Investigating on Incorporating Pretrained and Learnable Speaker Representations for Multi-Speaker Multi-Style Text-to-Speech

Chung-Ming Chien, Jheng-Hao Lin*, Chien-yu Huang*, Po-chun Hsu* and Hung-yi Lee

College of Electrical Engineering and Computer Science, National Taiwan University

國立臺灣大學
National Taiwan University

ICASSP 2021
TORONTO
Canada
June 6-11, 2021
Metro Toronto Convention Centre

IEEE ICASSP 2021

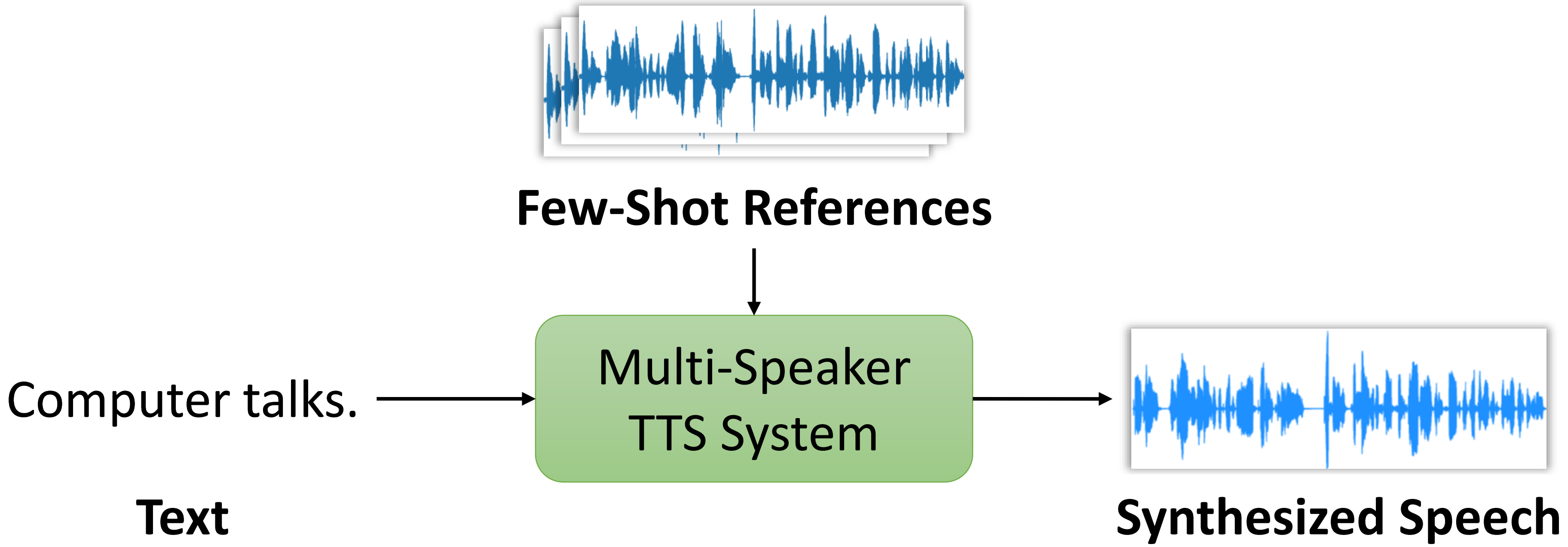
* These authors contributed equally.

Outline

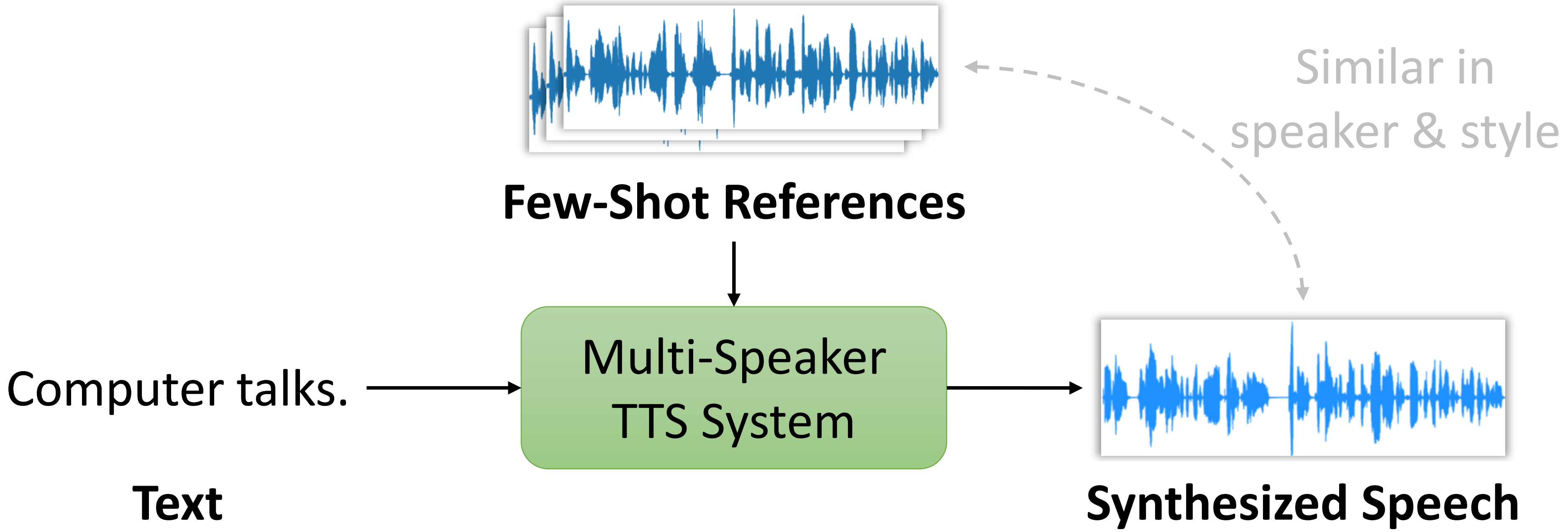
- Task Description
- Background & Motivation
- Methodology
- Experiments
- Conclusion

Task Description

Multi-Speaker Multi-Style Voice Cloning

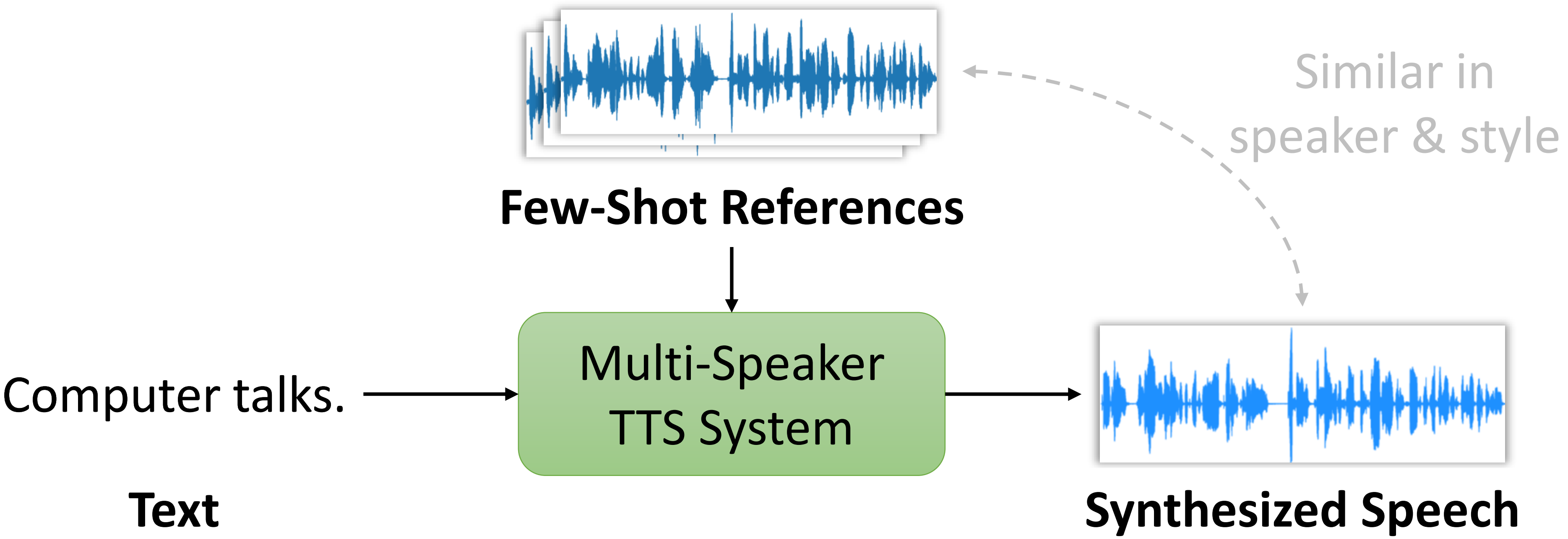


Multi-Speaker Multi-Style Voice Cloning



Task Description

Multi-Speaker Multi-Style Voice Cloning

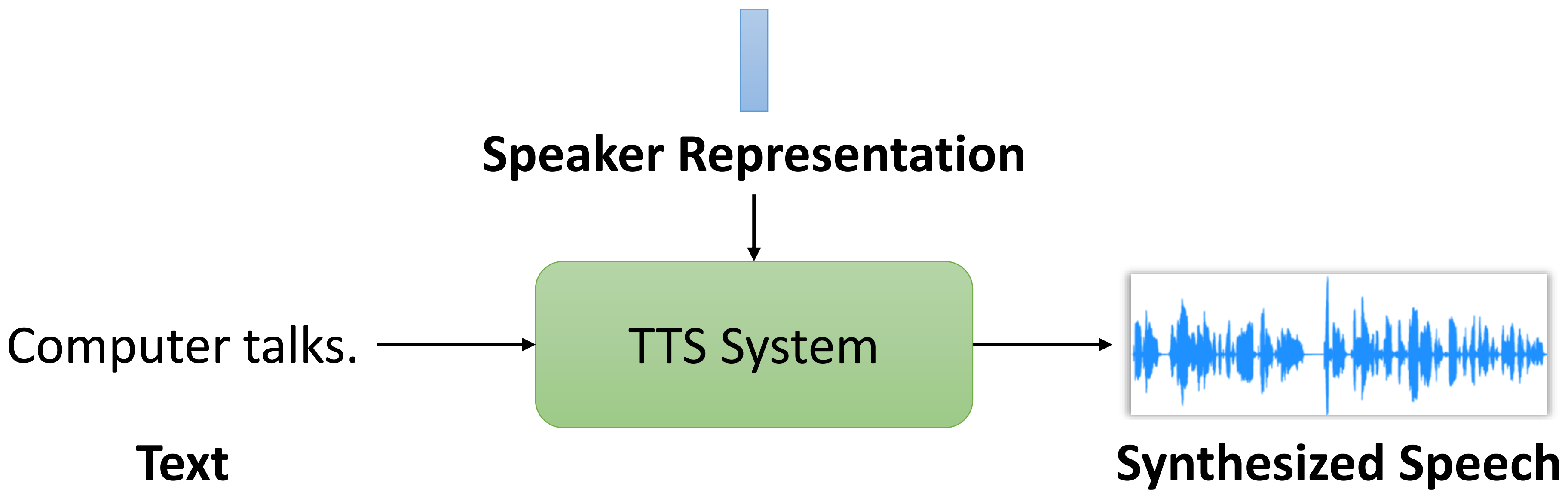


Challenge

- Extract speaker and style information from limited references
- Enable the TTS system to generalize to different speakers/styles

Background & Motivation

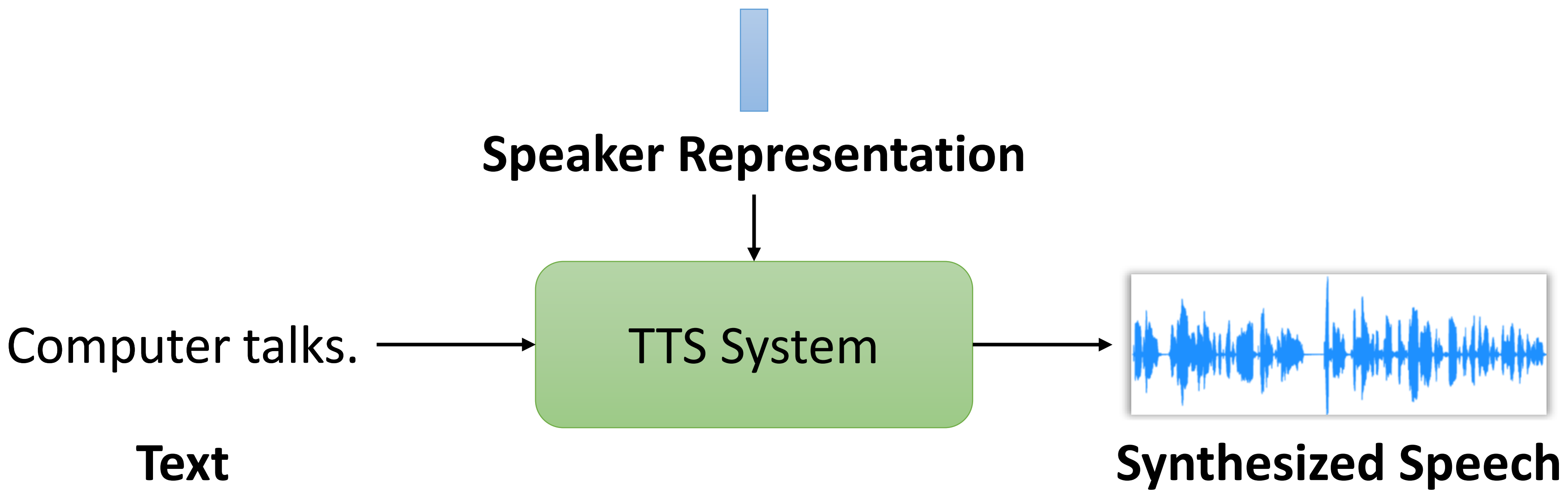
General Framework of Multi-Speaker TTS



General Framework of Multi-Speaker TTS

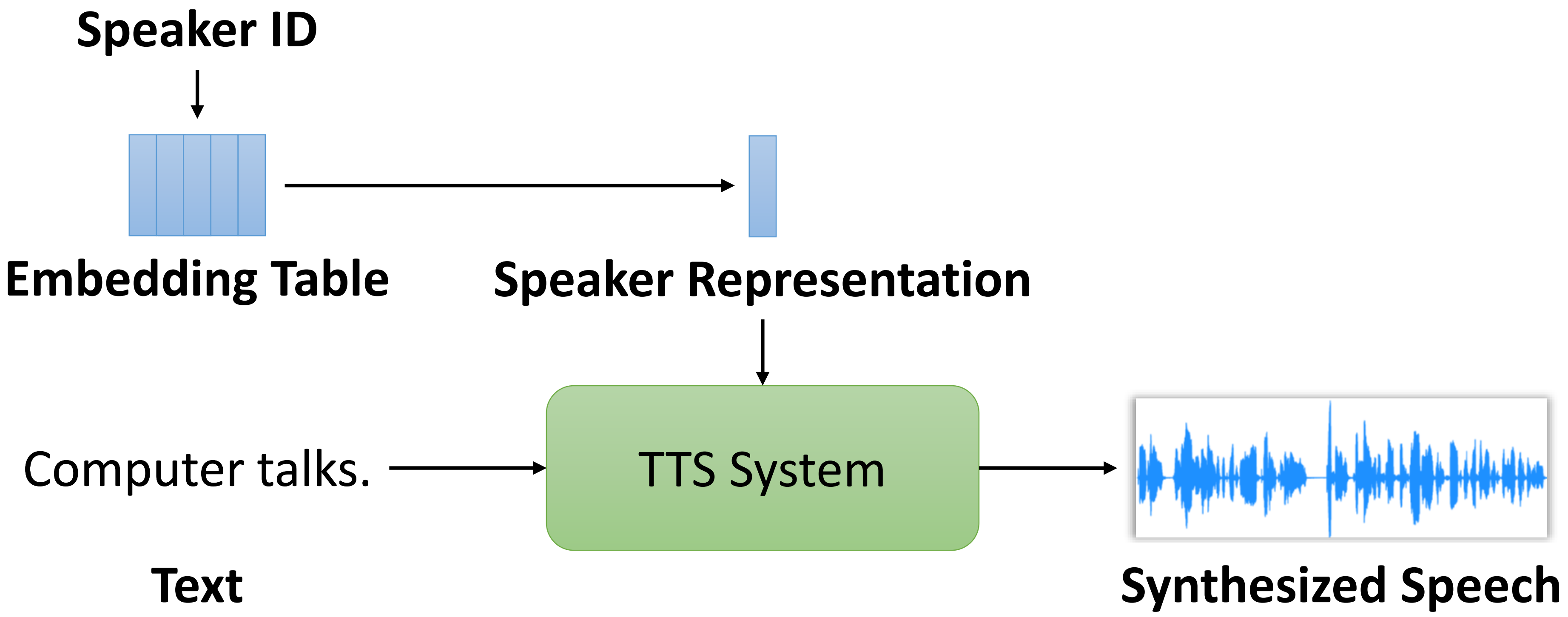
Learnable
- Embedding Table
- Trainable Speaker Encoder

Pretrained
- Pretrained Speaker Encoder



General Framework of Multi-Speaker TTS

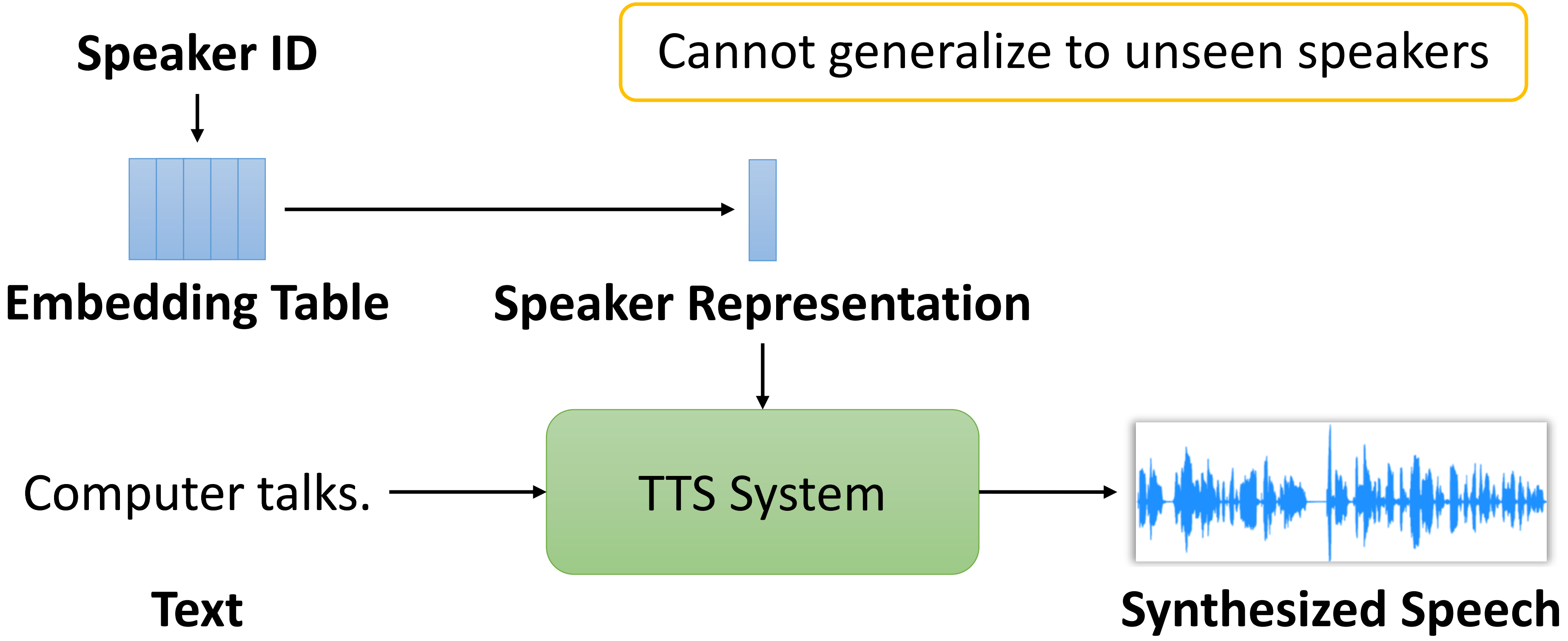
Learnable Speaker Representation



"Deep voice 3: Scaling text-to-speech with convolutional sequence learning", Ping, et. al, ICLR'18

General Framework of Multi-Speaker TTS

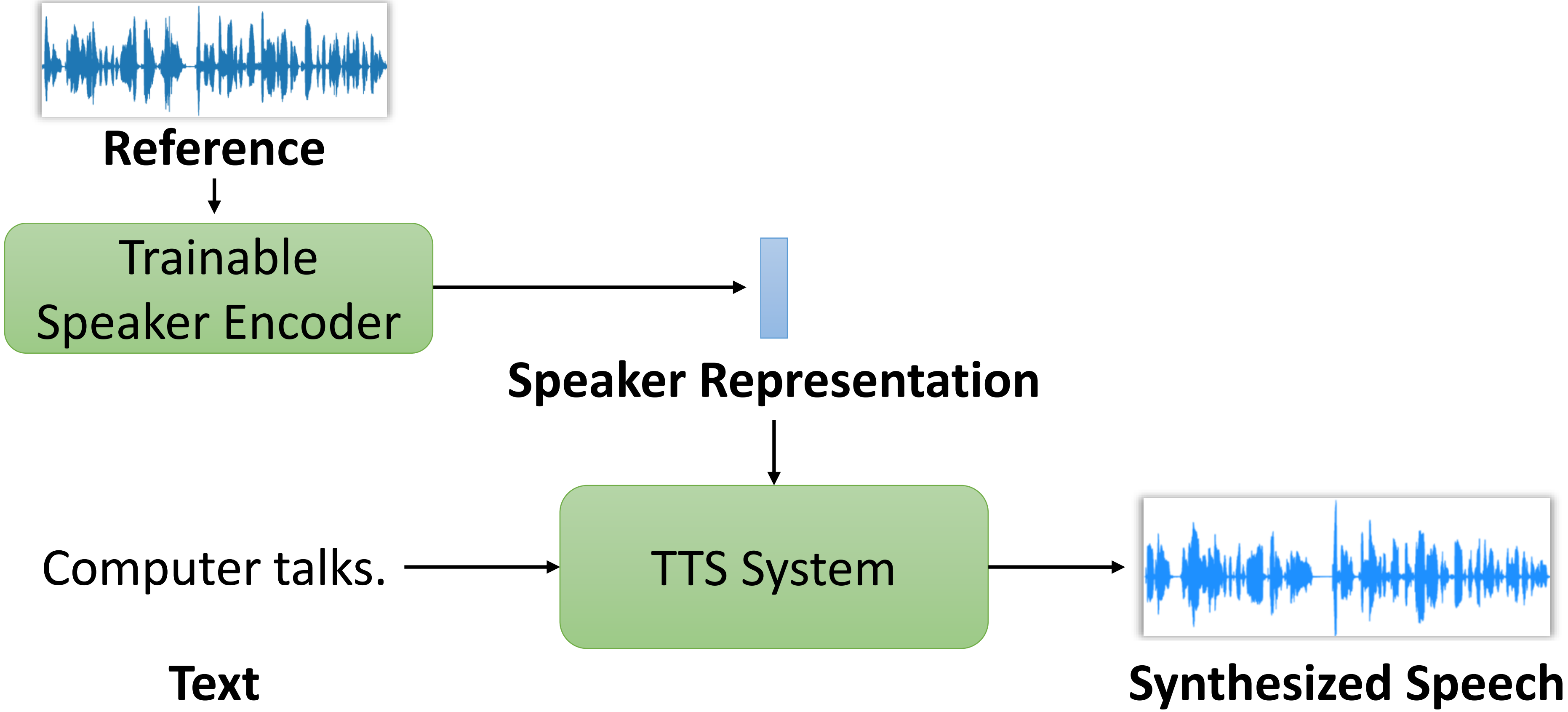
Learnable Speaker Representation



“Deep voice 3: Scaling text-to-speech with convolutional sequence learning”, Ping, et. al, ICLR’18

General Framework of Multi-Speaker TTS

Learnable Speaker Representation

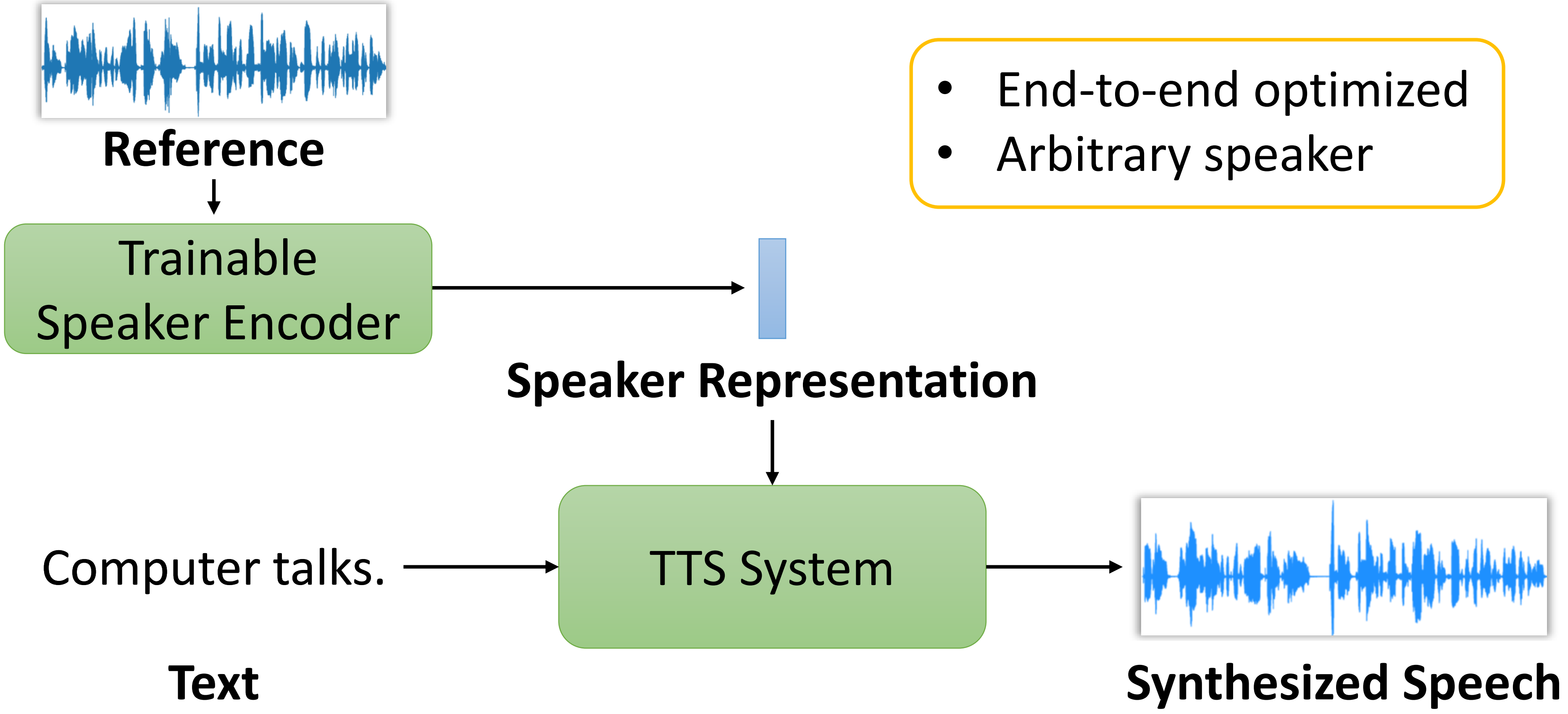


"Neural voice cloning with a few samples", Arik, et. al, NeurIPS'18
"Sample efficient adaptive text-to-speech", Chen, et. al, ICLR'19

General Framework of Multi-Speaker TTS

Learnable Speaker Representation

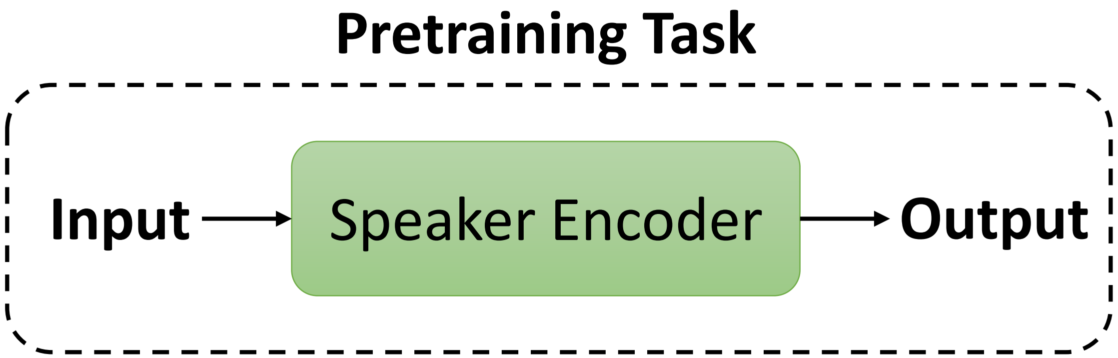
- End-to-end optimized
- Arbitrary speaker



"Neural voice cloning with a few samples", Arik, et. al, NeurIPS'18
"Sample efficient adaptive text-to-speech", Chen, et. al, ICLR'19

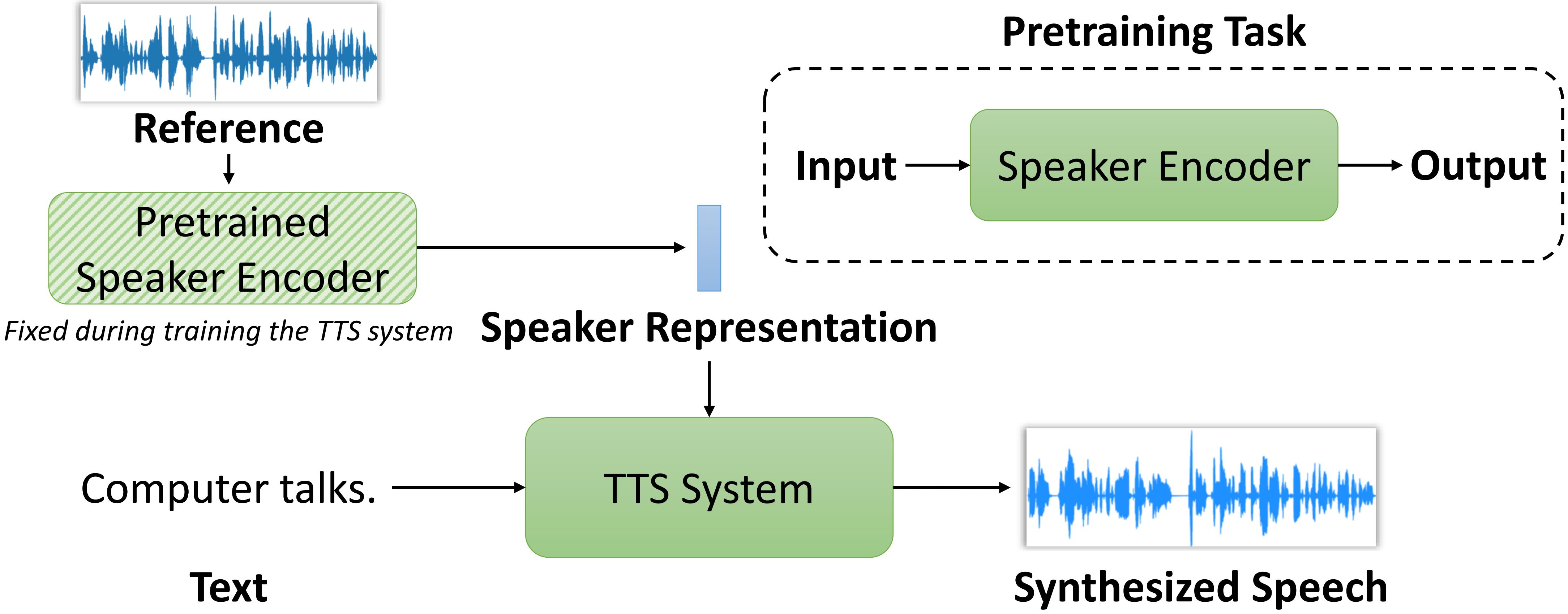
General Framework of Multi-Speaker TTS

Pretrained Speaker Representation



General Framework of Multi-Speaker TTS

Pretrained Speaker Representation

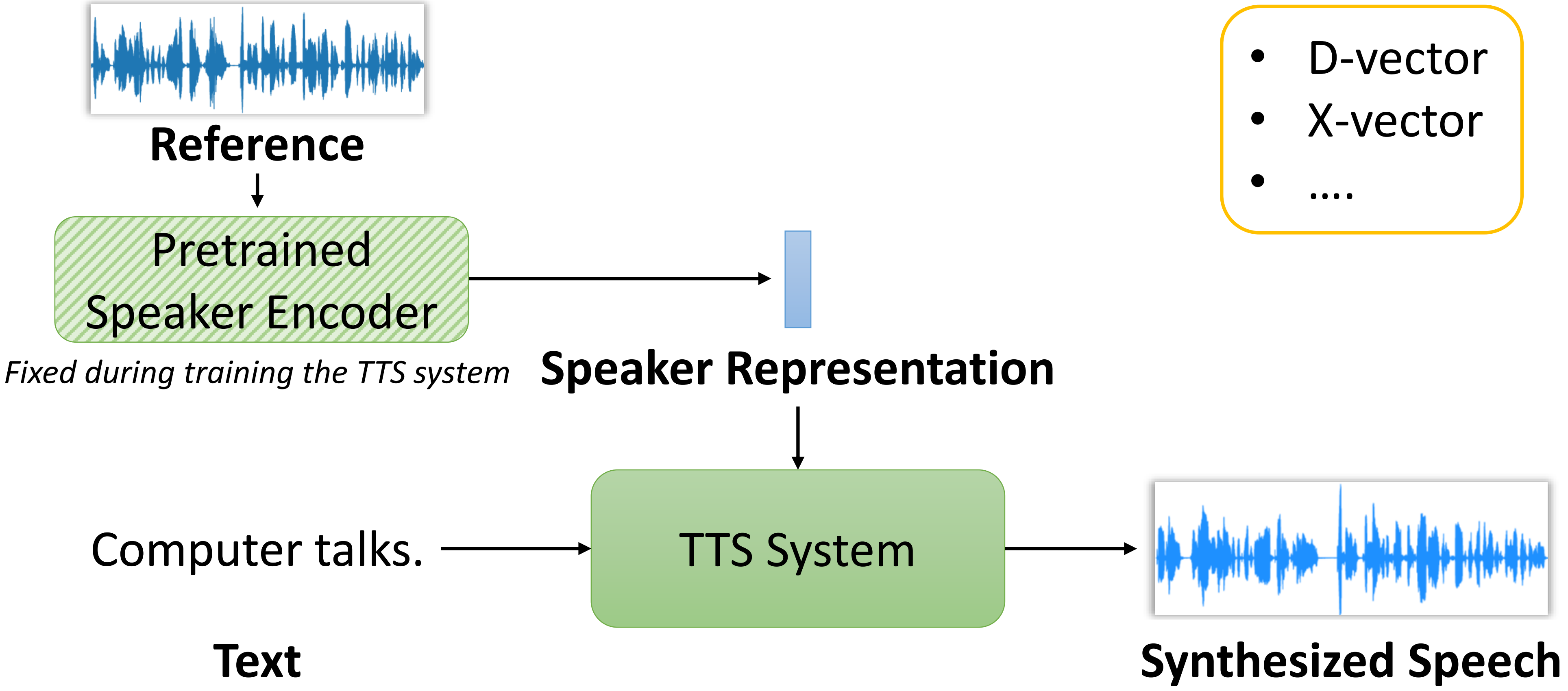


"Transfer learning from speaker verification to multi-speaker text-to-speech synthesis", Jia, et. al, NeurIPS'18

"Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings", Cooper, et. al, ICASSP'20

General Framework of Multi-Speaker TTS

Pretrained Speaker Representation

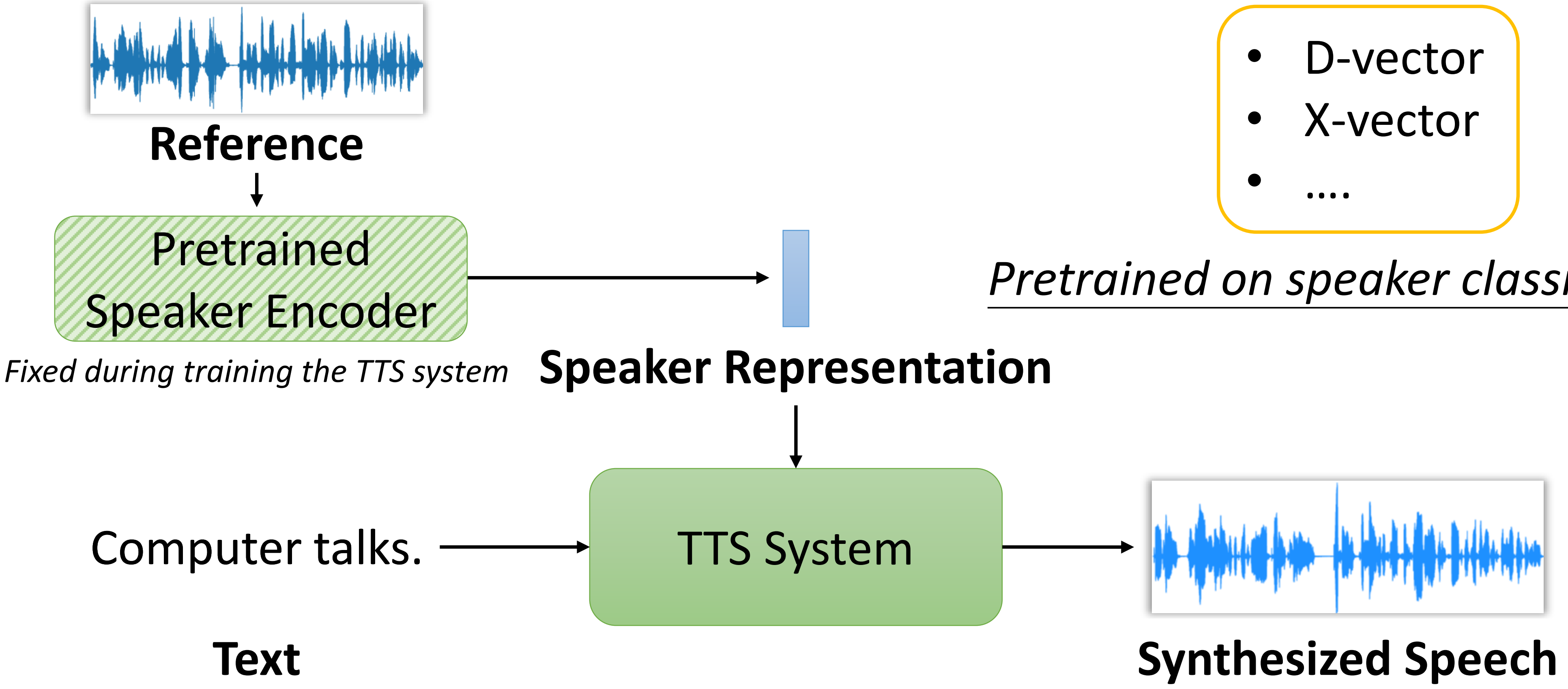


"Transfer learning from speaker verification to multi-speaker text-to-speech synthesis", Jia, et. al, NeurIPS'18

"Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings", Cooper, et. al, ICASSP'20

General Framework of Multi-Speaker TTS

Pretrained Speaker Representation



- D-vector
- X-vector
- ….

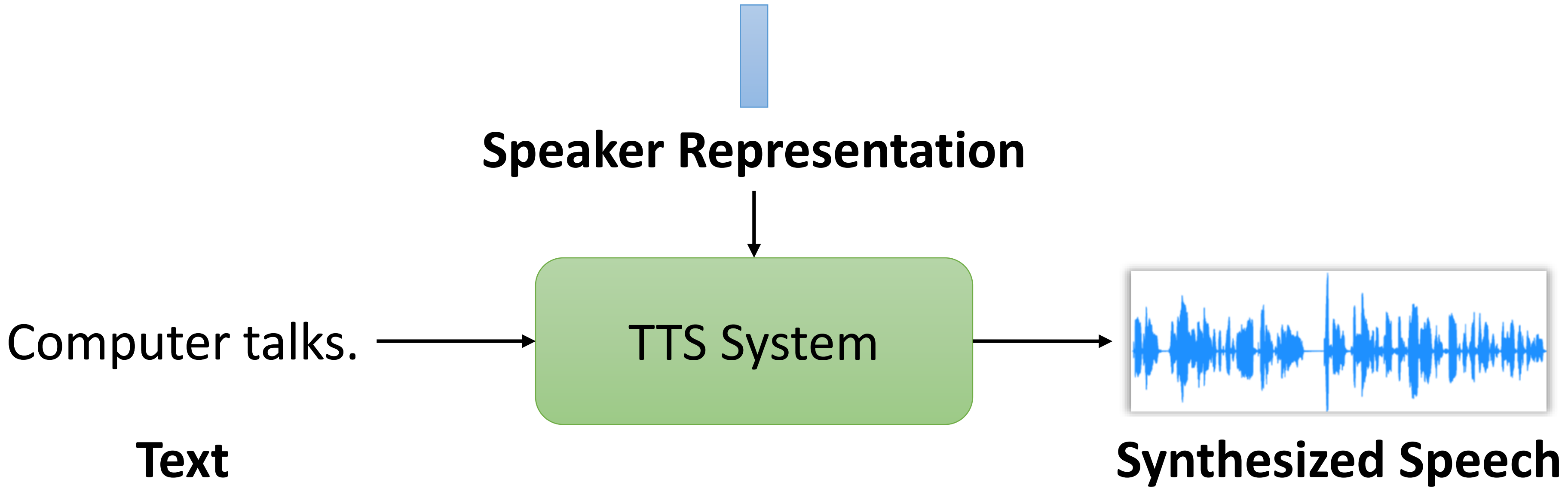
Pretrained on speaker classification!

"Transfer learning from speaker verification to multi-speaker text-to-speech synthesis", Jia, et. al, NeurIPS'18
"Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings", Cooper, et. al, ICASSP'20

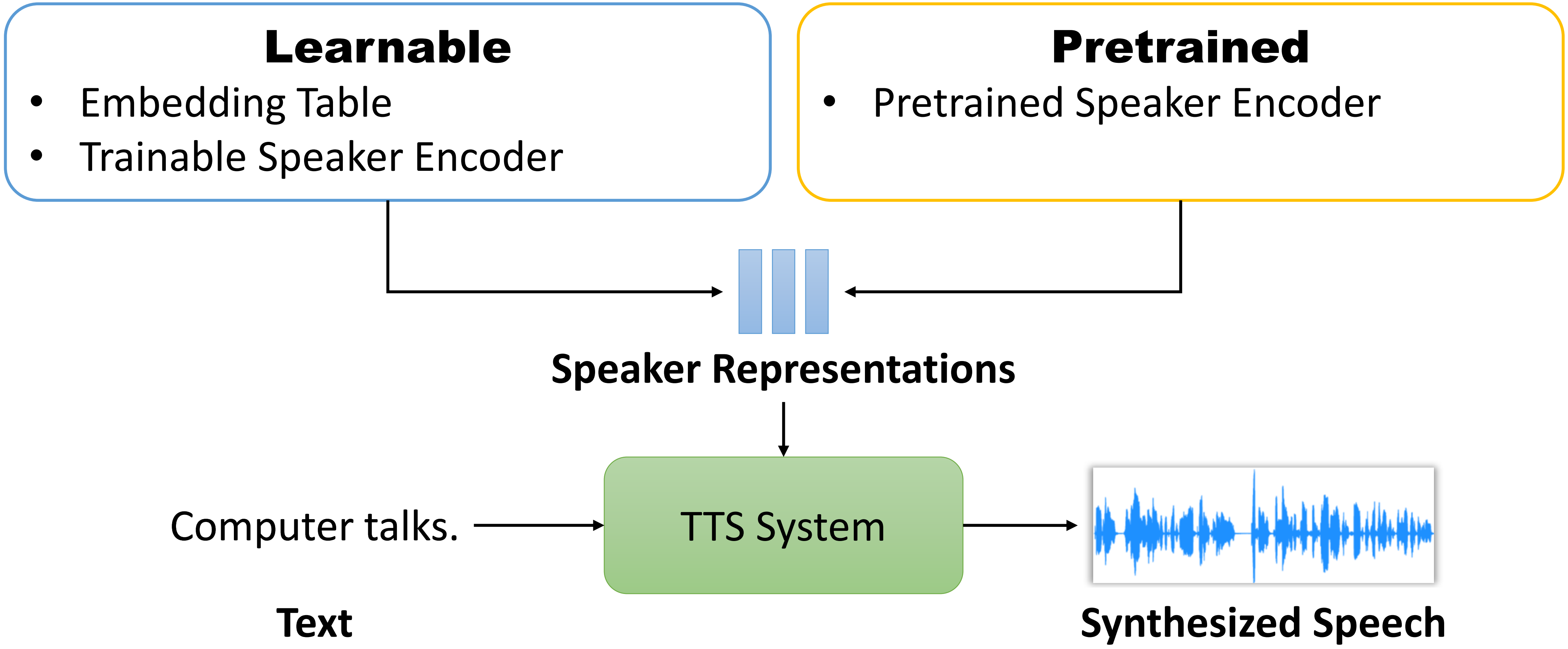
Motivation: Combining Different Representations

Learnable
- Embedding Table
- Trainable Speaker Encoder

Pretrained
- Pretrained Speaker Encoder



Motivation: Combining Different Representations



Motivation: Different Pretraining Tasks

- D-vector
- X-vector
- ….

Discriminative Pretraining Tasks

e.g. speaker classification

Motivation: Different Pretraining Tasks

- D-vector
- X-vector
- ….

VS

Discriminative Pretraining Tasks

e.g. speaker classification

Generative Pretraining Tasks?

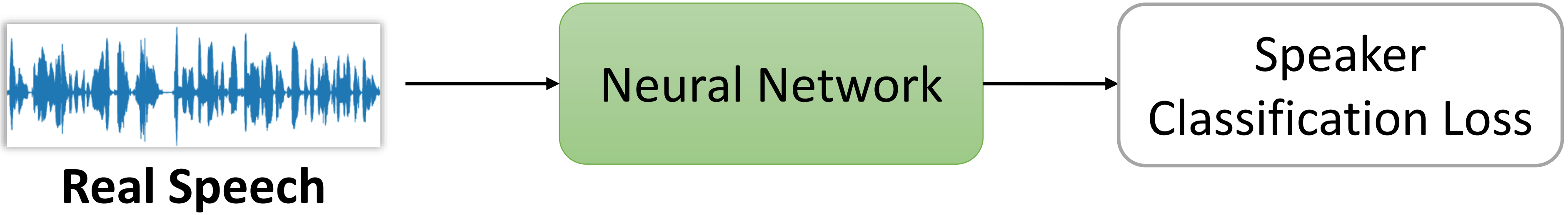
Methodology

Workflow


Speaker Representation Pretraining
TTS Training
TTS Inference

Speaker Representation Pretraining

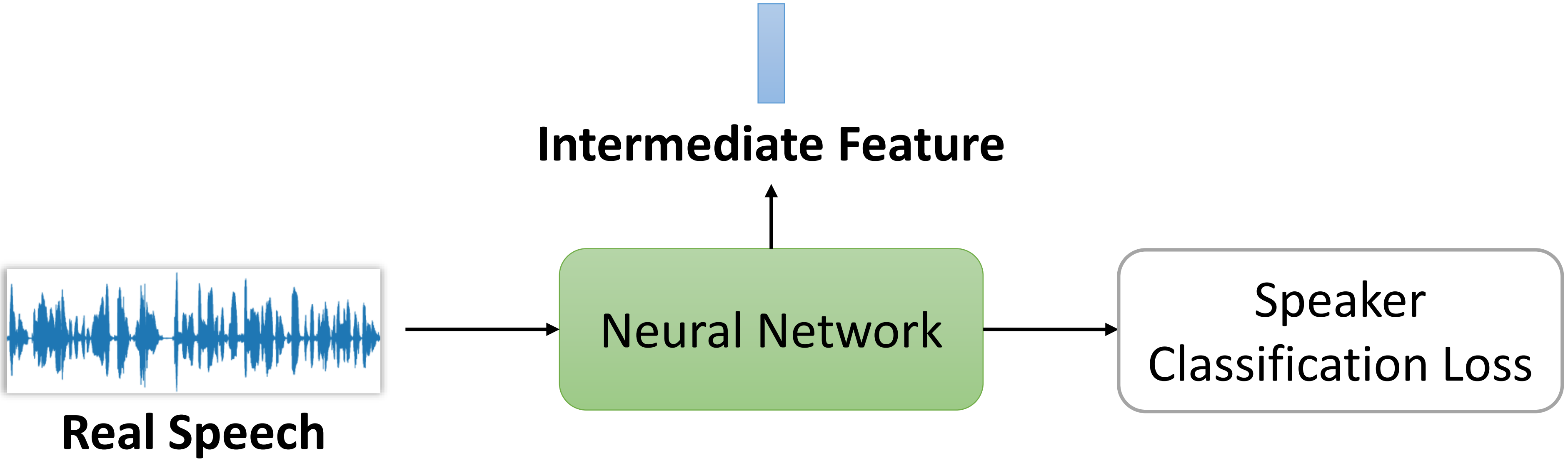
Discriminative Tasks: D-vec & X-vec



"Generalized end-to-end loss for speaker verification", Wan, et. al, ICASSP'18
"X-vectors: Robust dnn embeddings for speaker recognition", Snyder, et. al, ICASSP'18

Speaker Representation Pretraining

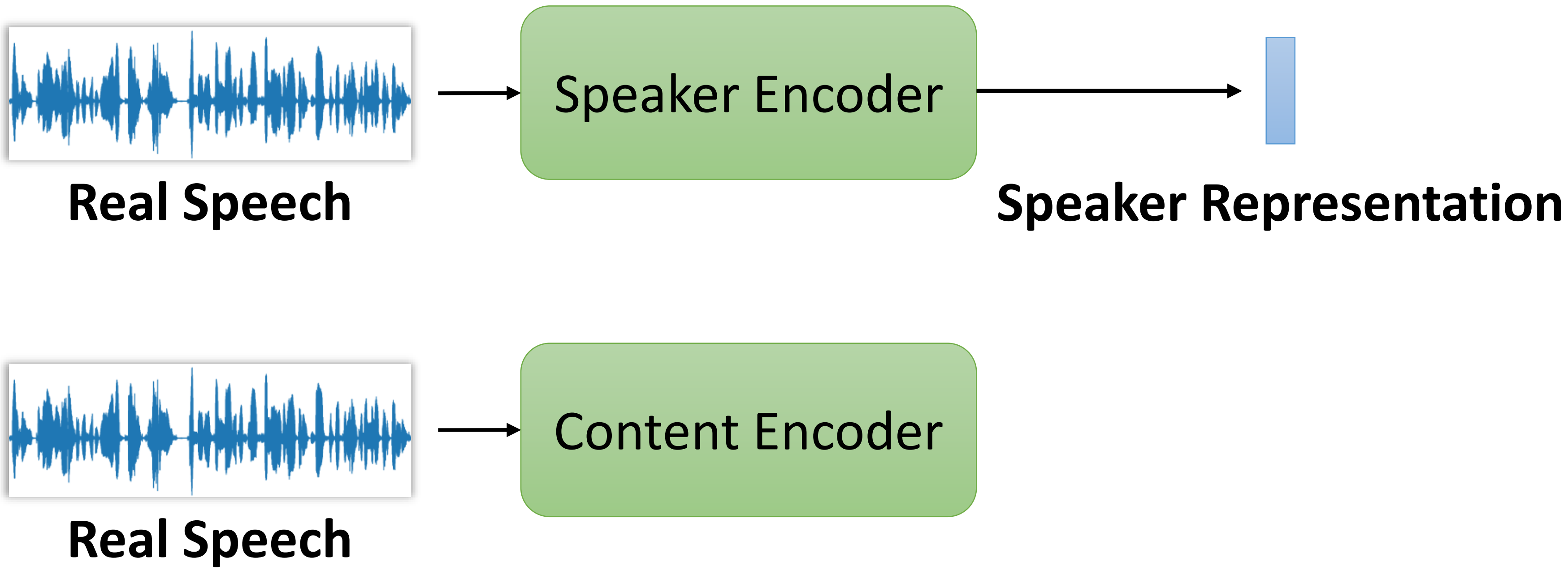
Discriminative Tasks: D-vec & X-vec



"Generalized end-to-end loss for speaker verification", Wan, et. al, ICASSP'18
"X-vectors: Robust dnn embeddings for speaker recognition", Snyder, et. al, ICASSP'18

Speaker Representation Pretraining

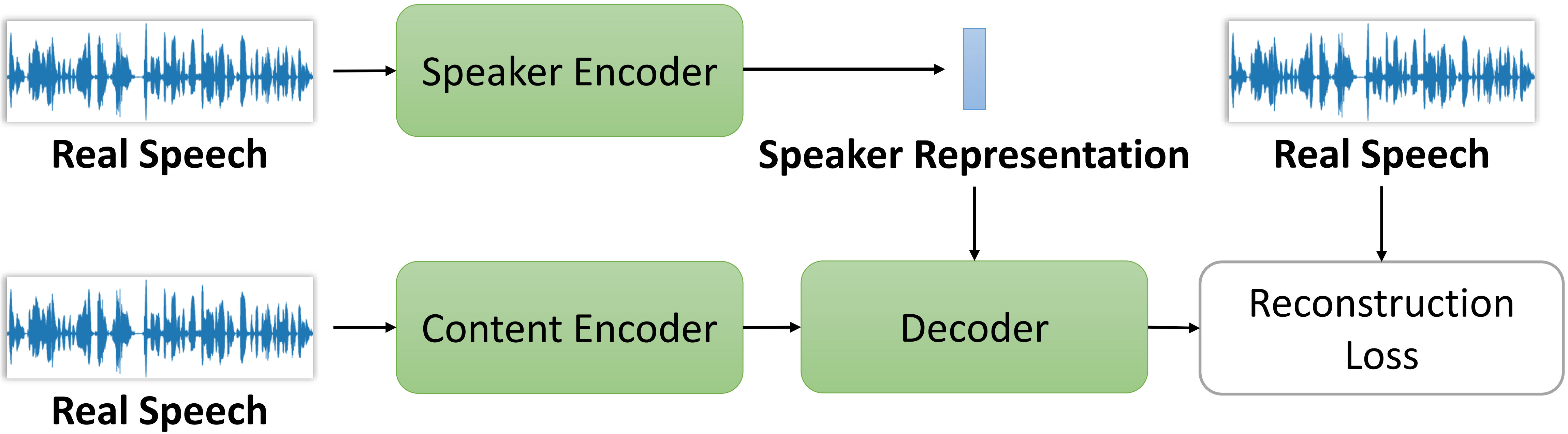
Generative Tasks: AdaIN-VC (One-Shot)



"One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization", Chou, et. al, InterSpeech'19

Speaker Representation Pretraining

Generative Tasks: AdaIN-VC (One-Shot)

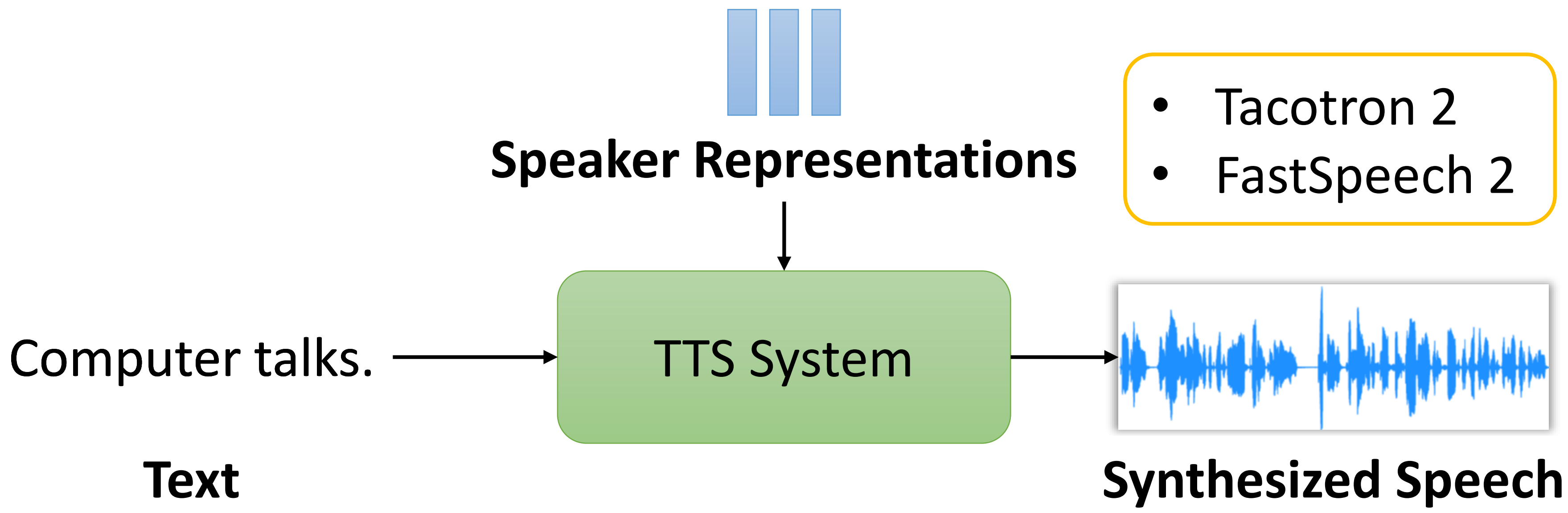


"One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization", Chou, et. al, InterSpeech'19

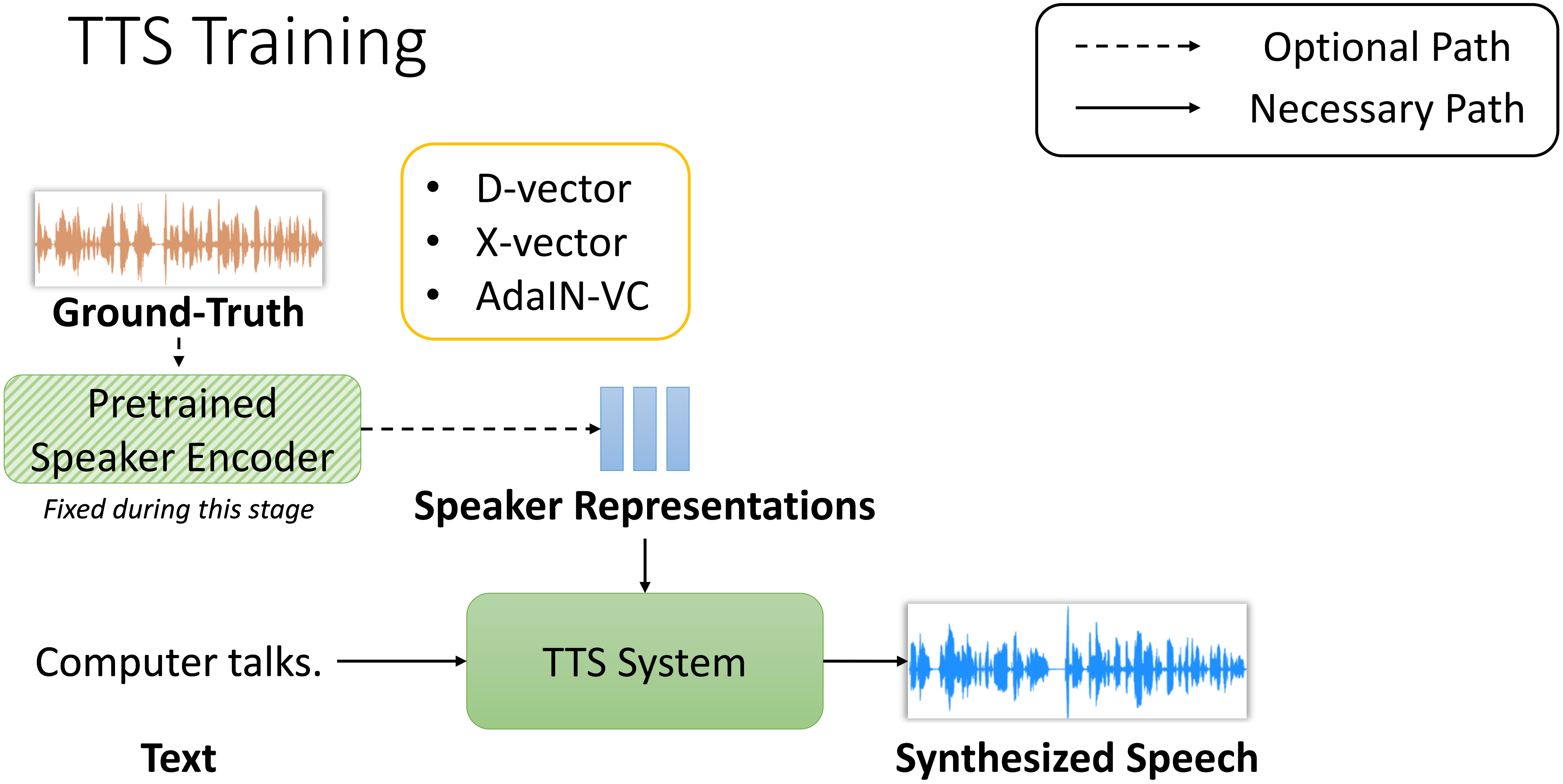
TTS Training

Optional Path

Necessary Path

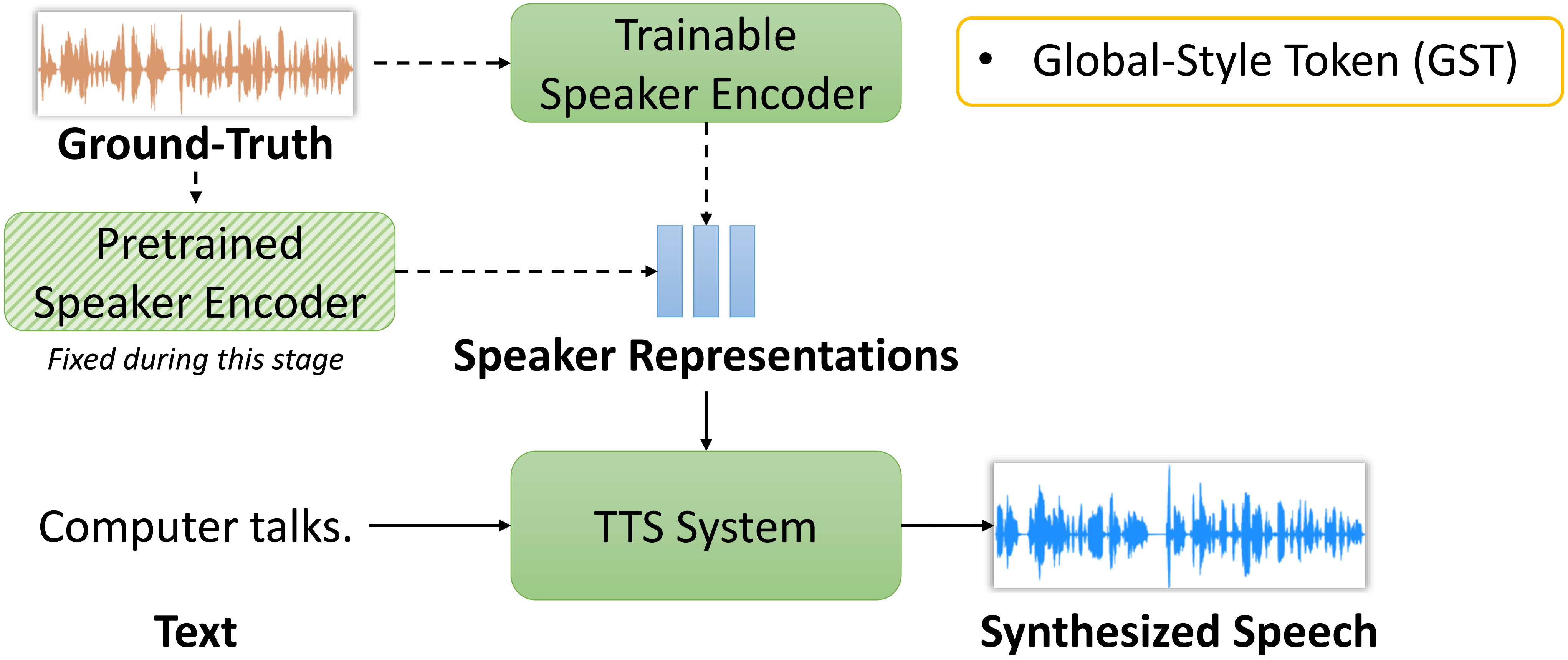


TTS Training



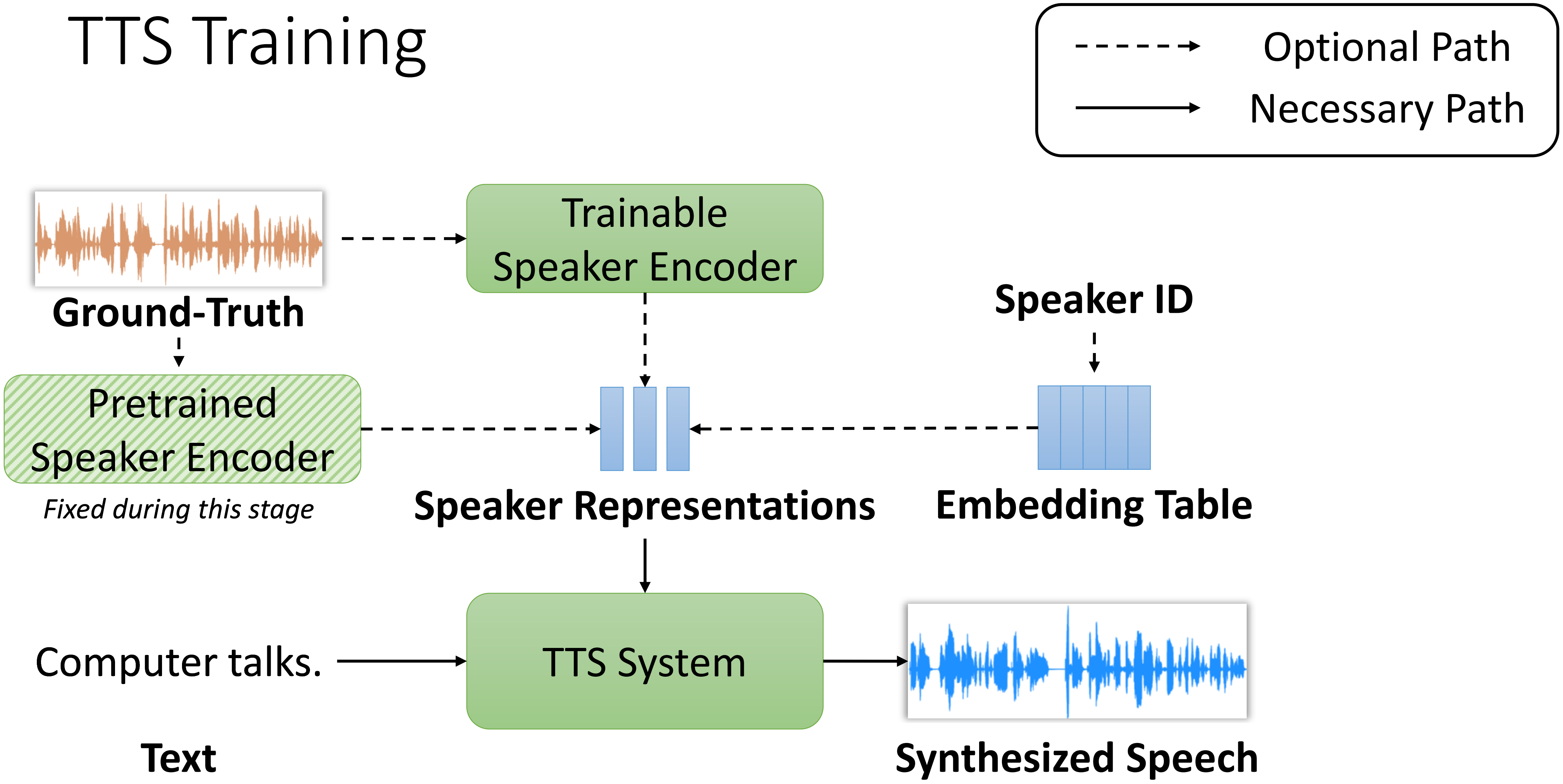
TTS Training

- - - - → Optional Path
——→ Necessary Path

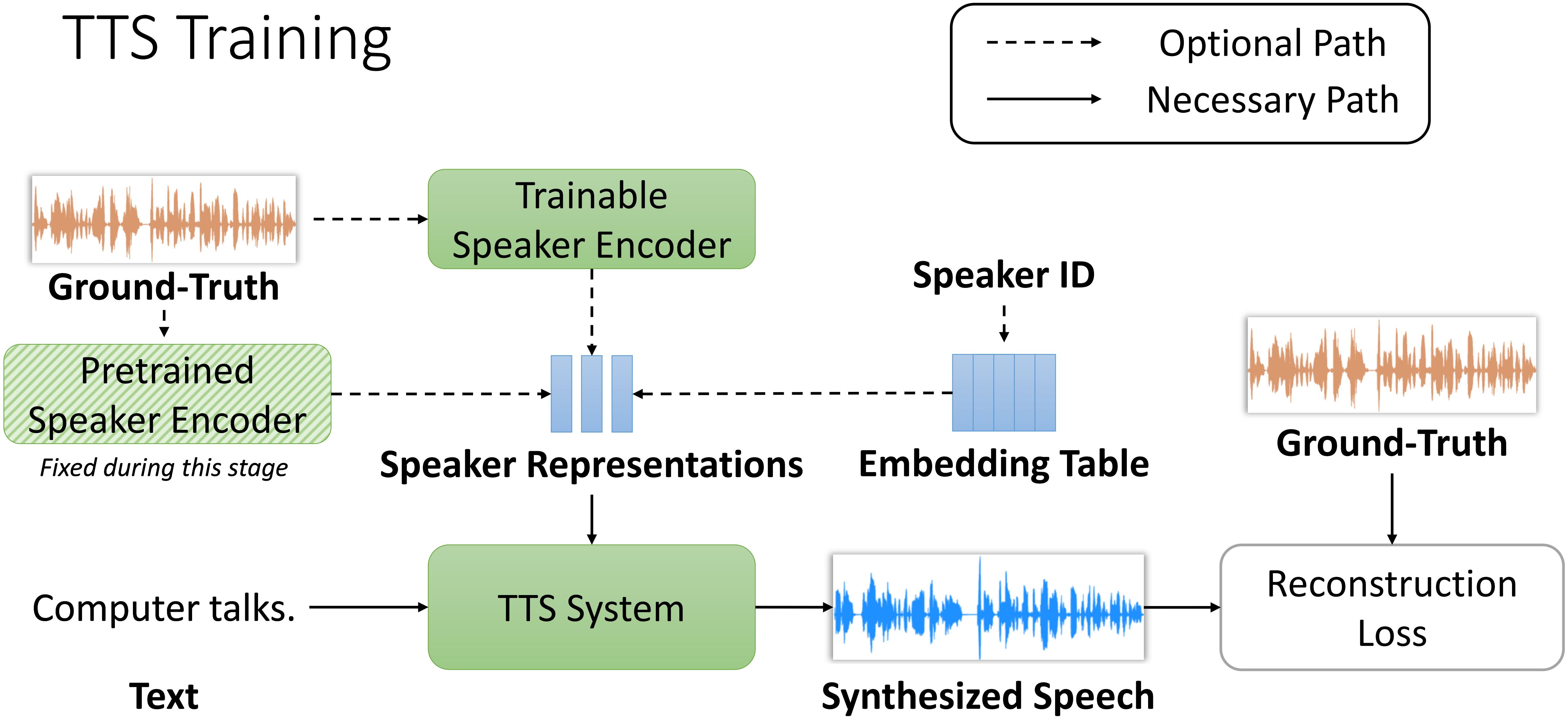


““Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis”, Wang, et. al, ICML’18

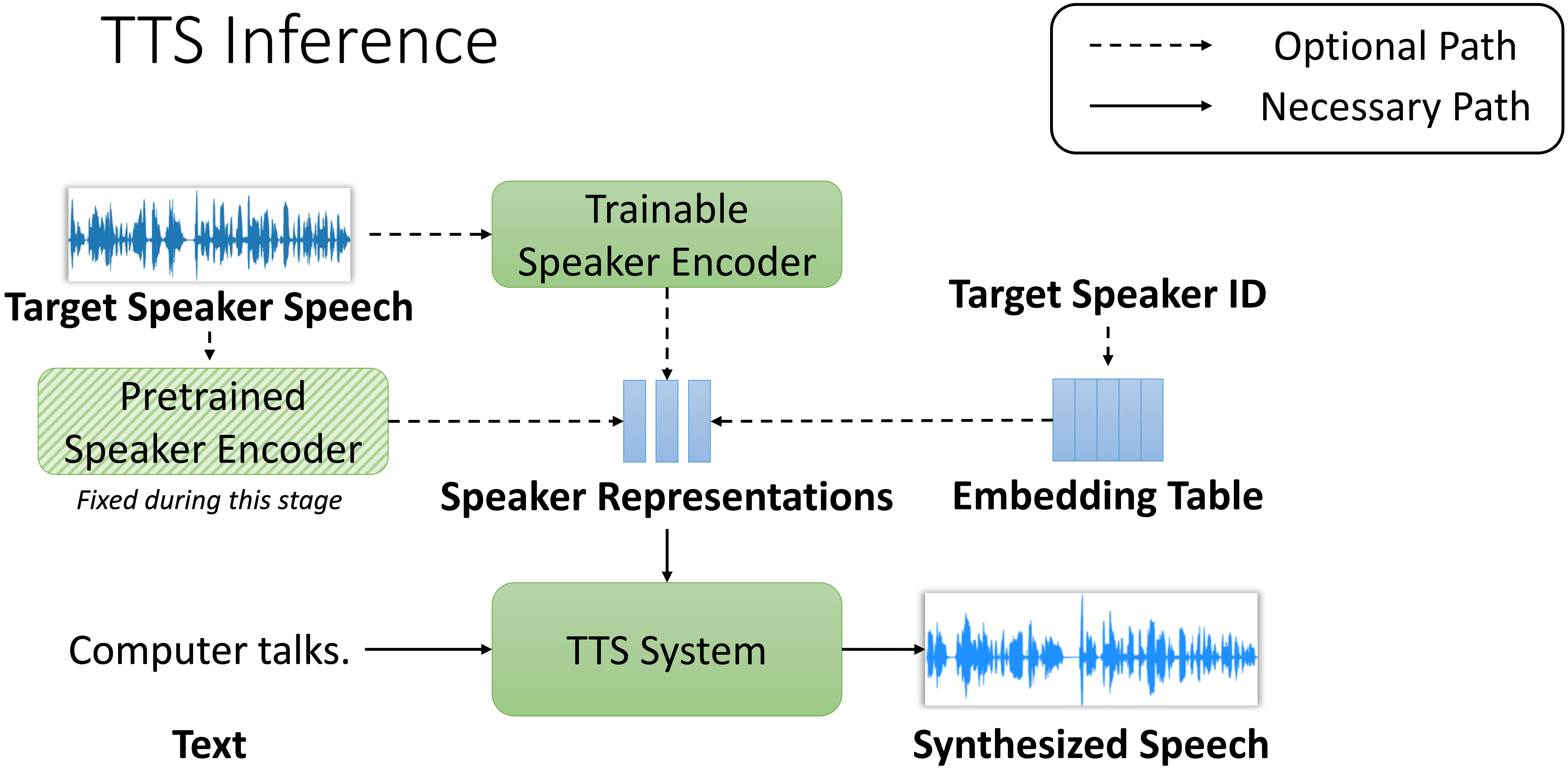
TTS Training



TTS Training



TTS Inference



Experiments

Dataset

- Training: 96 hours of Mandarin speech by 230 speakers with transcriptions
  - AIShell-3
  - M2VoC dataset

Dataset

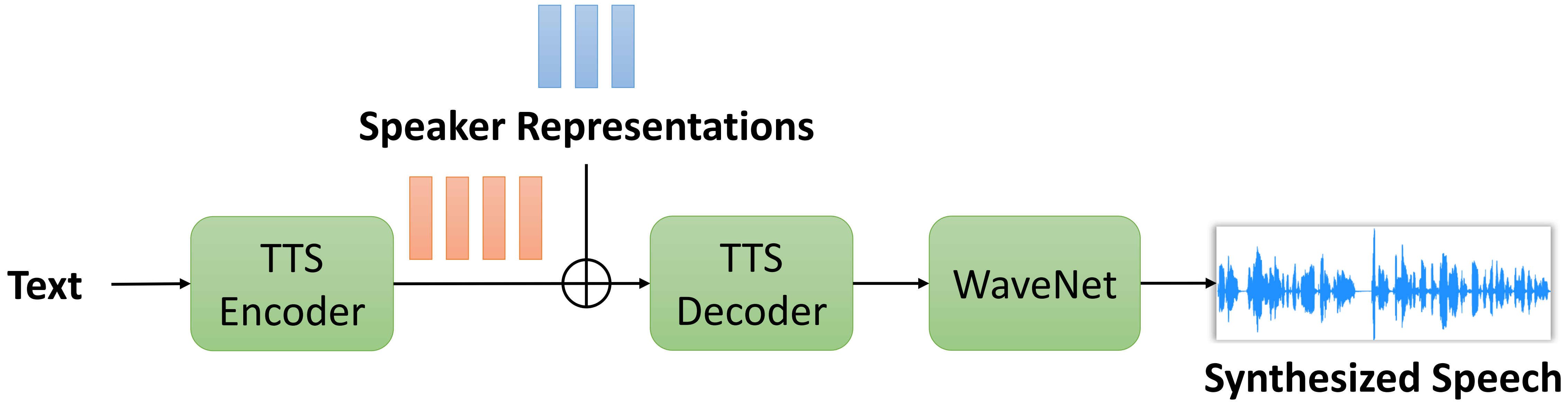
- Training: 96 hours of Mandarin speech by 230 speakers with transcriptions
  - AIShell-3
  - M2VoC dataset
- 6 few-shot target speakers
  - Track 1: 3 speakers with 100 recordings
  - Track 2: 3 speakers with 5 recordings

Dataset

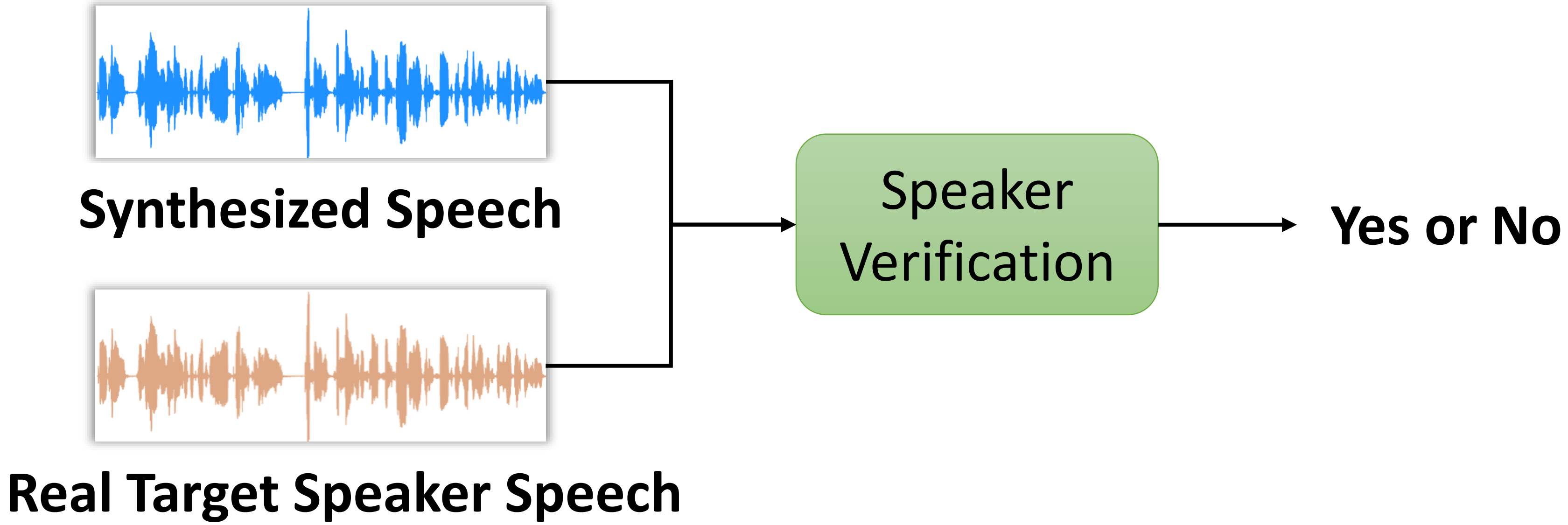
- Training: 96 hours of Mandarin speech by 230 speakers with transcriptions
  - AIShell-3
  - M2VoC dataset
- 6 few-shot target speakers
  - Track 1: 3 speakers with 100 recordings
  - Track 2: 3 speakers with 5 recordings
- The few shot speakers are also used to train the speaker representation models and the TTS models

TTS Model Setup

- Tacotron 2 & FastSpeech 2
  - Speaker representations are added to encoder outputs
- WaveNet vocoder



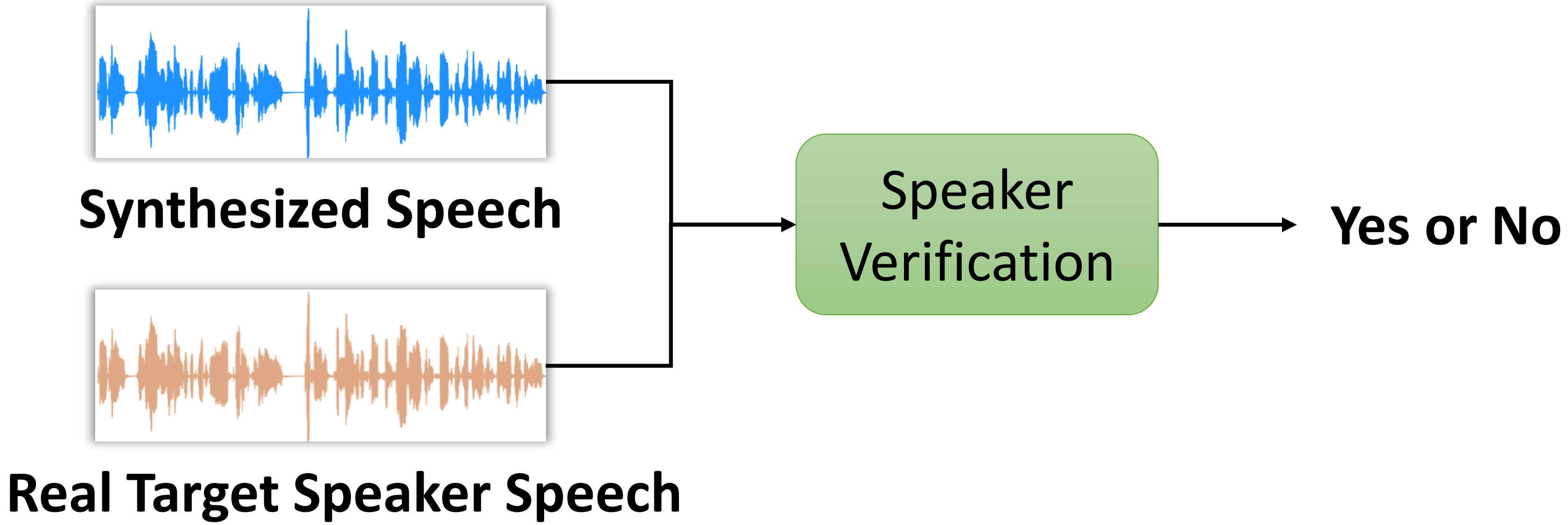
Automatic Speaker Similarity Evaluation



Automatic Speaker Similarity Evaluation

Metrics | Speaker Verification Accuracy

Scale: 0 ~ 1, the larger the better



Automatic Speaker Similarity Evaluation

Metrics | Speaker Verification Accuracy

Scale: 0 ~ 1, the larger the better

Model	Speaker Representation: Pretrained d-vec	Pretrained x-vec	Pretrained VC	Learnable embed	Learnable GST	Results SV Accuracy: Track 1	Track 2
(a) Tacotron 2	✓					.772	.367
		✓				.785	.377
			✓			.942	.727
				✓		.630	.703
					✓	.102	.050
(b) FastSpeech2	✓					.977	.323
		✓				.973	.623
			✓			.980	.837
				✓		.988	.490
					✓	.778	.340

Generative Pretraining > Others

Automatic Speaker Similarity Evaluation

Metrics | Speaker Verification Accuracy

Scale: 0 ~ 1, the larger the better

Model	Speaker Representation					Results	
	Pretrained			Learnable		SV Accuracy	
	d-vec	x-vec	VC	embed	GST	Track 1	Track 2
(a) Tacotron 2	✓					.772	.367
		✓				.785	.377
			✓			.942	.727
				✓		.630	.703
					✓	.102	.050
(b) FastSpeech2	✓					.977	.323
		✓				.973	.623
			✓			.980	.837
				✓		.988	.490
					✓	.778	.340

Audio samples (Track 2, 5 references)

Target Speaker

d-vec

x-vec

VC

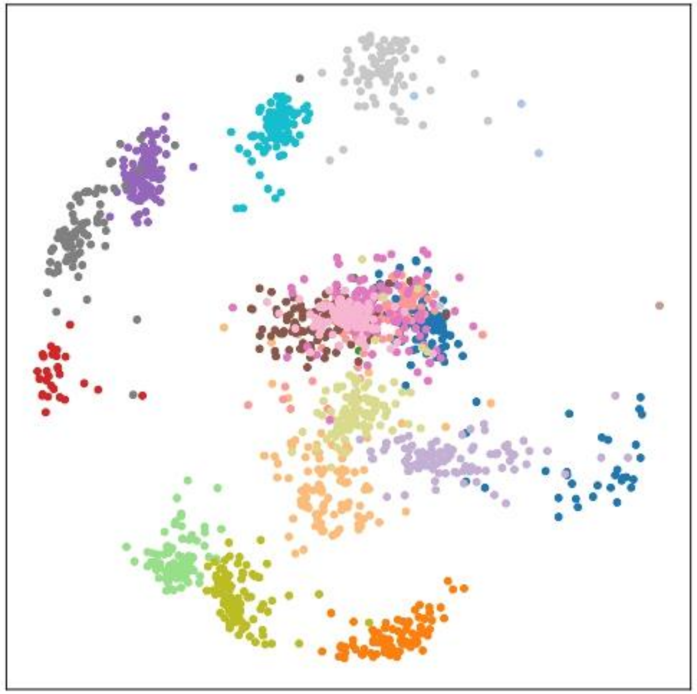
embed

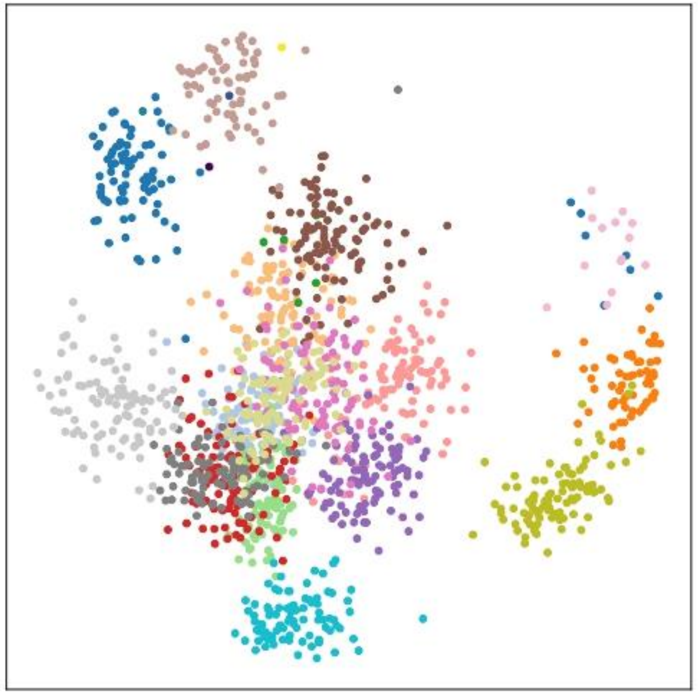
GST

Automatic Speaker Similarity Evaluation

Metrics | Speaker Verification Accuracy

Scale: 0 ~ 1, the larger the better


(a) d-vector


(b) x-vector


(c) VC

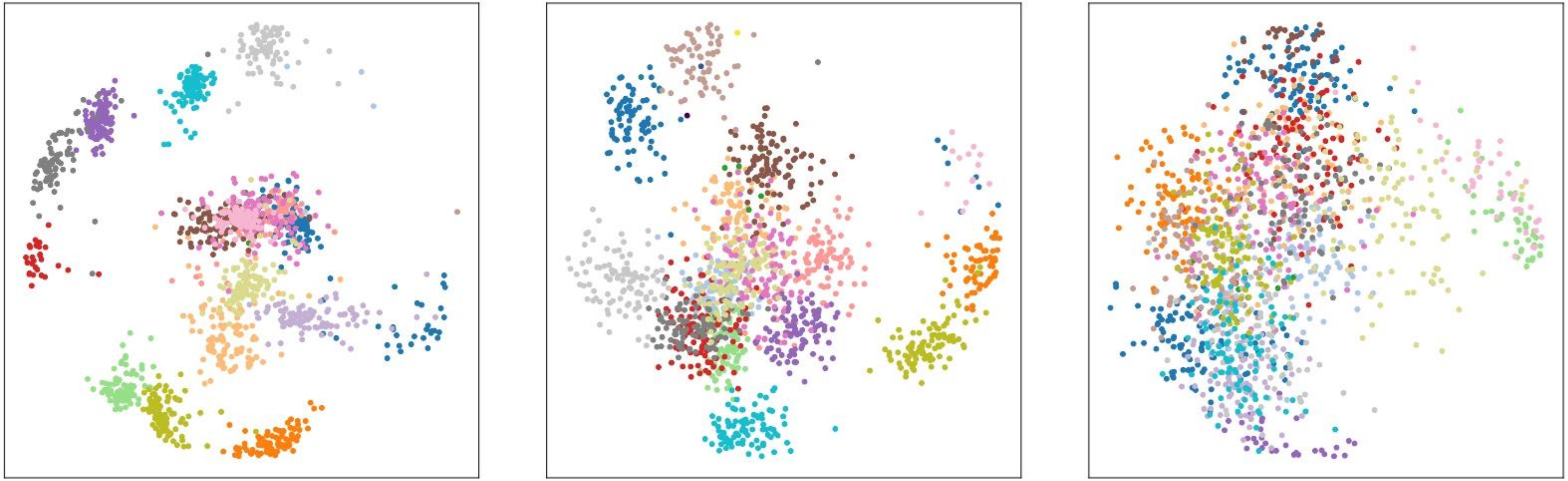
Experiments

Automatic Speaker Similarity Evaluation

Metrics | Speaker Verification Accuracy

Scale: 0 ~ 1, the larger the better

More Continuous



(a) d-vector (b) x-vector (c) VC

Automatic Speaker Similarity Evaluation

Metrics | Speaker Verification Accuracy

Scale: 0 ~ 1, the larger the better

Model	Speaker Representation					Results	
	Pretrained			Learnable		SV Accuracy	
	d-vec	x-vec	VC	embed	GST	Track 1	Track 2
(b) FastSpeech2			✓			.980	.837
(c) FastSpeech2	✓		✓			.978	.747
		✓	✓			**.992**	.860
			✓	✓		.983	**.937**
			✓		✓	.982	.783
			✓	✓	✓	.988	.897
	✓	✓	✓	✓	✓	.990	.887

* The colored row is the model used for the final submission to the ICASSP 2021 M2VoC challenge. Due to the time limitation, we did not submit our best model.

Multiple speaker representations

Track 1 (100 references):
No obvious difference

Automatic Speaker Similarity Evaluation

Metrics | Speaker Verification Accuracy

Scale: 0 ~ 1, the larger the better

Model	Speaker Representation					Results	
	Pretrained			Learnable		SV Accuracy	
	d-vec	x-vec	VC	embed	GST	Track 1	Track 2
(b) FastSpeech2			✓			.980	.837
(c) FastSpeech2	✓		✓			.978	.747
		✓	✓			**.992**	.860
			✓	✓		.983	**.937**
			✓		✓	.982	.783
			✓	✓	✓	.988	.897
	✓	✓	✓	✓	✓	.990	.887

* The colored row is the model used for the final submission to the ICASSP 2021 M2VoC challenge. Due to the time limitation, we did not submit our best model.

Multiple speaker representations

Track 1 (100 references):
No obvious difference

Track 2 (5 references):
Multiple Representations > Single Representation

Subjective Evaluation (FastSpeech 2, Track 2)

Metrics | Quality MOS | Speaker Similarity MOS

Scale: 1 ~ 5, the larger the better

Subjective Evaluation (FastSpeech 2, Track 2)

Metrics | Quality MOS | Speaker Similarity MOS

Scale: 1 ~ 5, the larger the better

Model	Speaker Representation			
	x-vec	VC	Embed	VC+Embed
$\mathrm{MOS}_{\mathrm{quality}}$	$3.47 \pm .13$	$3.61 \pm .13$	$\mathbf{3.65} \pm .13$	$3.55 \pm .12$
$\mathrm{MOS}_{\mathrm{similarity}}$	$3.25 \pm .13$	$3.19 \pm .14$	$3.27 \pm .13$	$\mathbf{3.38} \pm .14$

Speaker Similarity: Multiple Representations > Single Representation

Subjective Evaluation (FastSpeech 2, Track 2)

Metrics | Quality MOS | Speaker Similarity MOS

Scale: 1 ~ 5, the larger the better

Model	Speaker Representation			
	x-vec	VC	Embed	VC+Embed
$\mathrm{MOS}_{\mathrm{quality}}$	$3.47 \pm .13$	$3.61 \pm .13$	$\mathbf{3.65} \pm .13$	$3.55 \pm .12$
$\mathrm{MOS}_{\mathrm{similarity}}$	$3.25 \pm .13$	$3.19 \pm .14$	$3.27 \pm .13$	$\mathbf{3.38} \pm .14$

Audio samples (Track 2, 5 references)

Target Speaker | VC | VC+Embed

Official Evaluation Results

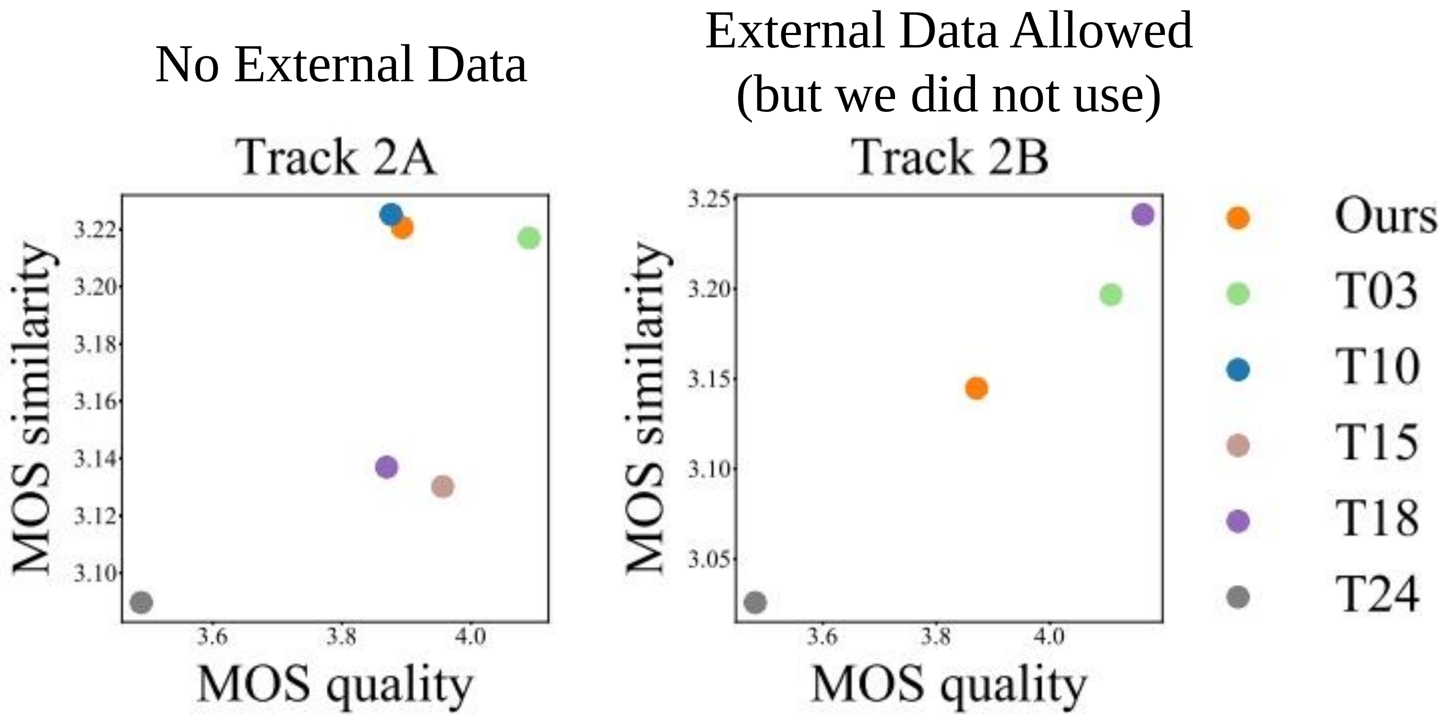


Fig. 3: The official subjective evaluation results of Track 2.

Conclusion

Conclusion

- Pretrained speaker representation + learnable speaker representations > single representation

Conclusion

- Pretrained speaker representation + learnable speaker representations > single representation
- Generative pretraining > discriminative pretraining

Resources

- Audio Samples: https://ming024.github.io/M2VoC/
- Code: https://github.com/ming024/FastSpeech2/tree/M2VoC
- Paper: https://arxiv.org/abs/2103.04088