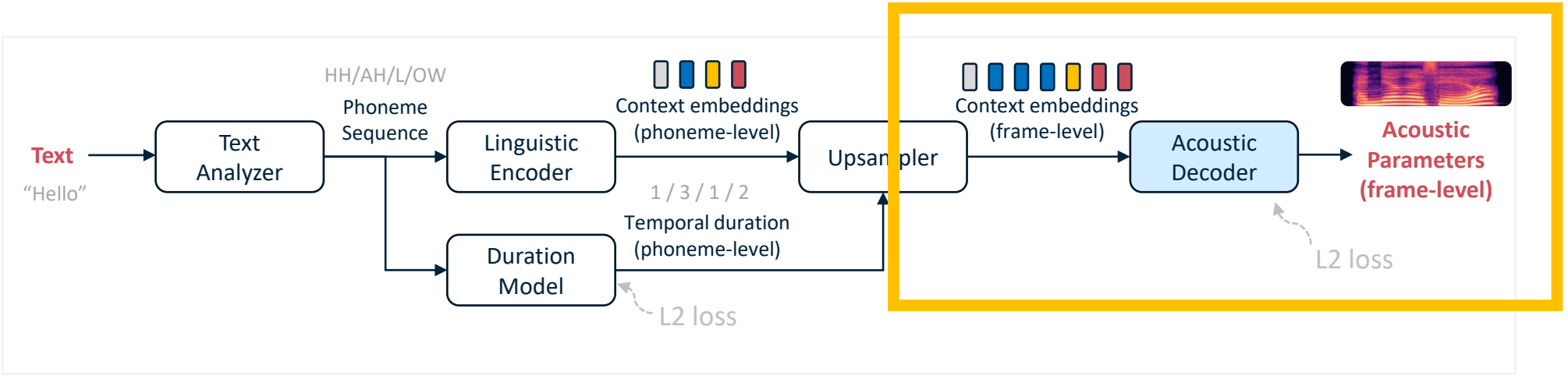
"Unlabeled"

대규모 데이터를 이용하여 pretrained 모델을 구축하고

소량의 데이터로 downstream task 에 fine-tuning 할 수 있음

Recall

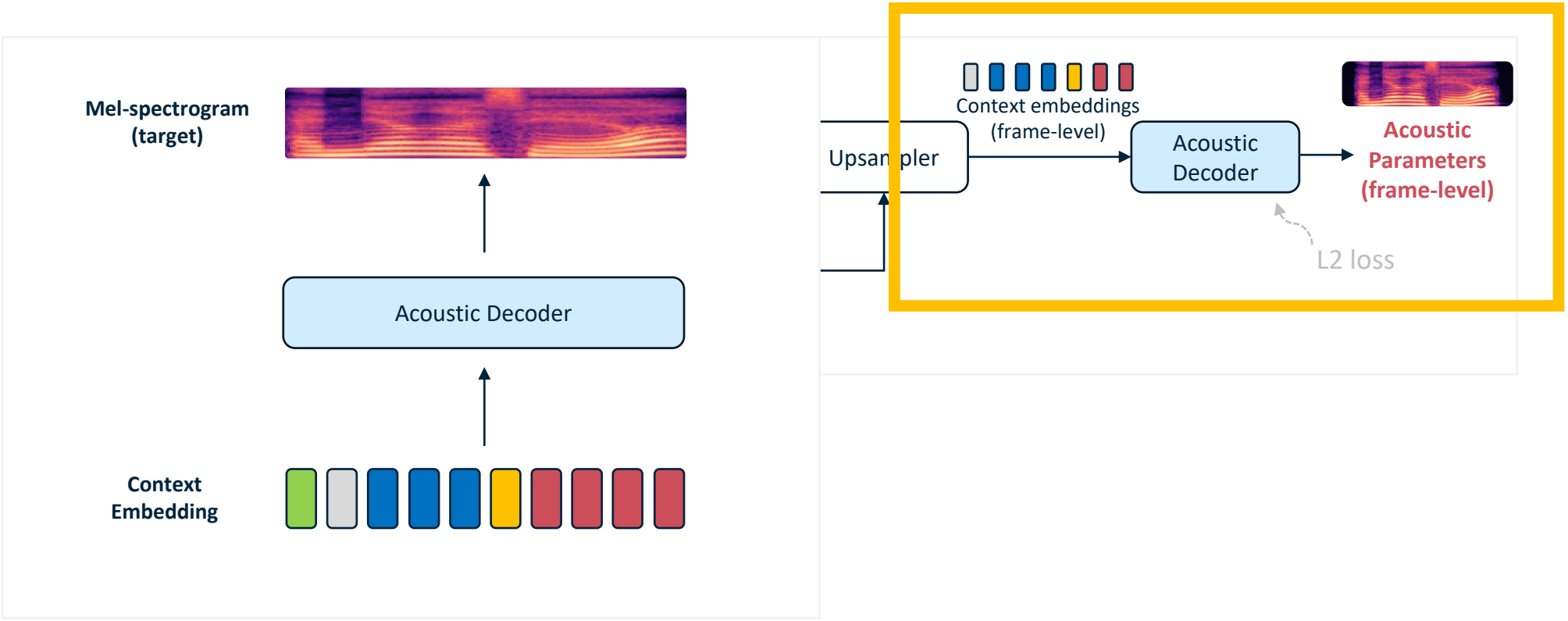
TTS acoustic model



Acoustic decoder predicts masked melspectrogram from the given context embeddings + melspectrograms

Recall

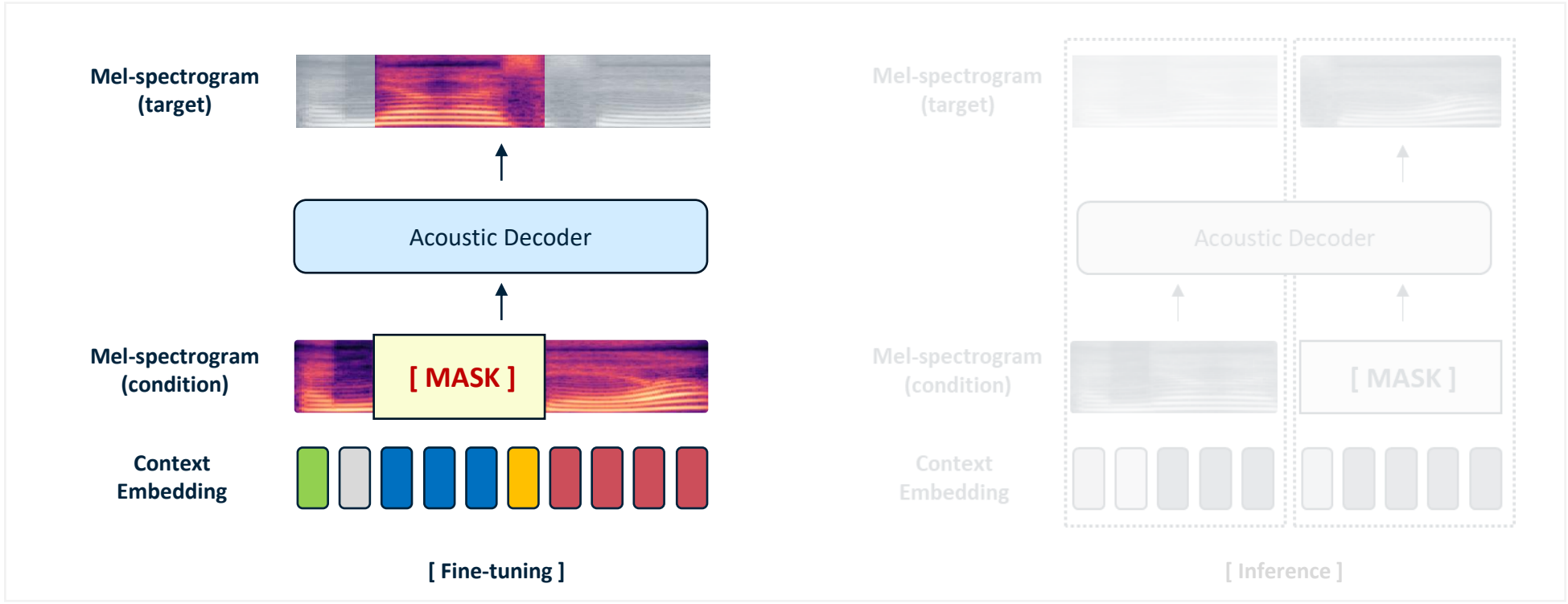
TTS acoustic model



Acoustic decoder predicts masked melspectrogram from the given context embeddings + melspectrograms

Scaling TTS model

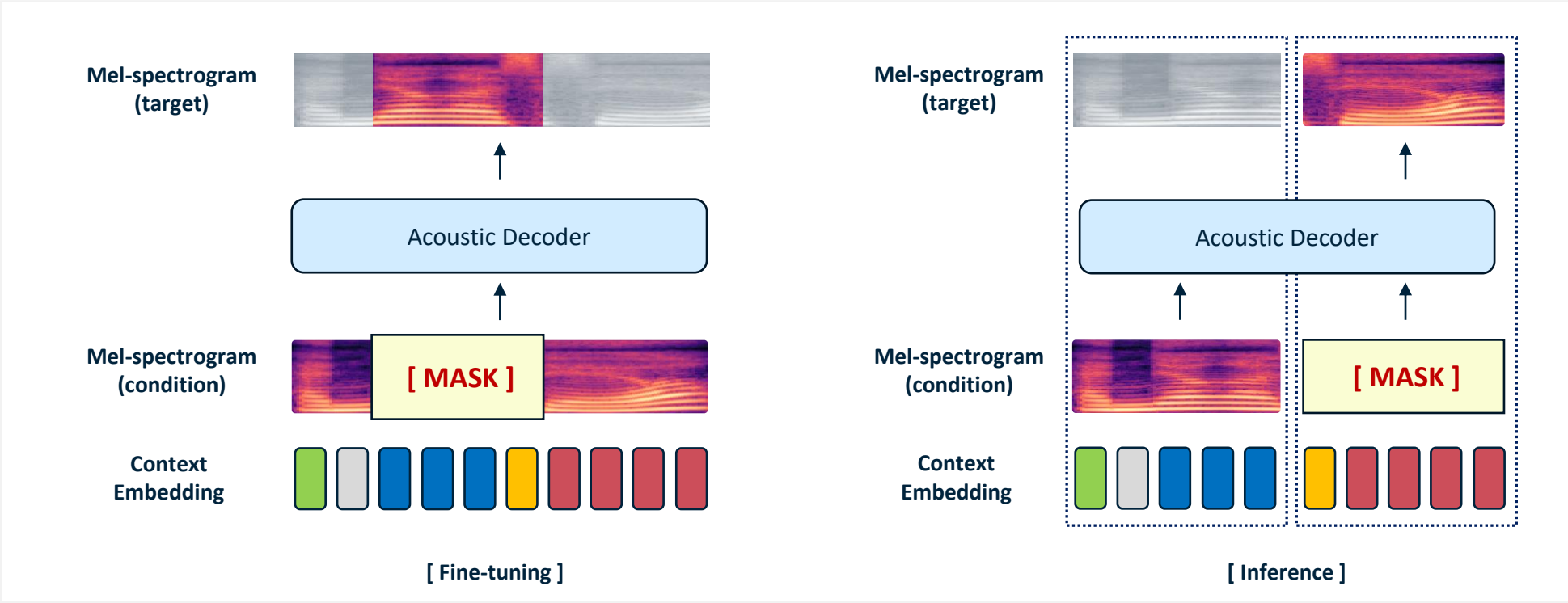
Voice cloning



Acoustic decoder predicts masked melspectrogram from the given context embeddings + melspectrograms

Scaling TTS model

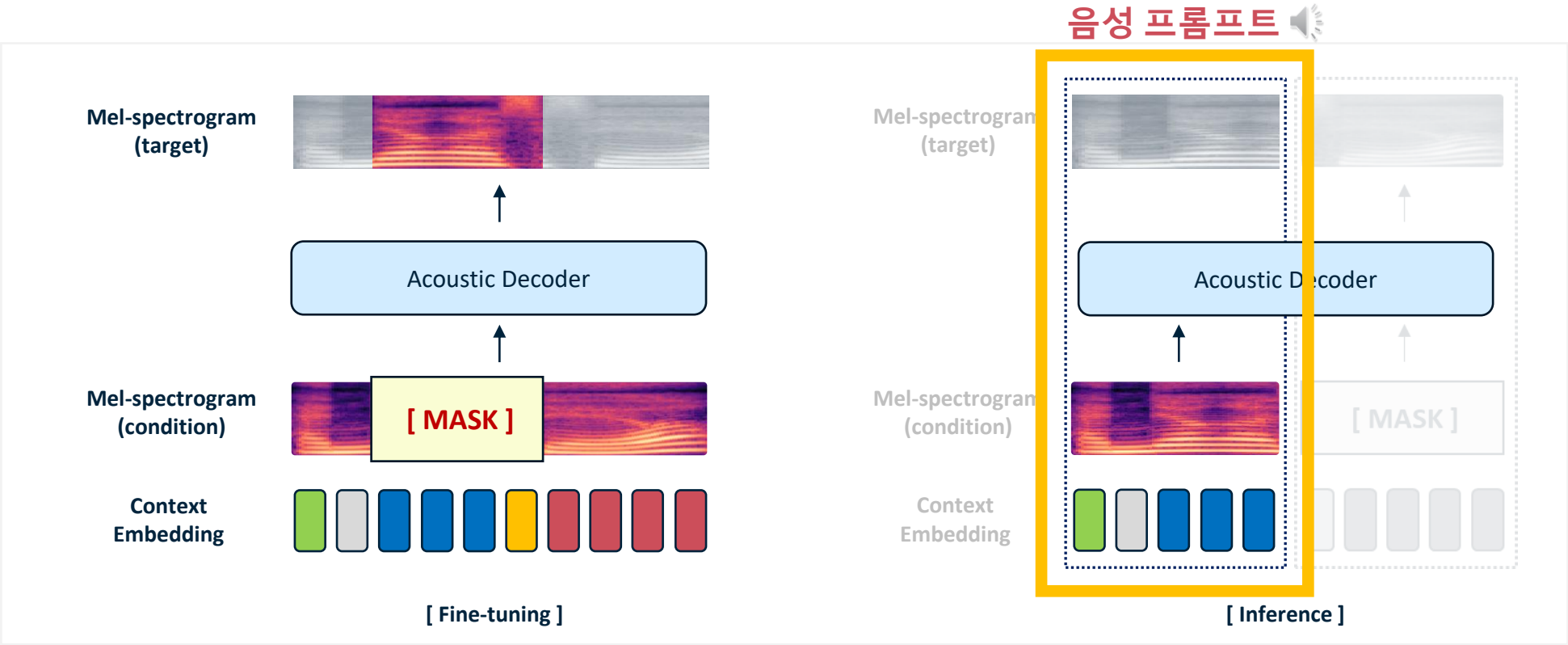
Voice cloning



Acoustic decoder predicts masked melspectrogram from the given context embeddings + melspectrograms

Scaling TTS model

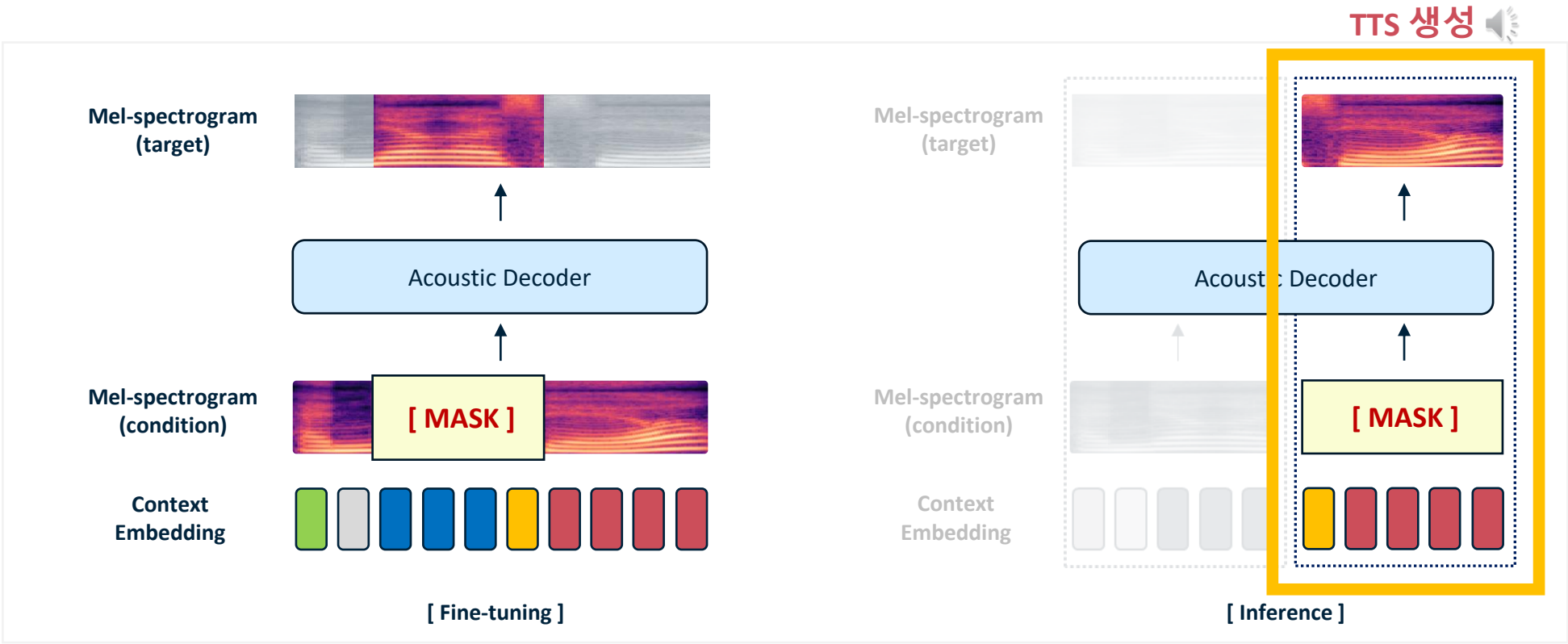
Voice cloning



디코더는 음성 프롬프트의 특성을 반영하여 TTS 생성을 할 수 있음

Scaling TTS model

Voice cloning



디코더는 음성 프롬프트의 특성을 반영하여 TTS 생성을 할 수 있음

Scaling TTS model

Voice cloning

	Conventional TTS	Custom TTS
Speaker	Professional	Non-professional
Environment	Clean studio	Anywhere
Amount	> 30~60 min	< Few seconds
Speaking type	Script reading	Spontaneous
Synthetic quality	Very natural	Natural ~~Unnatural~~

Scaling TTS model

Voice cloning

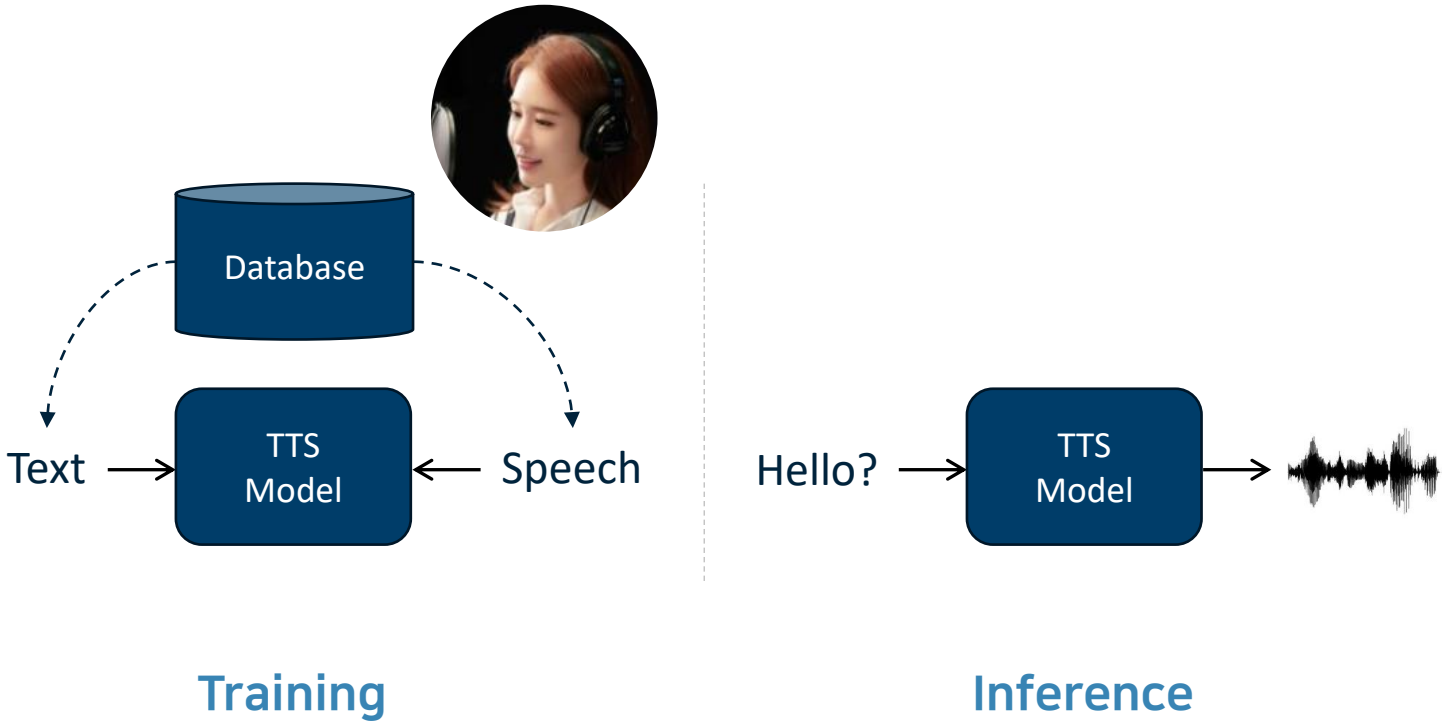
	Conventional TTS	Custom TTS
Speaker	Professional	Non-professional
Environment	Clean studio	Anywhere
Amount	> 3	< Few seconds
Speaking type	Scr	Spontaneous
Synthetic quality	Very natural	Natural ~~Unnatural~~

Ethical problem ?

Summary

Summary

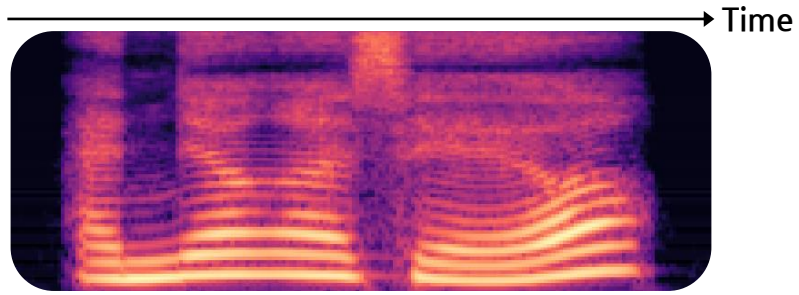
Deep learning-based TTS system



Human-like voice quality ☺

Summary

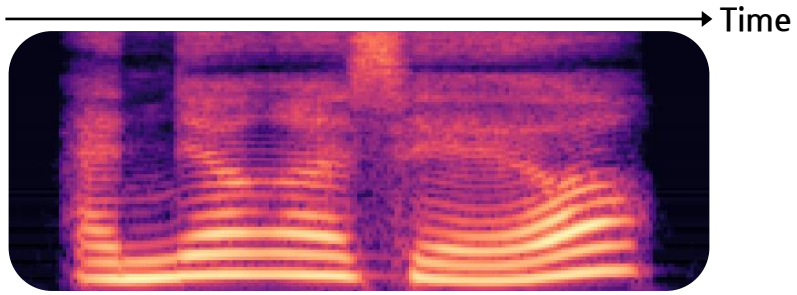
Deep learning-based TTS system



Acoustic model + Vocoding model

Summary

Deep learning-based TTS system



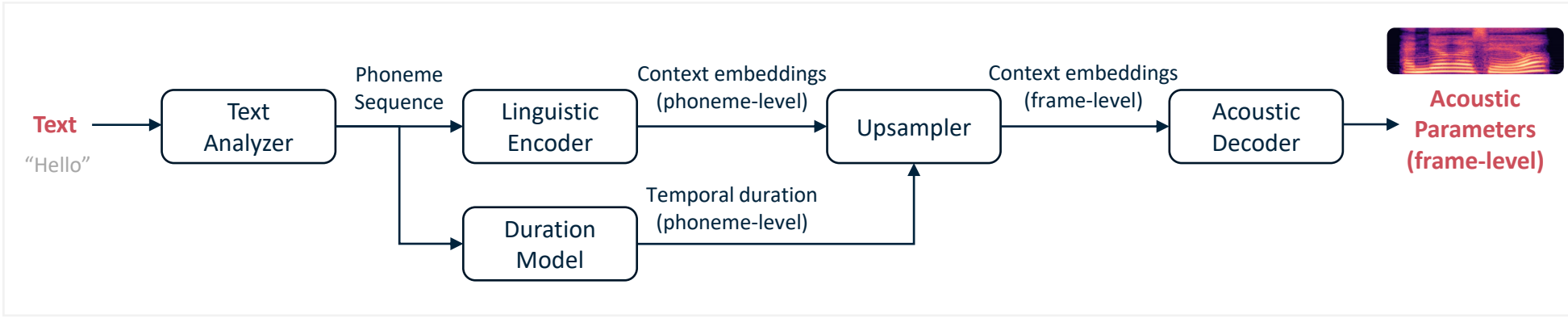
Estimating acoustic parameters from text inputs

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

Acoustic model + Vocoding model

Summary

Deep learning-based TTS system



Estimating acoustic parameters from text inputs

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

Acoustic model + Vocoding model

Summary

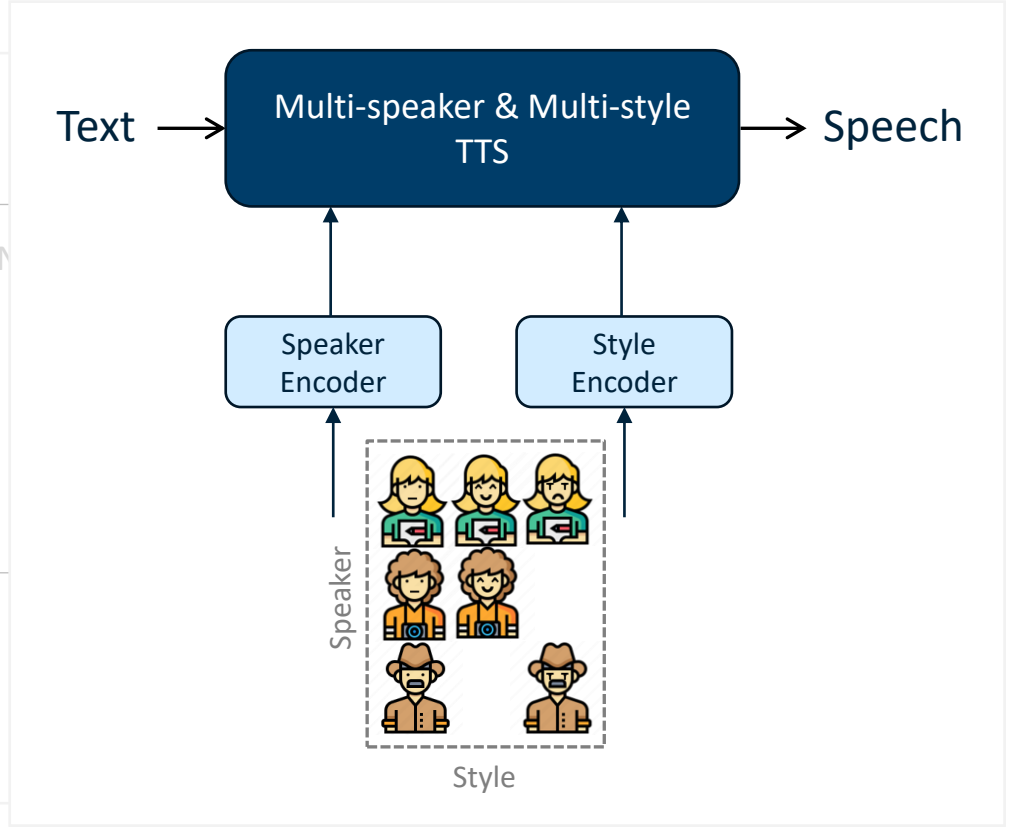
Data collection

	Conventional TTS	Custom TTS
Speaker	Professional	Non-professional
Environment	Clean studio	Anywhere
Amount	> 30~60 min	< Few seconds
Speaking type	Script reading	Spontaneous
Synthetic quality	Very natural	Natural ~~Unnatural~~

Summary

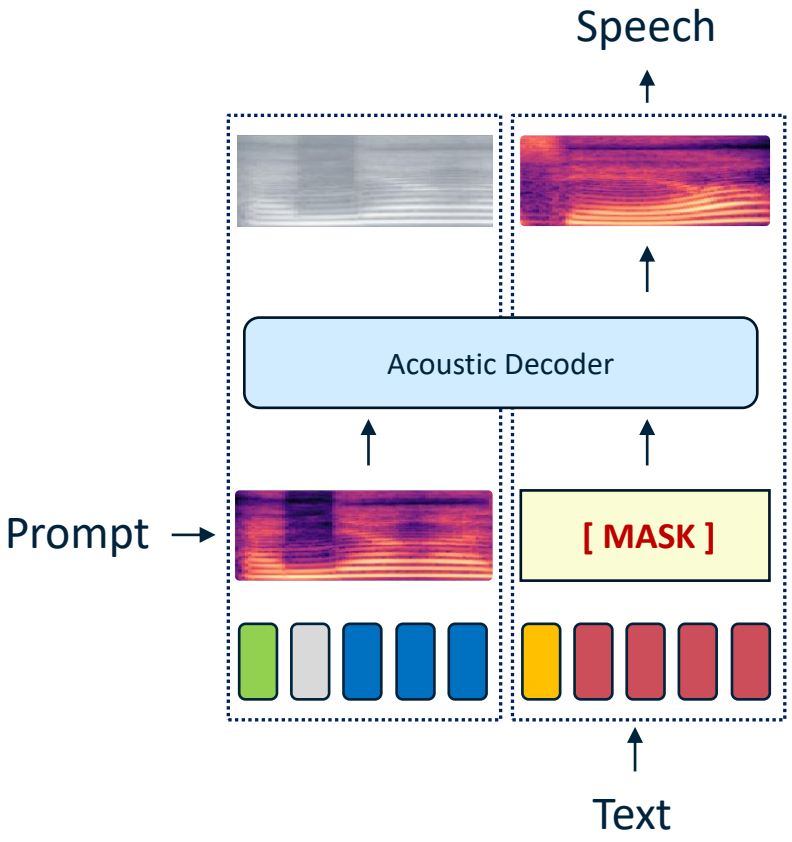
Multi-speaker & Multi-style TTS

	Conventional TTS
Speaker	Professional
Environment	Clean studio
Amount	> 30~60 min
Speaking type	Script reading
Synthetic quality	Very natural



Summary

Large-scale TTS



	Custom TTS
	Non-professional
	Anywhere
	< Few seconds
	Spontaneous
	Natural
	~~Unnatural~~

Q / A

gregorio.song@gmail.com