ENGINEERING DAY

음성 합성 모델로 음성 합성 모델 만들기

송은우 / HDTS

NAVER

ENGINEERING DAY

Data-selective TTS augmentation

송은우 / HDTS

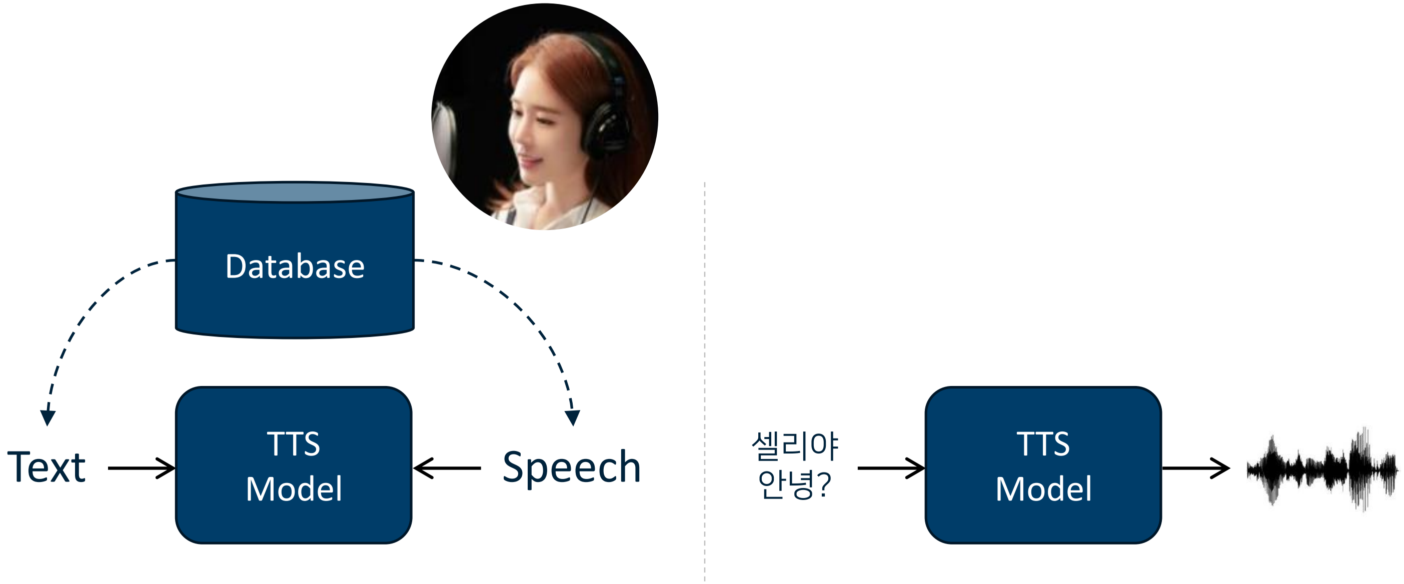
NAVER

Introduction

Text-to-speech (TTS): Synthesize speech signal from text input

Introduction

Deep learning-based TTS system



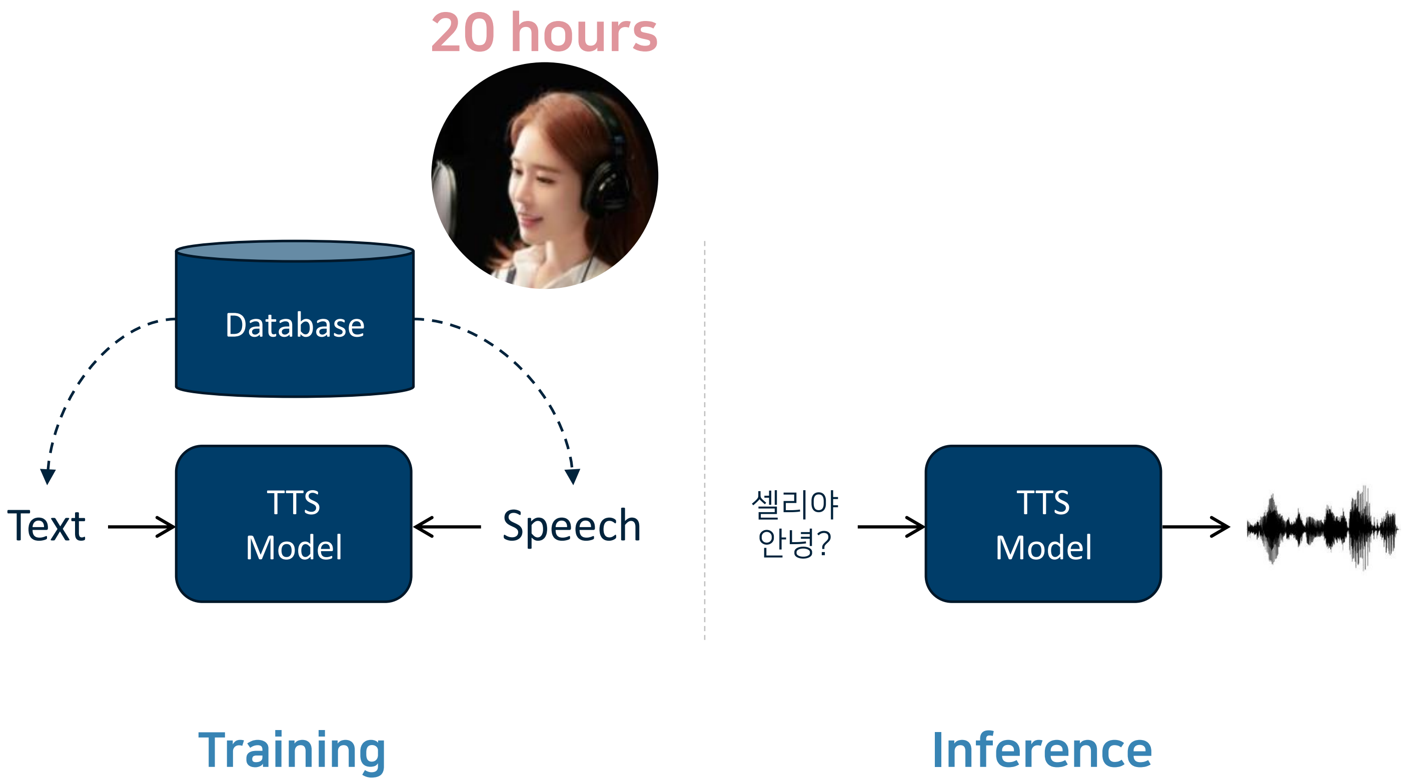
Training

Inference

Human-like voice quality

Introduction

Deep learning-based TTS system



Require huge amount of speech recordings

Introduction

Solutions

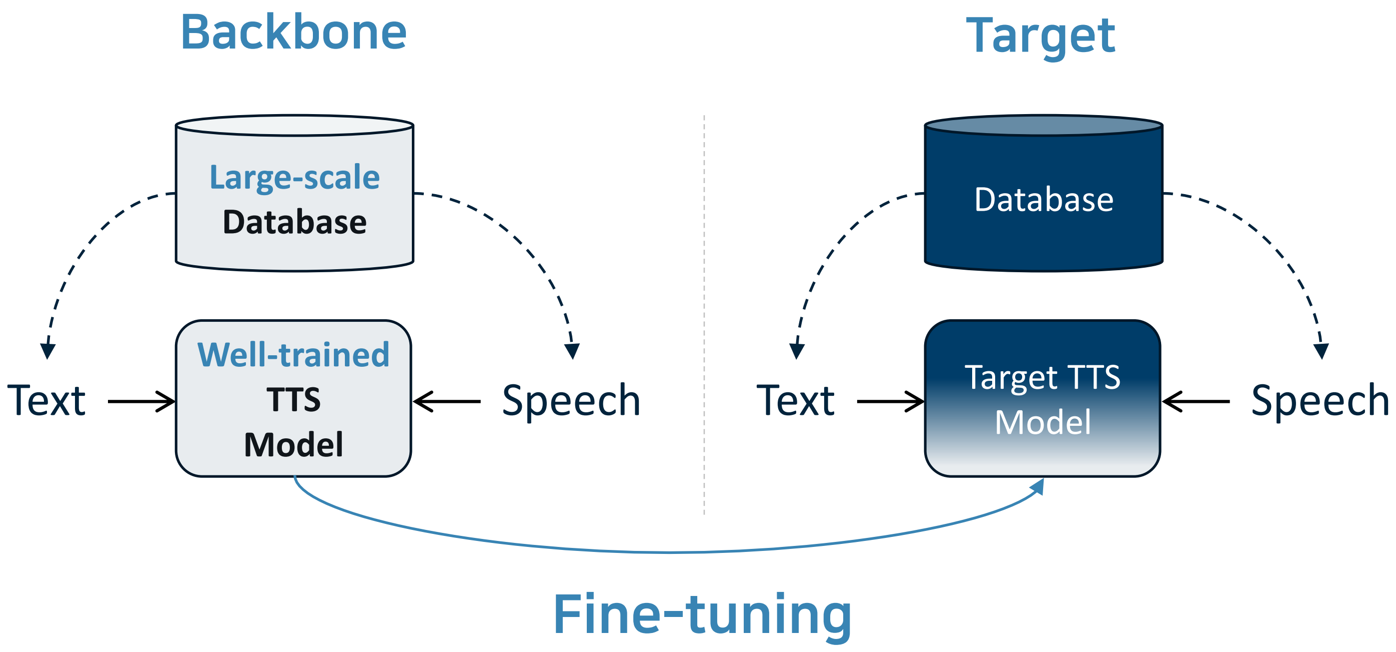
①: Speaker adaptation

②: Data augmentation

Introduction

Deep learning-based TTS system

Solution ①: Speaker adaptation

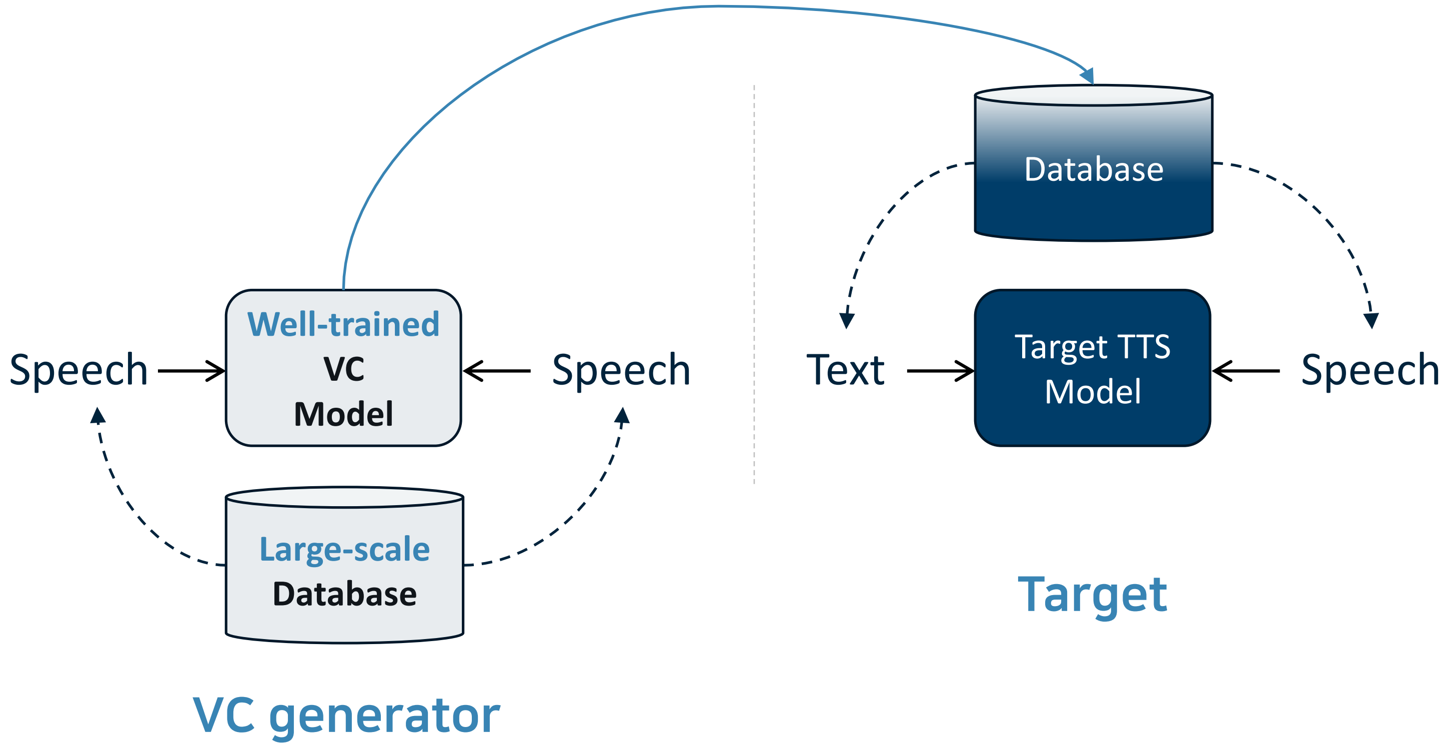


"안녕하세요, 당신의 똑똑한 비서 클로바 입니다. 무엇을 도와드릴까요?"

Introduction

Deep learning-based TTS system

Solution ②: Data augmentation using voice conversion (VC)

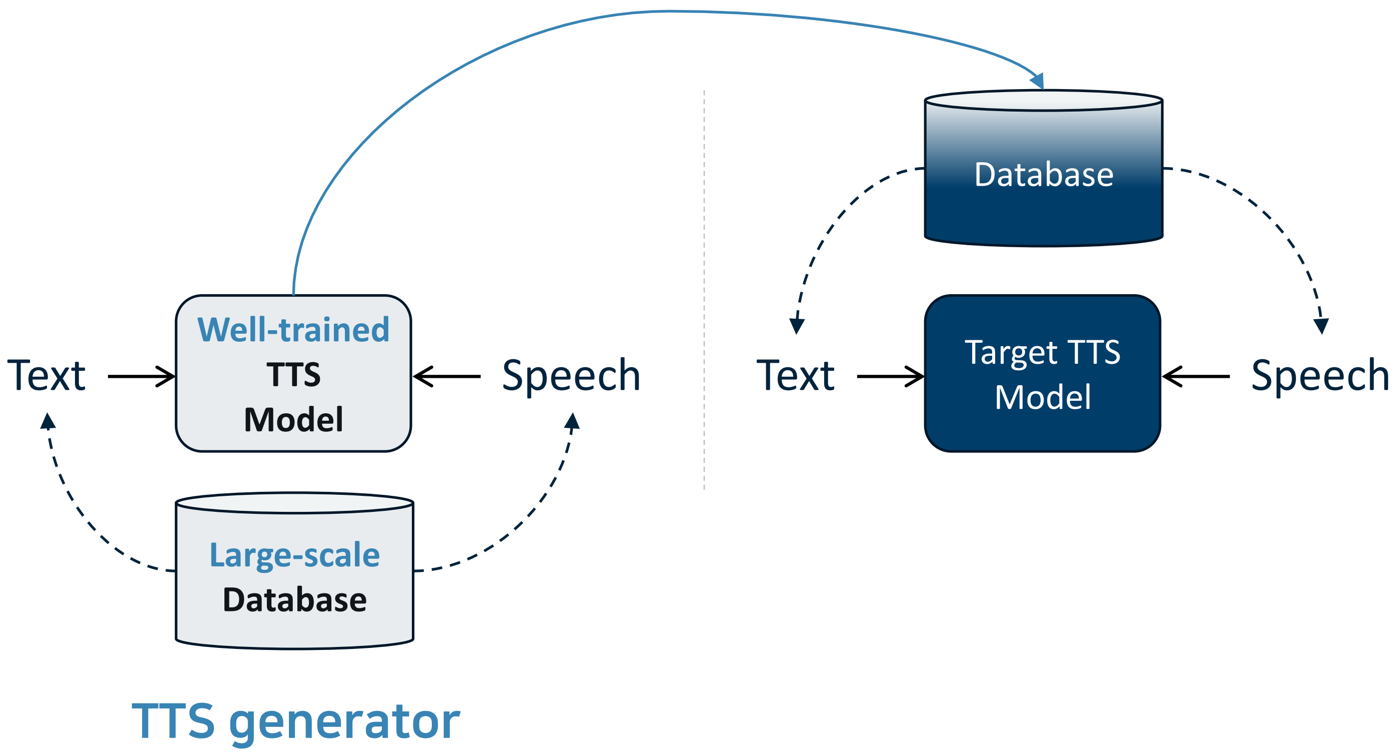


"난 약한 친구들을 괴롭히는 심술쟁이는 딱 질색이야!"

Introduction

Deep learning-based TTS system

Solution ②: Data augmentation using TTS



Data-selective TTS augmentation

TTS-by-TTS

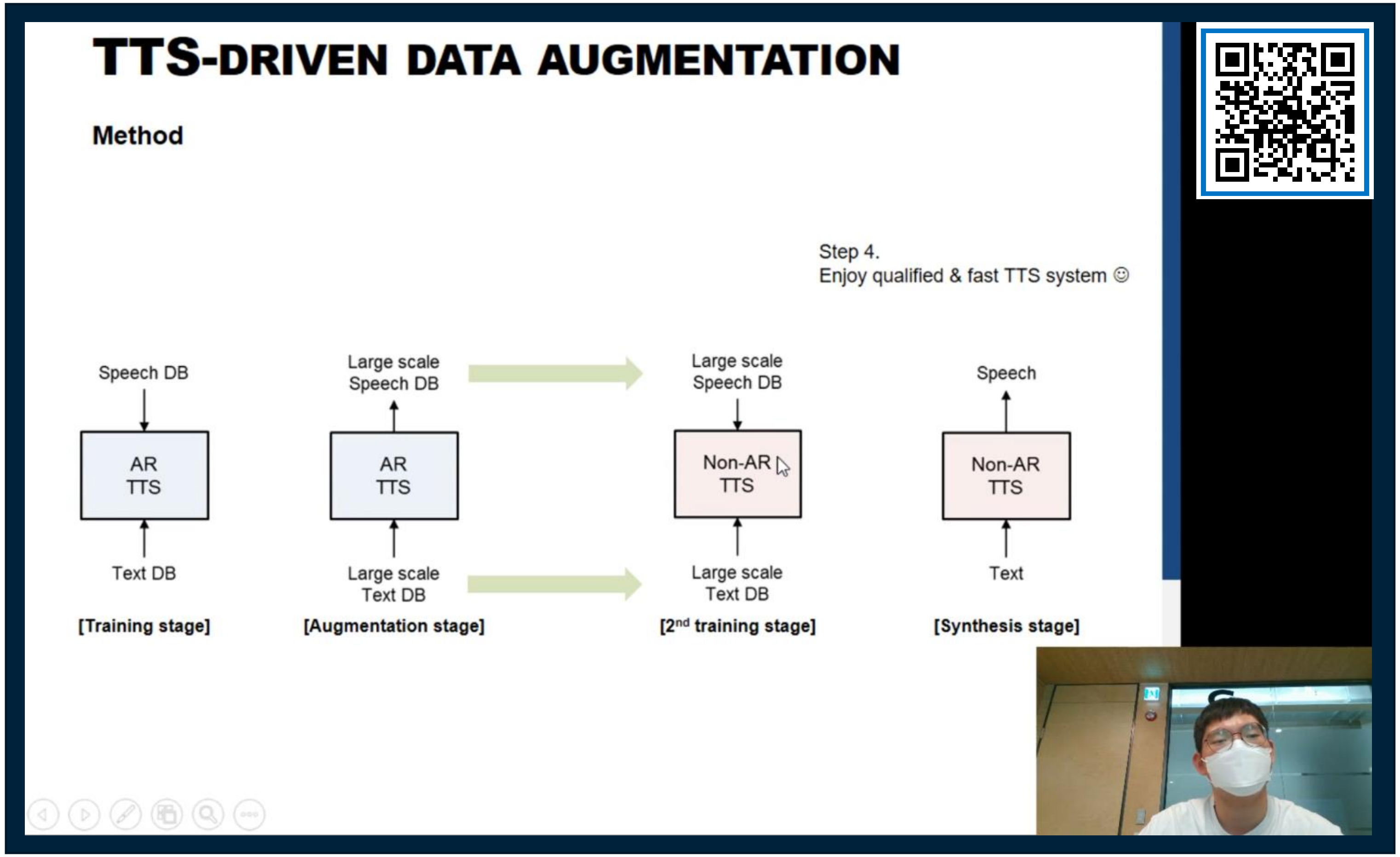
TTS-by-TTS

Train target TTS model

via large-scale corpora synthesized by TTS model

TTS-by-TTS

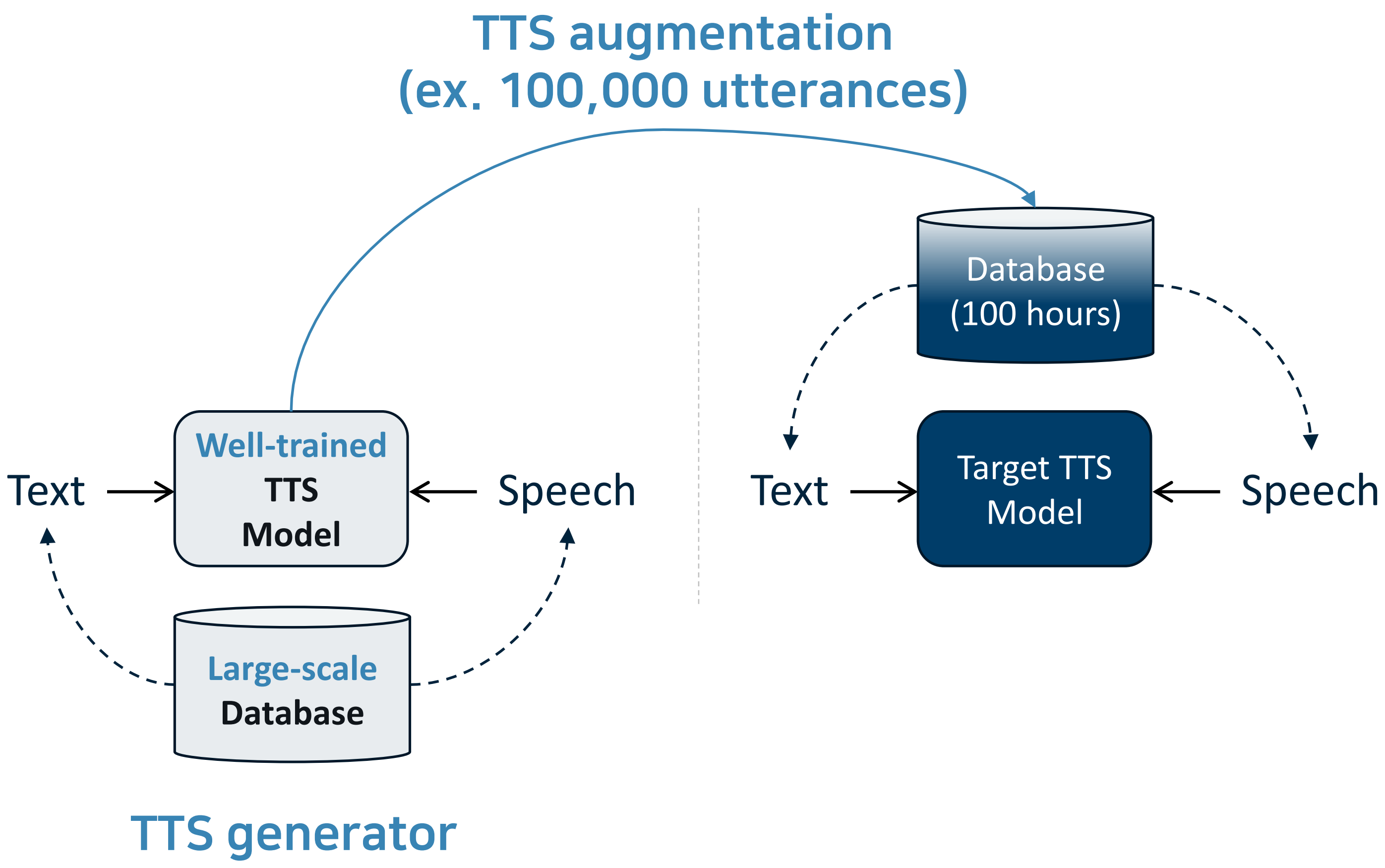
2020 Engineering day: 가짜 목소리 DB로 고품질 음성합성기를 만들어보자 (HDTS 황민제님)



https://share.navercorp.com/neday2020/lecture/245259

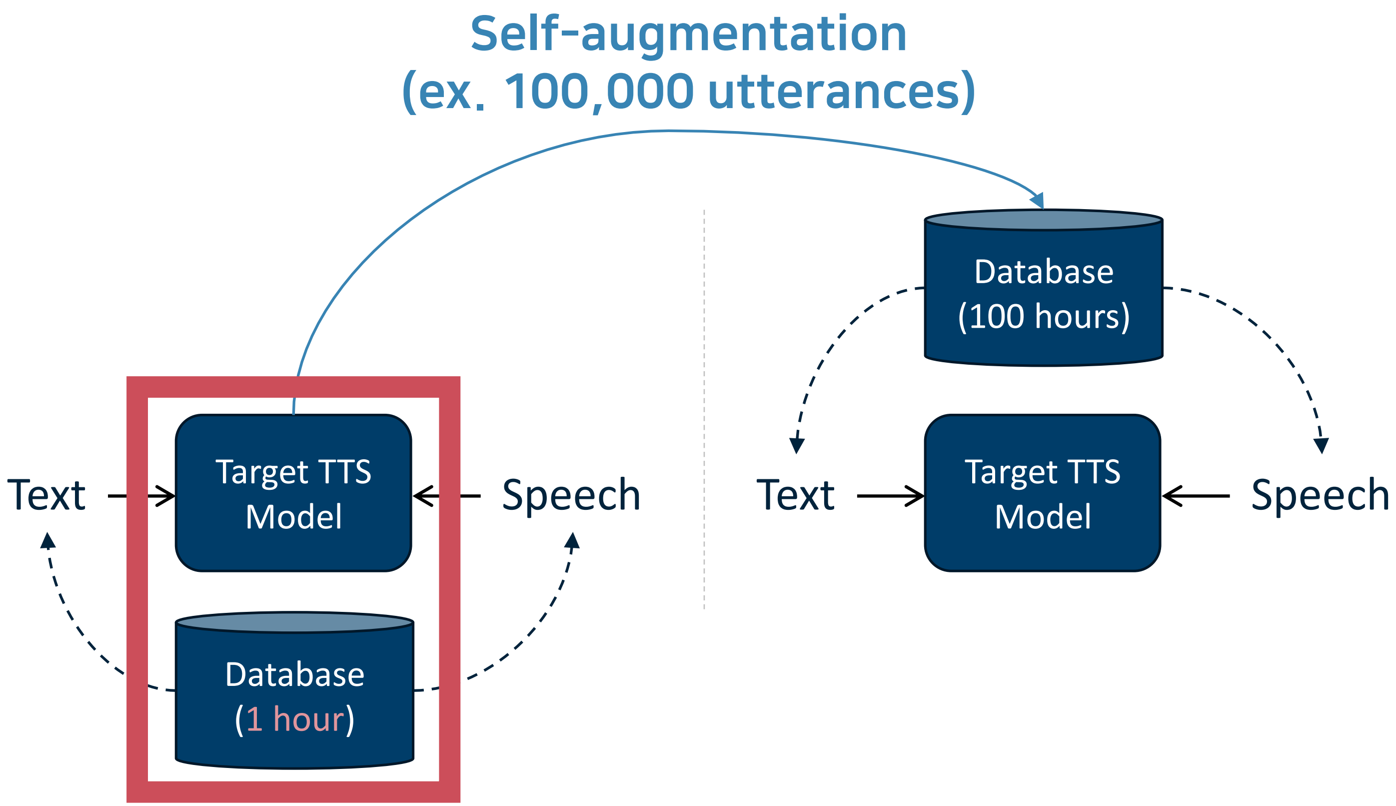
TTS-by-TTS

Train target TTS model via large-scale corpora synthesized by TTS model



TTS-by-TTS

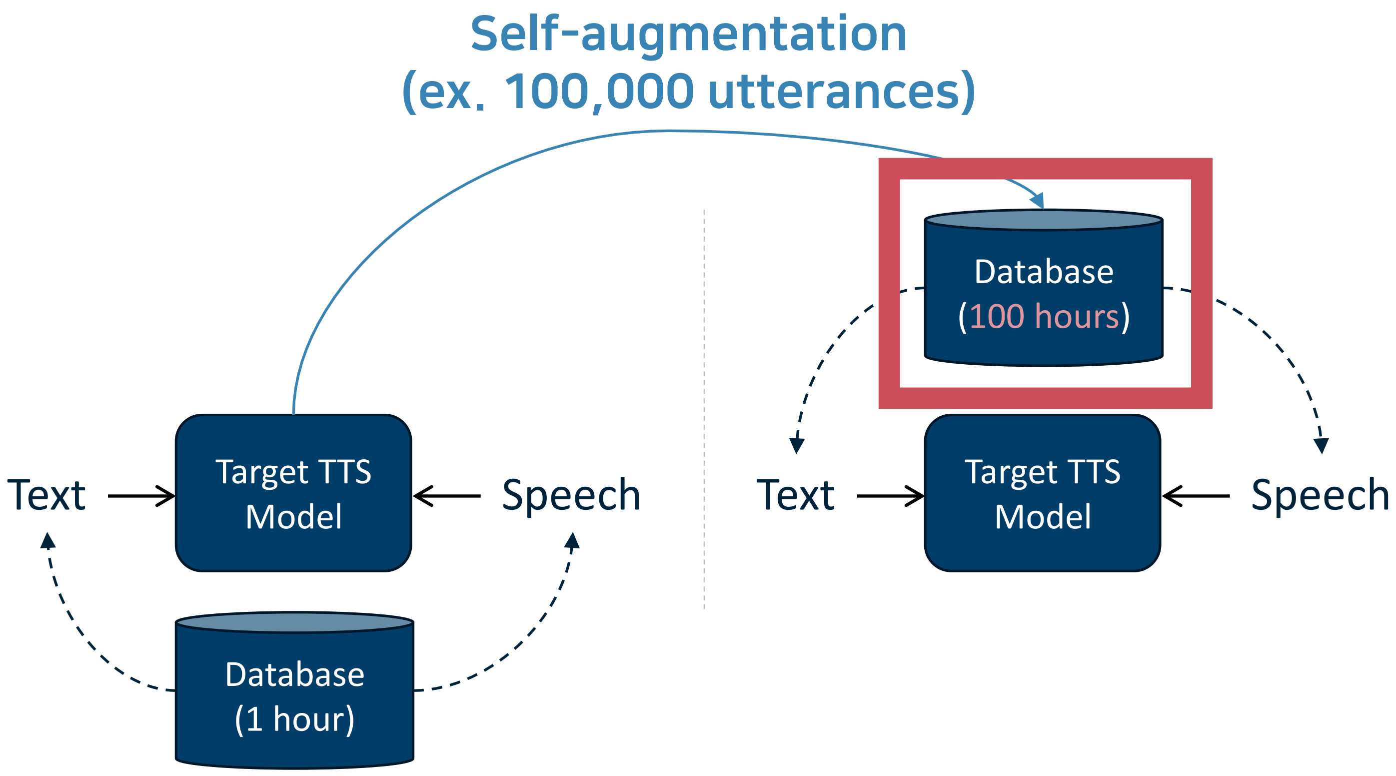
Train target TTS model via large-scale corpora synthesized by TTS model



If the amount of training data is not enough…

TTS-by-TTS

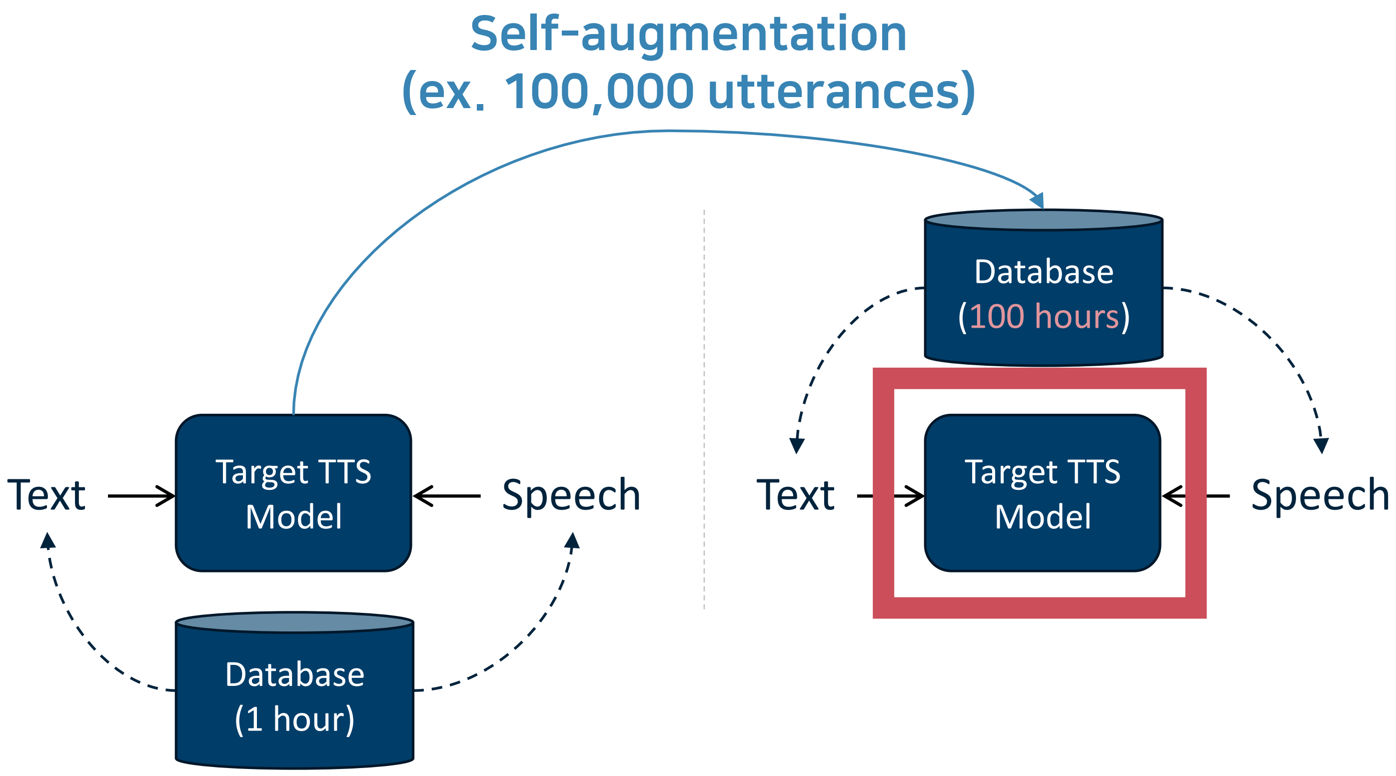
Train target TTS model via large-scale corpora synthesized by TTS model



Many of synthetic corpora contain poorly generated speech samples ☹

TTS-by-TTS

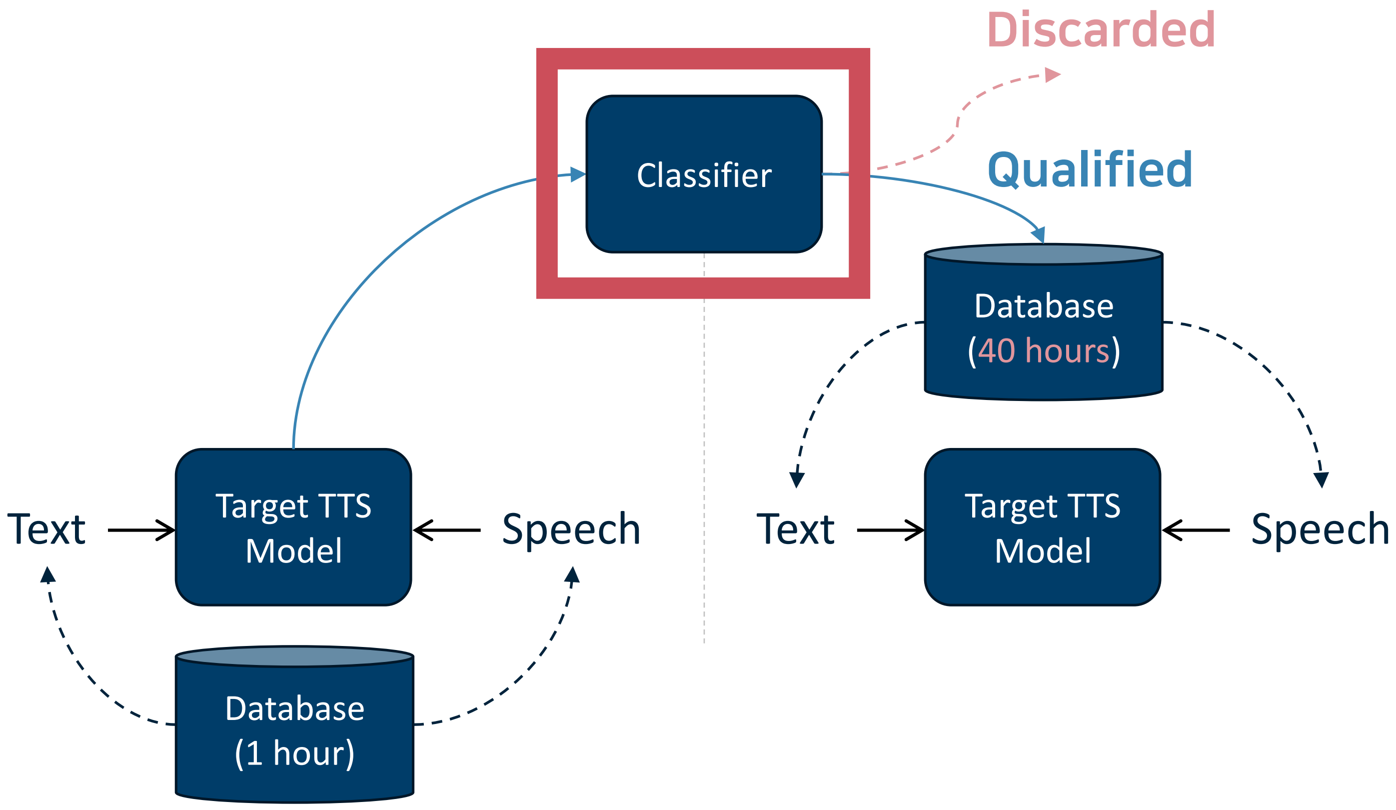
Train target TTS model via large-scale corpora synthesized by TTS model



Merely increasing synthetic data is not always advantageous ☹

TTS-by-TTS

Train target TTS model via large-scale corpora synthesized by TTS model



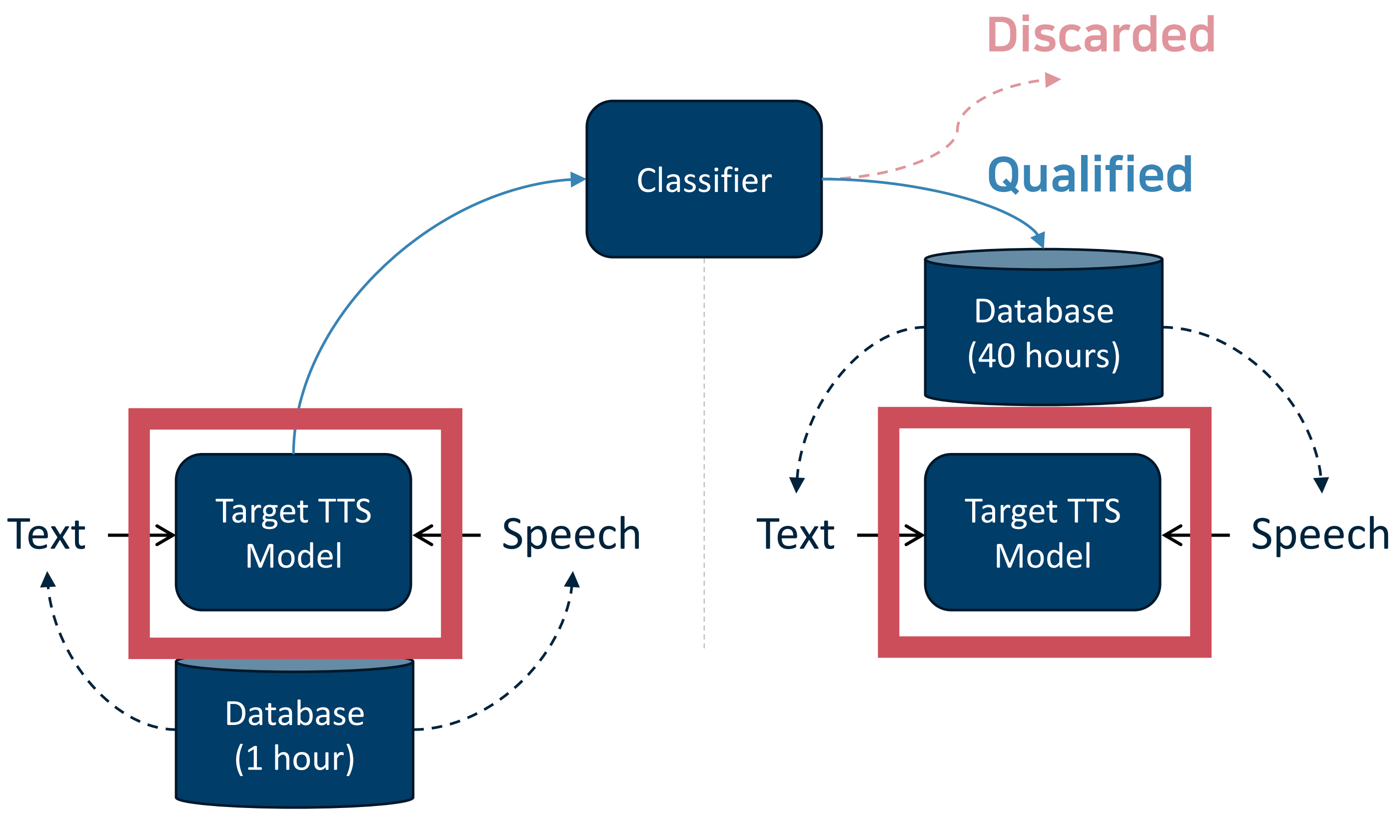
It is very important to selectively choose synthetic data that are beneficial to training process

Data-selective TTS augmentation

Method

Target TTS model

Duration informed Tacotron 2 with variational autoencoder (VAE)



It is crucial to design a well-structured TTS model to synthesize high-quality speech database

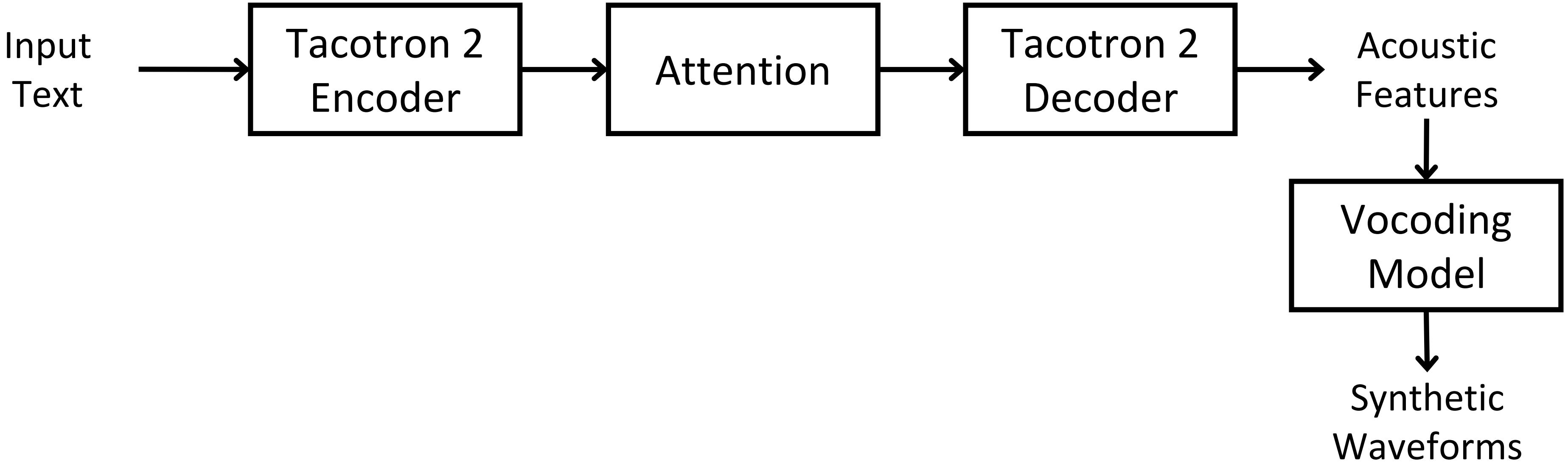
Target TTS model

Duration informed Tacotron 2 with variational autoencoder (VAE)

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`



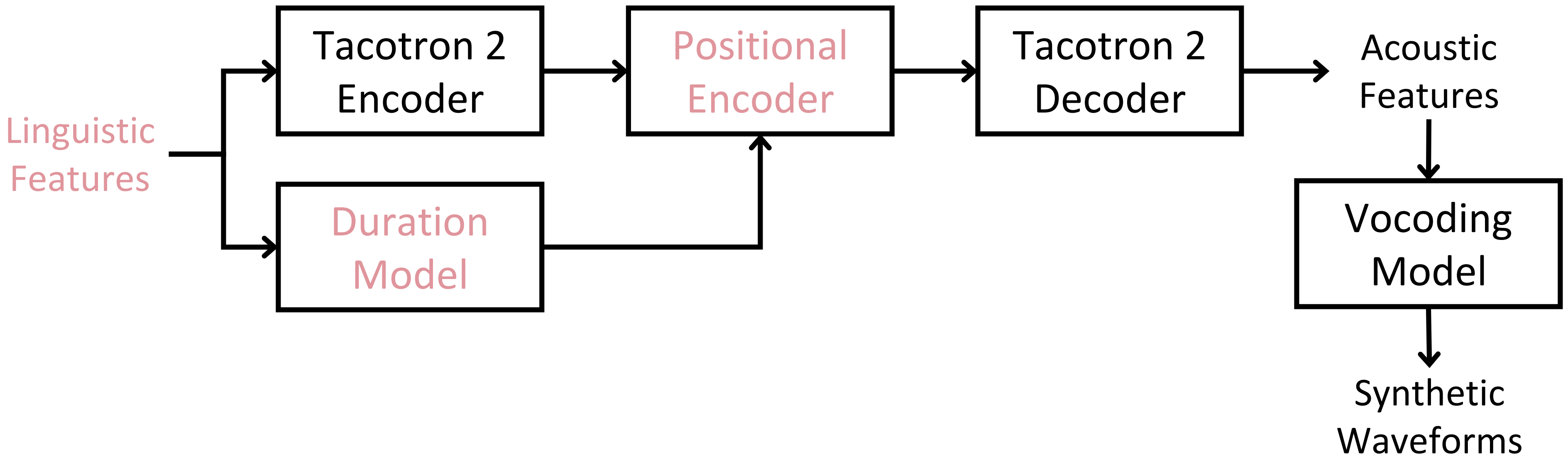
Target TTS model

Duration informed Tacotron 2 with variational autoencoder (VAE)

TACOTRON-BASED ACOUSTIC MODEL USING PHONEME ALIGNMENT FOR PRACTICAL NEURAL TEXT-TO-SPEECH SYSTEMS

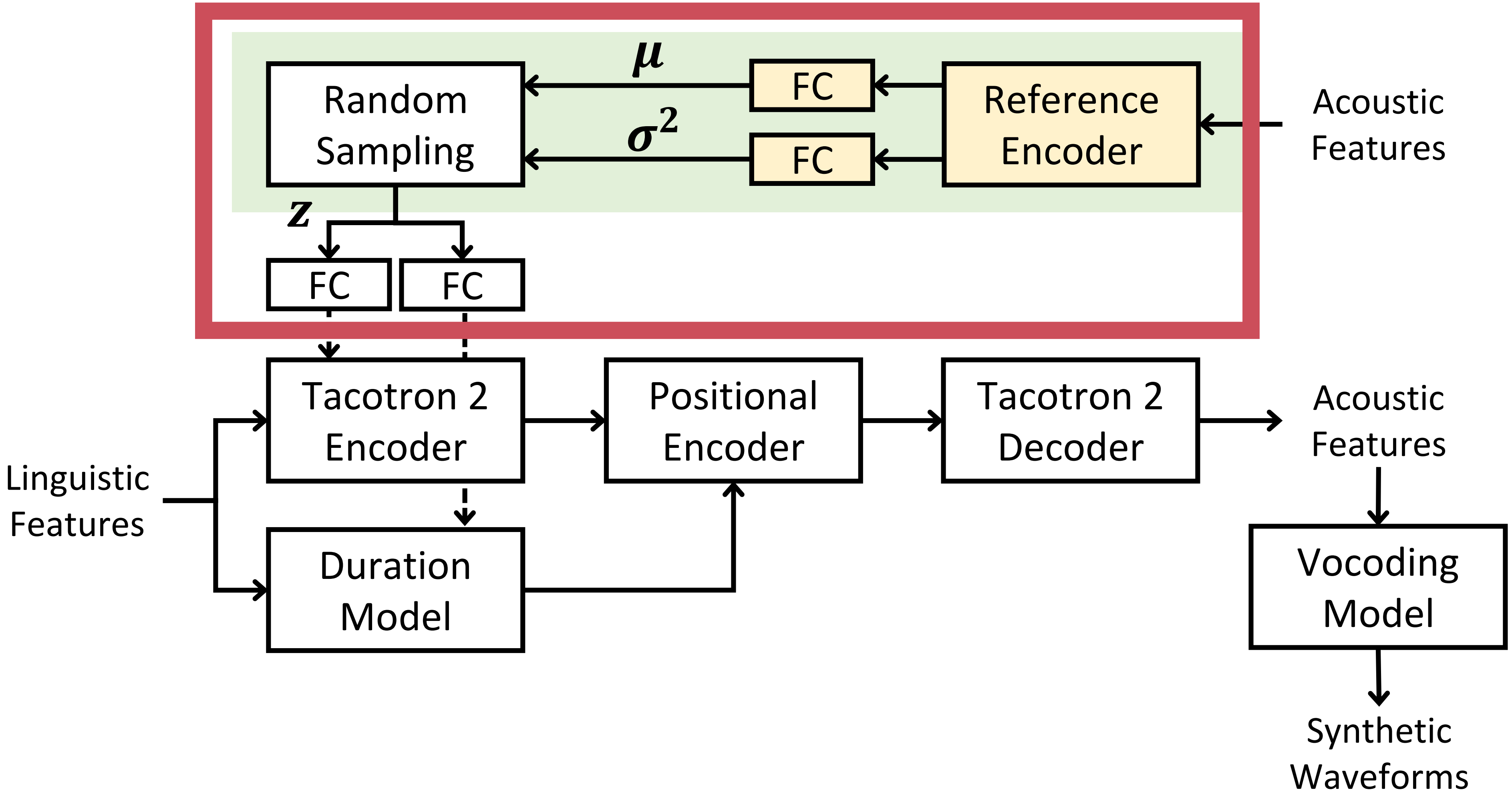
Takuma Okamoto[1], *Tomoki Toda*[2,1], *Yoshinori Shiga*[1], *and Hisashi Kawai*[1]

[1]National Institute of Information and Communications Technology, Japan
[2]Information Technology Center, Nagoya University, Japan



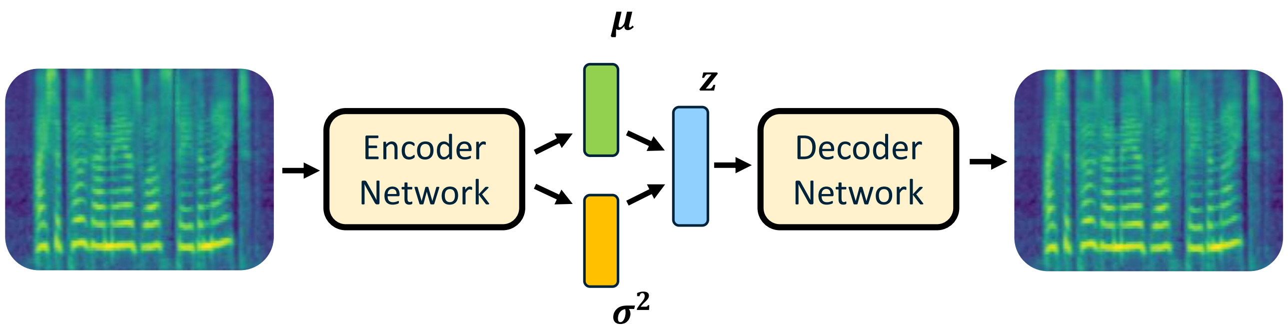
Target TTS model

Duration informed Tacotron 2 with variational autoencoder (VAE)



VAE

Duration informed Tacotron 2 with variational autoencoder (VAE)



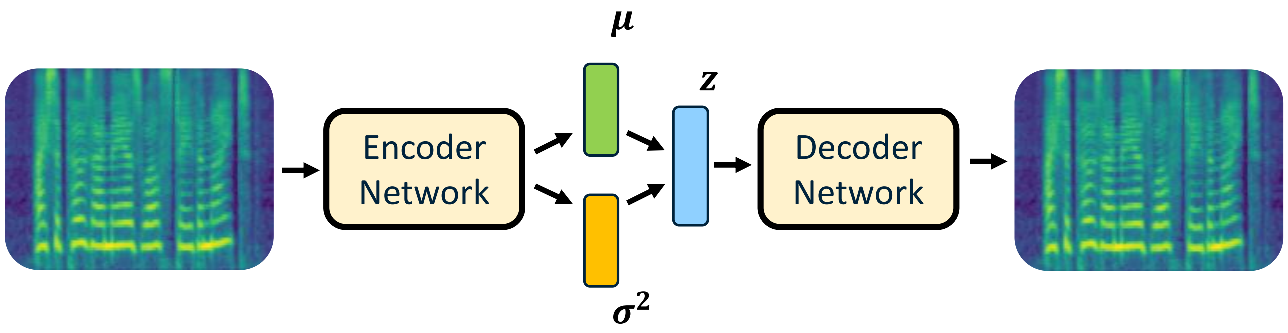
$\boldsymbol{\mu}$ Mean vector

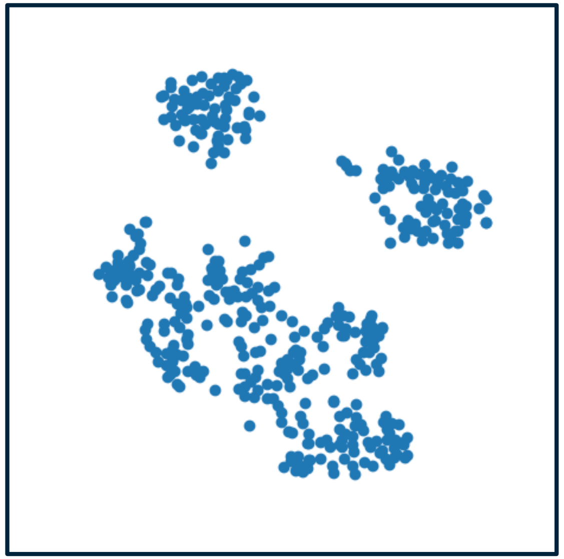
$\boldsymbol{\sigma}^2$ Variance vector

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

VAE

Duration informed Tacotron 2 with variational autoencoder (VAE)





μ Mean vector

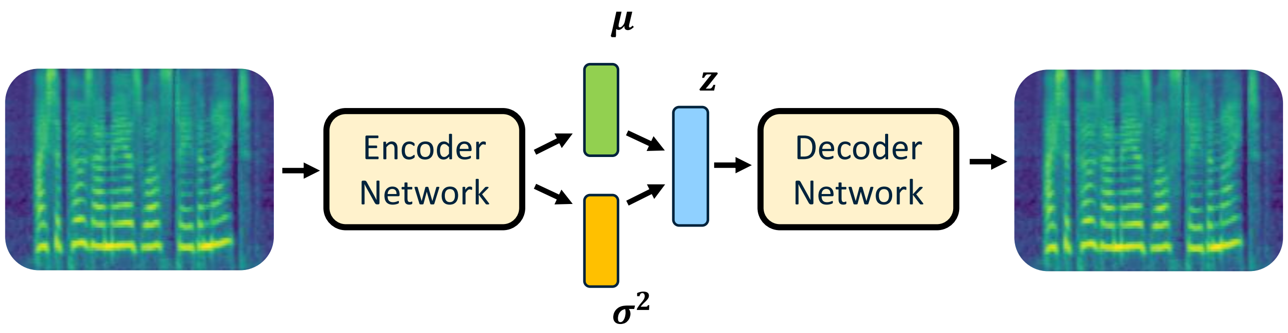
σ^2 Variance vector

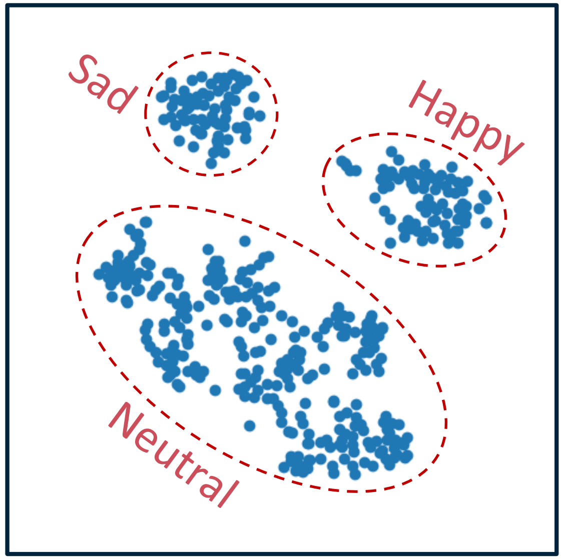
Unsupervised feature representation

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

VAE

Duration informed Tacotron 2 with variational autoencoder (VAE)





μ Mean vector

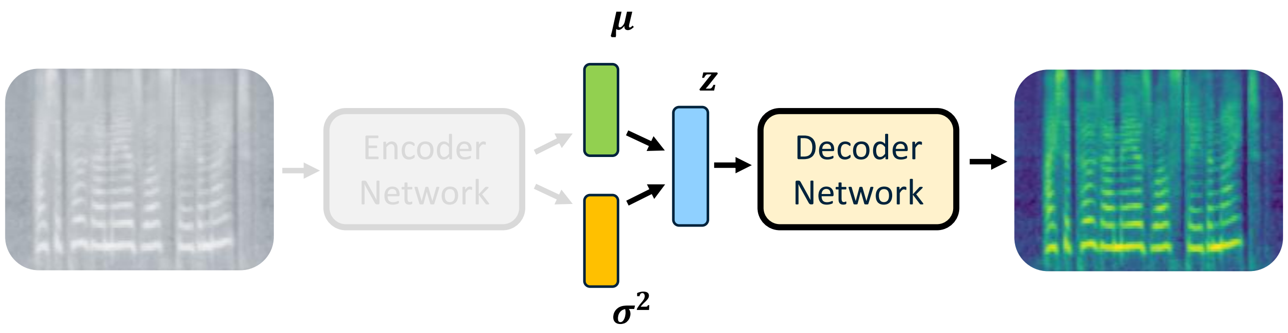
σ^2 Variance vector

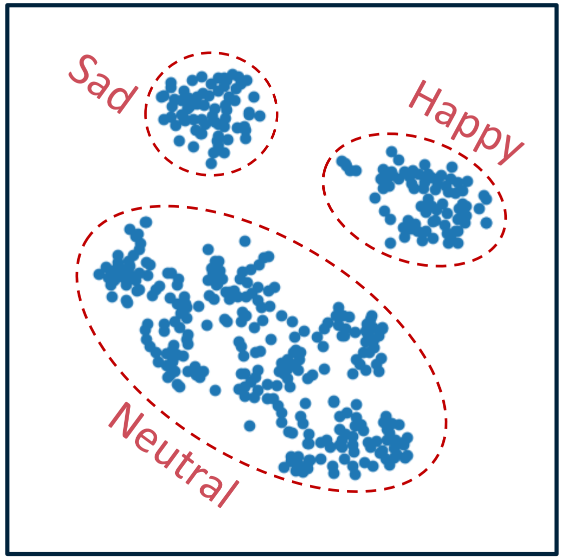
Unsupervised feature representation

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

VAE

Duration informed Tacotron 2 with variational autoencoder (VAE)





μ Mean vector

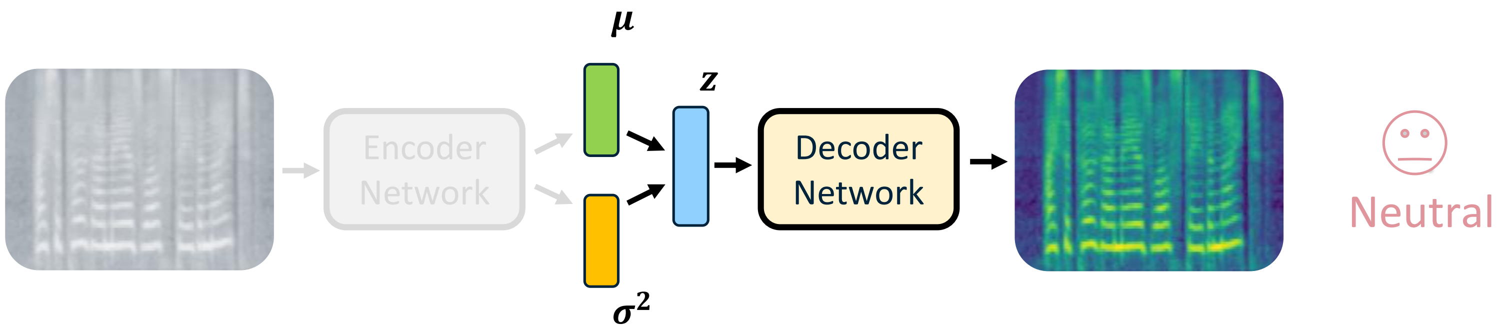
σ^2 Variance vector

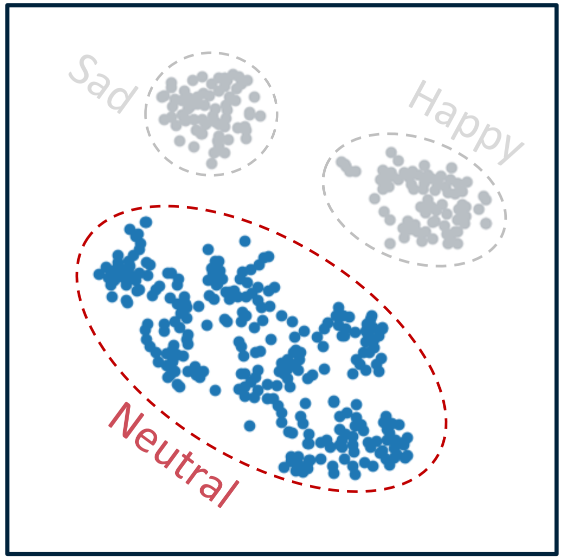
Unsupervised feature representation

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

VAE

Duration informed Tacotron 2 with variational autoencoder (VAE)





μ Mean vector

σ^2 Variance vector

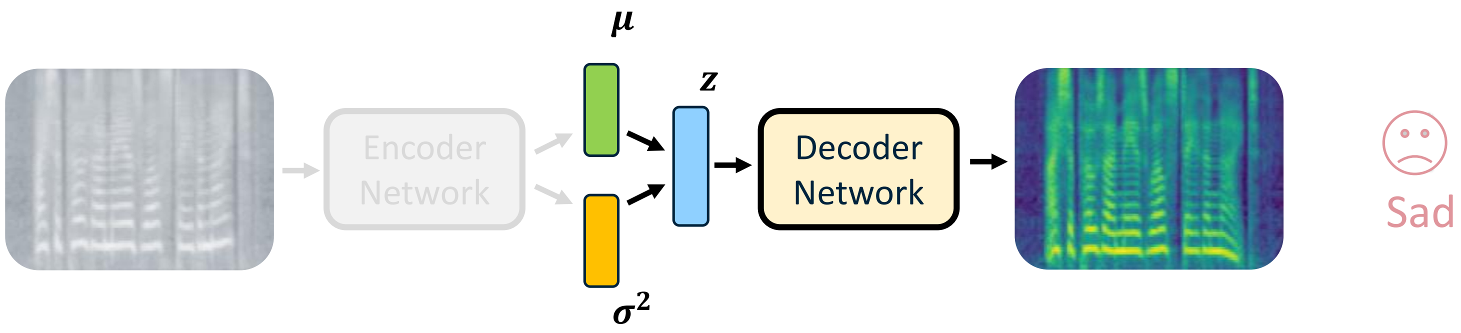
Unsupervised feature representation

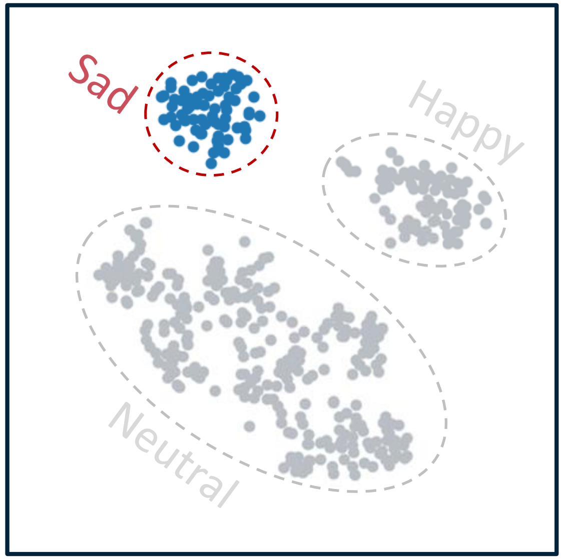
z Latent vector ~ $N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ Variations to output

"제가 당신의 위로가 되고 싶어요. 기분 쳐지지 말고 파이팅 하세요."

VAE

Duration informed Tacotron 2 with variational autoencoder (VAE)





μ Mean vector

σ^2 Variance vector

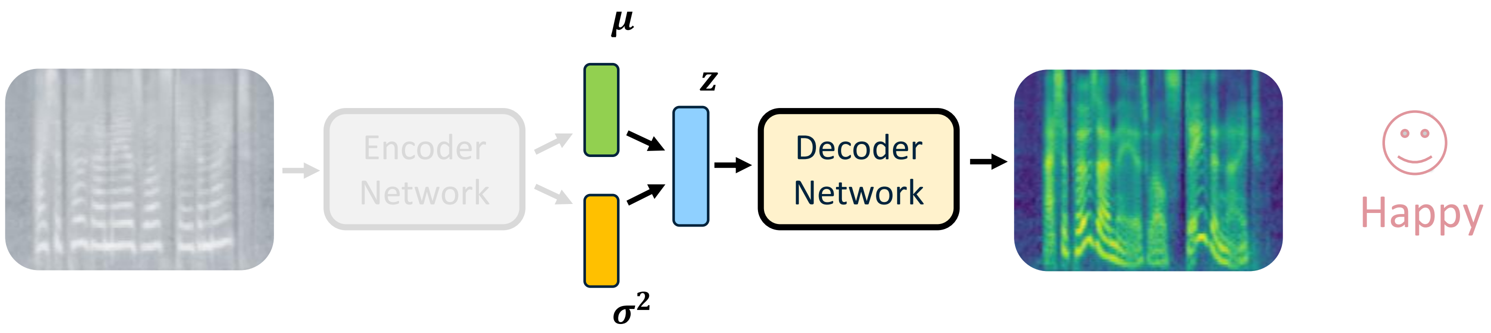
Unsupervised feature representation

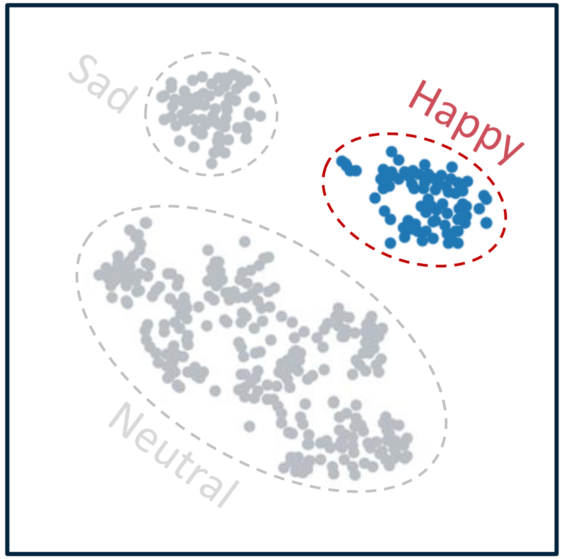
z Latent vector ~ $N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ Variations to output

"제가 당신의 위로가 되고 싶어요. 기분 쳐지지 말고 파이팅 하세요."

VAE

Duration informed Tacotron 2 with variational autoencoder (VAE)





μ Mean vector

σ^2 Variance vector

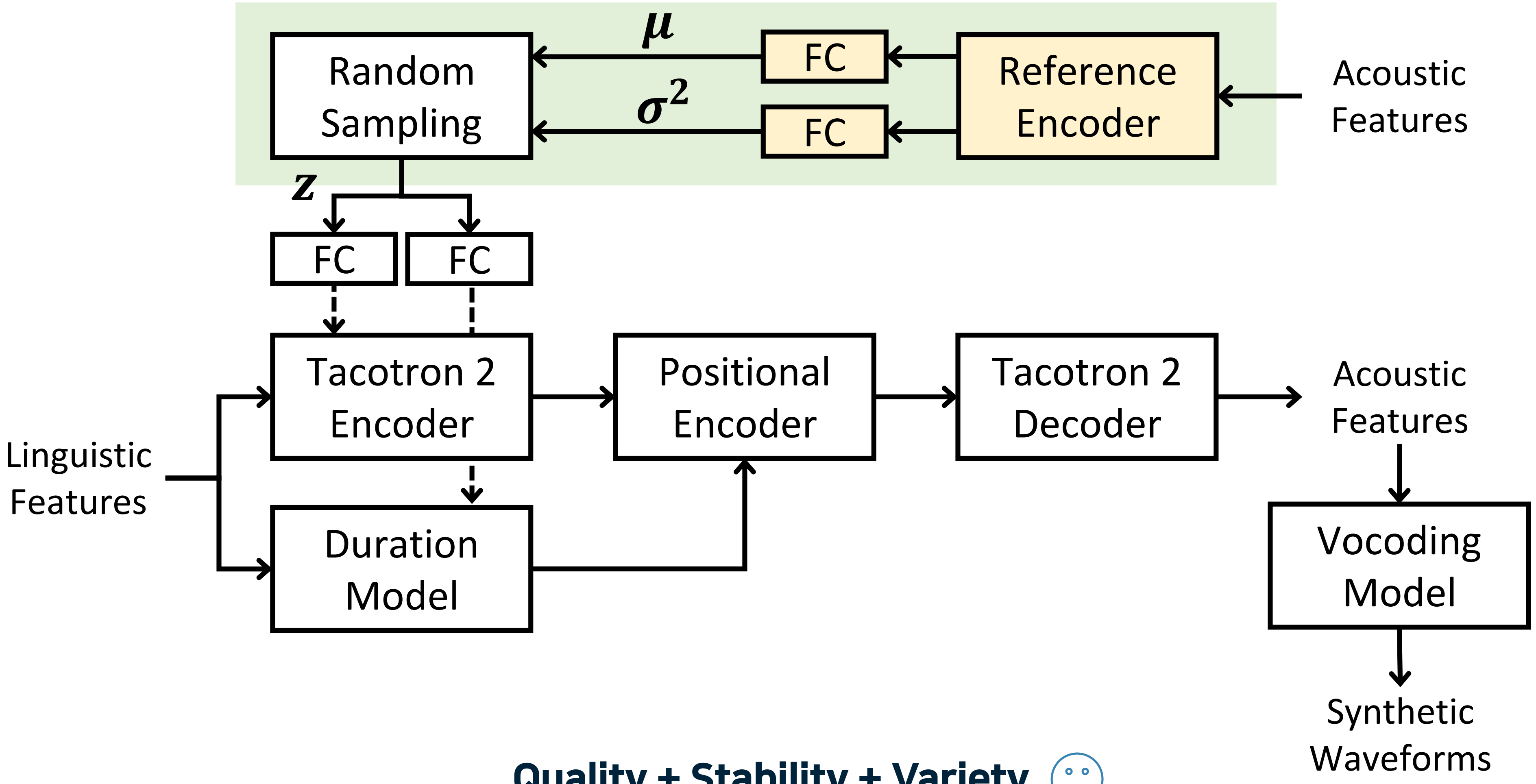
Unsupervised feature representation

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

“제가 당신의 위로가 되고 싶어요. 기분 쳐지지 말고 파이팅 하세요.”

Target TTS model

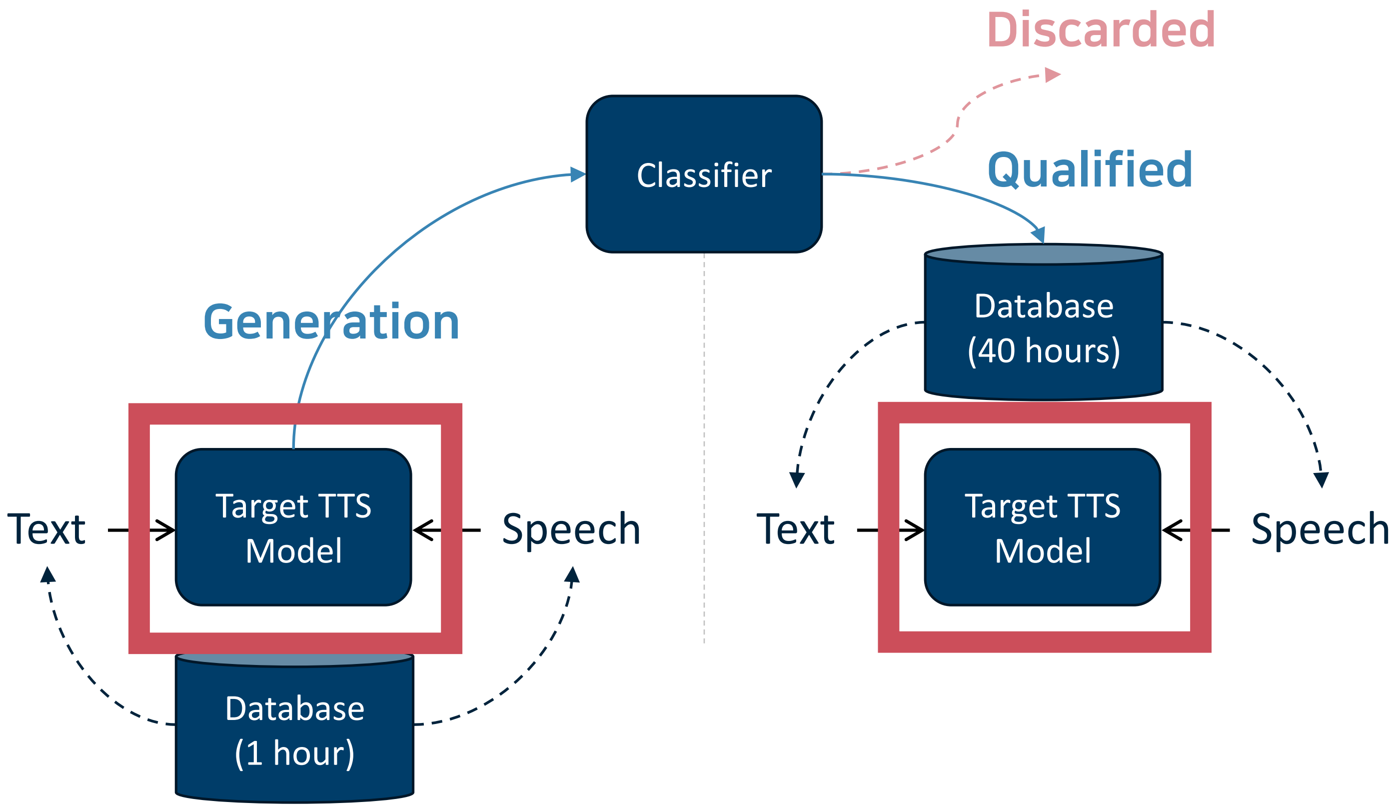
Duration informed Tacotron 2 with variational autoencoder (VAE)



Quality + Stability + Variety ☺

Target TTS model

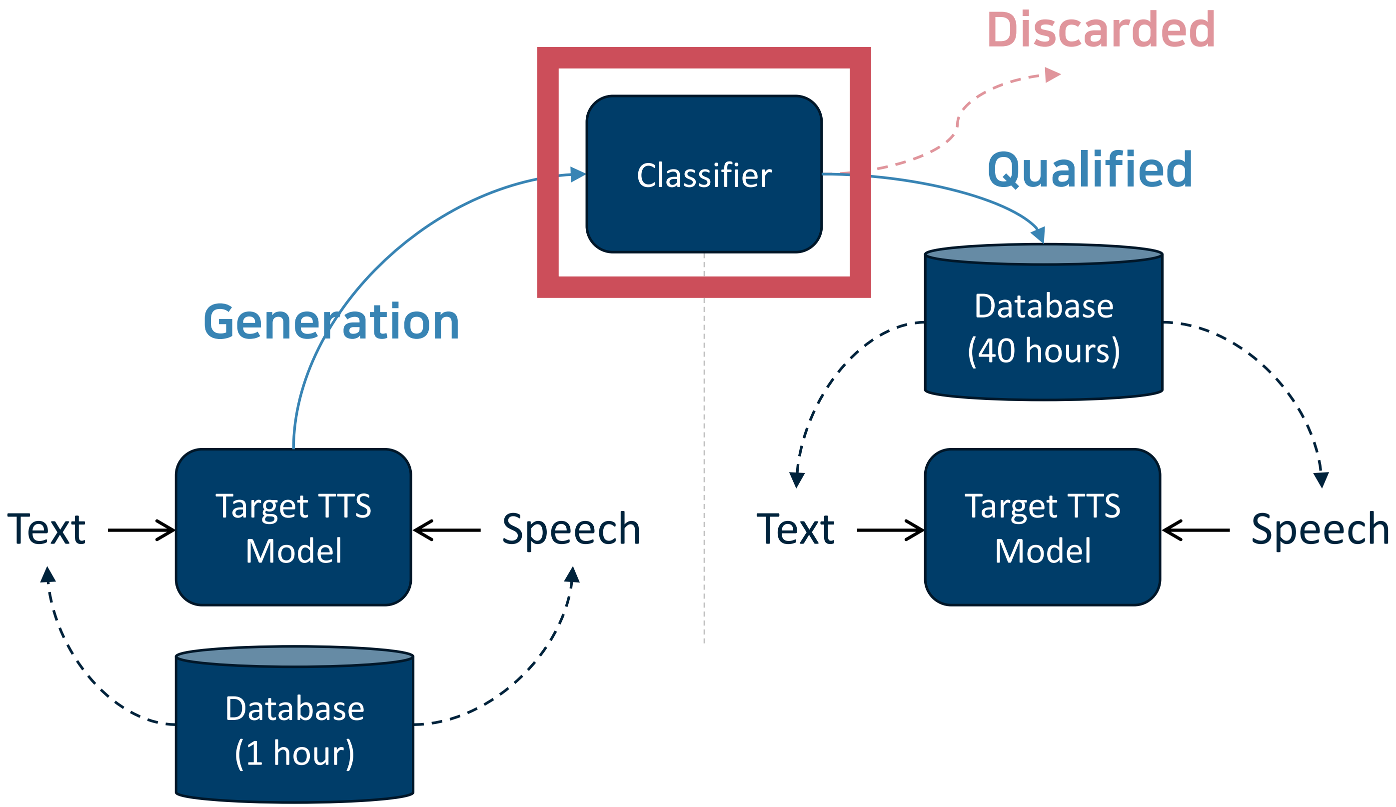
Duration informed Tacotron 2 with variational autoencoder (VAE)



It is crucial to design a well-structured TTS model to synthesize high-quality speech database

Classifier

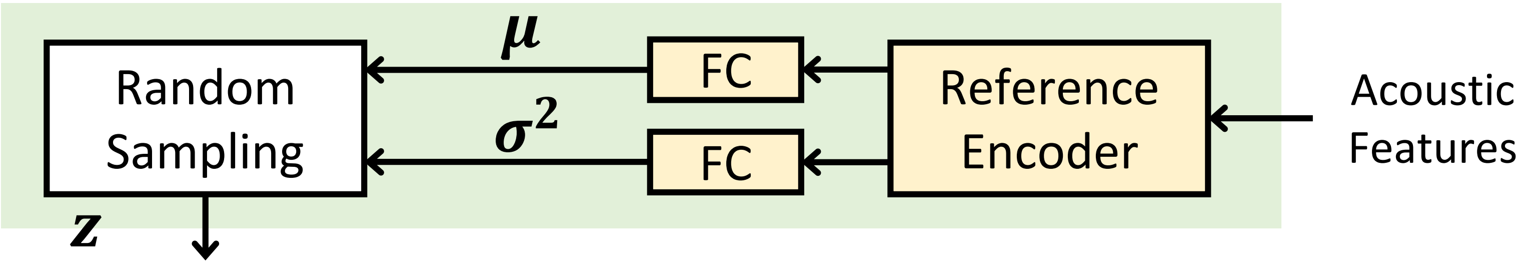
Ranking support vector machine (RankSVM) with VAE's posterior distribution

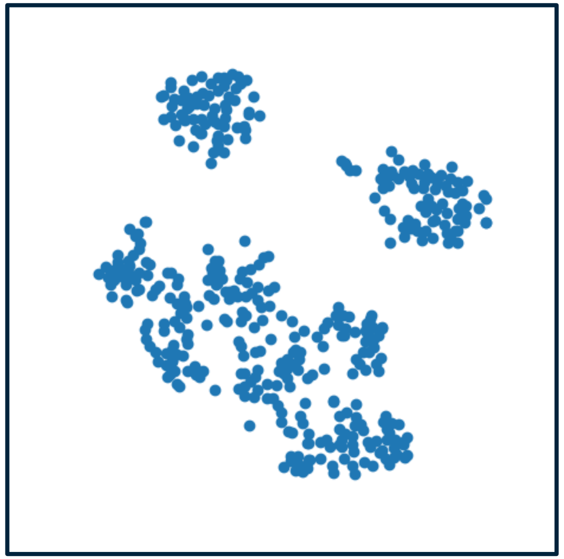


Compare and score similarities between synthetic and recorded samples

Classifier

Ranking support vector machine (RankSVM) with VAE's posterior distribution





μ Mean vector

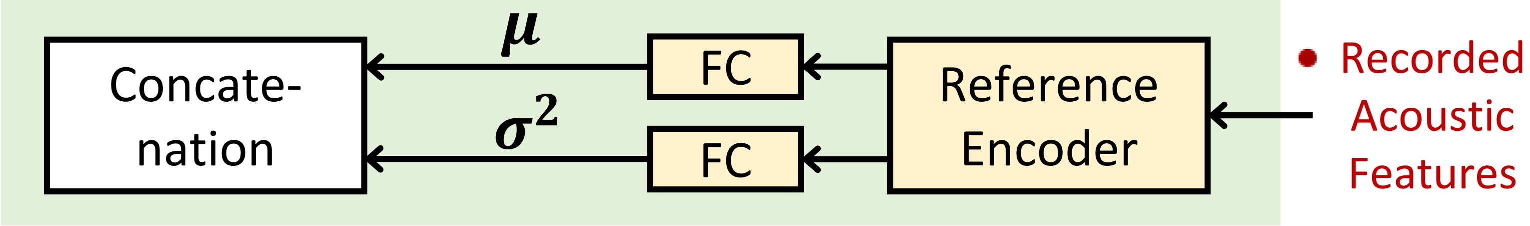
σ^2 Variance vector

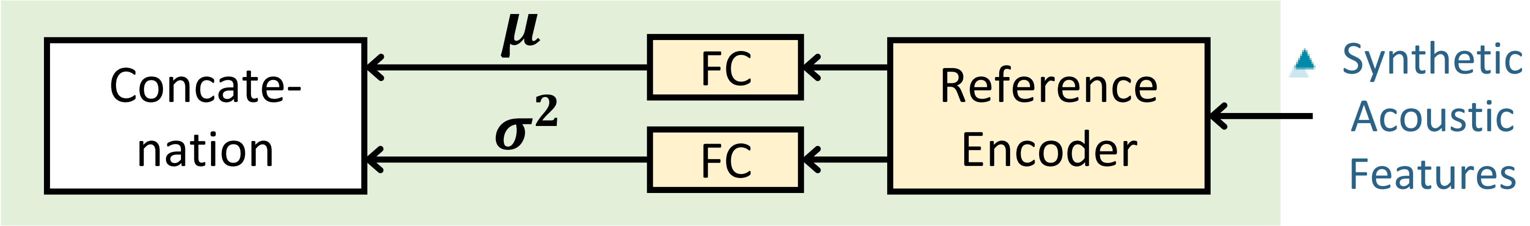
Unsupervised feature representation

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

Classifier

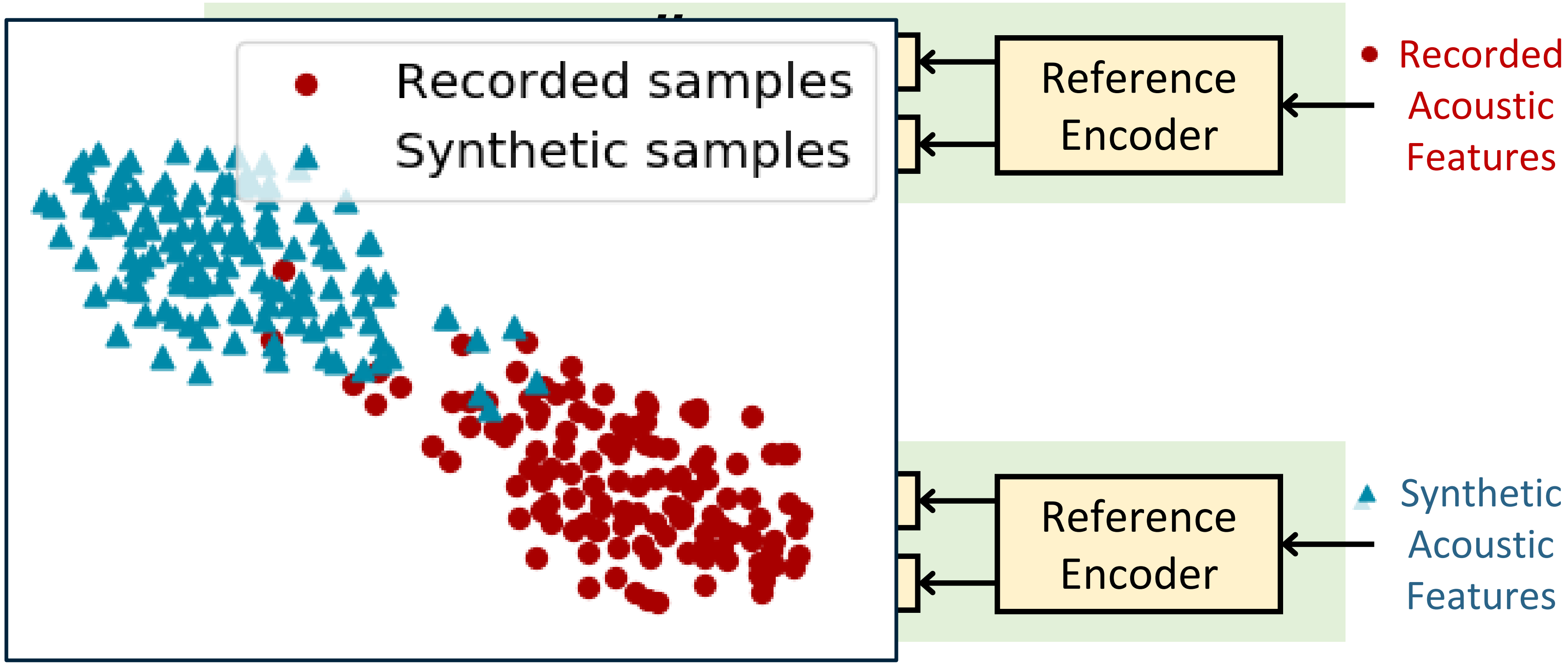
Ranking support vector machine (RankSVM) with VAE's posterior distribution





Classifier

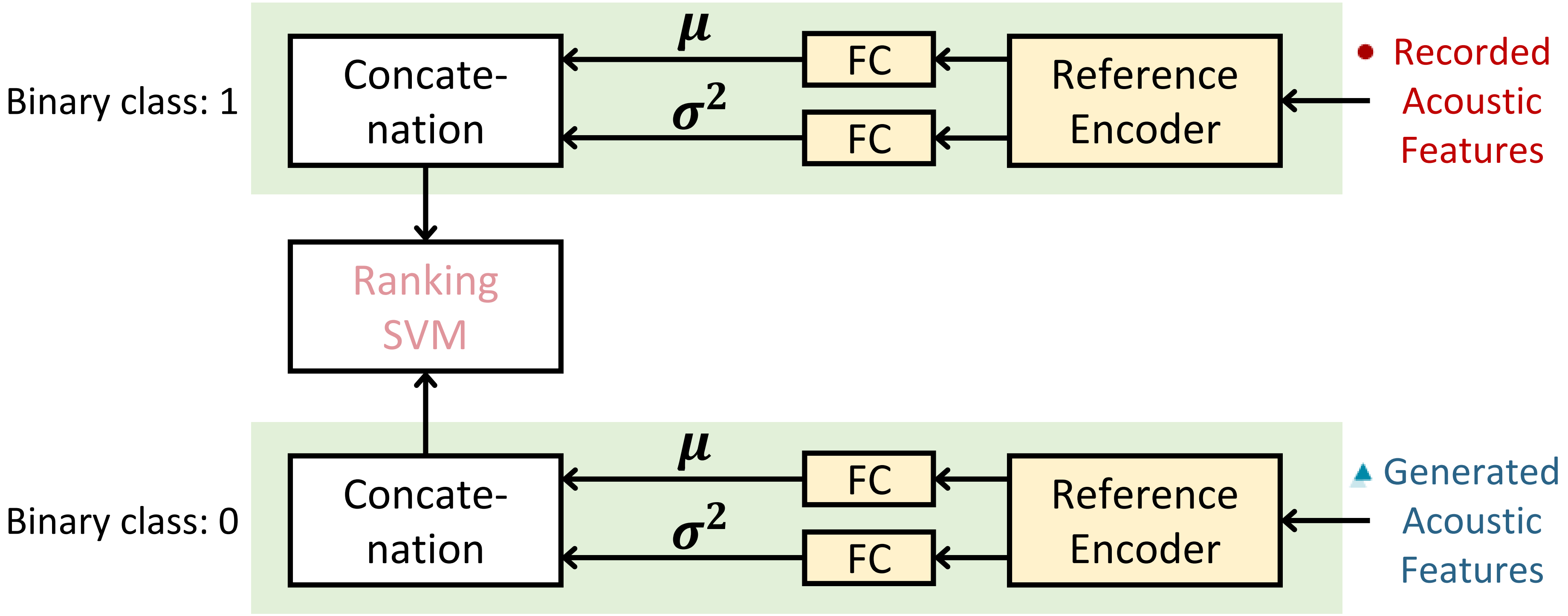
Ranking support vector machine (RankSVM) with VAE's posterior distribution



VAE can be a good feature representation between synthetic and recorded samples

Classifier

Ranking support vector machine (RankSVM) with VAE's posterior distribution



RankSVM

Score relative attributes between binary classes

Relative Attributes

Devi Parikh
Toyota Technological Institute Chicago (TTIC)
dparikh@ttic.edu

Kristen Grauman
University of Texas at Austin
grauman@cs.utexas.edu

(a) Smiling (b) ? (c) Not smiling

(d) Natural (e) ? (f) Manmade

Figure 1. Binary attributes are an artificially restrictive way to describe images. While it is clear that (a) is smiling, and (c) is not, the more informative and intuitive description for (b) is via *relative* attributes: he is smiling more than (a) but less than (c). Similarly, scene (e) is less natural than (d), but more so than (f). Our main idea is to model relative attributes via learned ranking functions, and then demonstrate their impact on novel forms of zero-shot learning and generating image descriptions.

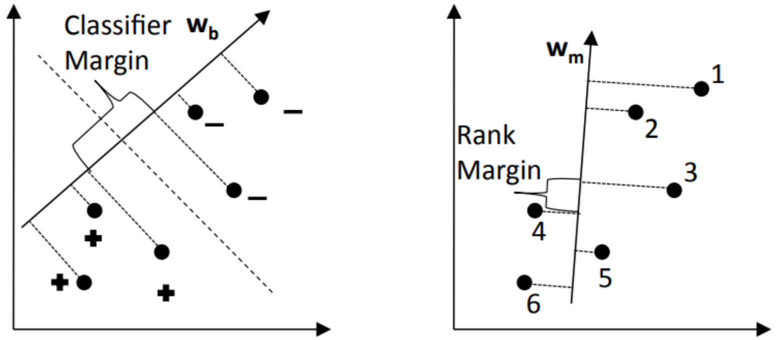


Figure 2. Distinction between learning a wide-margin ranking function (right) that enforces the desired ordering on training points (1-6) , and a wide-margin binary classifier (left) that only separates the two classes (+ and −), and does not necessarily preserve a desired ordering on the points.

RankSVM

Score relative attributes between binary classes

Relative Attributes

Devi Parikh
Toyota Technological Institute Chicago (TTIC)
dparikh@ttic.edu

Kristen Grauman
University of Texas at Austin
grauman@cs.utexas.edu

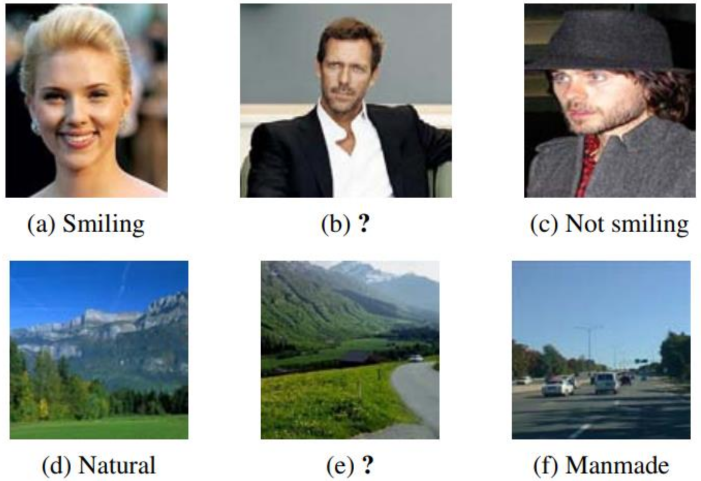
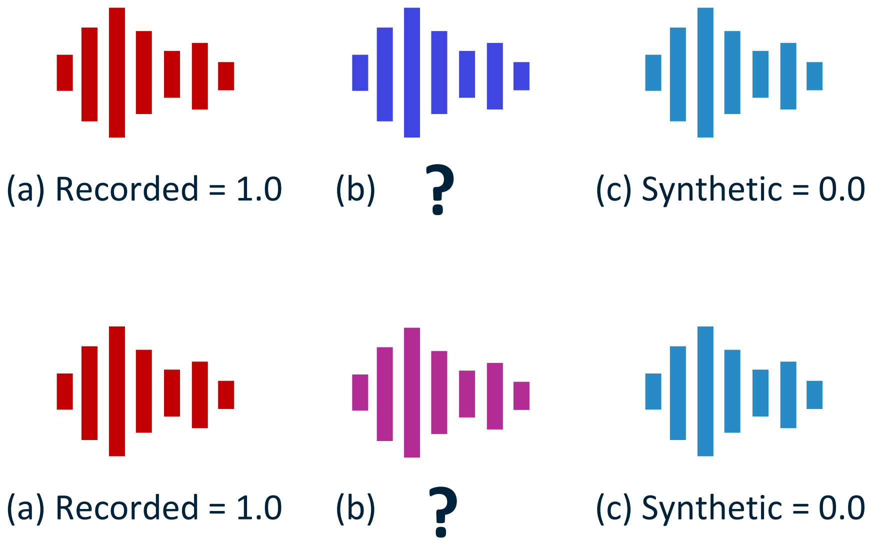


Figure 1. Binary attributes are an artificially restrictive way to describe images. While it is clear that (a) is smiling, and (c) is not, the more informative and intuitive description for (b) is via *relative* attributes: he is smiling more than (a) but less than (c). Similarly, scene (e) is less natural than (d), but more so than (f). Our main idea is to model relative attributes via learned ranking functions, and then demonstrate their impact on novel forms of zero-shot learning and generating image descriptions.



RankSVM

Score relative attributes between binary classes

Relative Attributes

Devi Parikh
Toyota Technological Institute Chicago (TTIC)
dparikh@ttic.edu

Kristen Grauman
University of Texas at Austin
grauman@cs.utexas.edu

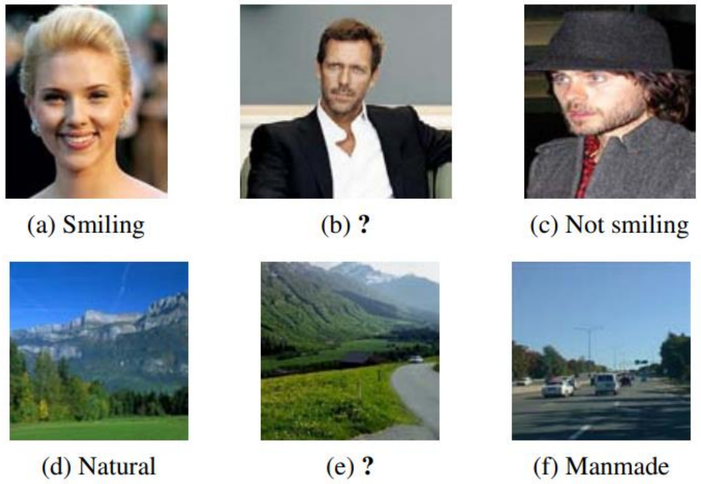
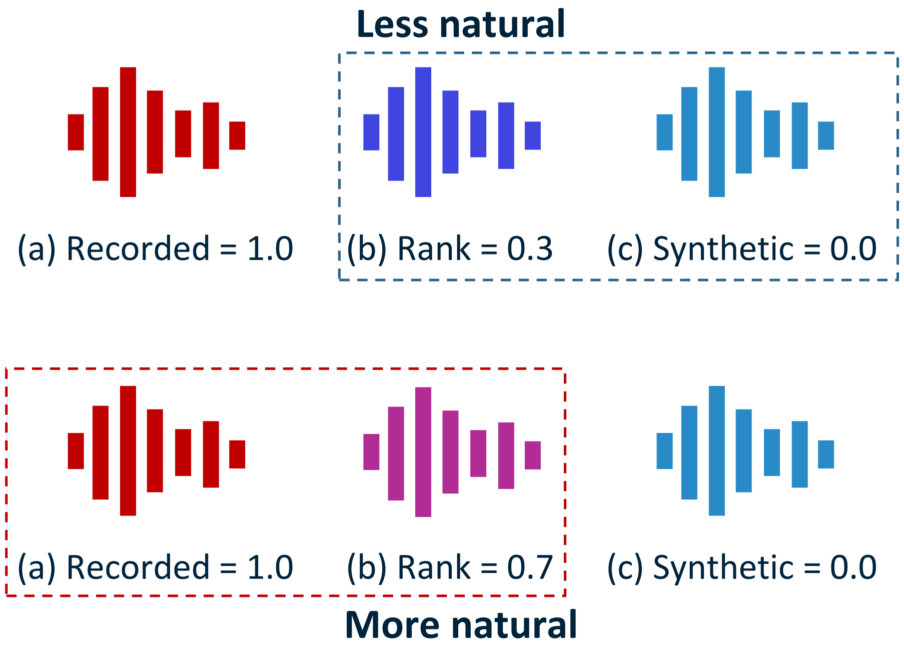
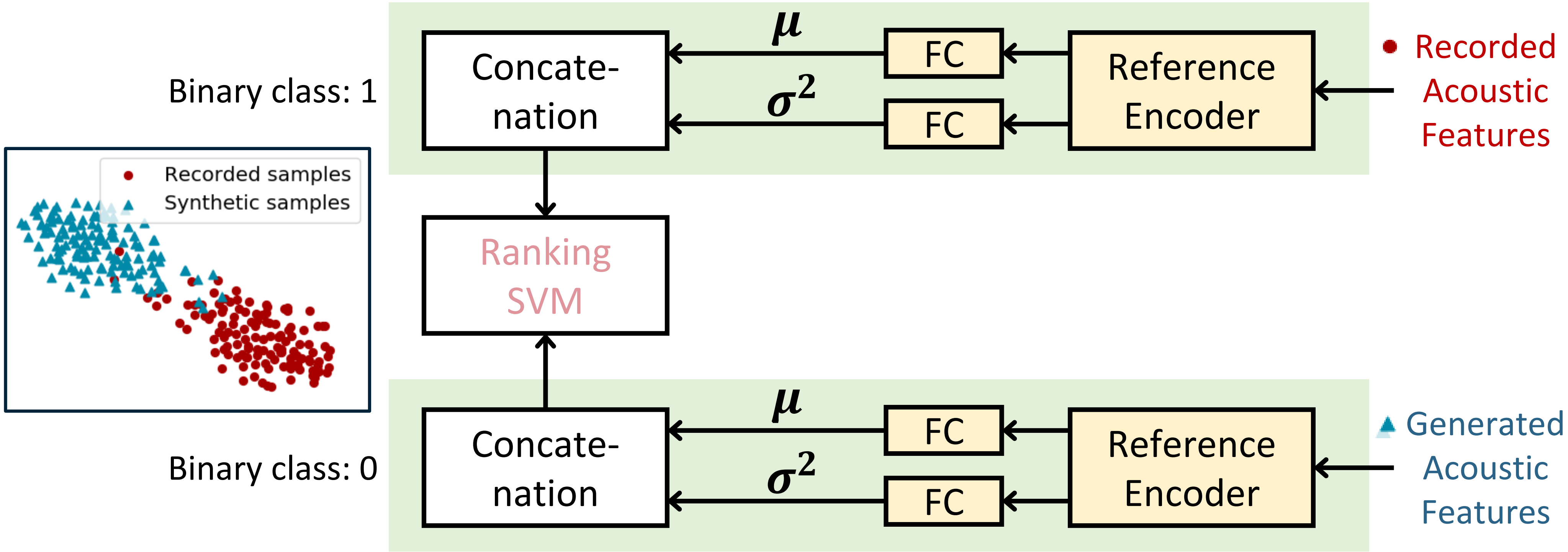


Figure 1. Binary attributes are an artificially restrictive way to describe images. While it is clear that (a) is smiling, and (c) is not, the more informative and intuitive description for (b) is via *relative* attributes: he is smiling more than (a) but less than (c). Similarly, scene (e) is less natural than (d), but more so than (f). Our main idea is to model relative attributes via learned ranking functions, and then demonstrate their impact on novel forms of zero-shot learning and generating image descriptions.



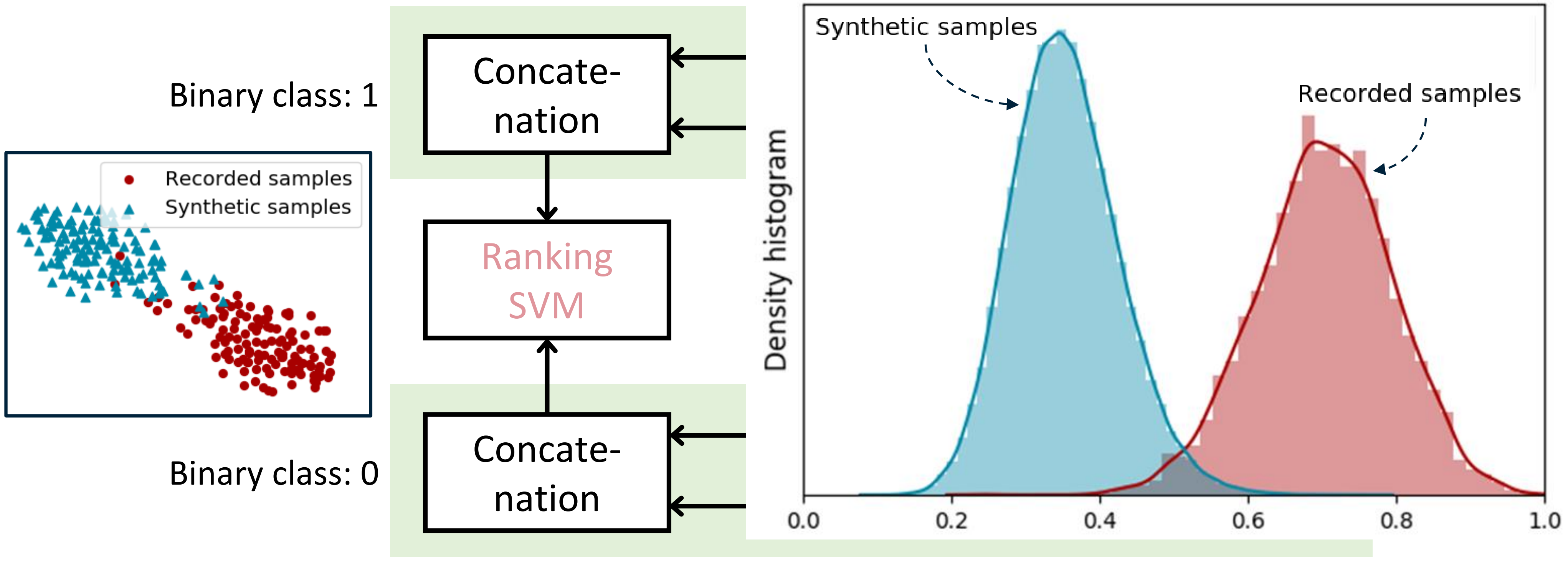
Classifier

Ranking support vector machine (RankSVM) with VAE's posterior distribution



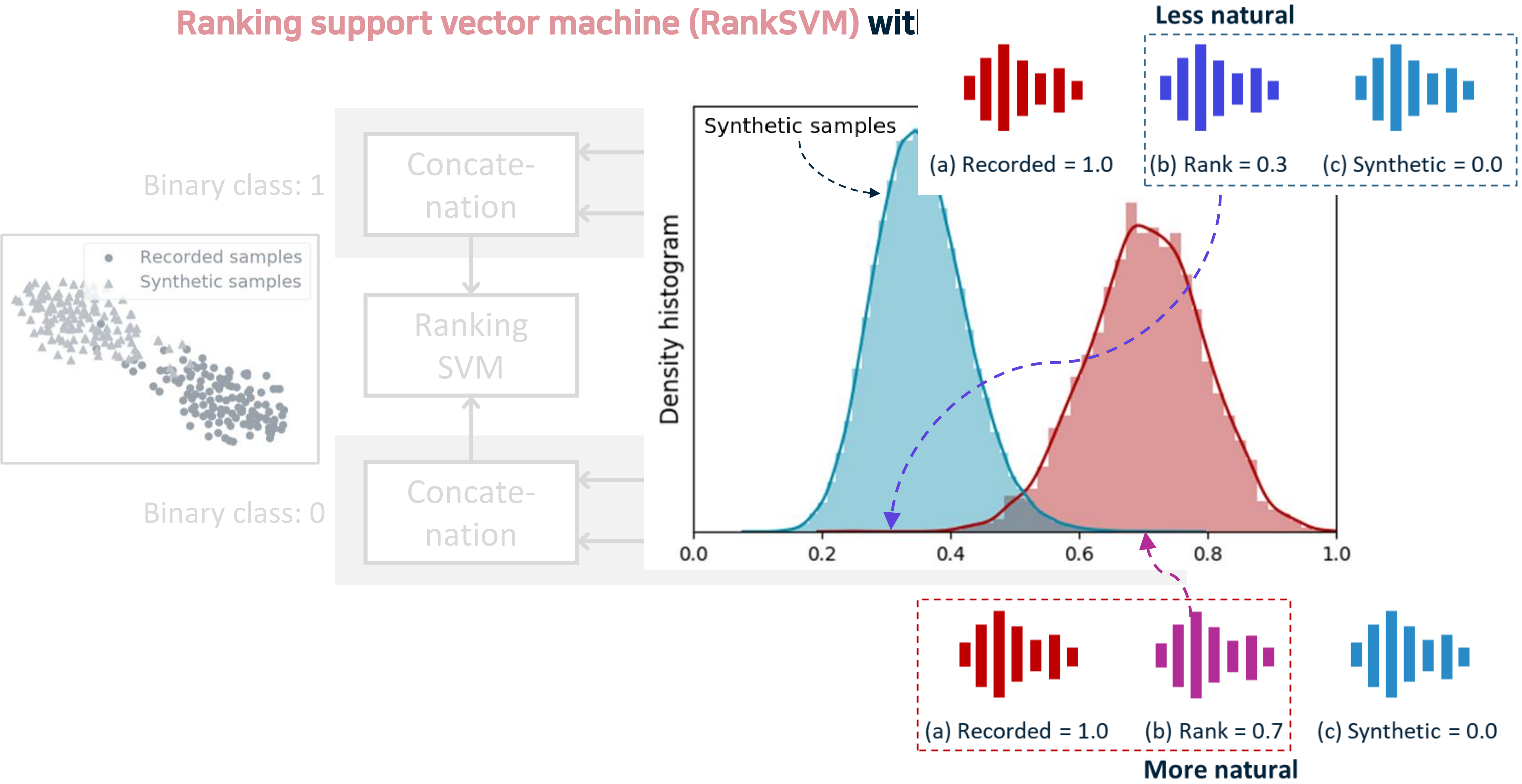
Classifier

Ranking support vector machine (RankSVM) with VAE's posterior distribution



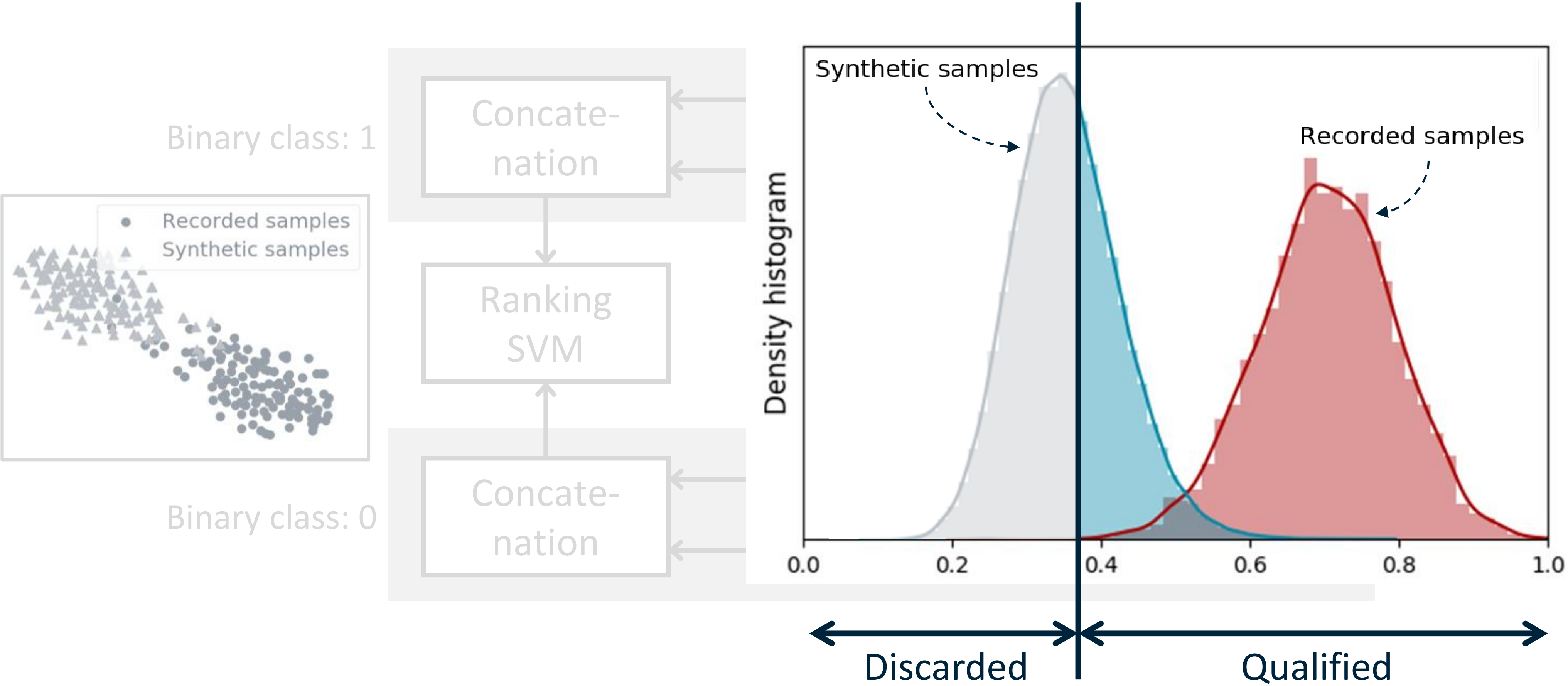
Classifier

Ranking support vector machine (RankSVM) wit



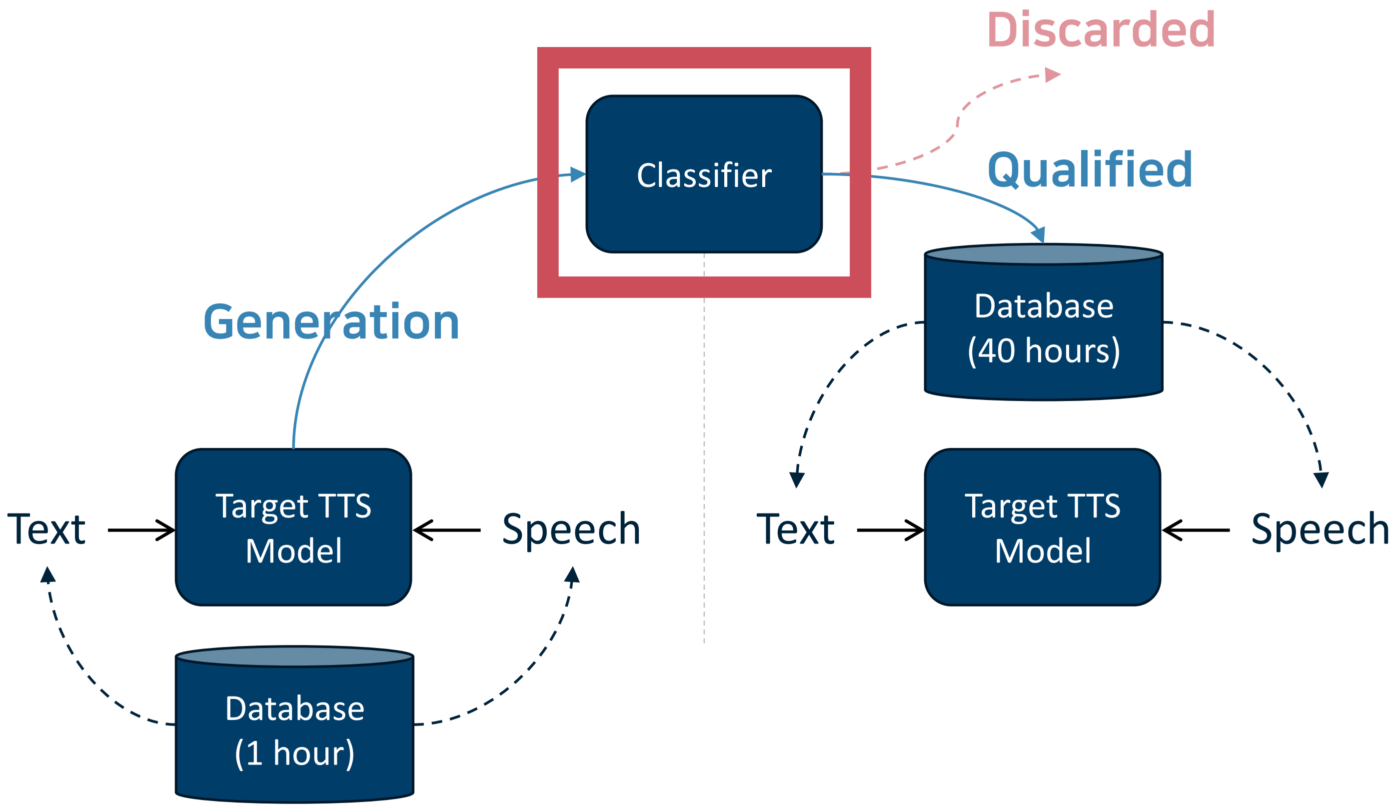
Classifier

Ranking support vector machine (RankSVM) with VAE's posterior distribution



Classifier

Ranking support vector machine (RankSVM) with VAE's posterior distribution



Classifier

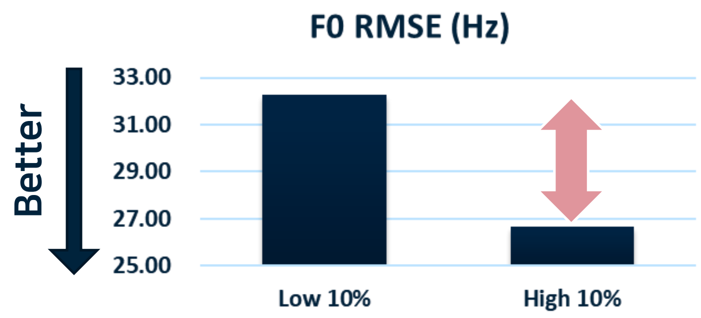
Questions

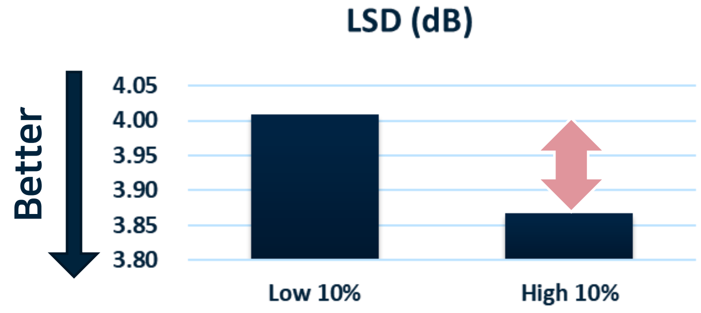
①: Ranking score vs Speech quality ?

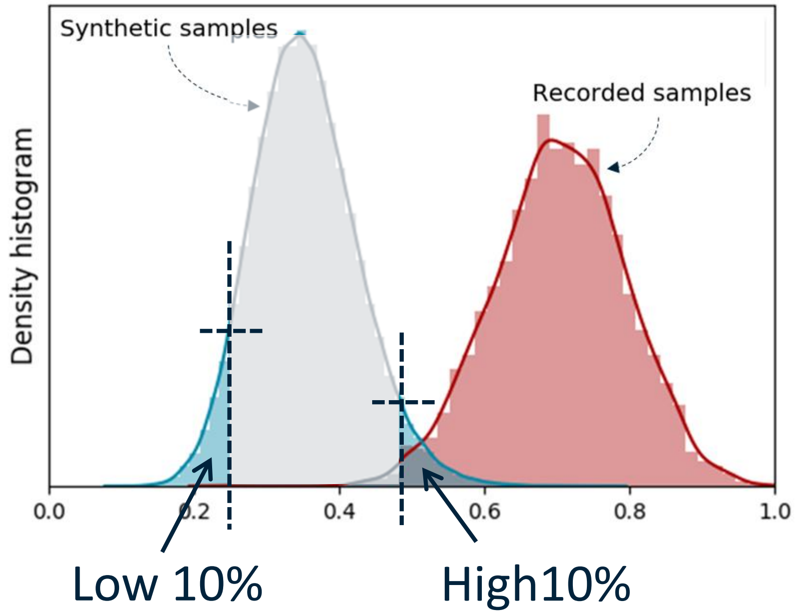
②: How to determine decision criteria ?

Verification

Ranking support vector machine (RankSVM) with VAE's posterior distribution





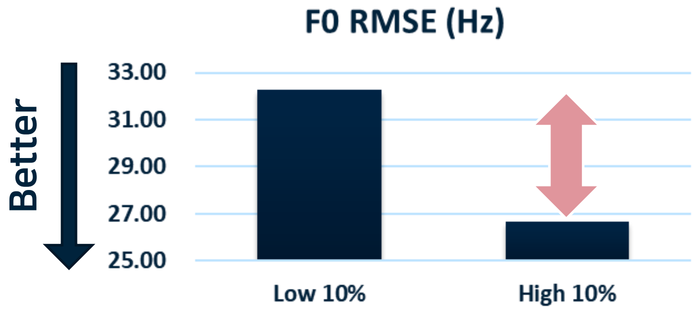


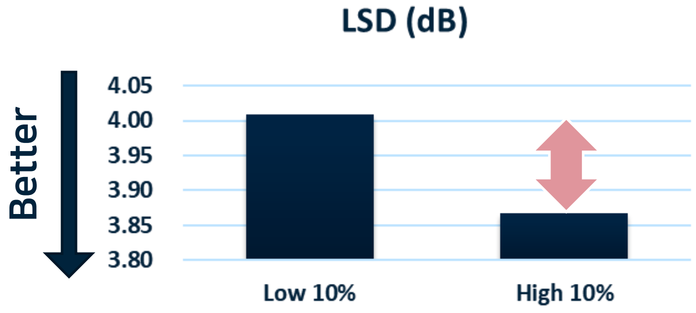
1. Ranking score vs Speech quality
2. How to determine decision criteria

Verification

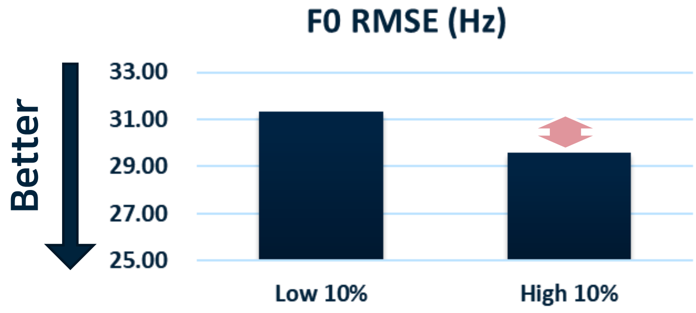
Ranking support vector machine (RankSVM) with VAE's posterior distribution

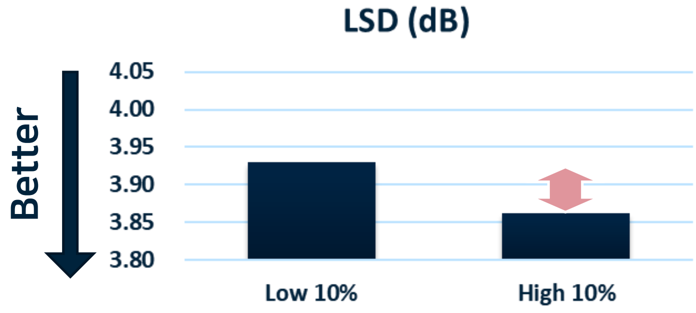
VAE features





OpenSMILE features





1. Ranking score vs Speech quality
2. How to determine decision criteria

openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor

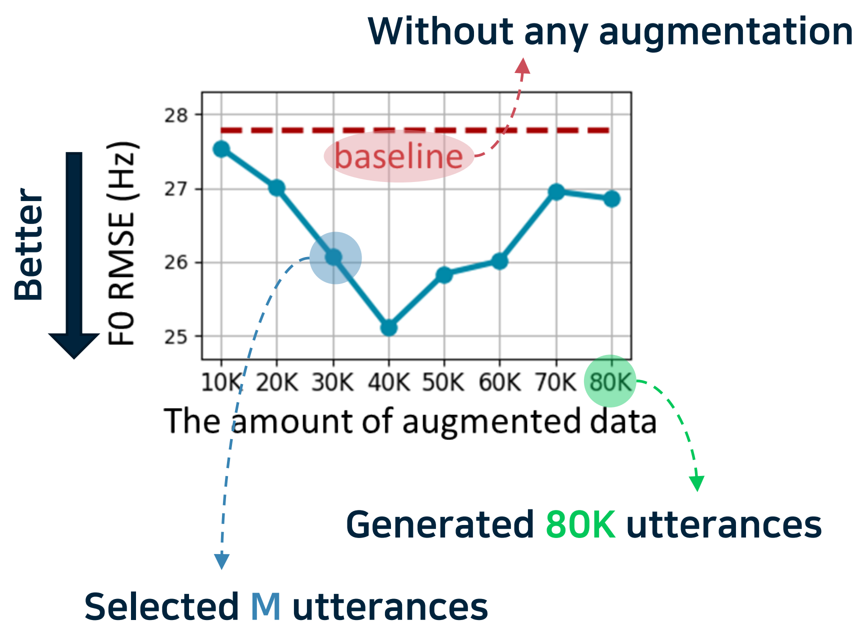
Florian Eyben
Institute for Human-Machine Communication
Technische Universität München
80290 München, Germany
eyben@tum.de

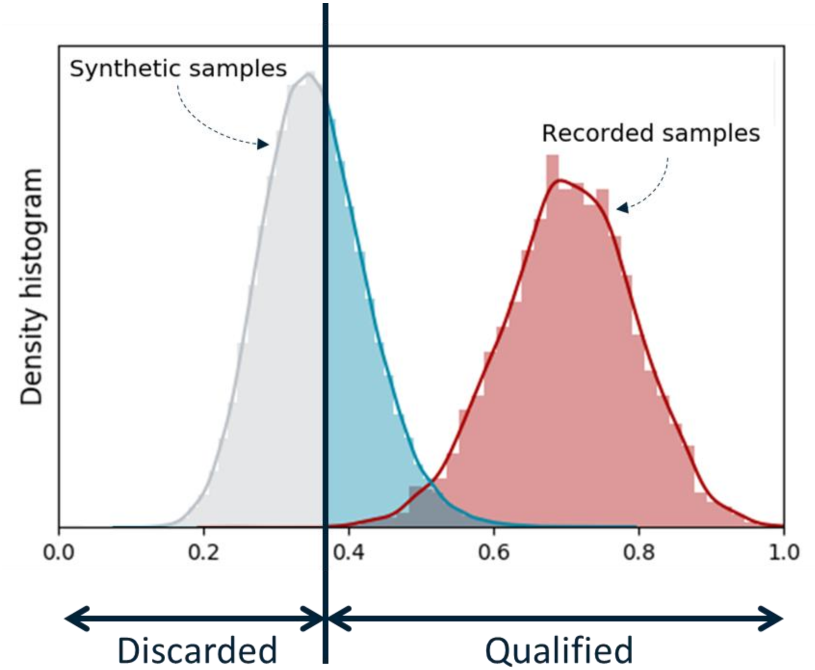
Martin Wöllmer
Institute for Human-Machine Communication
Technische Universität München
80290 München, Germany
woellmer@tum.de

Björn Schuller
Institute for Human-Machine Communication
Technische Universität München
80290 München, Germany
schuller@tum.de

Verification

Ranking support vector machine (RankSVM) with VAE's posterior distribution

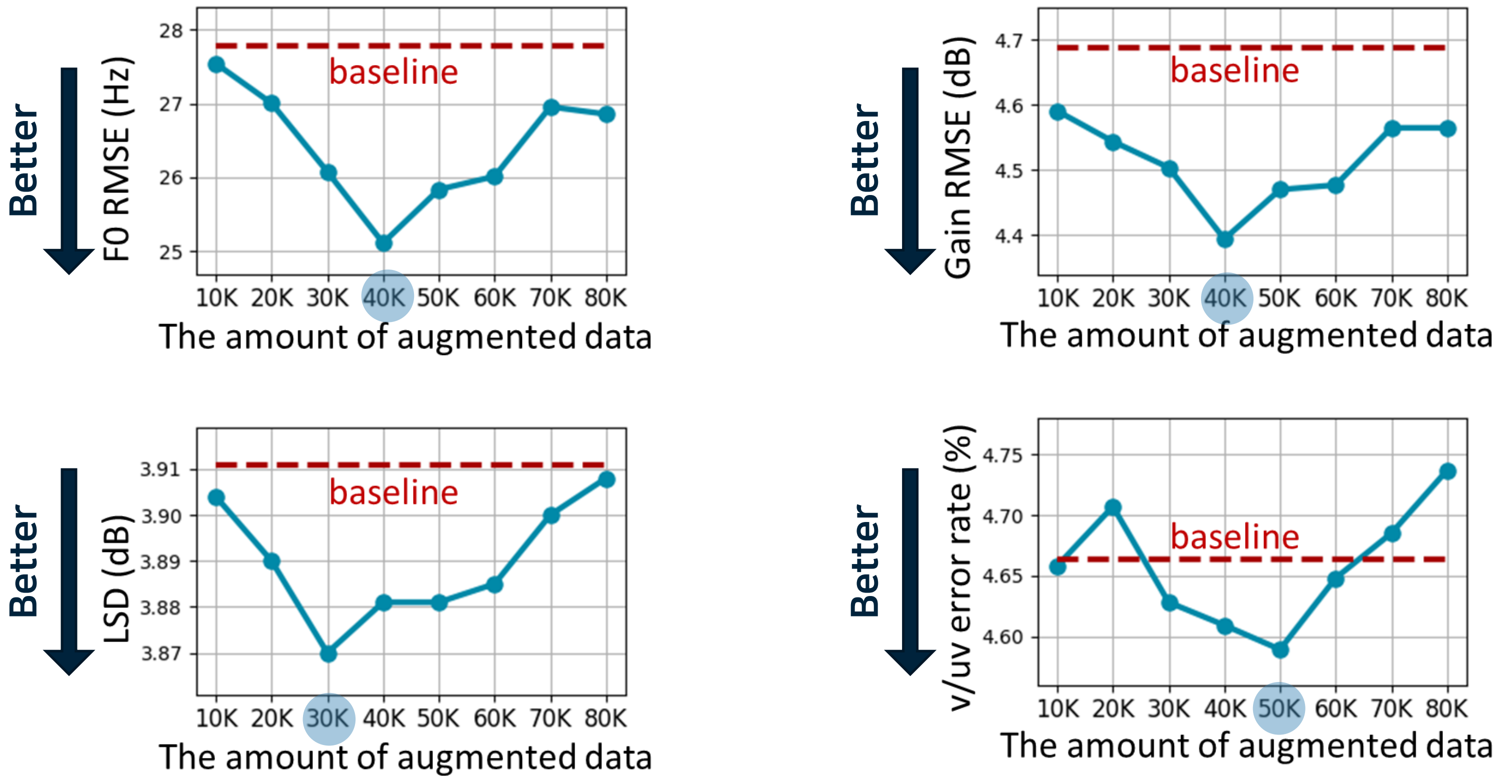




1. **Ranking score vs Speech quality**
2. **How to determine decision criteria**

Verification

Ranking support vector machine (RankSVM) with VAE's posterior distribution

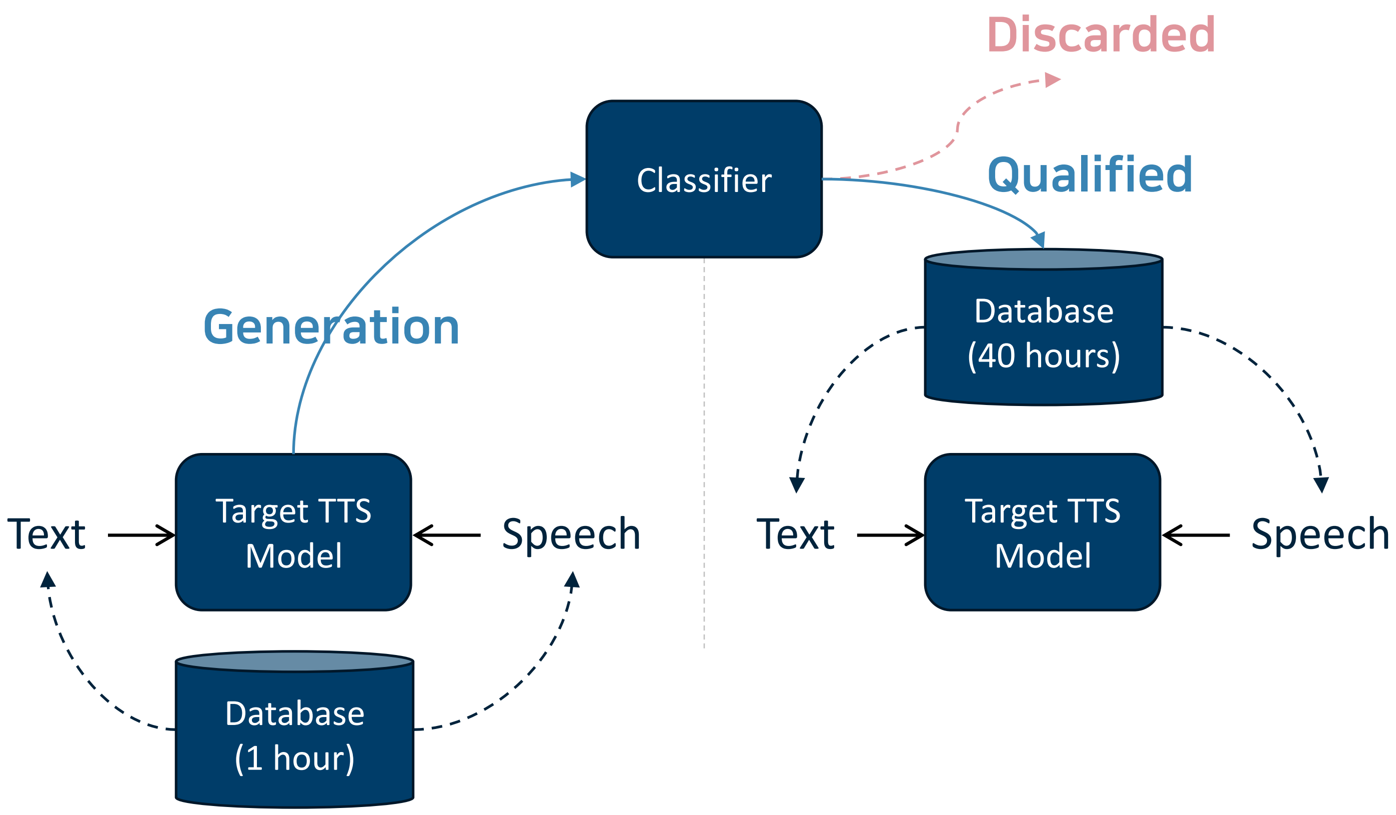


1. **Ranking score vs Speech quality**
2. **How to determine decision criteria** → 40K would be the best

Data-selective TTS augmentation

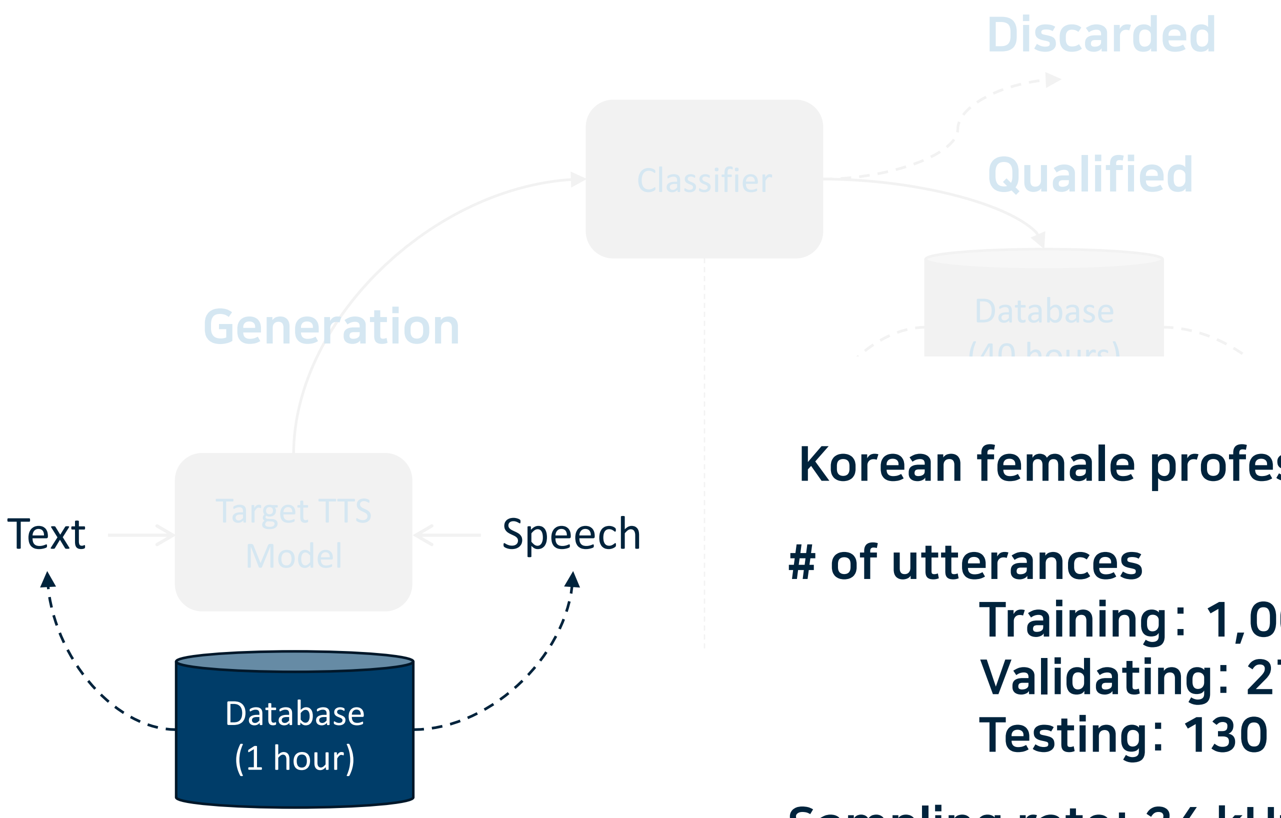
Evaluations

Experiment


Discarded
Classifier
Qualified
Generation
Database
(40 hours)
Text
Target TTS
Model
Speech
Text
Target TTS
Model
Speech
Database
(1 hour)

Experiment

Database



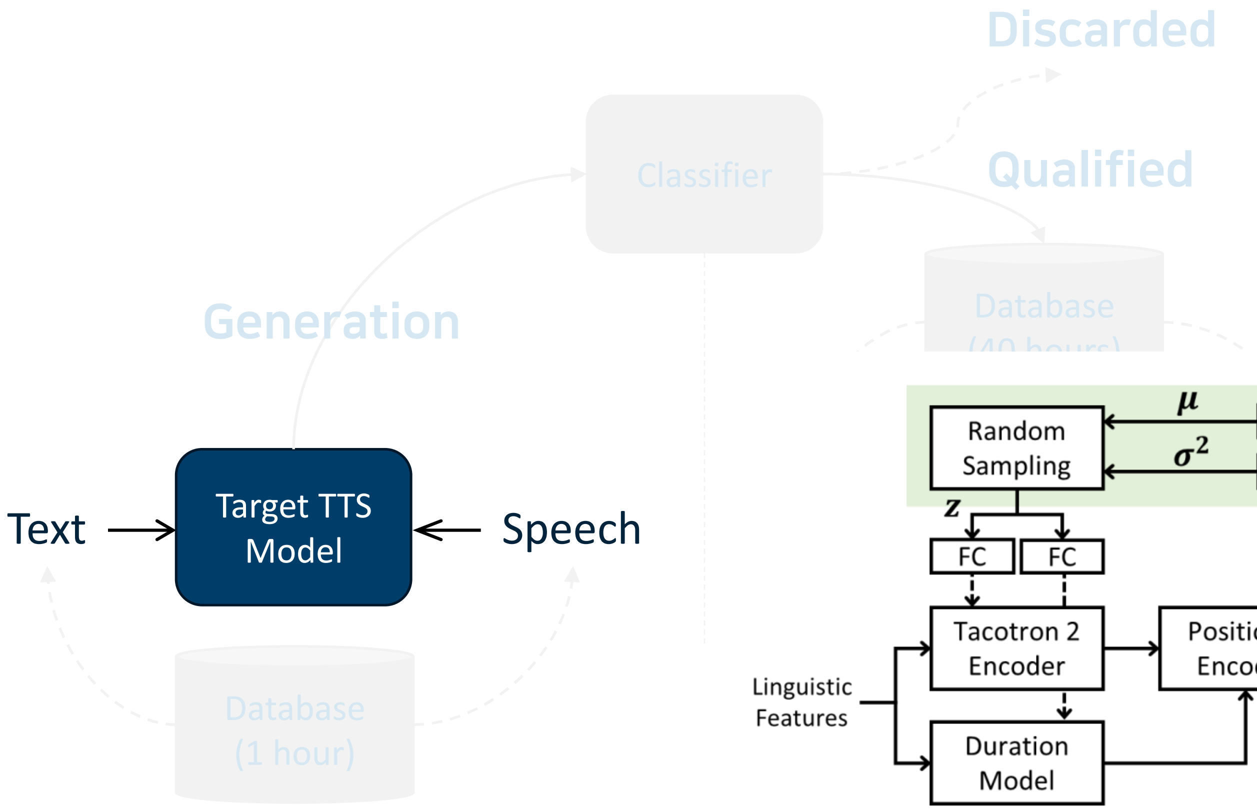
Korean female professional speaker ☺

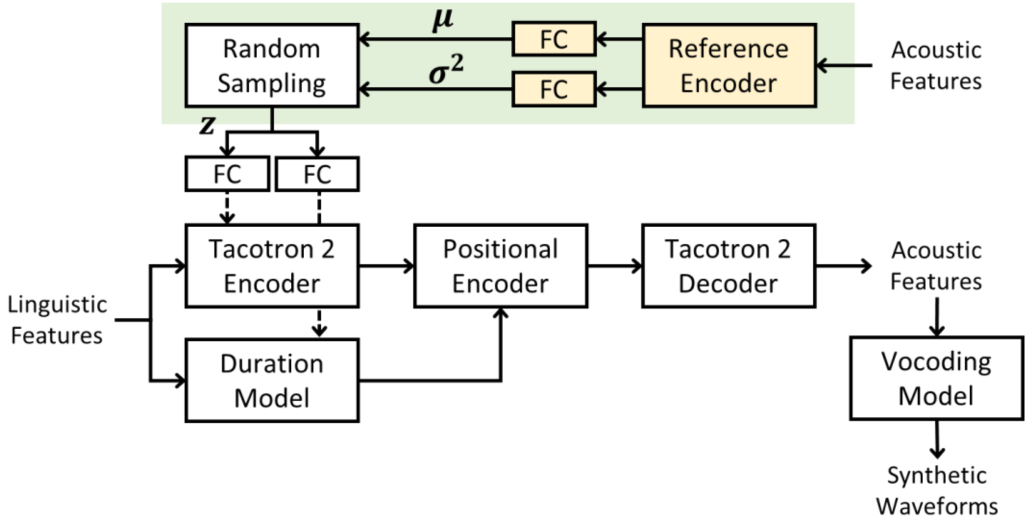
# of utterances
Training: 1,000
Validating: 270
Testing: 130

Sampling rate: 24 kHz

Experiment

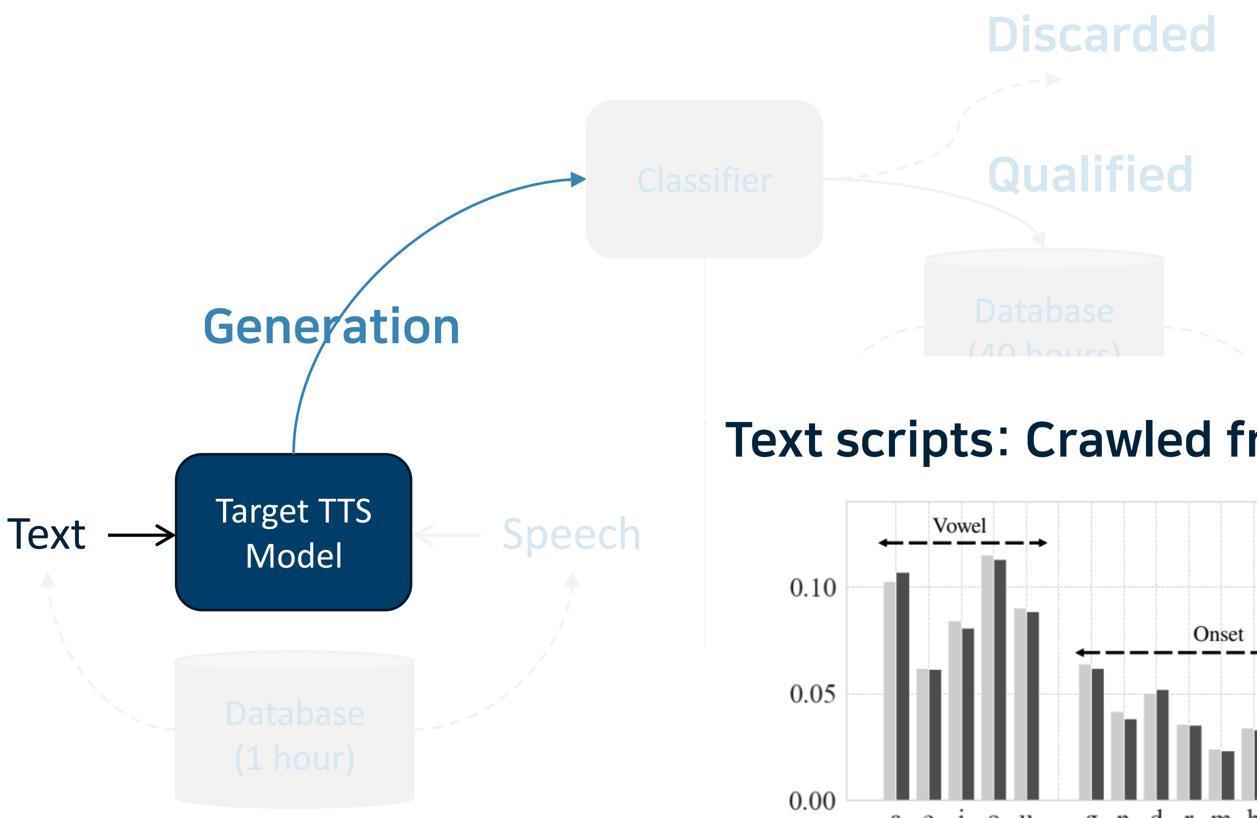
TTS model: Duration informed Tacotron2 with VAE



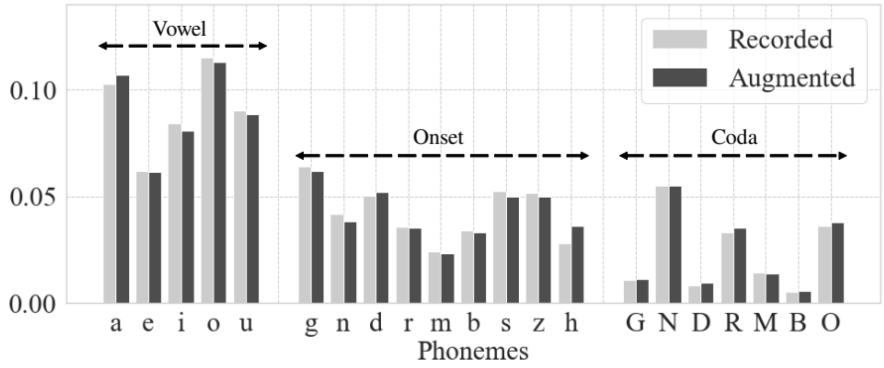


Experiment

TTS-based data augmentation



Text scripts: Crawled from news articles

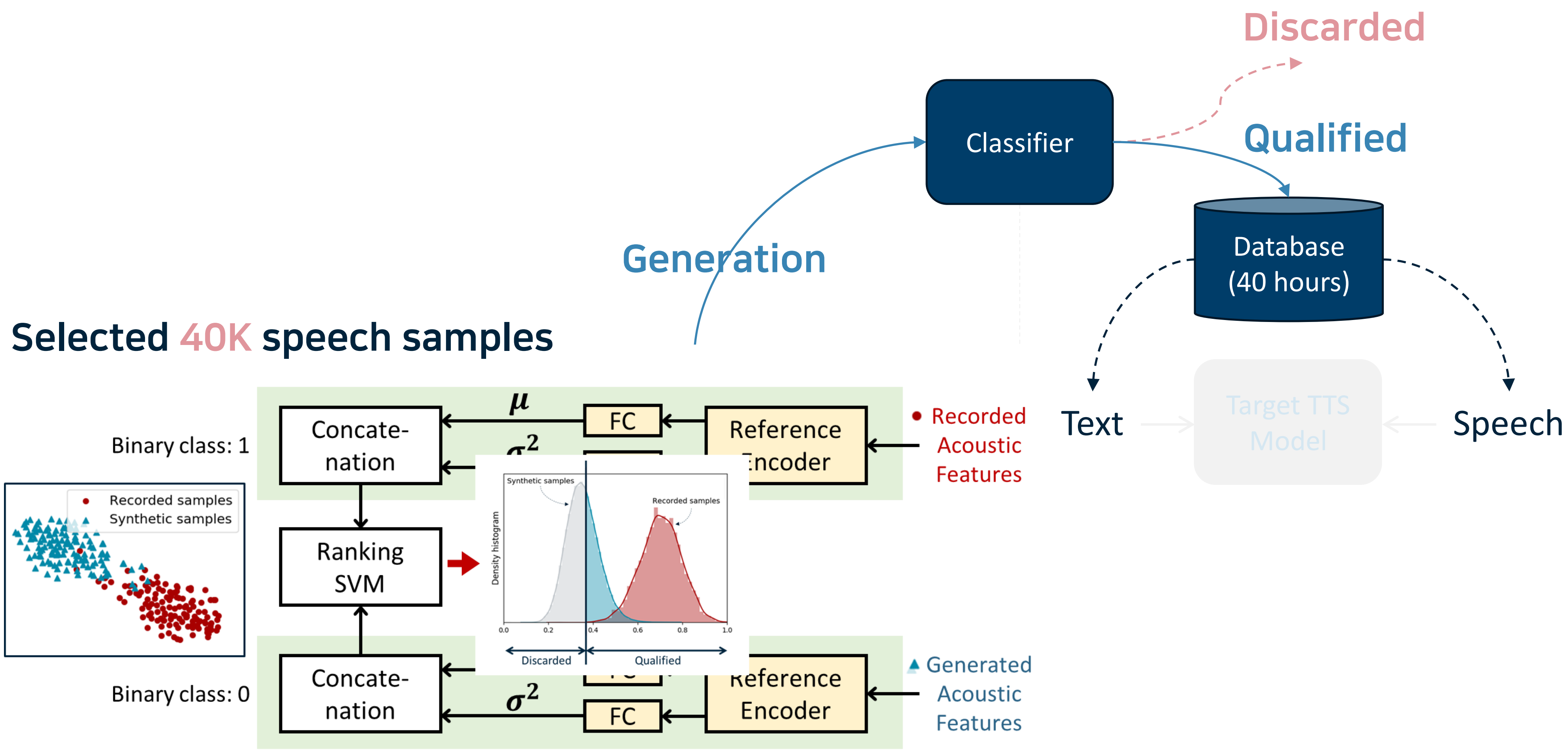


2020 Engineering day: 가짜 목소리 DB로 고품질 음성합성기를 만들어보자 (HDTS 황민제님)

Generated 80K speech samples

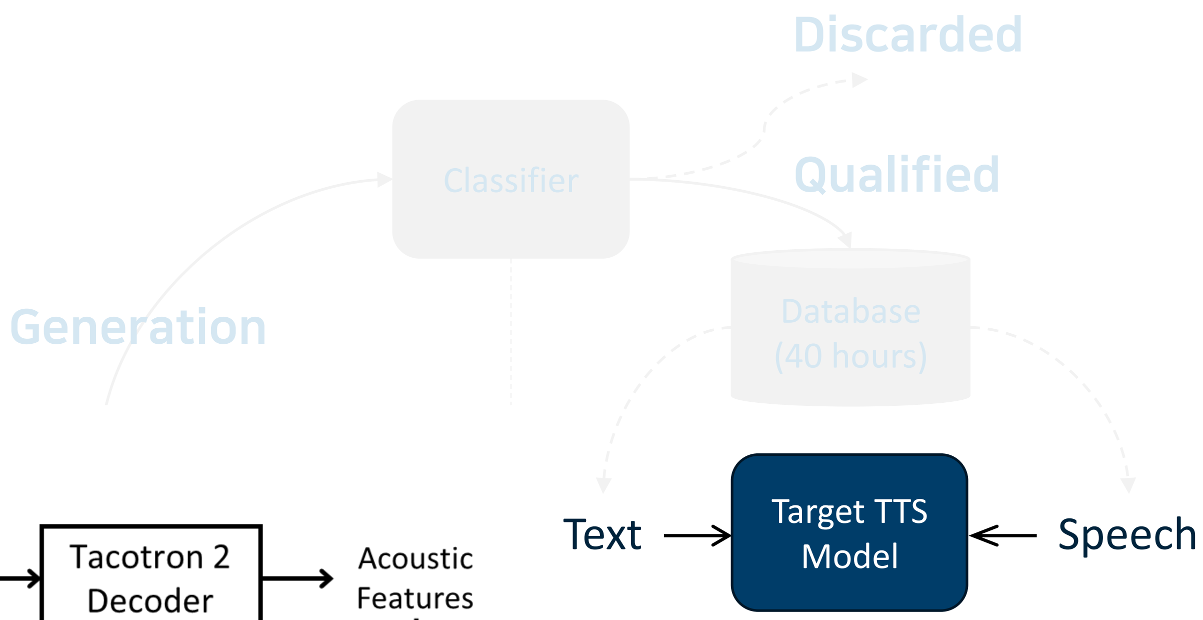
Experiment

RankSVM-based data selection

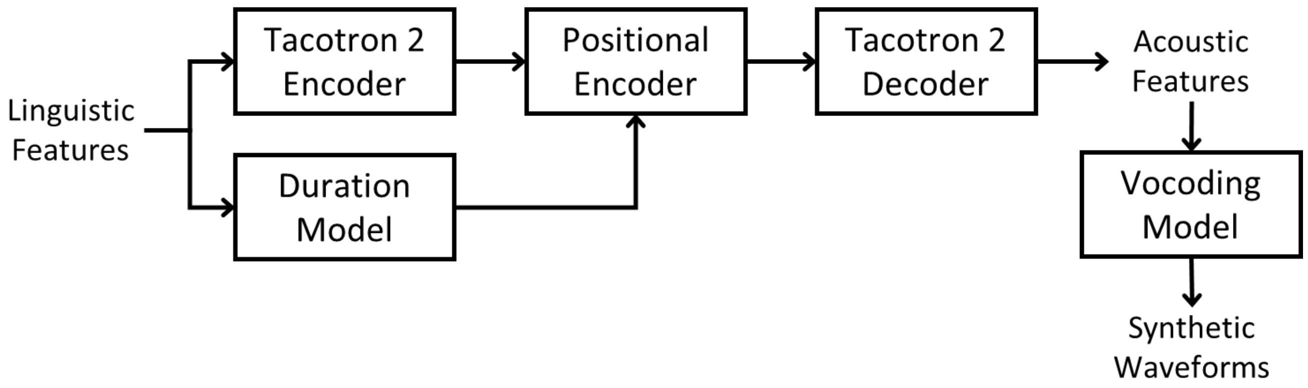


Experiment

TTS retraining with large-scale synthetic corpora

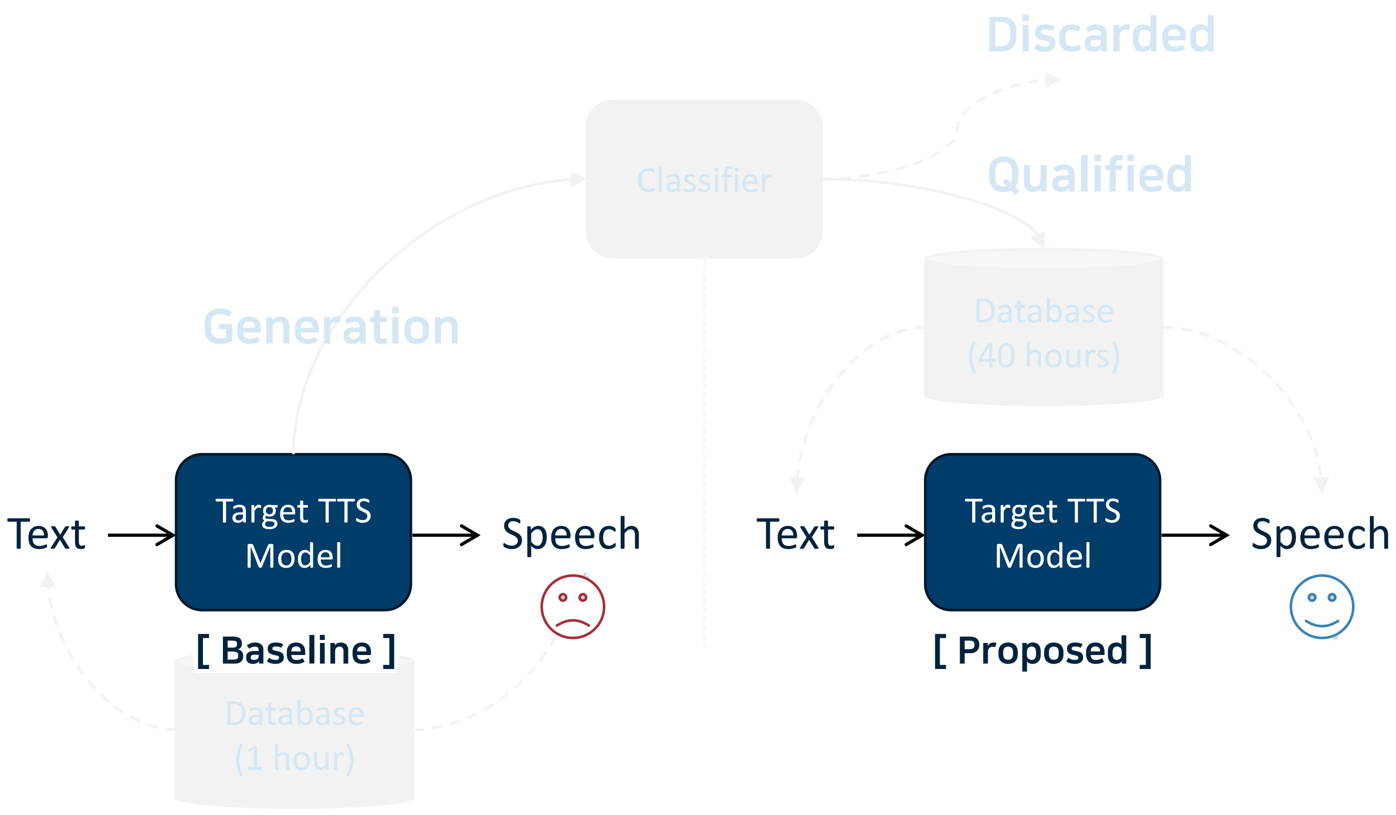


Retrained target TTS model



Experiment

Subjective mean opinion score (MOS) tests



“찜나라는 아구나라와 풍년해물탕 아귀찜나라 못난이 아구나라 등이 있네요. 자세한 결과는 네이버 클로바 앱에서 확인하세요.”

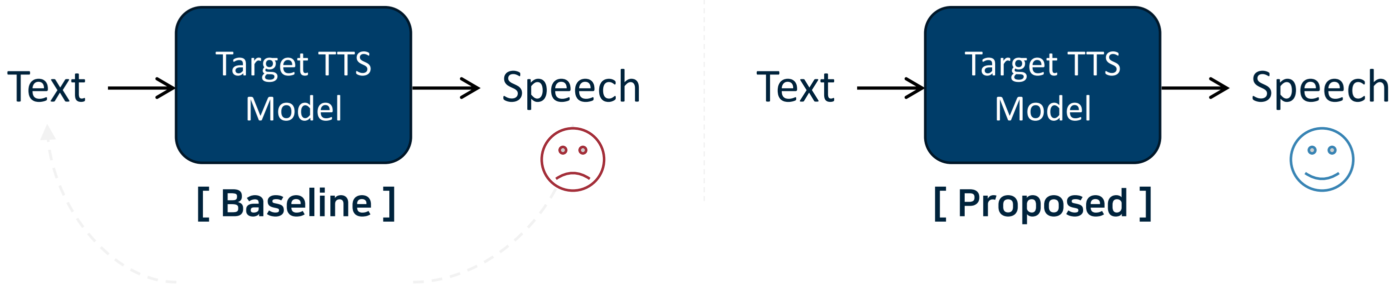
Experiment

Subjective mean opinion score (MOS) tests

20 listeners / 20 samples for each system

5-point naturalness responses

- 5: Excellent
- 4: Good
- 3: Fair
- 2: Poor
- 1: Bad

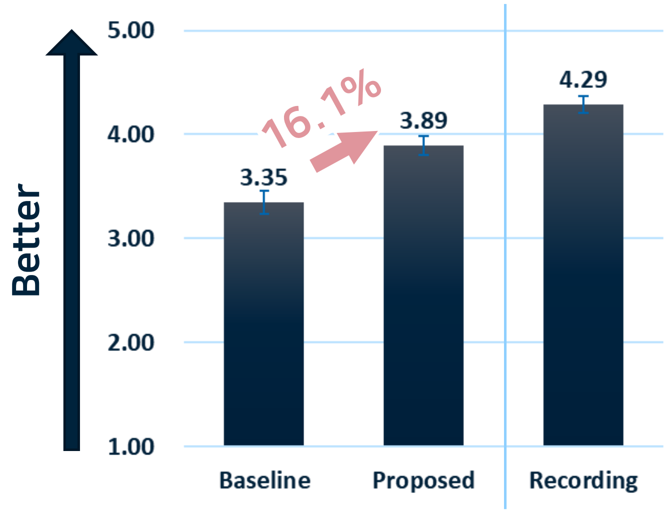


Experiment

Subjective mean opinion score (MOS) tests

20 listeners / 20 samples for each system

5-point naturalness responses
- 5: Excellent
- 4: Good
- 3: Fair
- 2: Poor
- 1: Bad

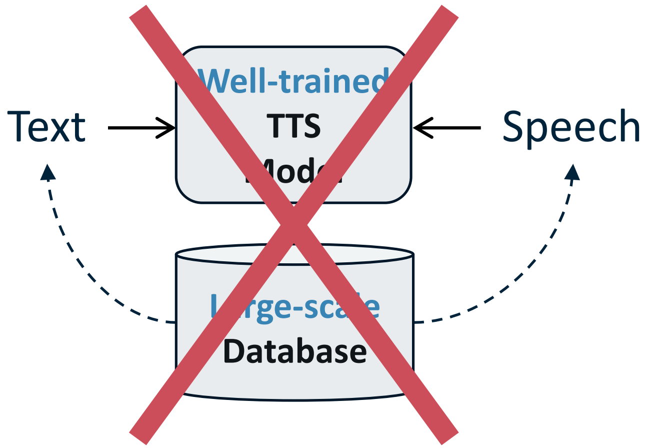


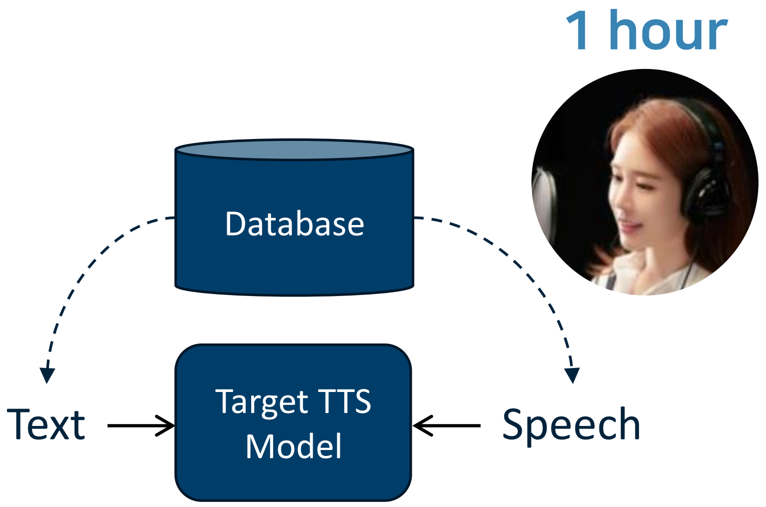
Data-selective TTS augmentation

Summary

Summary

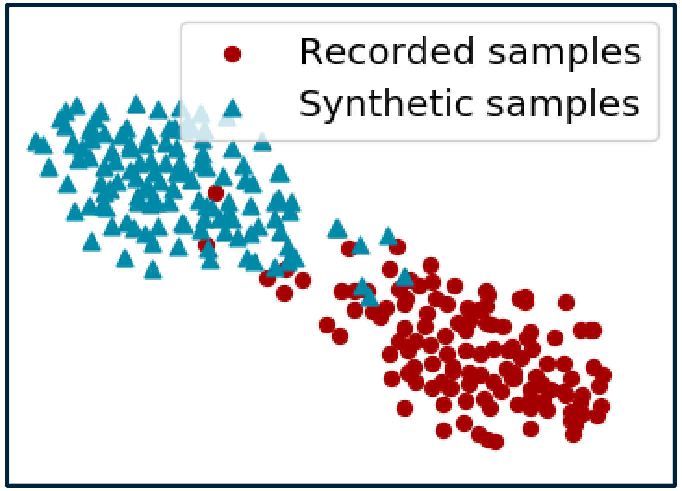
How to design TTS model with limited amount of training data ?



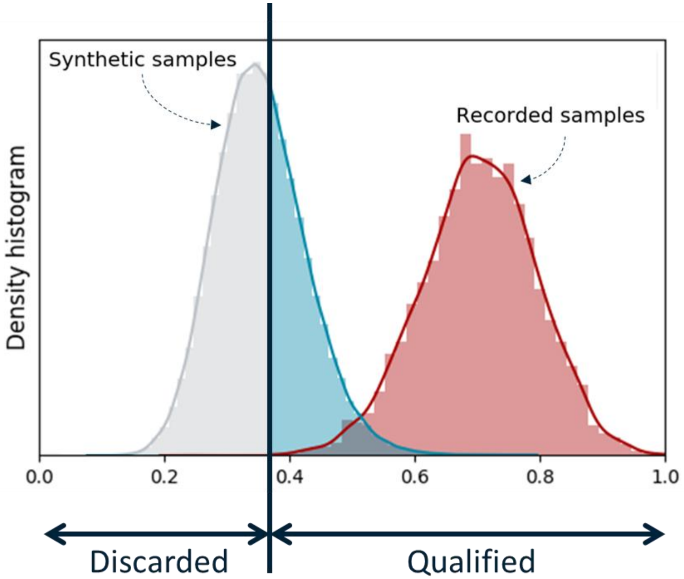


Summary

Ranking support vector machine (RankSVM) with VAE's posterior distribution



[VAE representation]

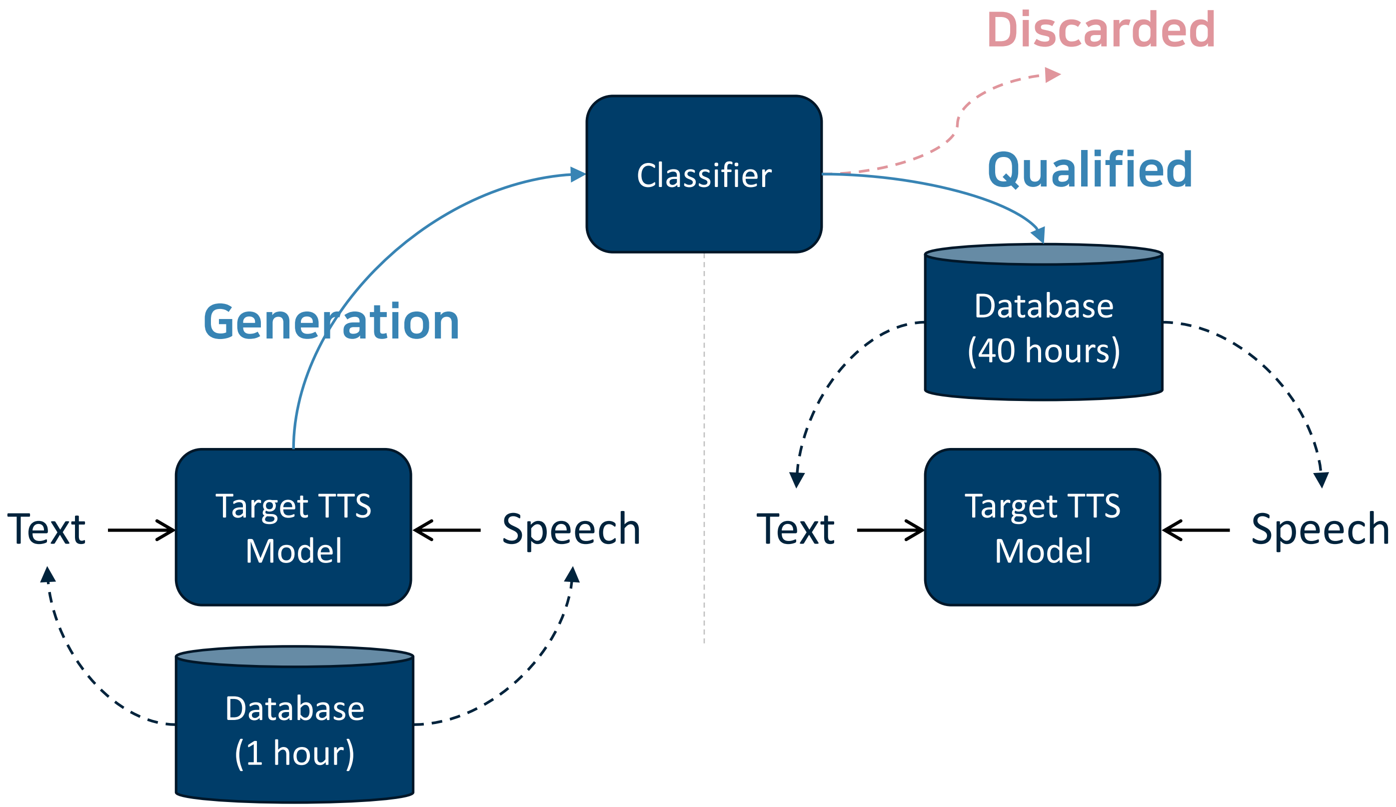


[RankSVM classifier]

We proposed a TTS-driven data-selective augmentation technique. From the large-scale synthetic corpora, a RankSVM with VAE's posterior distribution determined the originality that represents how the acoustic characteristics of the generated speech was similar to those of the natural recordings. By selectively including the synthetic data with the recorded one, the performance of the retrained TTS system has been improved significantly

Summary

Ranking support vector machine (RankSVM) with VAE's posterior distribution



We proposed a TTS-driven data-selective augmentation technique. From the large-scale synthetic corpora, a RankSVM with VAE's posterior distribution determined the originality that represents how the acoustic characteristics of the generated speech was similar to those of the natural recordings. By selectively including the synthetic data with the recorded one, the performance of the retrained TTS system has been improved significantly

Thanks! End of Documents.