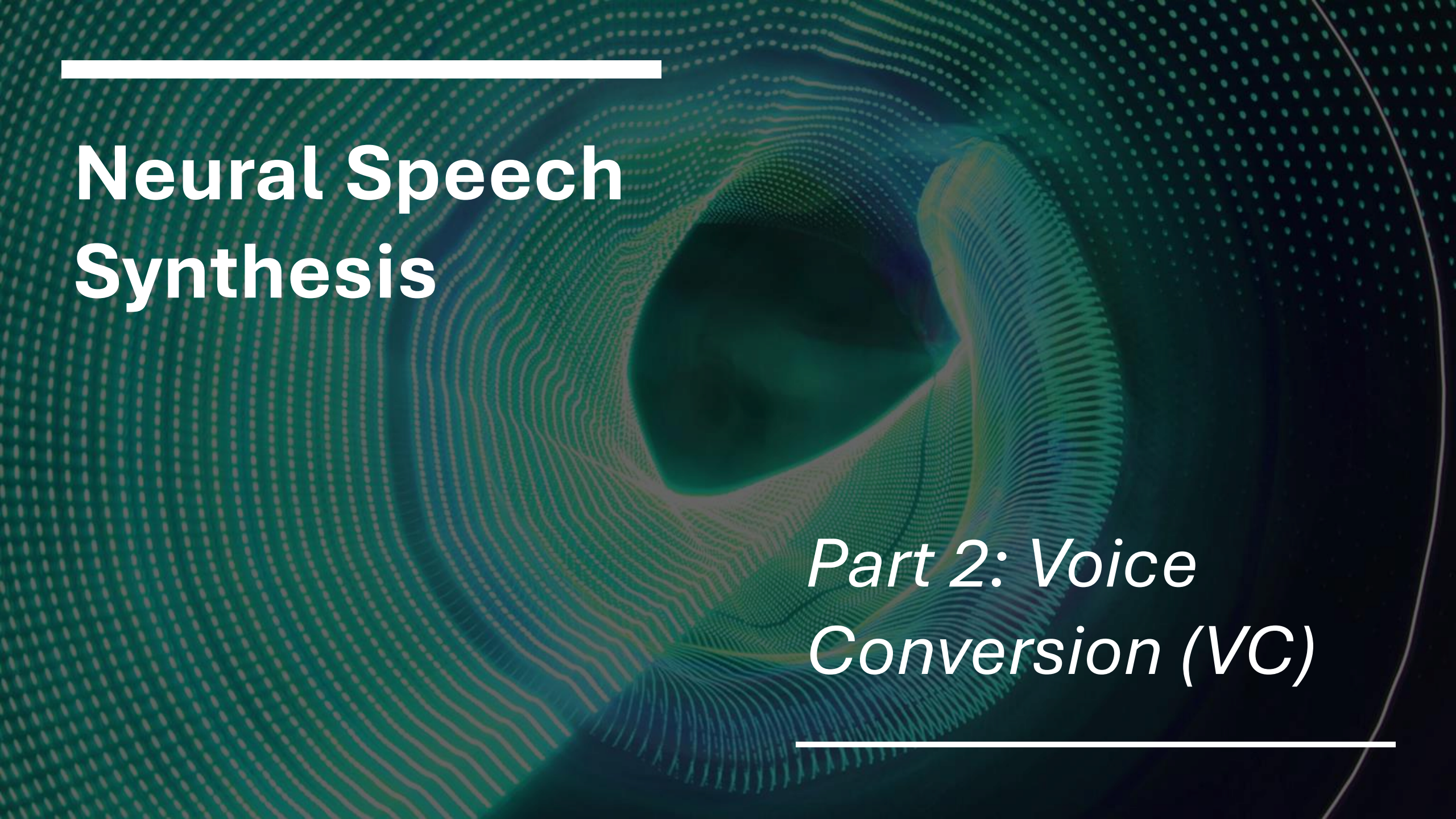


Neural Speech Synthesis

Part 2: Voice Conversion (VC)

Previous Tutorials

- Statistical voice conversion with direct waveform modeling, INTERSPEECH 2019

Tomoki Toda 

Kazuhiro Kobayashi 

Tomoki Hayashi 

- Theory and Practice of Voice Conversion, APSIPA 2020

Berrak Sisman 

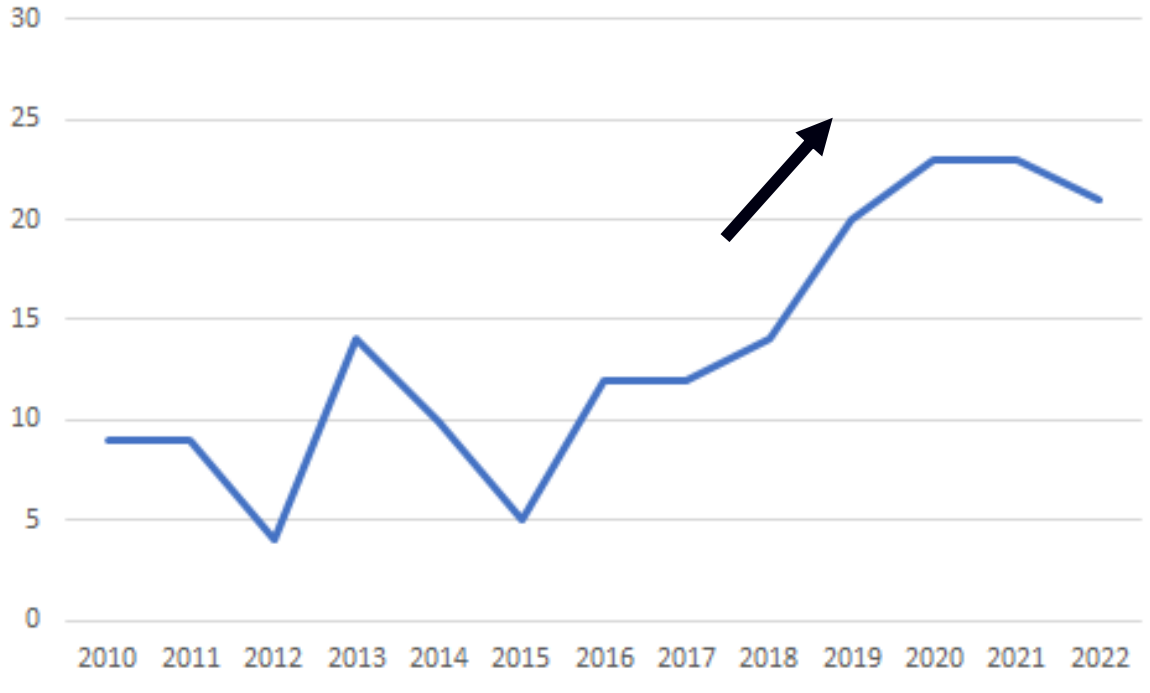
Yu Tsao 

Haizhou Li 

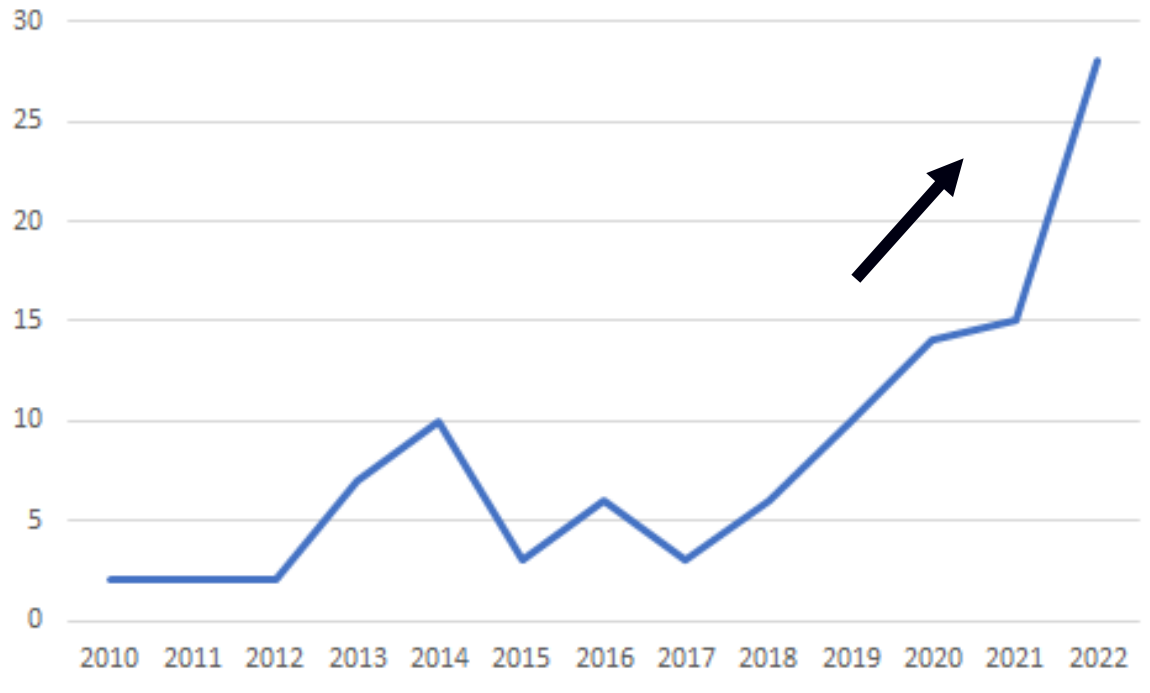
Trend

Number of papers with "*voice conversion*" in the titles

INTERSPEECH

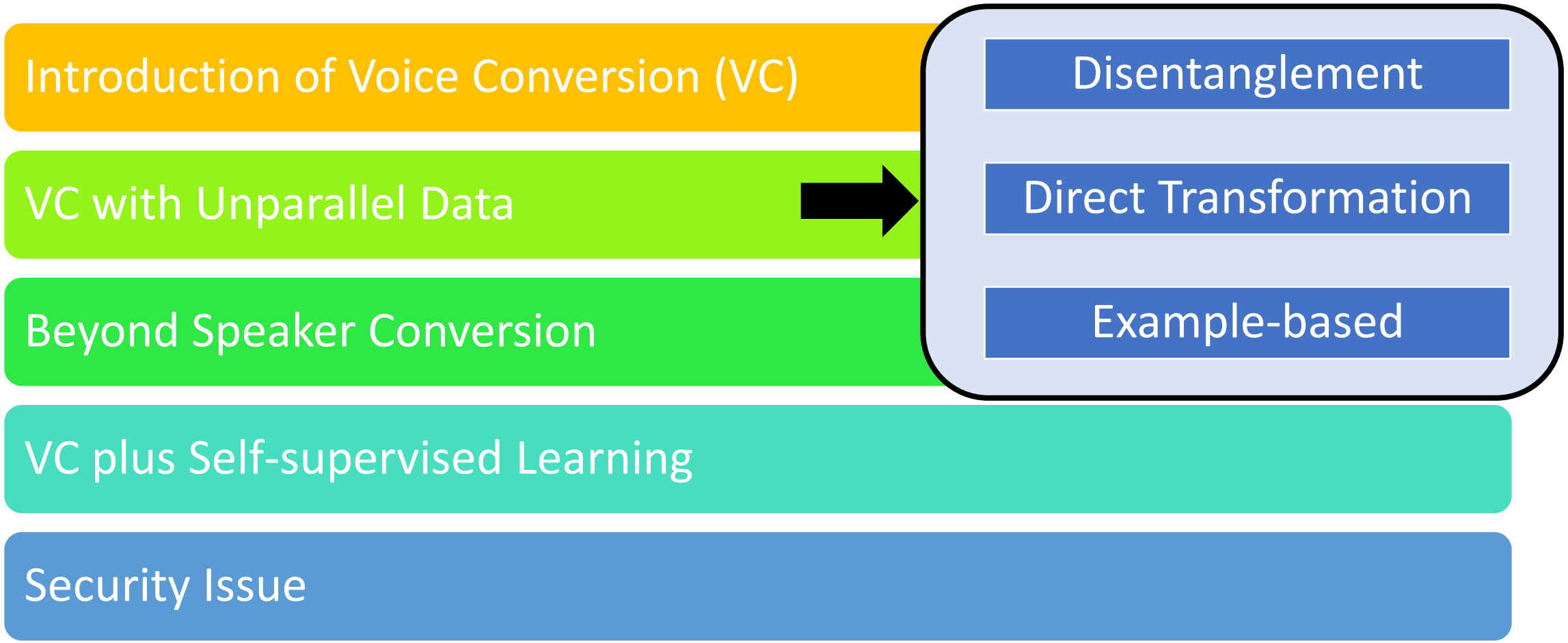


ICASSP



This tutorial focuses on developments over the past three years.

Outline


Introduction of Voice Conversion (VC)
VC with Unparallel Data
Beyond Speaker Conversion
VC plus Self-supervised Learning
Security Issue
Disentanglement
Direct Transformation
Example-based

Outline

Introduction of Voice Conversion (VC)

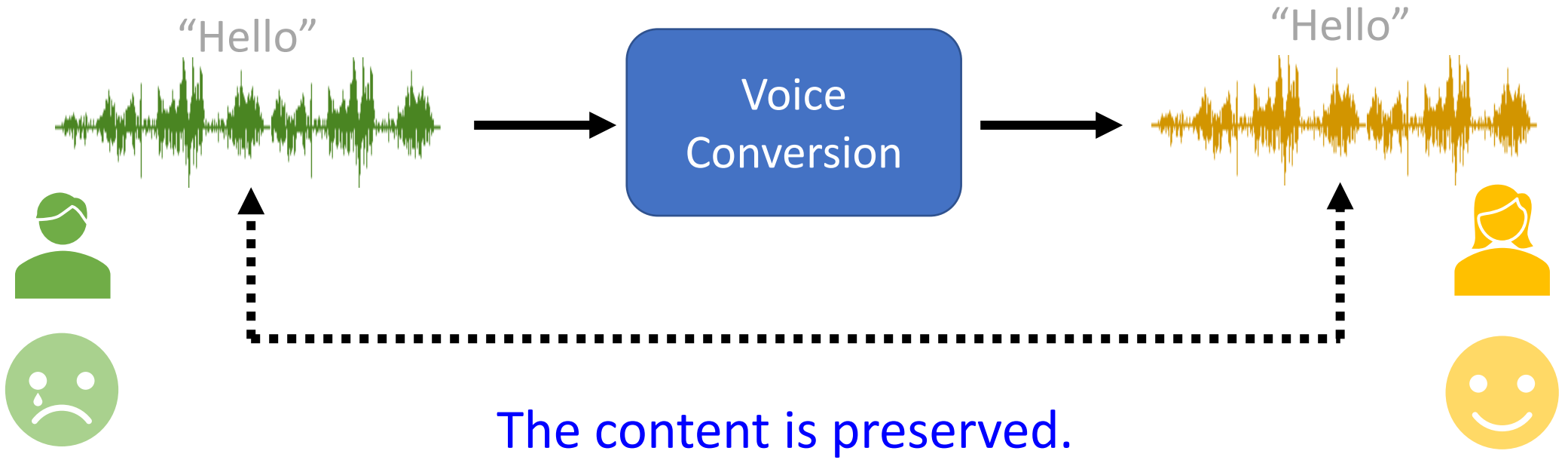
VC with Unparallel Data

Beyond Speaker Conversion

VC plus Self-supervised Learning

Security Issue

What is Voice Conversion (VC)?

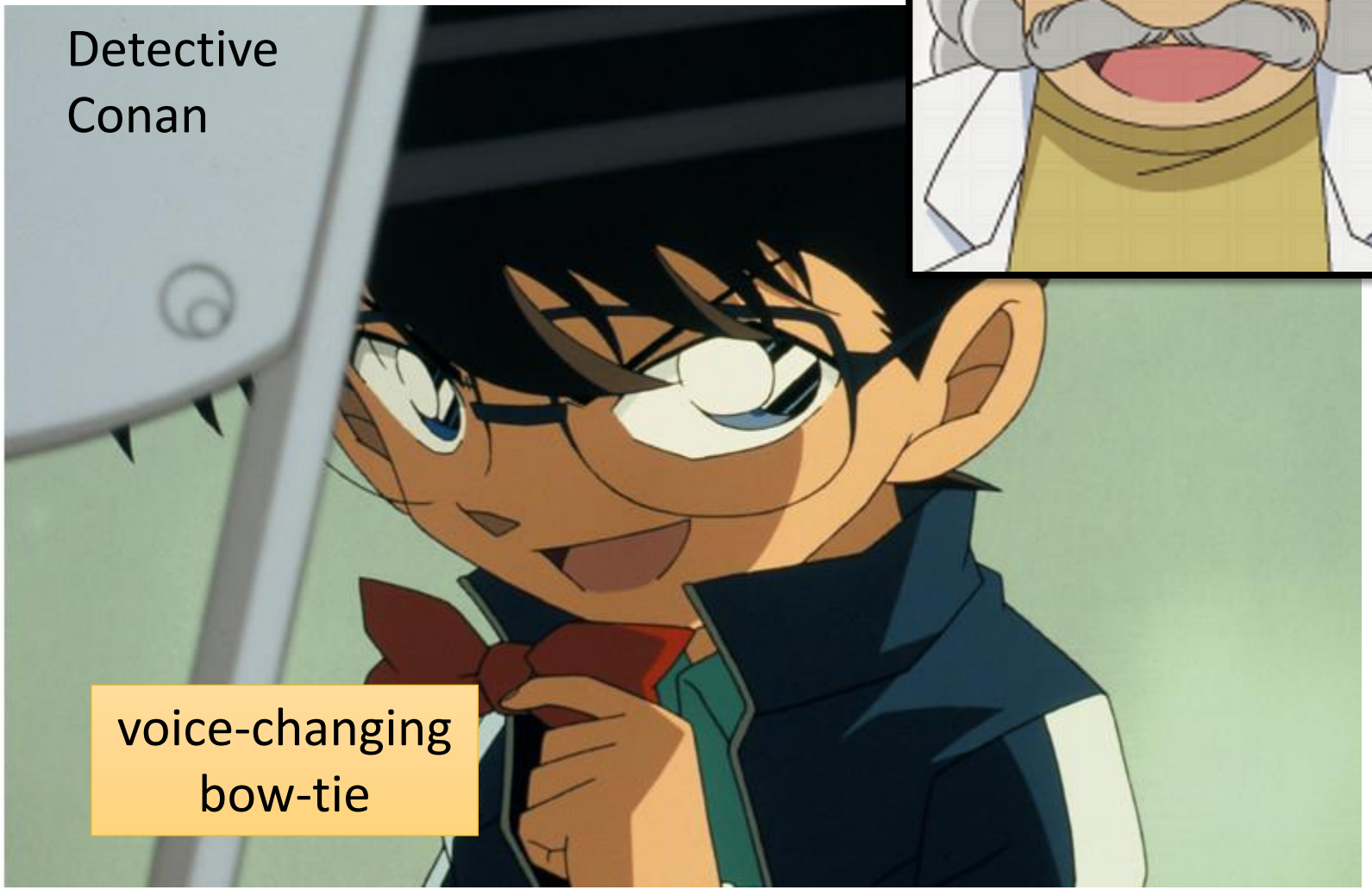


The content is preserved.

Many different aspects can be converted.

What is converted? **Speaker**

Agasa Hiroshi



What is converted? **Speaker**

- The same sentence said by different people has different effect.

Detective Conan

Kogoro Mouri

- Deep Fake: Fool humans / speaker verification system (Back to this issue at the end of the talk)

- Singing voice conversion (Not today)

[Nachmani, et al., INTERSPEECH'19]

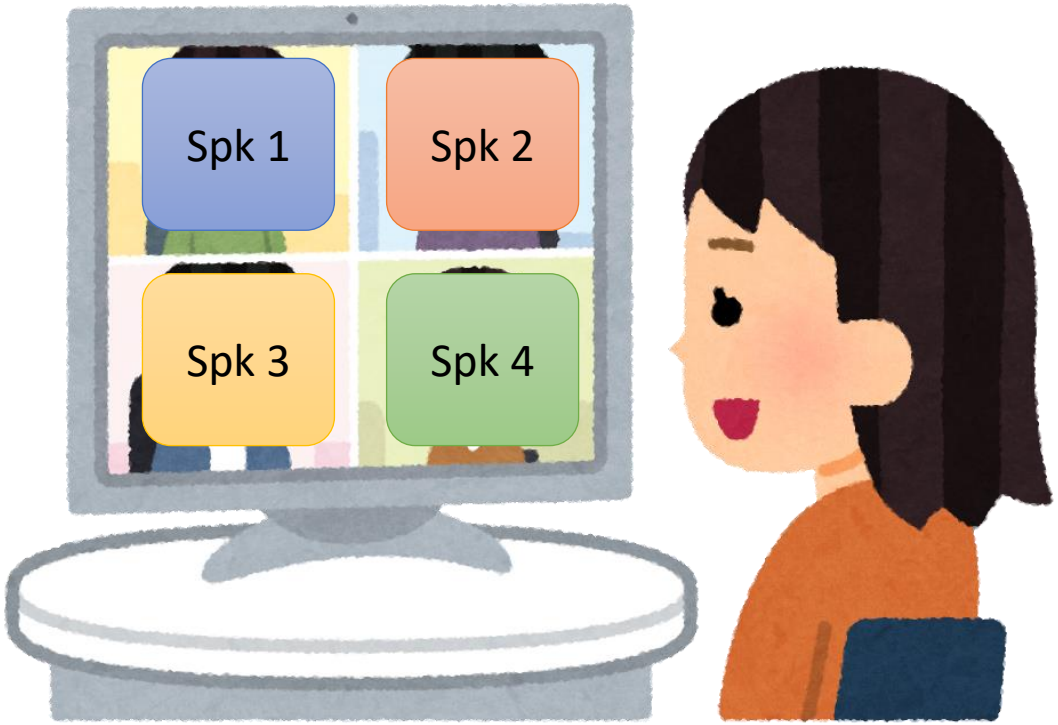
https://enk100.github.io/Unsupervised_Singing_Voice_Conversion/

[Deng, et al., ICASSP'20]

https://tencent-ailab.github.io/pitch-net/

What is converted? **Speaker**

- Privacy Preserving
  - Speech data conveys sensitive speaker attributes.
  - VC as an anonymization method.



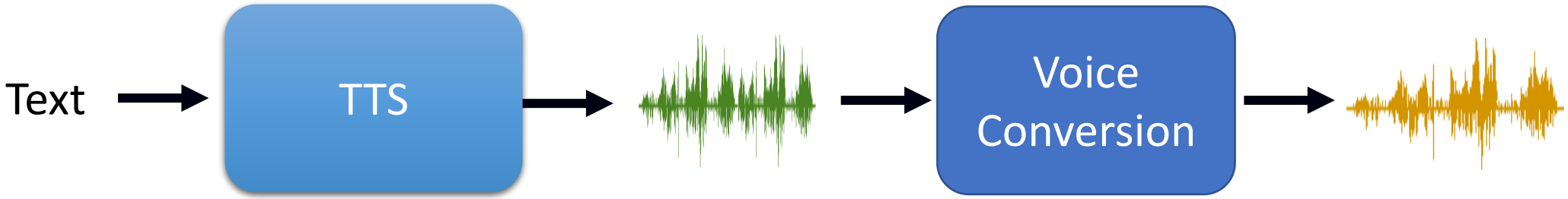
VoicePrivacy Challenge

https://www.voiceprivacychallenge.org/

What is converted? **Speaker**

- One simple way to achieve adaptive TTS

We already talk about adaptive TTS approaches in part 1. But these approaches need to modify TTS model.



Off-the-shelf

What is converted? **Speaking Style**

- Emotion

  [Gao, et al., INTERSPEECH'19]

- Normal-to-Lombard

  [Seshadri, et al., ICASSP'19]

- Whisper-to-Normal

  [Patel, et al., SSW'19]

- Singers vocal technique conversion

  [Luo, et al., ICASSP'20]

Lombard Effect

https://www.fohlio.com/blog/psychology-restaurant-interior-design-part-4-restaurant-acoustics

What is converted? **Speaking Style**

- Emotion

  [Gao, et al., INTERSPEECH'19]

- Normal-to-Lombard

  [Seshadri, et al., ICASSP'19]

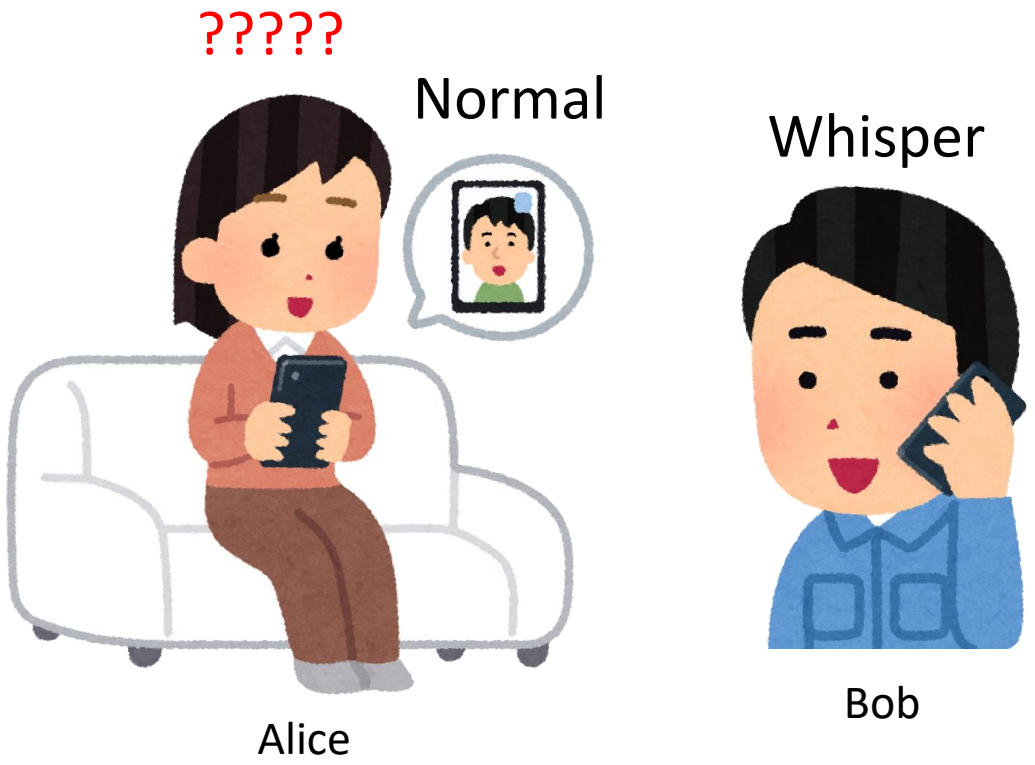
- Whisper-to-Normal

  [Patel, et al., SSW'19]

- Singers vocal technique conversion

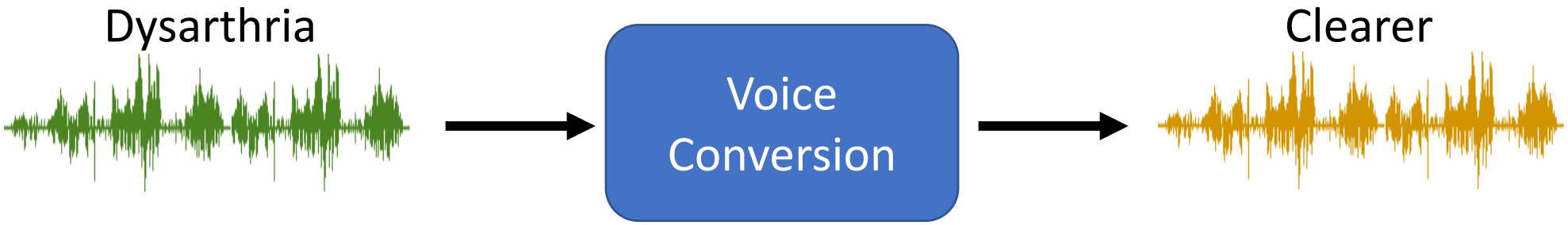
  [Luo, et al., ICASSP'20]

  e.g., 'lip thrill' or 'vibrato'



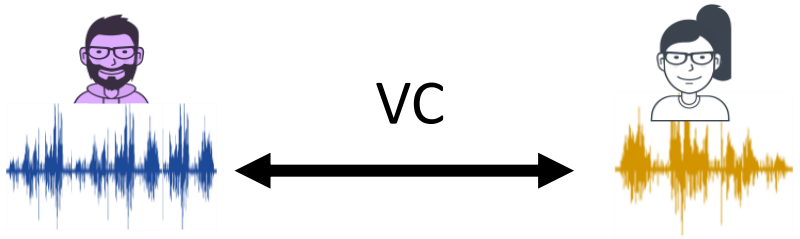
Improving Intelligibility

- Surgical patients who have had parts of their articulators removed
- Dysarthria: speech sound disorder resulting from neurological injury of the motor component of the motor-speech system.



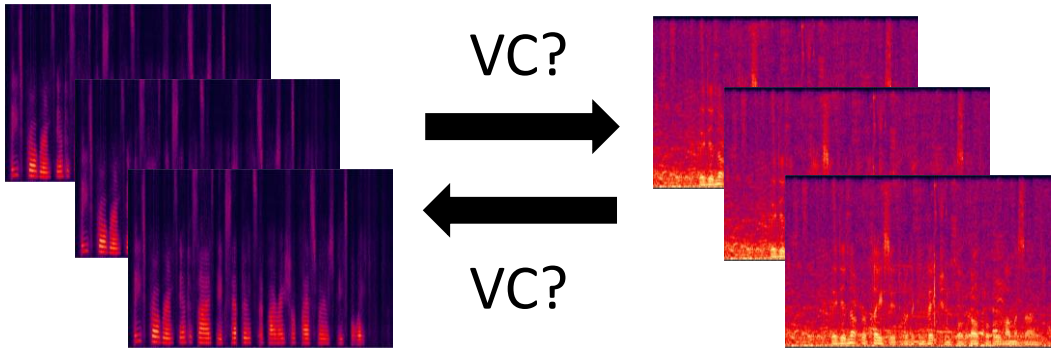
[Biadsy, et al., IS'19]
[Chen et al., IS'19]
[Huang, et al., IS'21]
[Huang, et al., ICASSP'22b]
[Wang, et al., ICASSP'22]

Data Augmentation



Training
Data x 2

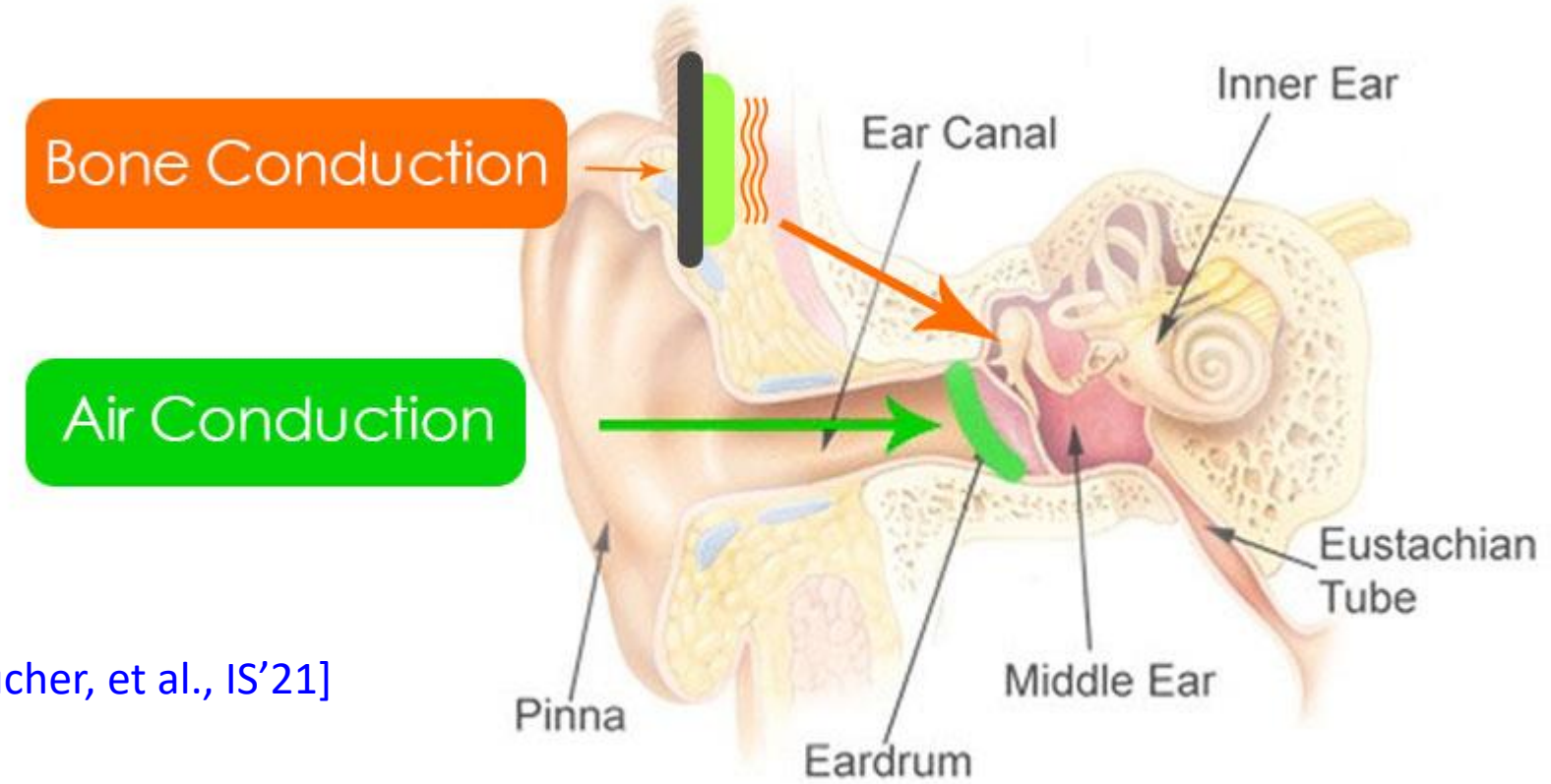
[Keskin, et al., ICML workshop'19]



[Mimura, et al., ASRU 2017]

Airborne to bone-conducted speech

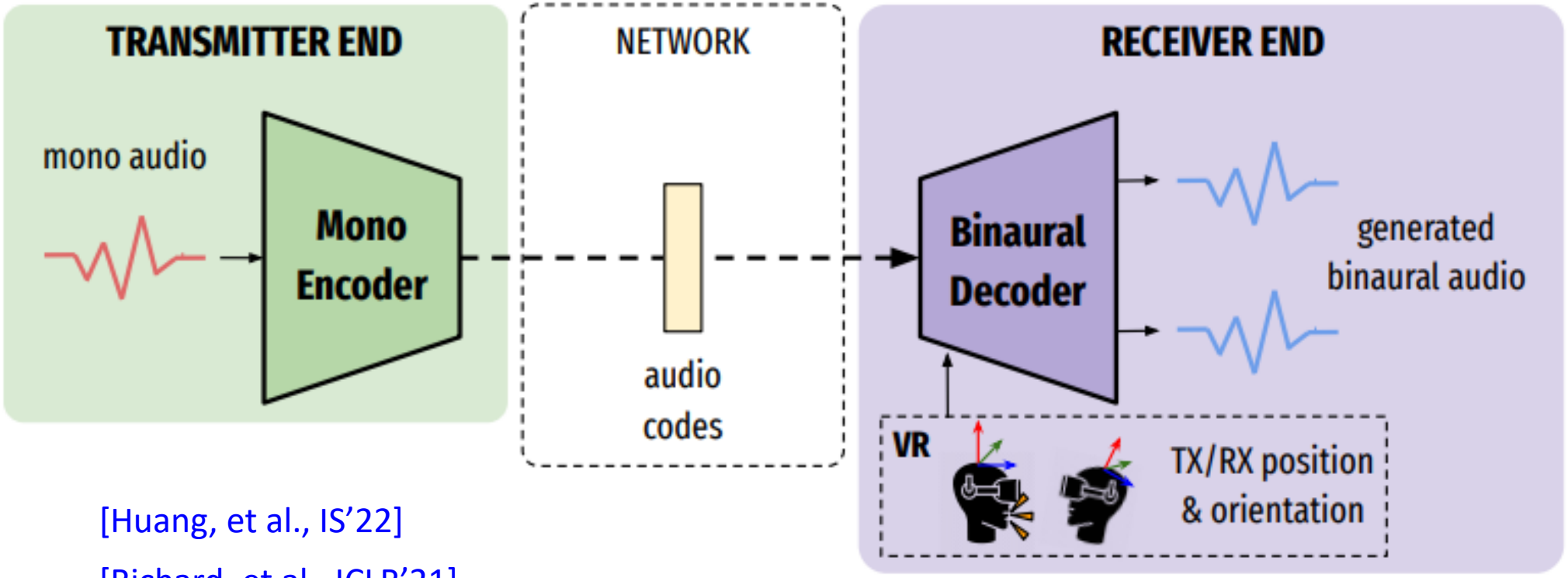
Conversion



[Pucher, et al., IS'21]

Binaural Speech Synthesis

- crucial for acoustic realism and depth perception

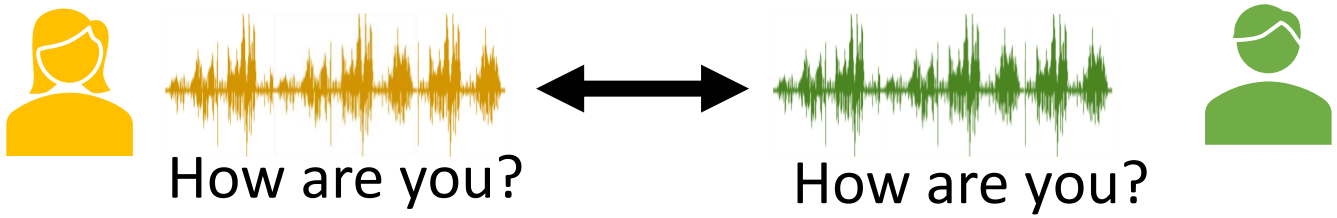


[Huang, et al., IS'22]

[Richard, et al., ICLR'21]

Data Available

Parallel Data



Lack of training data:

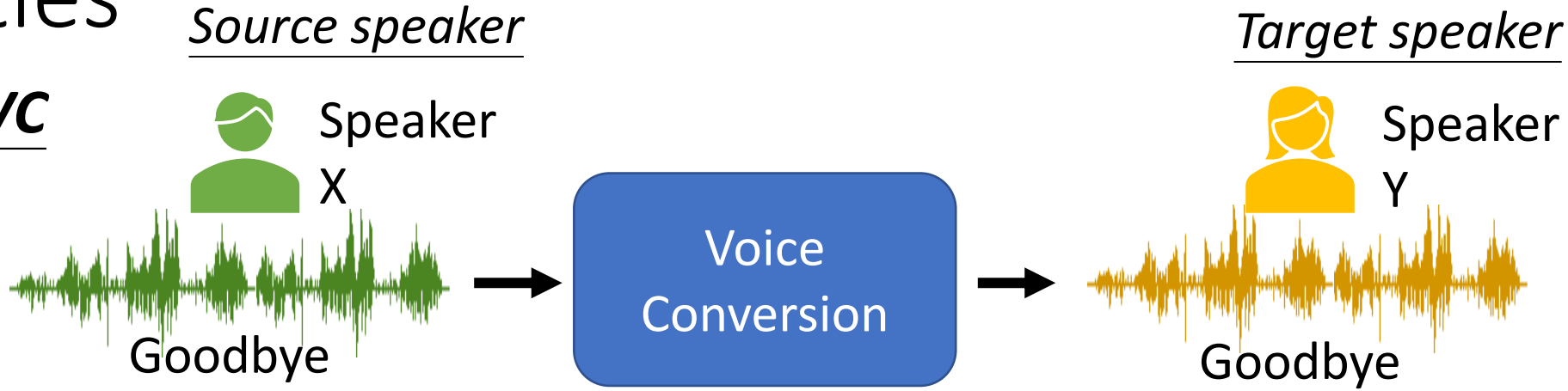
- Model Pre-training [Huang, et al., NTERSPEECH'20]
- Synthesized data! [Biadsy, et al., INTERSPEECH'19]

Unparallel Data

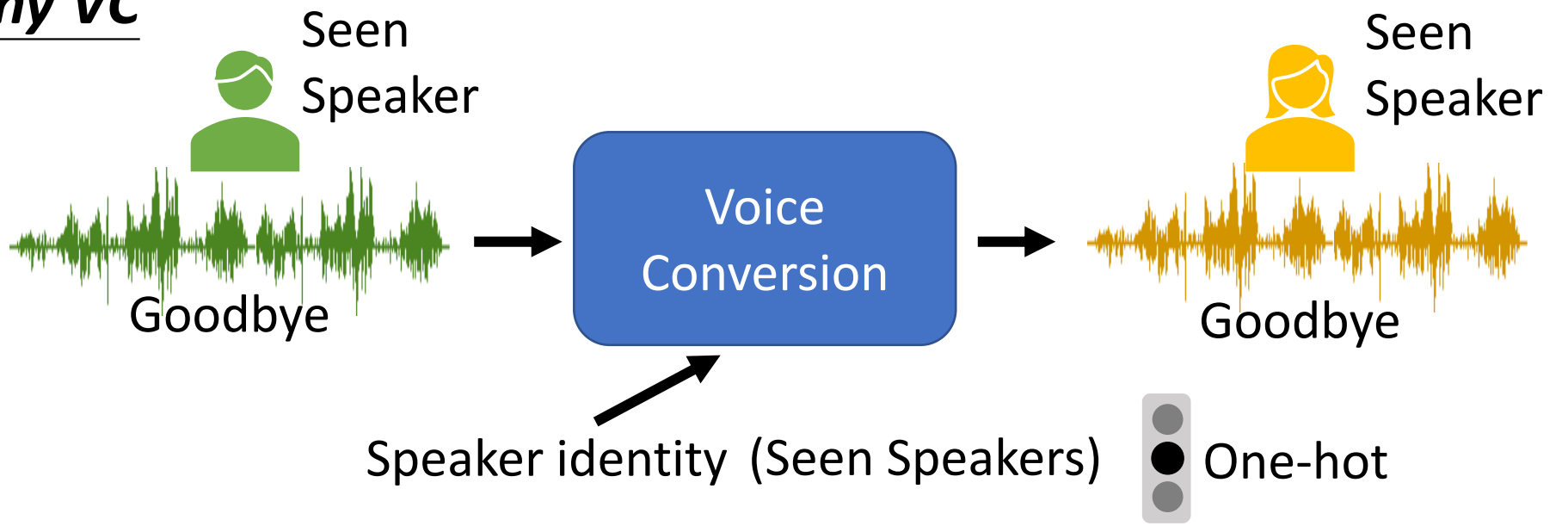
the focus of today's talk

Capabilities

One-to-one VC

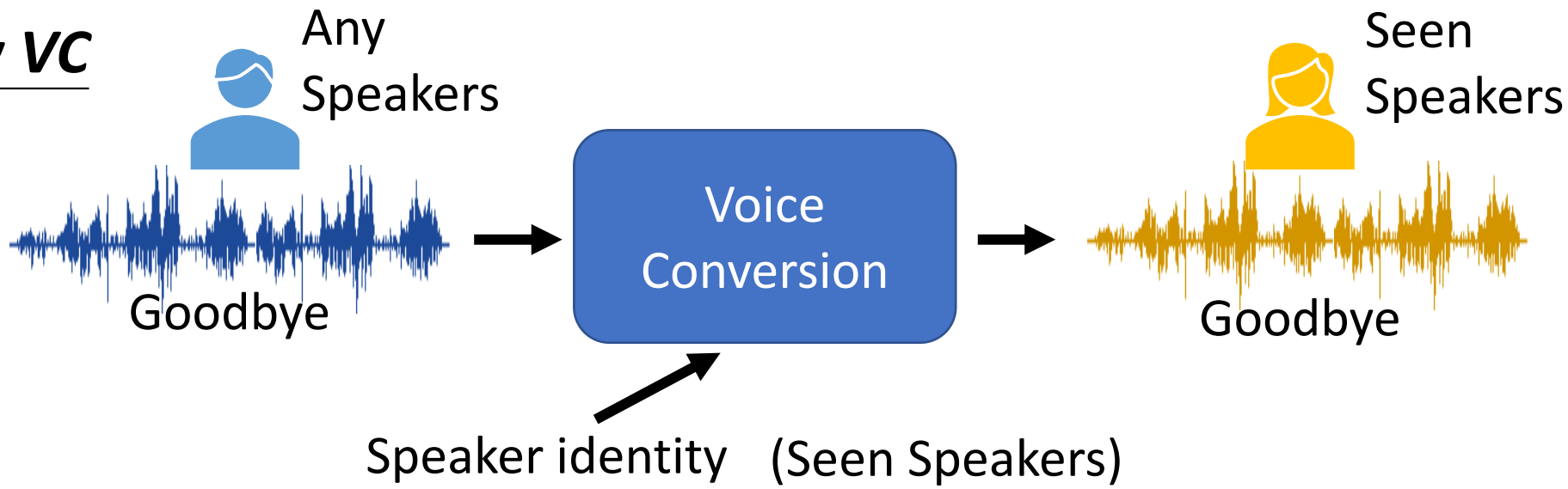


Many-to-many VC

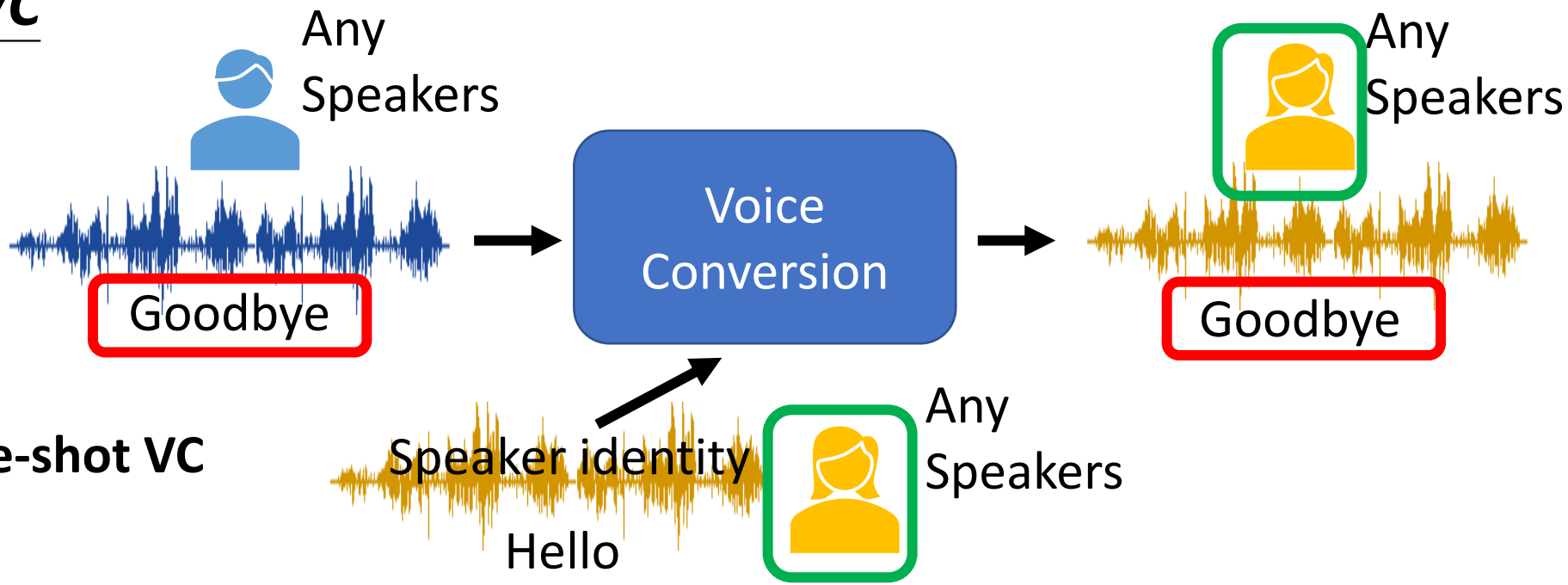


Seen Speakers
= in training data

Any-to-many VC



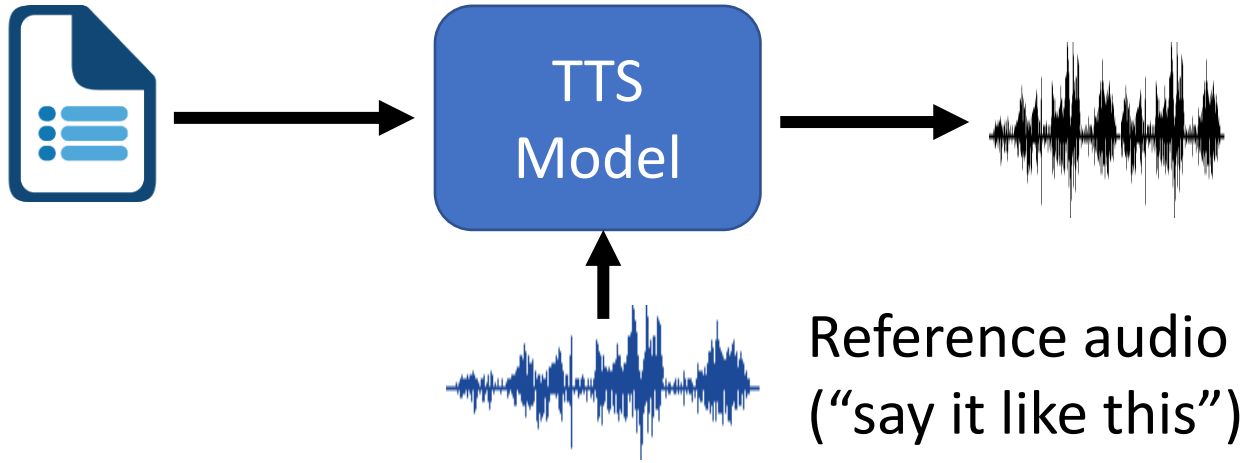
Any-to-any VC



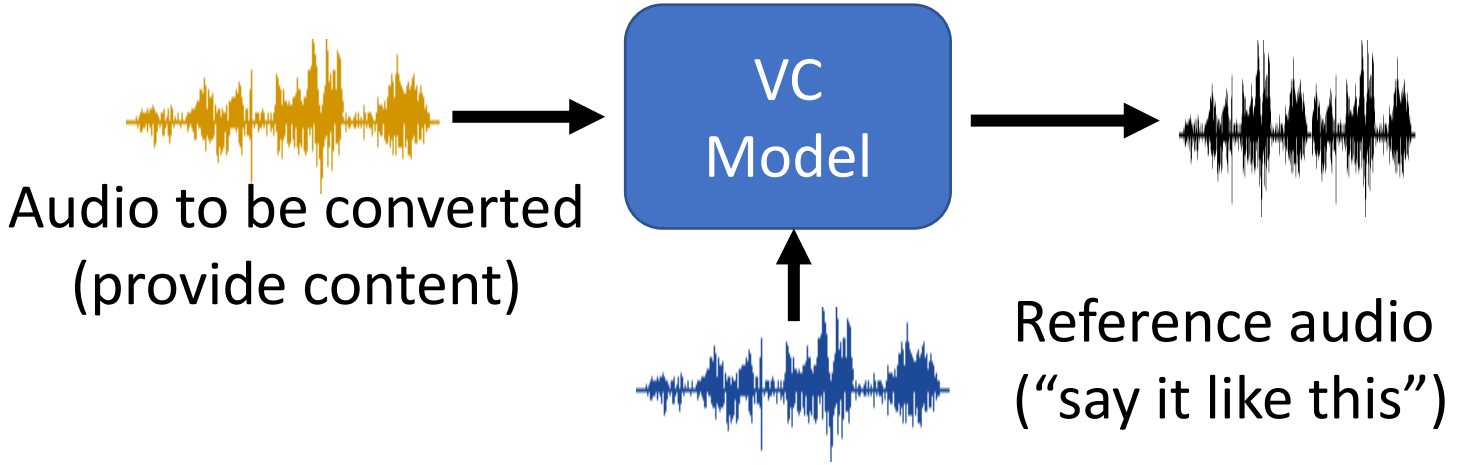
Also known as **one-shot VC**
(or zero-shot VC?)

Adaptive TTS vs. Any-to-any VC

Adaptive TTS



Any-to-any VC



Outline

Much of the discussion here is based on speaker conversion.

(the same idea can be applied to other types of conversions)

Introduction of Voice Conversion (VC)

VC with Unparallel Data

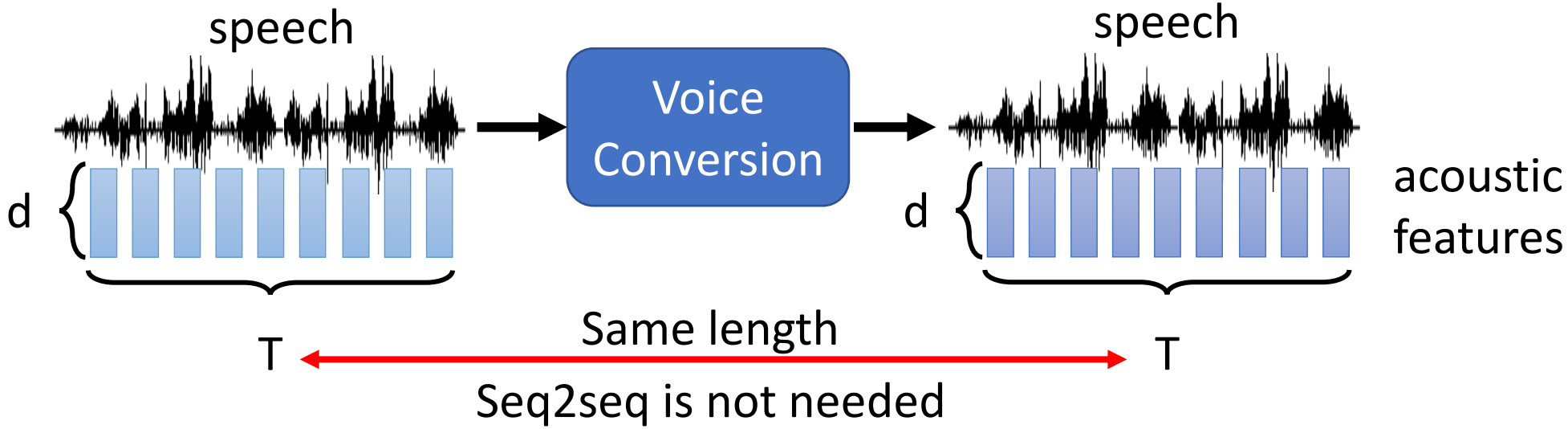
Beyond Speaker Conversion

VC plus Self-supervised Learning

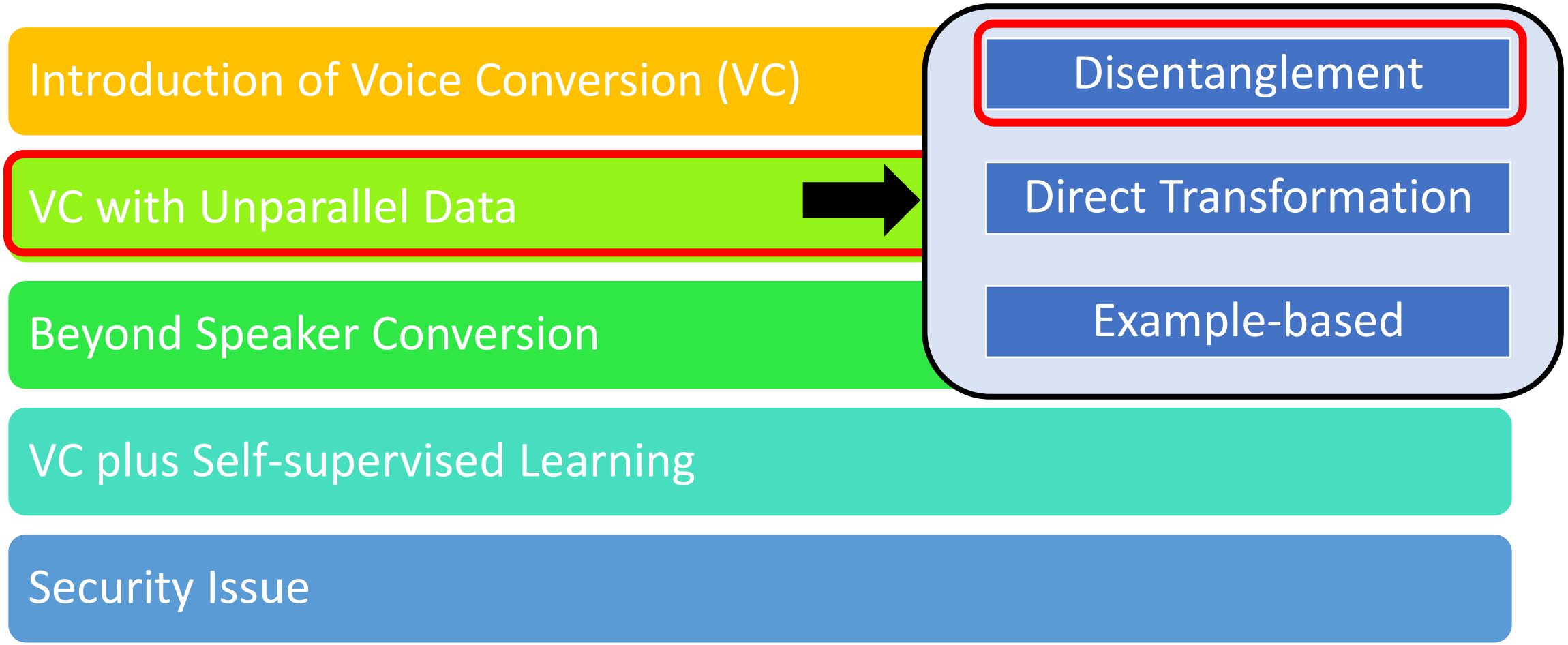
Security Issue

VC with Unparallel Data

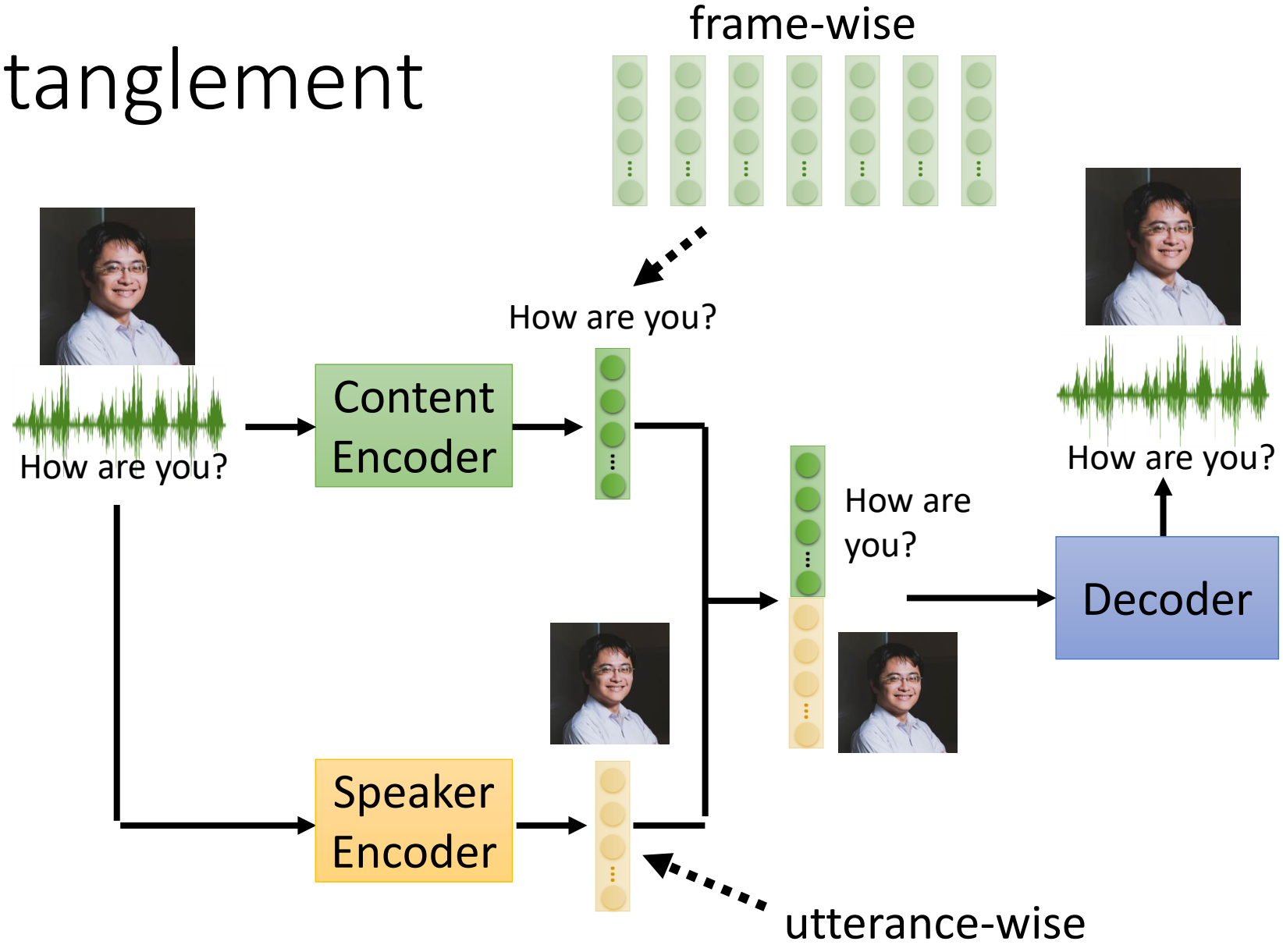
In most real implementations ……



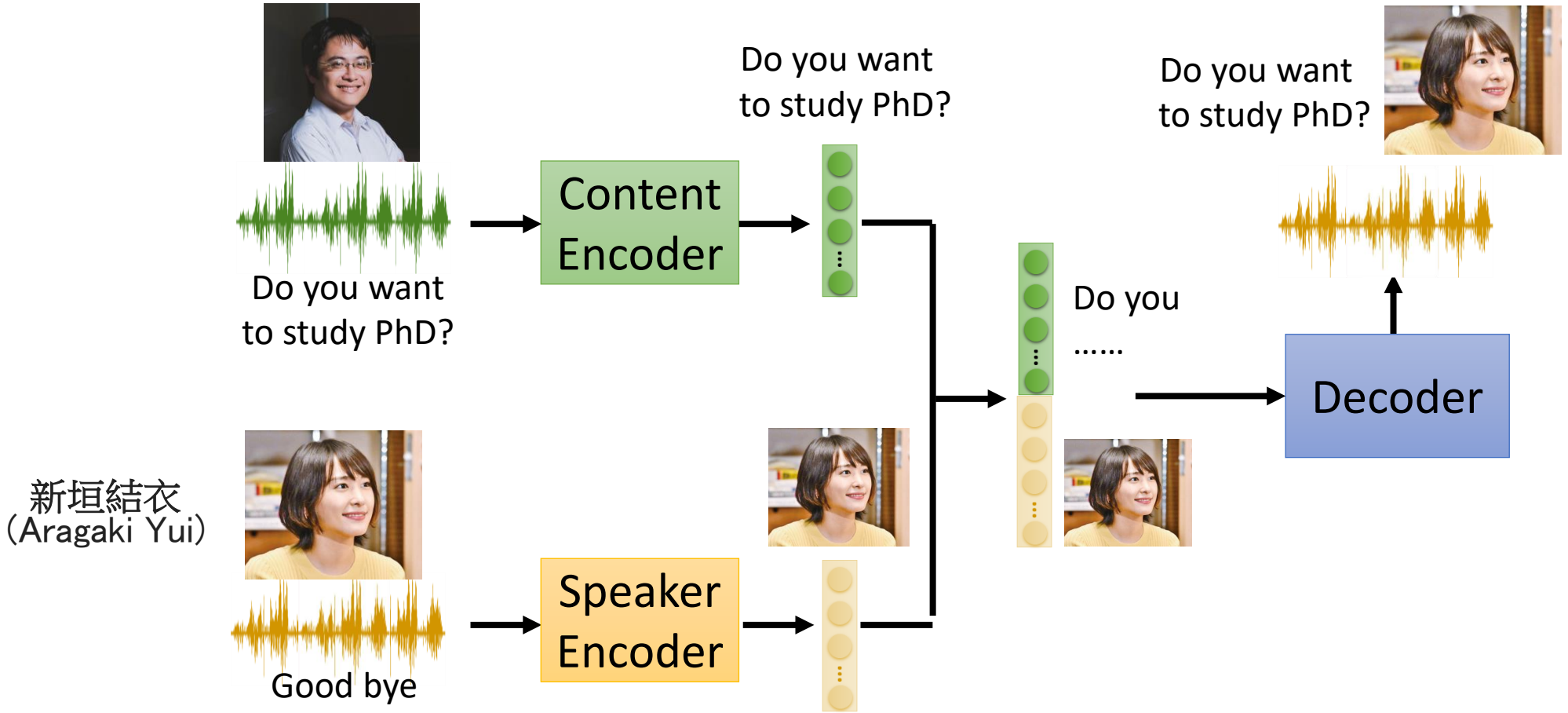
Outline


Introduction of Voice Conversion (VC)
VC with Unparallel Data
Beyond Speaker Conversion
VC plus Self-supervised Learning
Security Issue
Disentanglement
Direct Transformation
Example-based

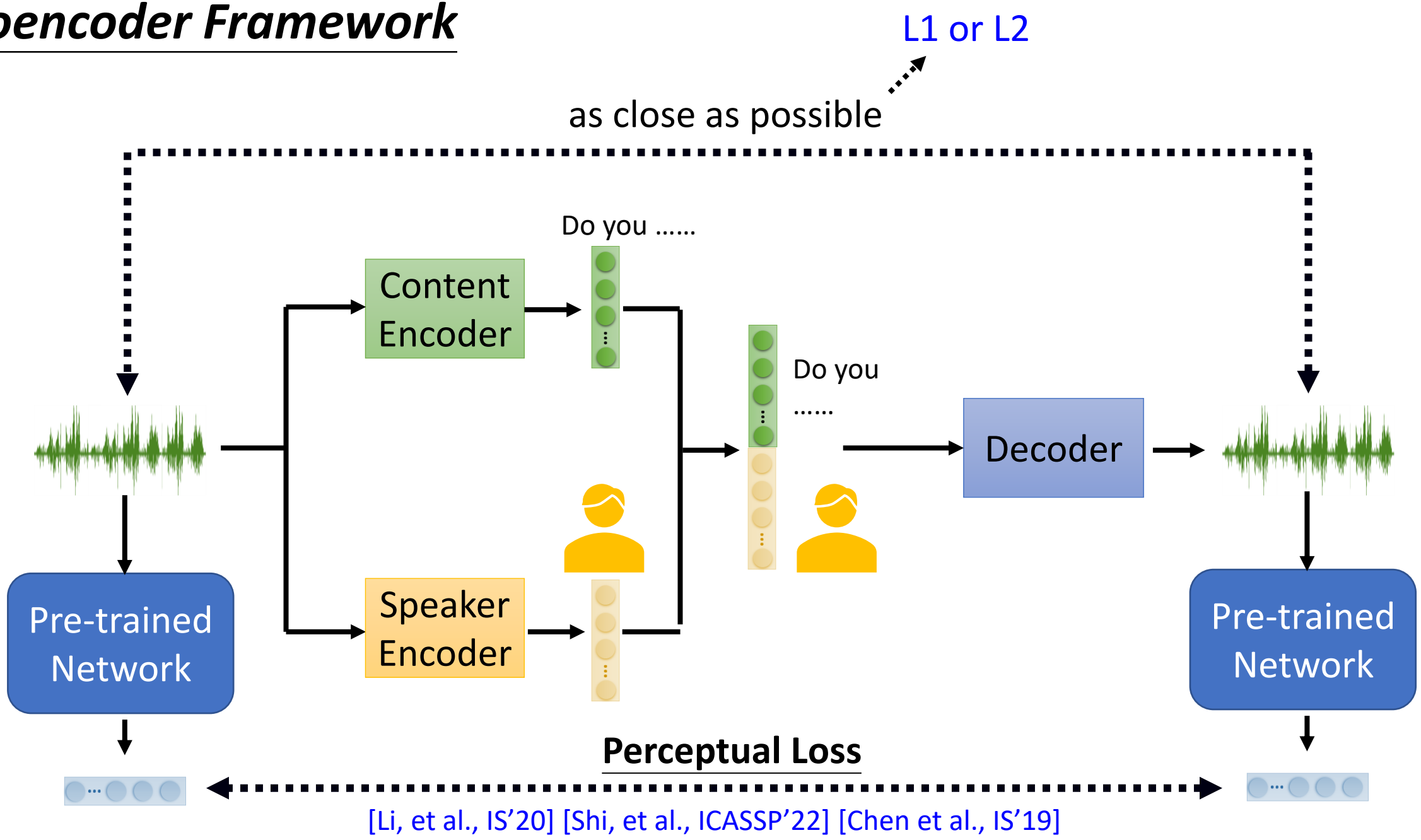
Disentanglement



Disentanglement

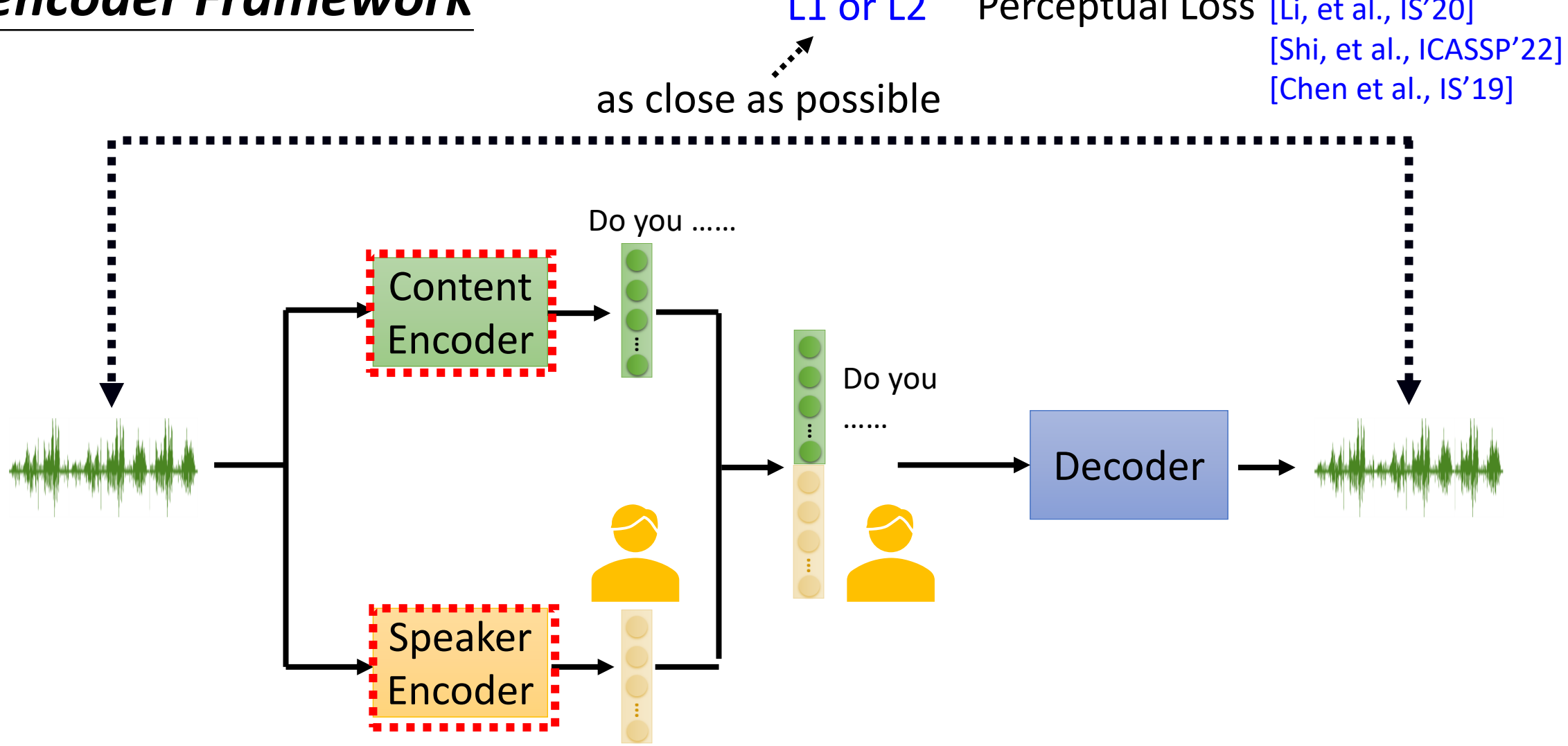


Autoencoder Framework



[Li, et al., IS'20] [Shi, et al., ICASSP'22] [Chen et al., IS'19]

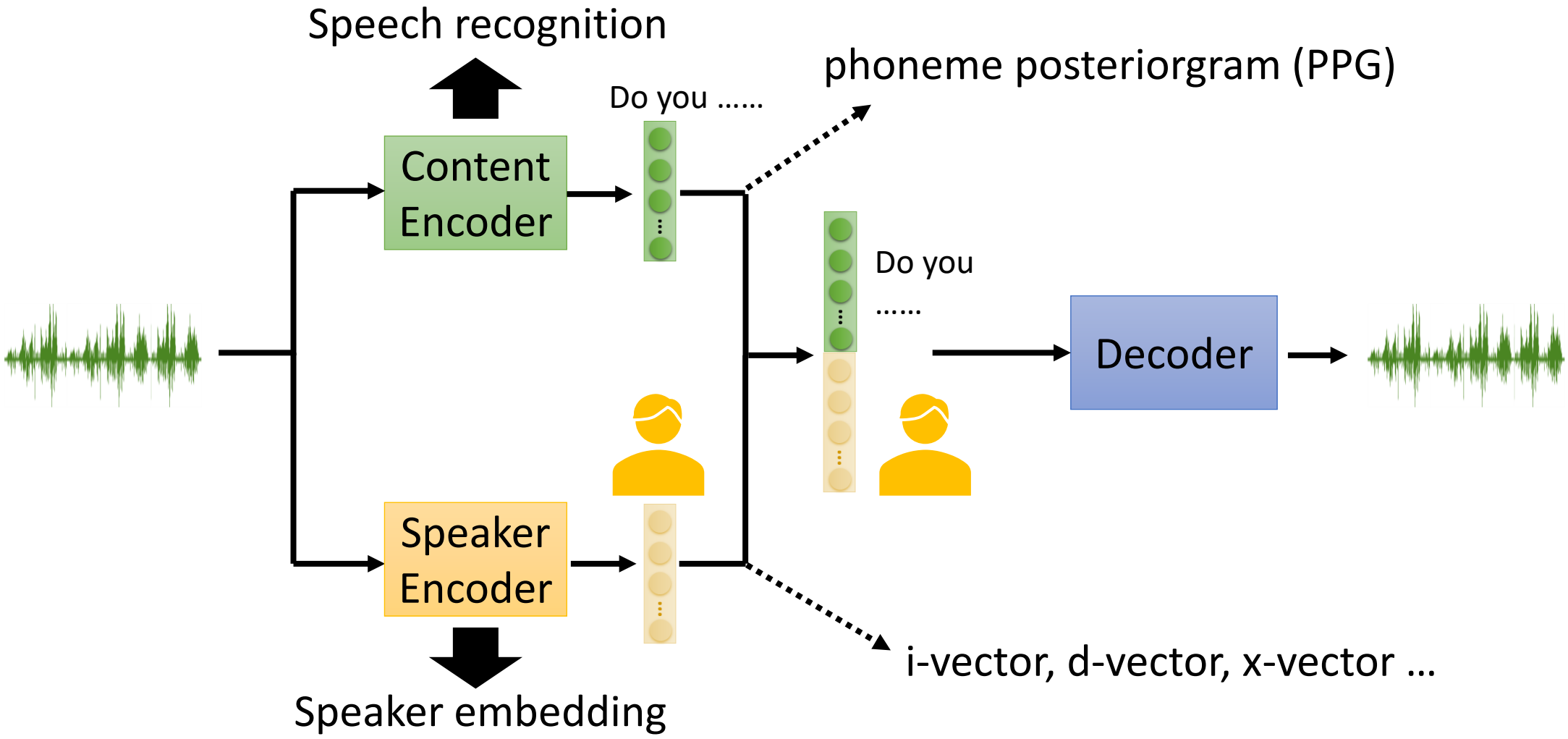
Autoencoder Framework



How can you make one encoder for content and one for speaker?

Initializing Encoders Properly

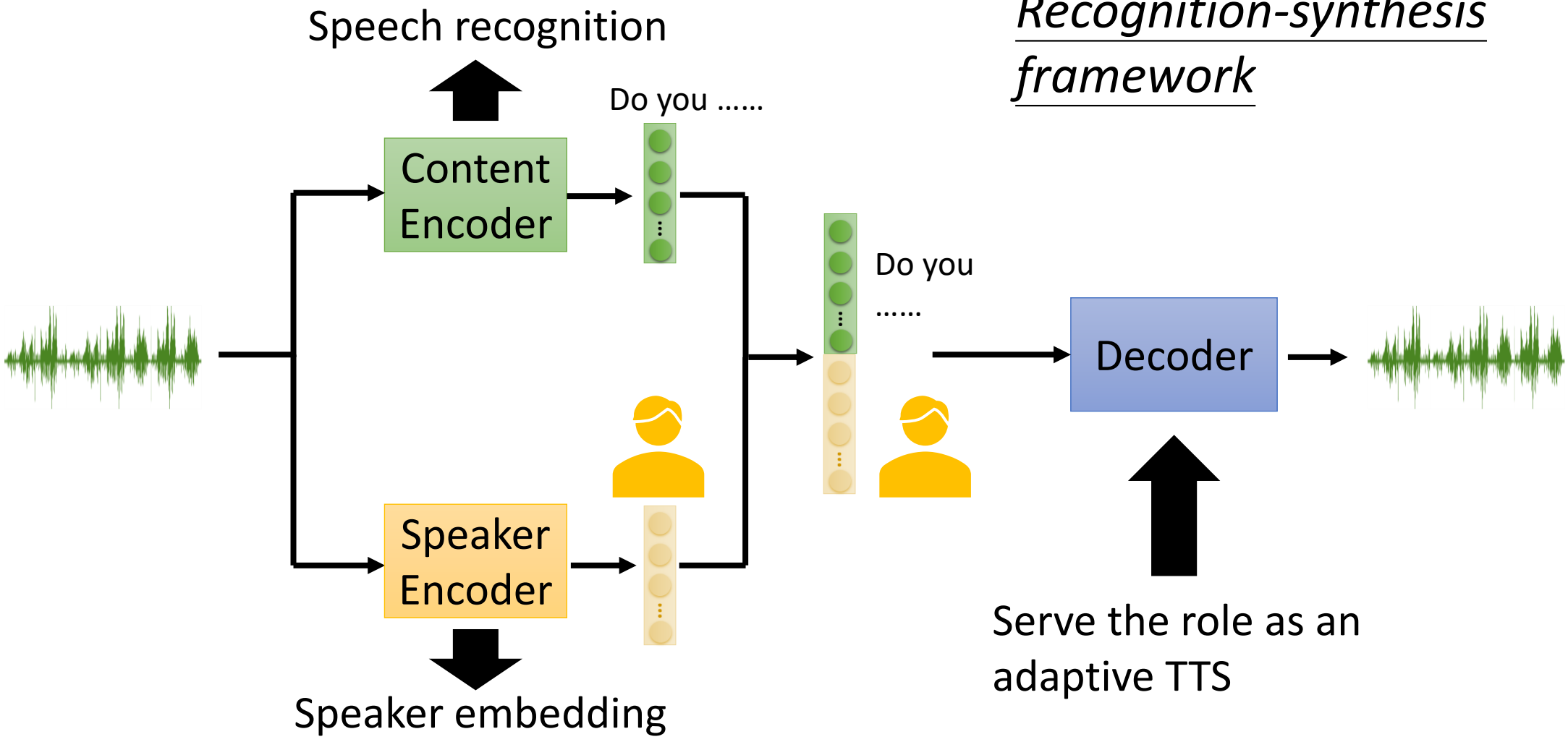
[Sun, et al., ICME'16] [Liu, et al., INTERSPEECH'18]



[Hsu, et al., APSIPA'16] [Qian, et al., ICML'19] [Liu, et al., INTERSPEECH'18]

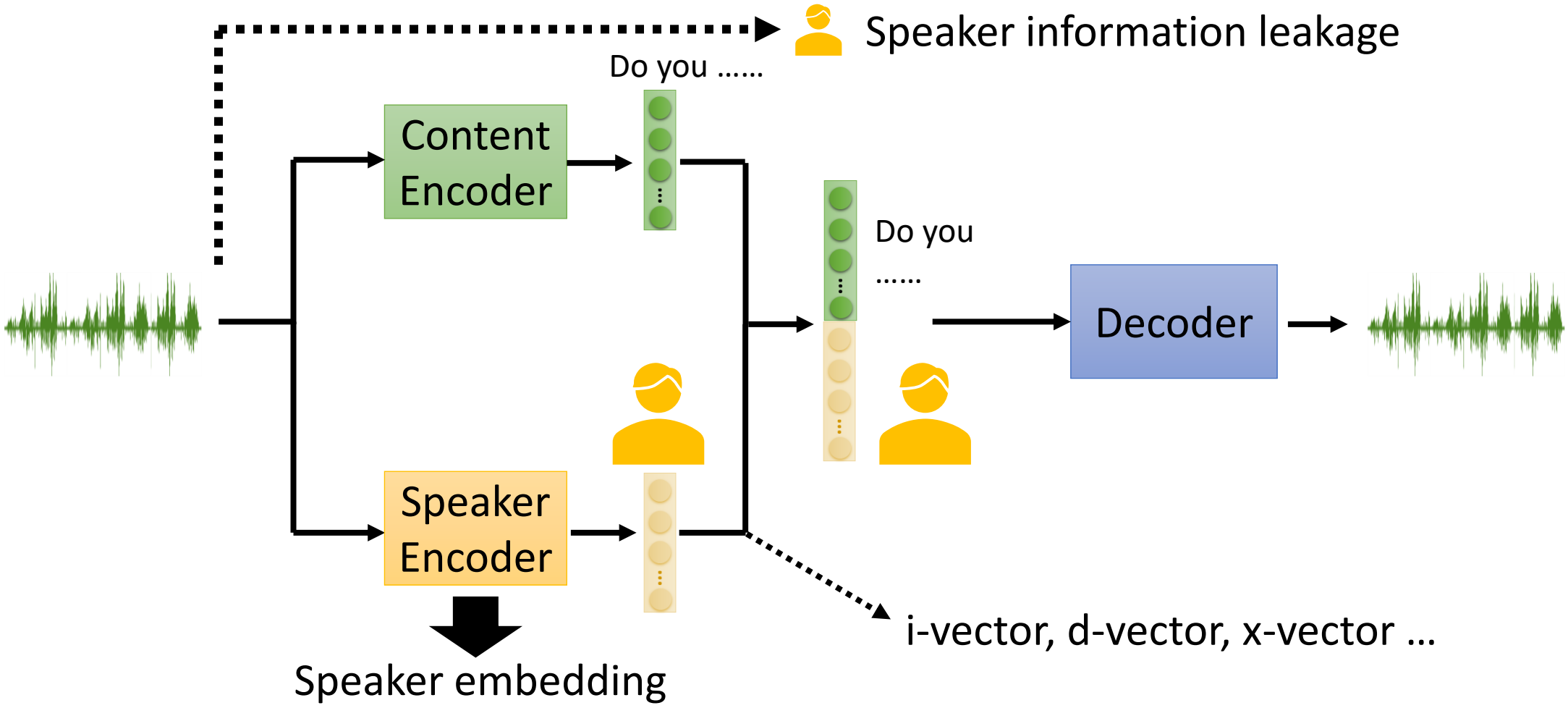
Initializing Encoders Properly

Also known as Recognition-synthesis framework



Initializing Encoders Properly

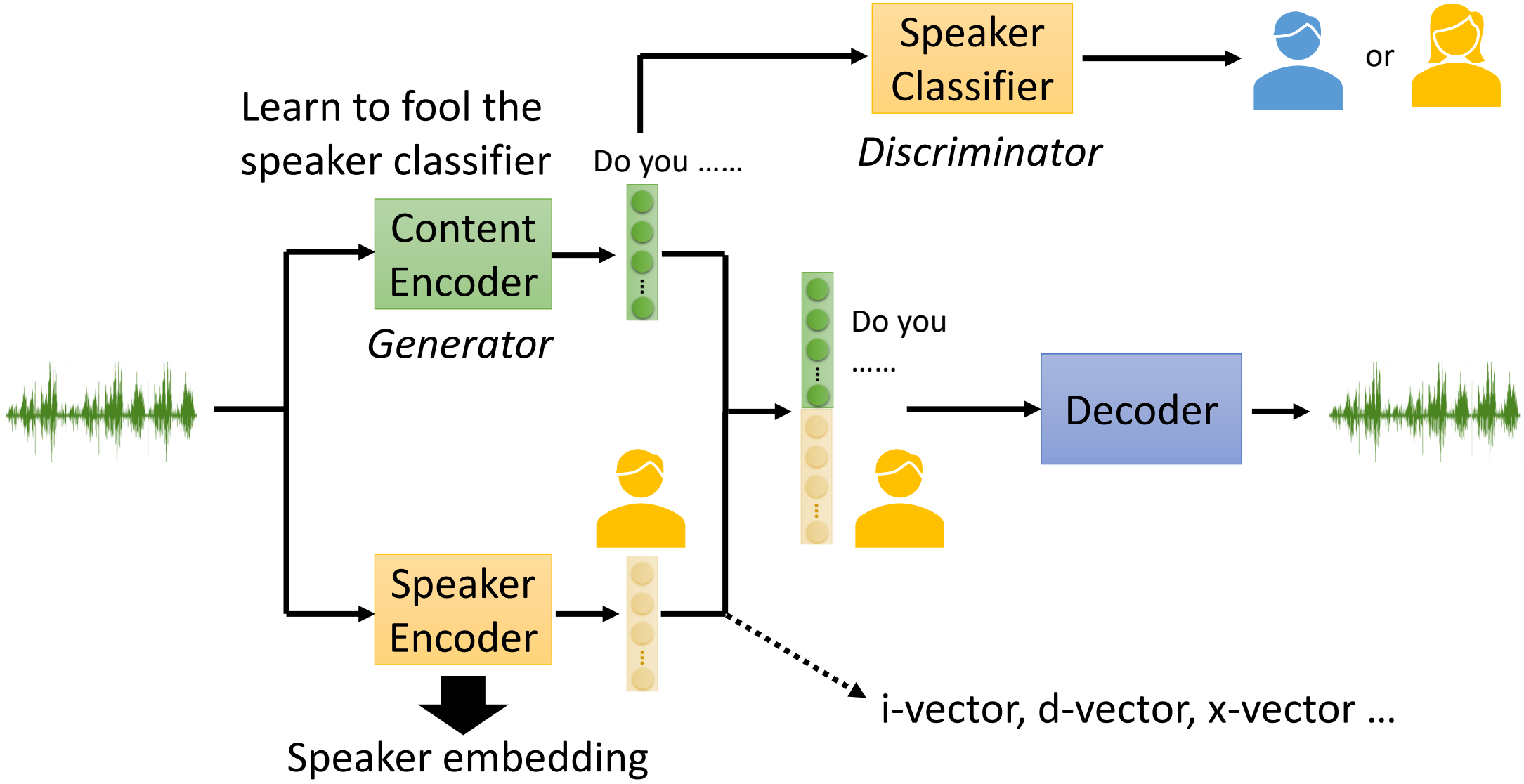
If speech recognizer is not available ……



[Hsu, et al., APSIPA'16] [Qian, et al., ICML'19] [Liu, et al., INTERSPEECH'18]

Adversarial Training

Speaker classifier and encoder are learned iteratively



[Hsu, et al., APSIPA'16] [Qian, et al., ICML'19] [Liu, et al., INTERSPEECH'18]

Information Bottleneck

Auto VC: control dimension

[Qian, et al., ICML'19]

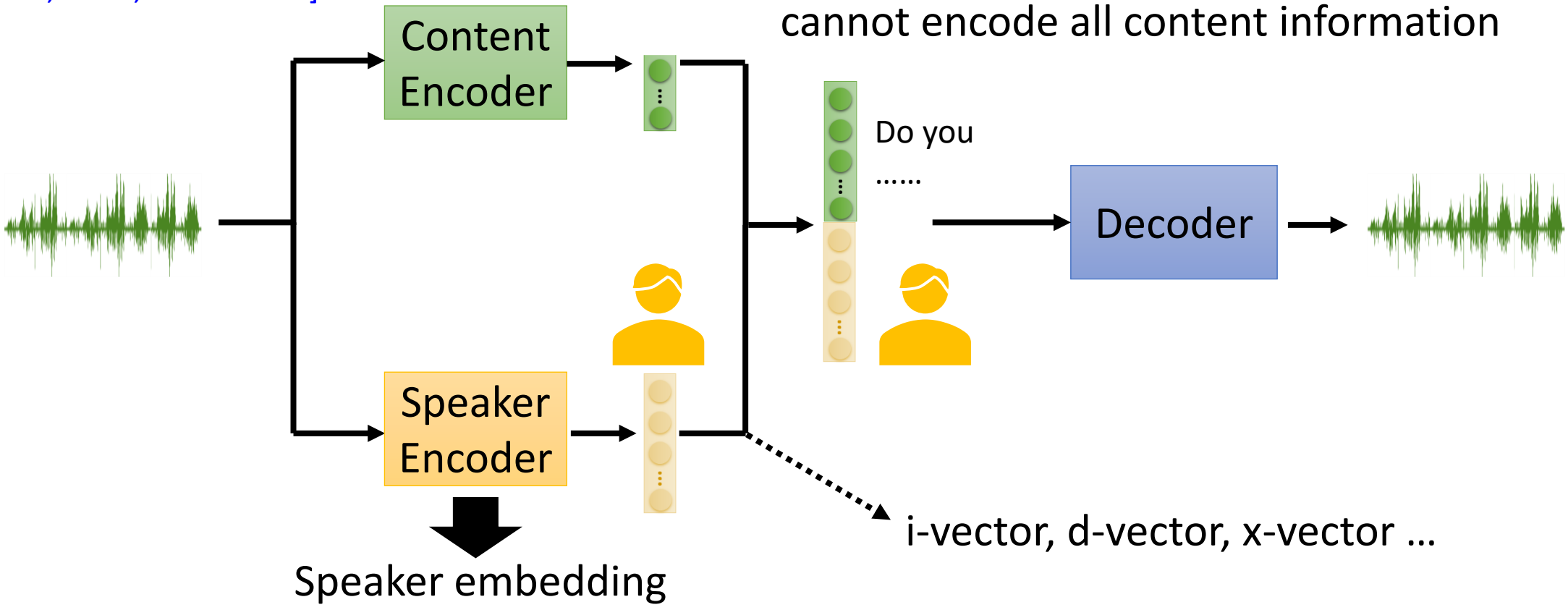
Again VC: Activation function

[Chen, et al., ICASSP'21a]

Too wide dimension: content encoder also encode speaker information

Decrease dimension: squeeze out speaker information

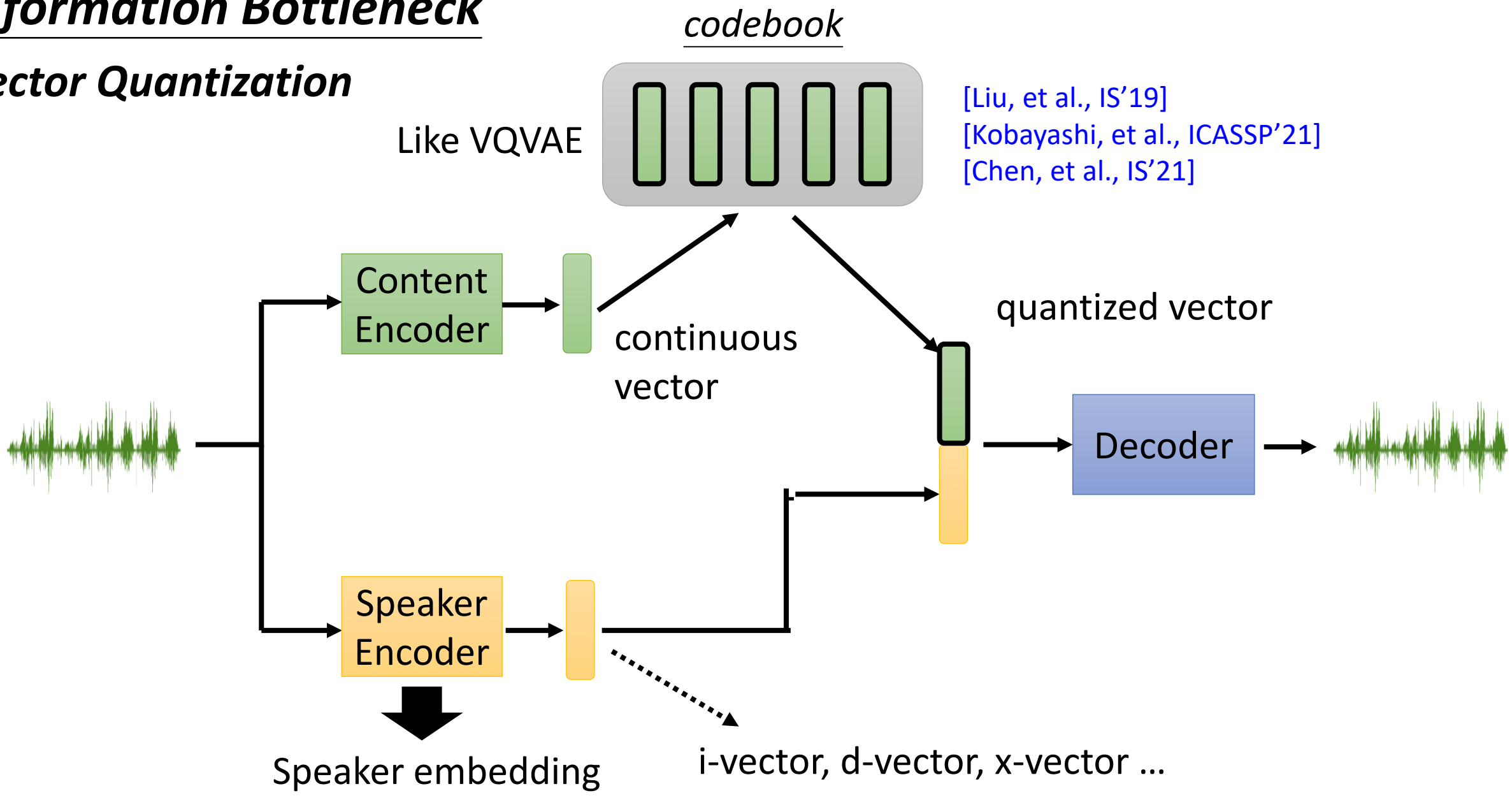
Too narrow dimension: Content encoder cannot encode all content information



[Hsu, et al., APSIPA'16] [Qian, et al., ICML'19] [Liu, et al., INTERSPEECH'18]

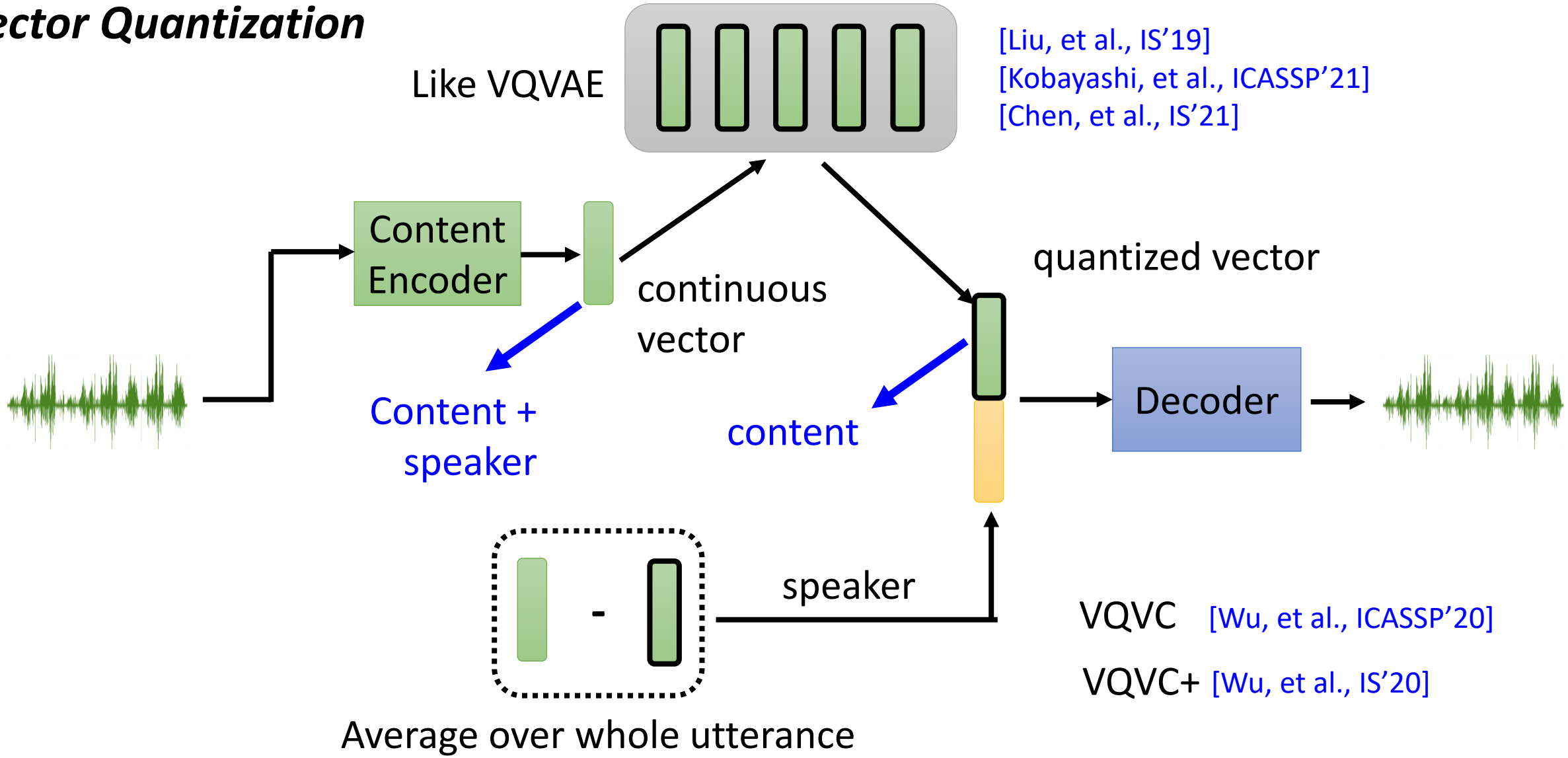
Information Bottleneck

Vector Quantization

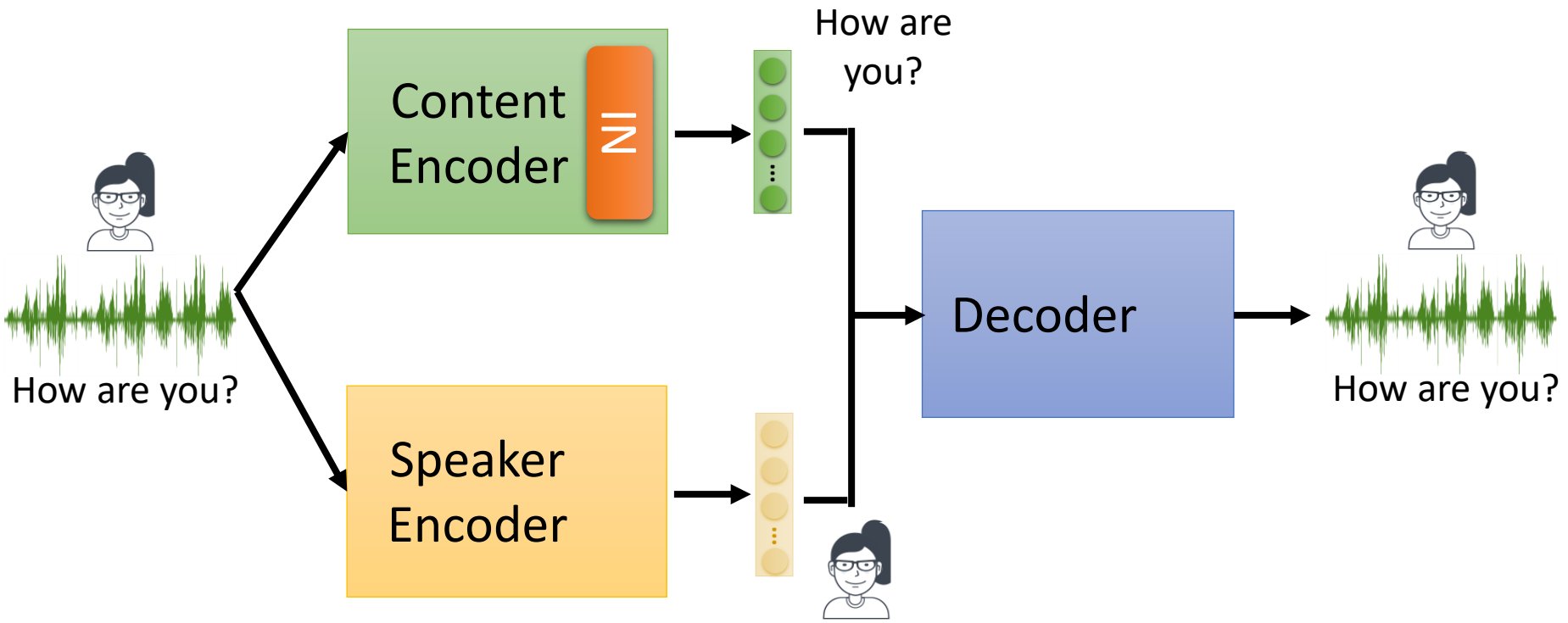


Information Bottleneck

Vector Quantization



Designing network architecture

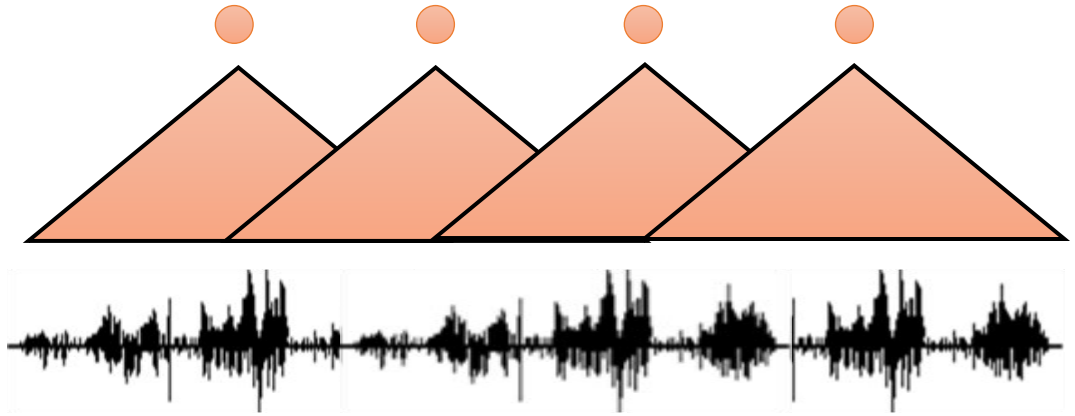


IN = instance normalization (remove speaker information)

Designing network architecture

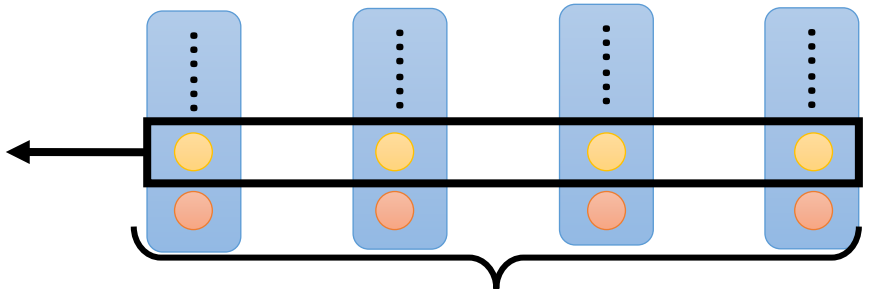
IN = instance normalization (remove speaker information)

Content Encoder



Designing network architecture

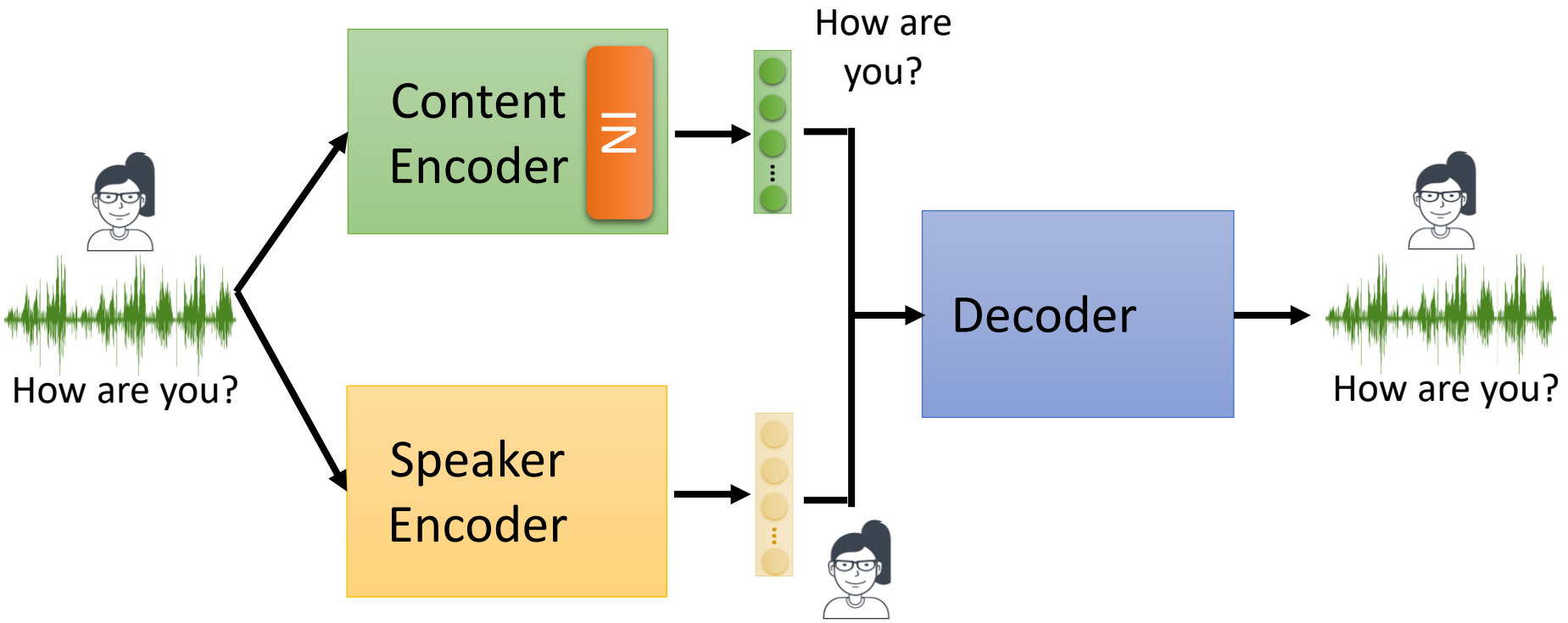
Each channel has zero mean and unit variance



Normalize for each channel

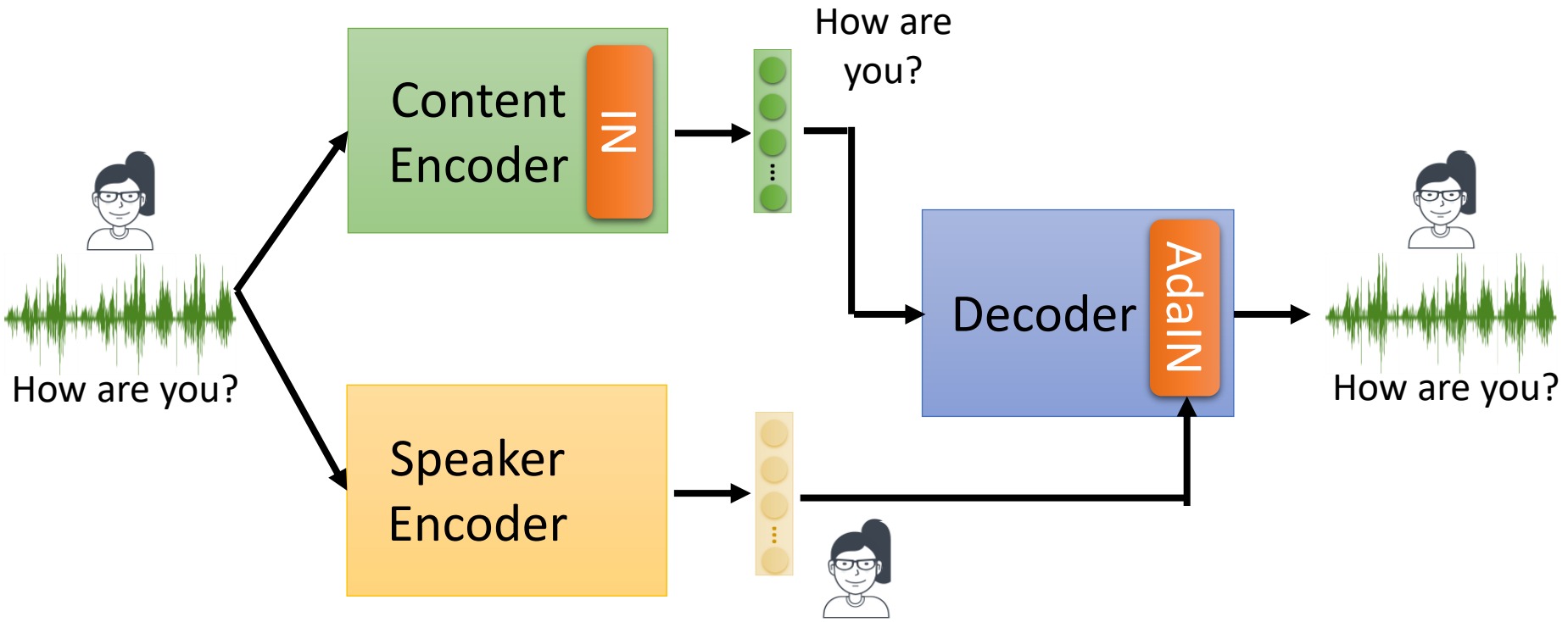
Content Encoder

Designing network architecture



IN = instance normalization (remove speaker information)

Designing network architecture



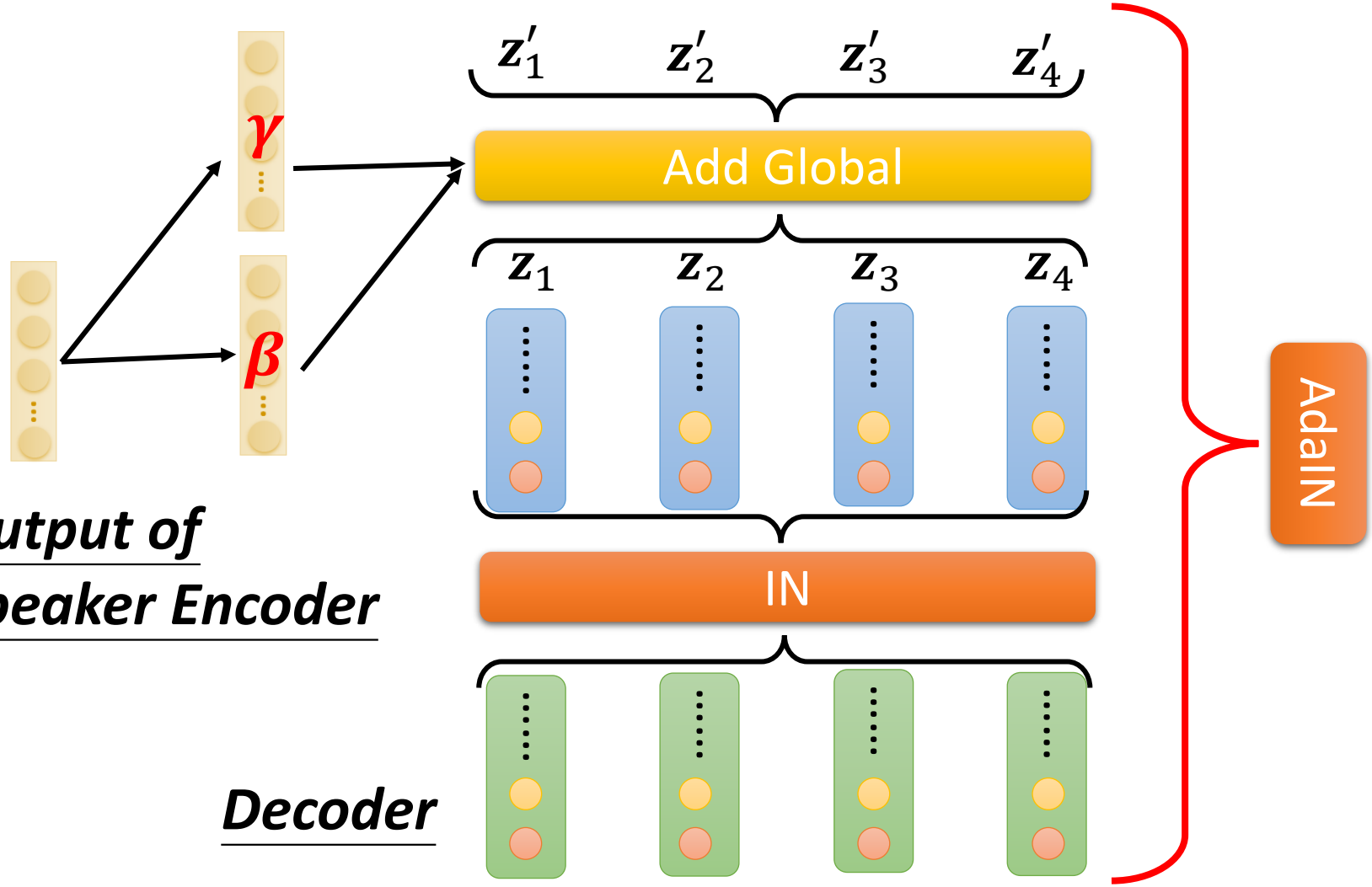
IN = instance normalization (remove speaker information)

AdaIN = adaptive instance normalization
(only influence speaker information)

AdaIN = adaptive instance normalization

(only influence speaker information)

$$z_i' = \gamma \odot z_i + \beta$$

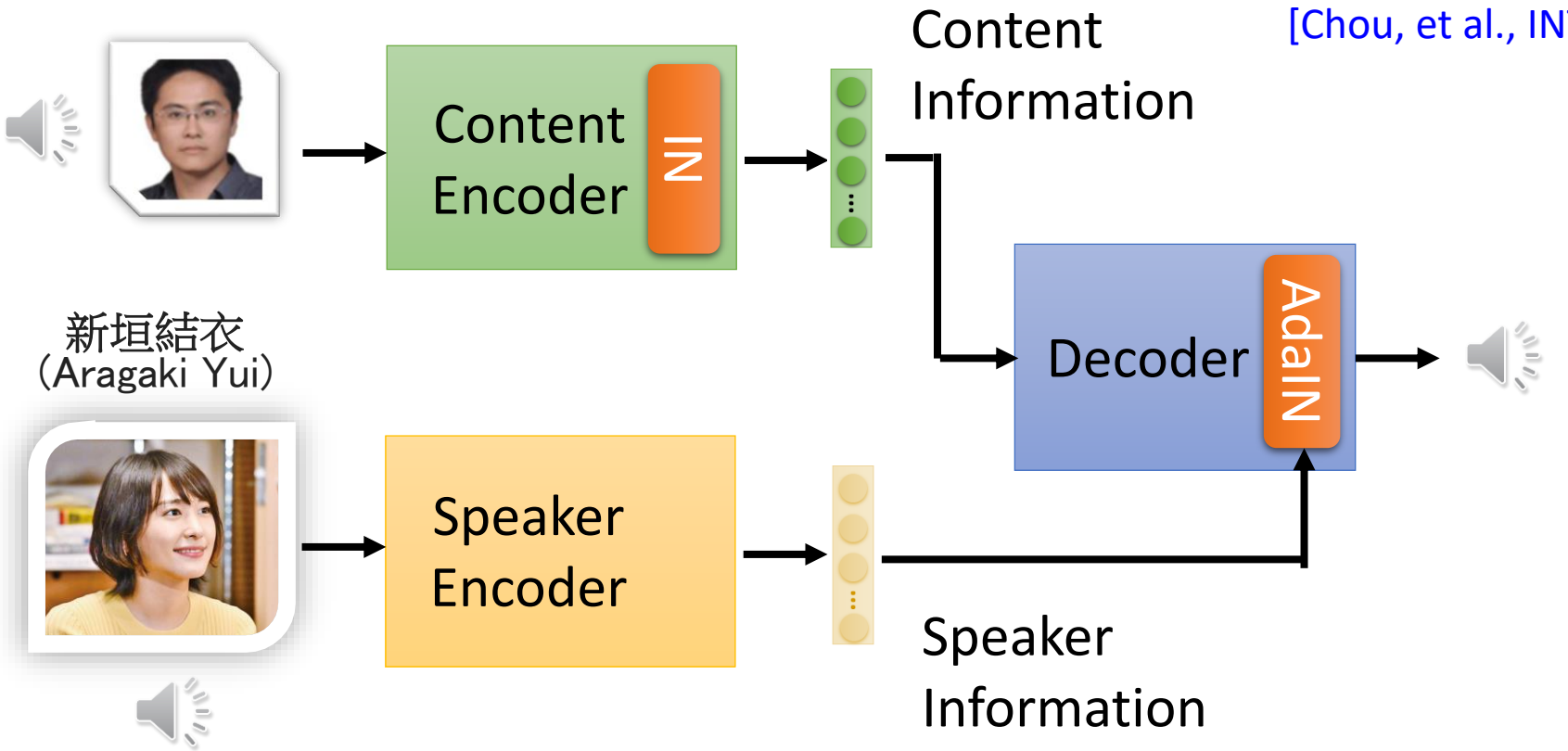


Designing network architecture

The speakers are **unseen** during training (**one-shot VC**).

Training from VCTK

For more results
[Chou, et al., INTERSPEECH 2019]



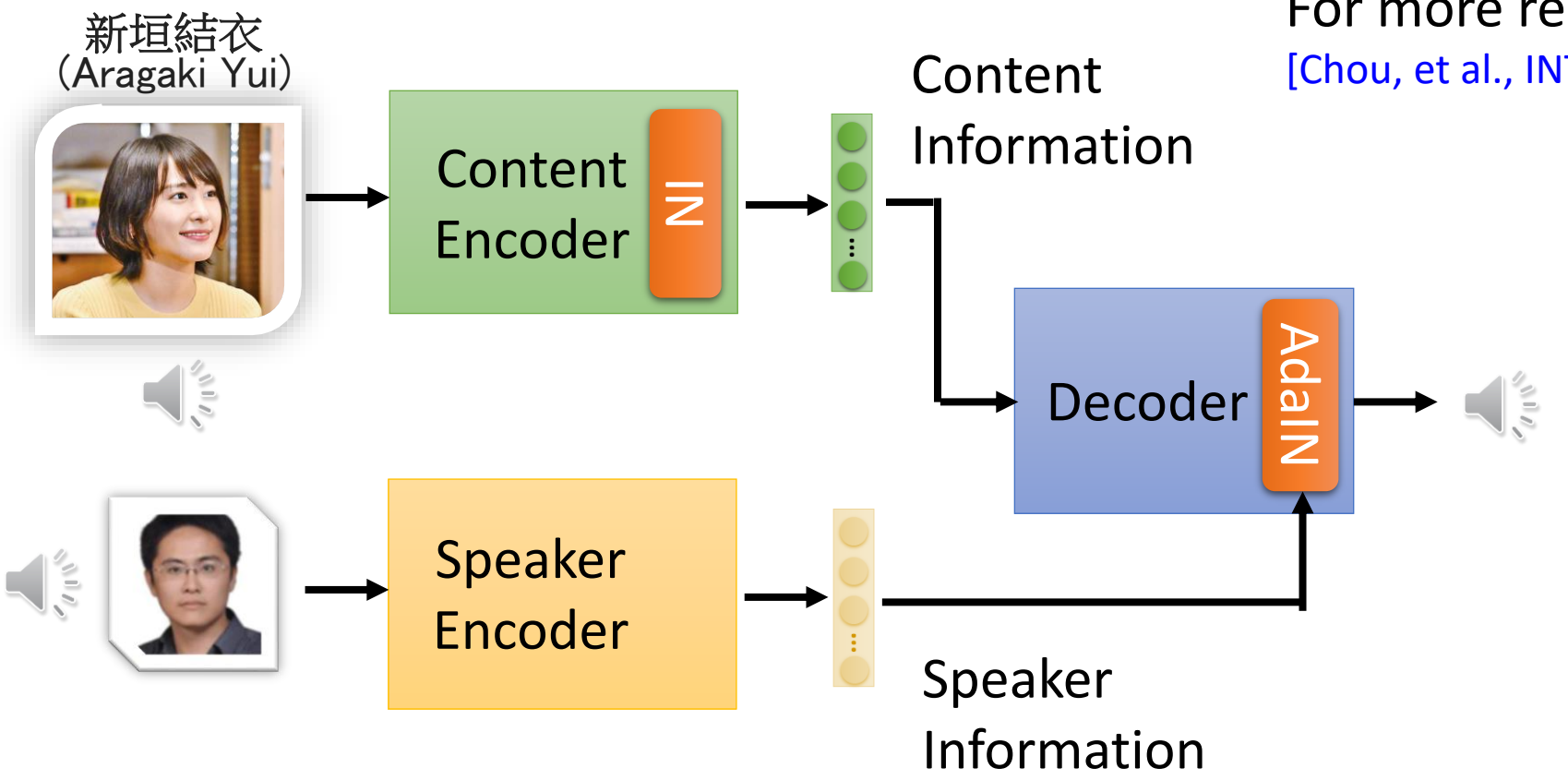
Designing network architecture

The speakers are **unseen** during training (**one-shot VC**).

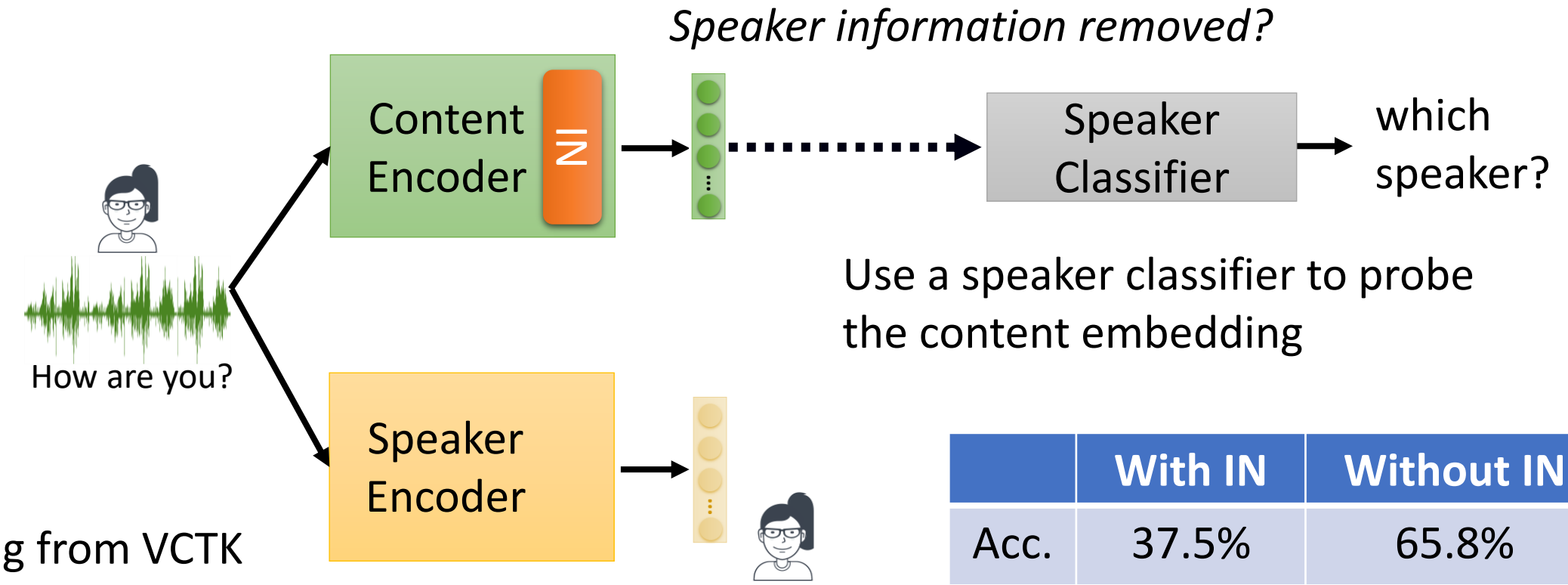
Training from VCTK

For more results

[Chou, et al., INTERSPEECH 2019]



Designing network architecture



Use a speaker classifier to probe the content embedding

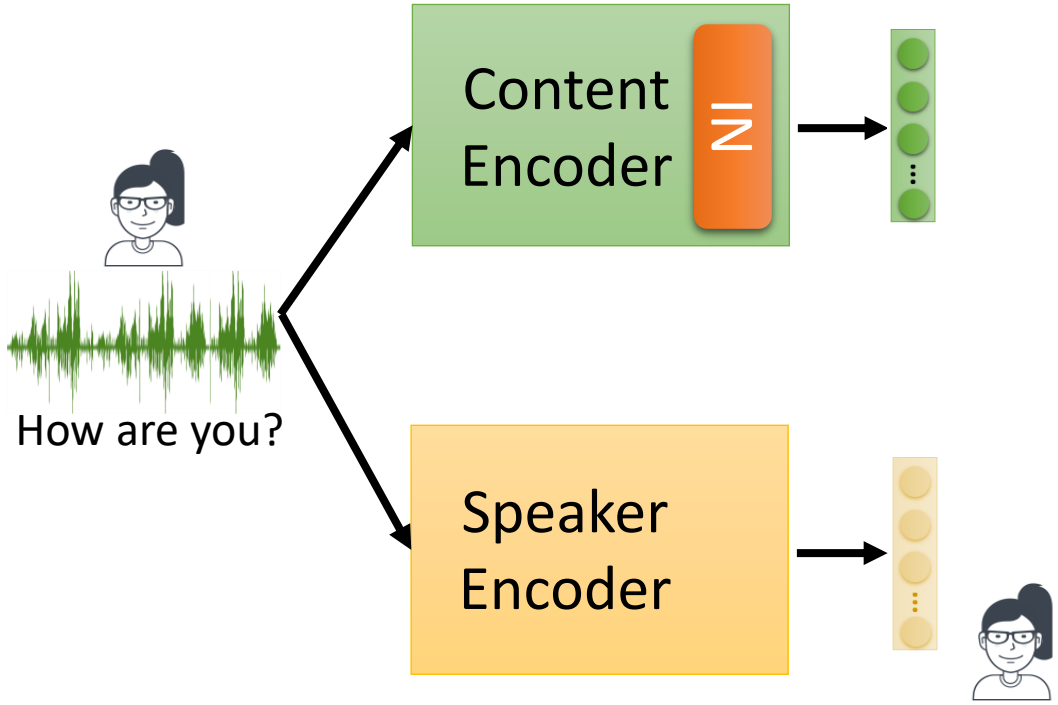
	With IN	Without IN
Acc.	37.5%	65.8%

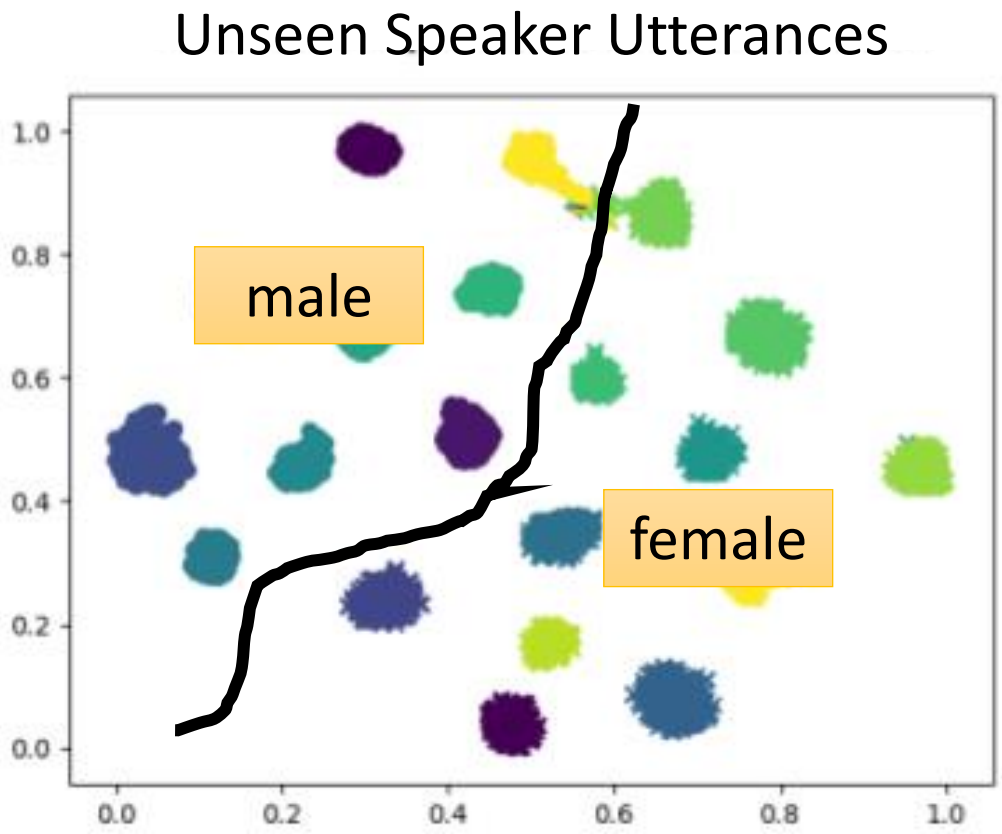
Training from VCTK

For more results

[Chou, et al., INTERSPEECH 2019]

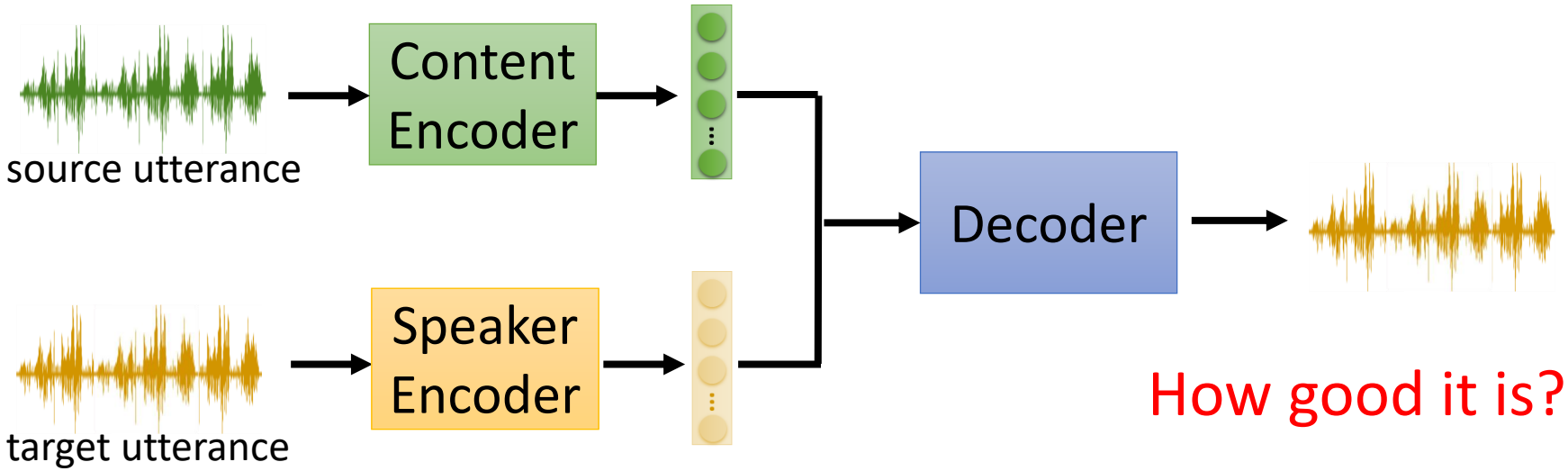
Designing network architecture





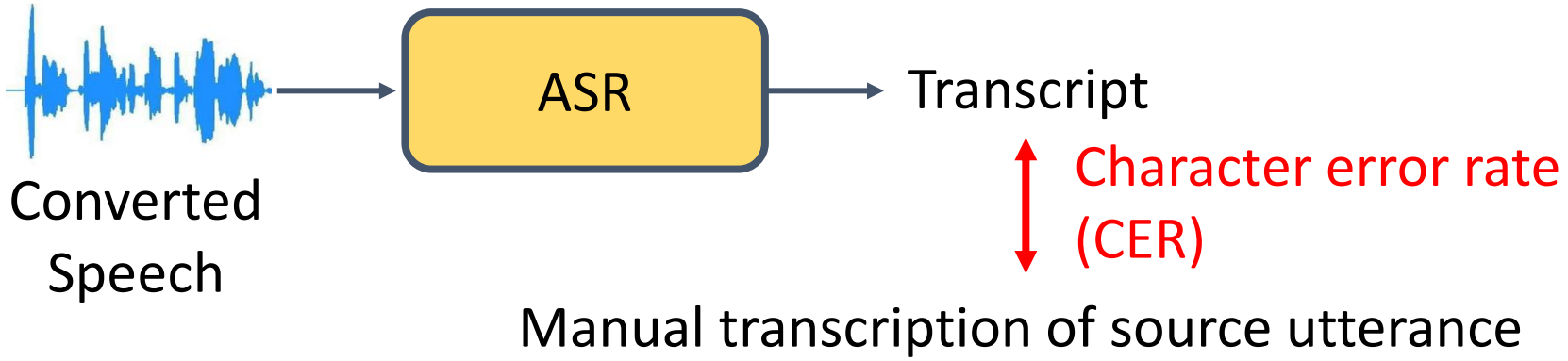
Comparison of VC approaches

[Huang, et al., SLT'21b]

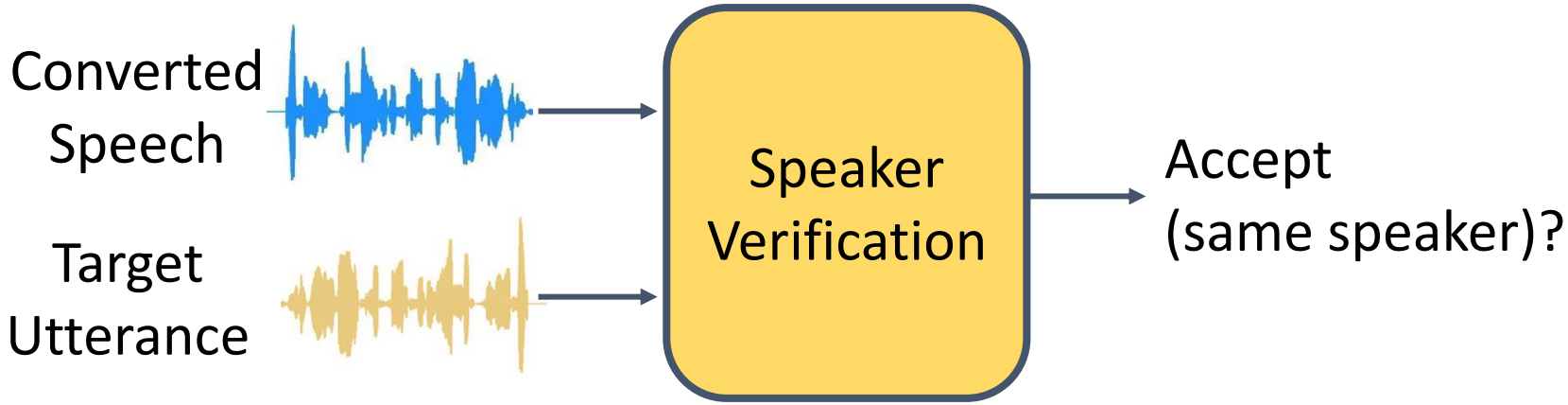


- Two aspects: **content preserving** and **target speaker similarity**
- Human evaluation is the best choice (Mean opinion score, MOS).
- But there are some acceptable automatic evaluation methods.

How to automatically evaluate **content preserving**?



How to automatically evaluate **target speaker similarity**?



Comparison of VC approaches

[Huang, et al., SLT'21b]

Training on VCTK

Testing on:

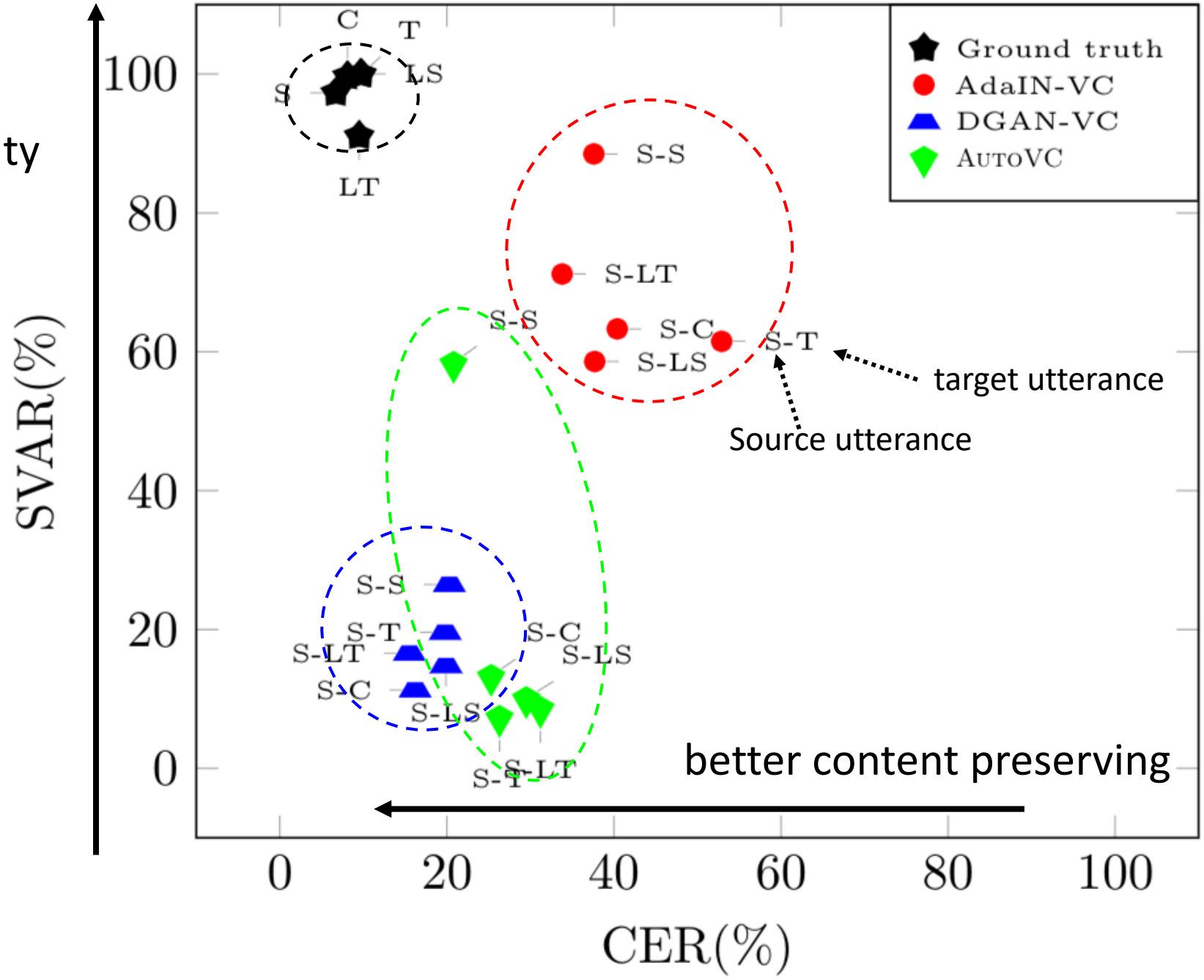
Dataset	Abbr.	
VCTK	S	→ In domain
LibriTTS	LT	→ out of domain
LibriSpeech	LS	
CMU	C	
THCHS-30	T	→ different language

Higher target speaker similarity

Training on VCTK

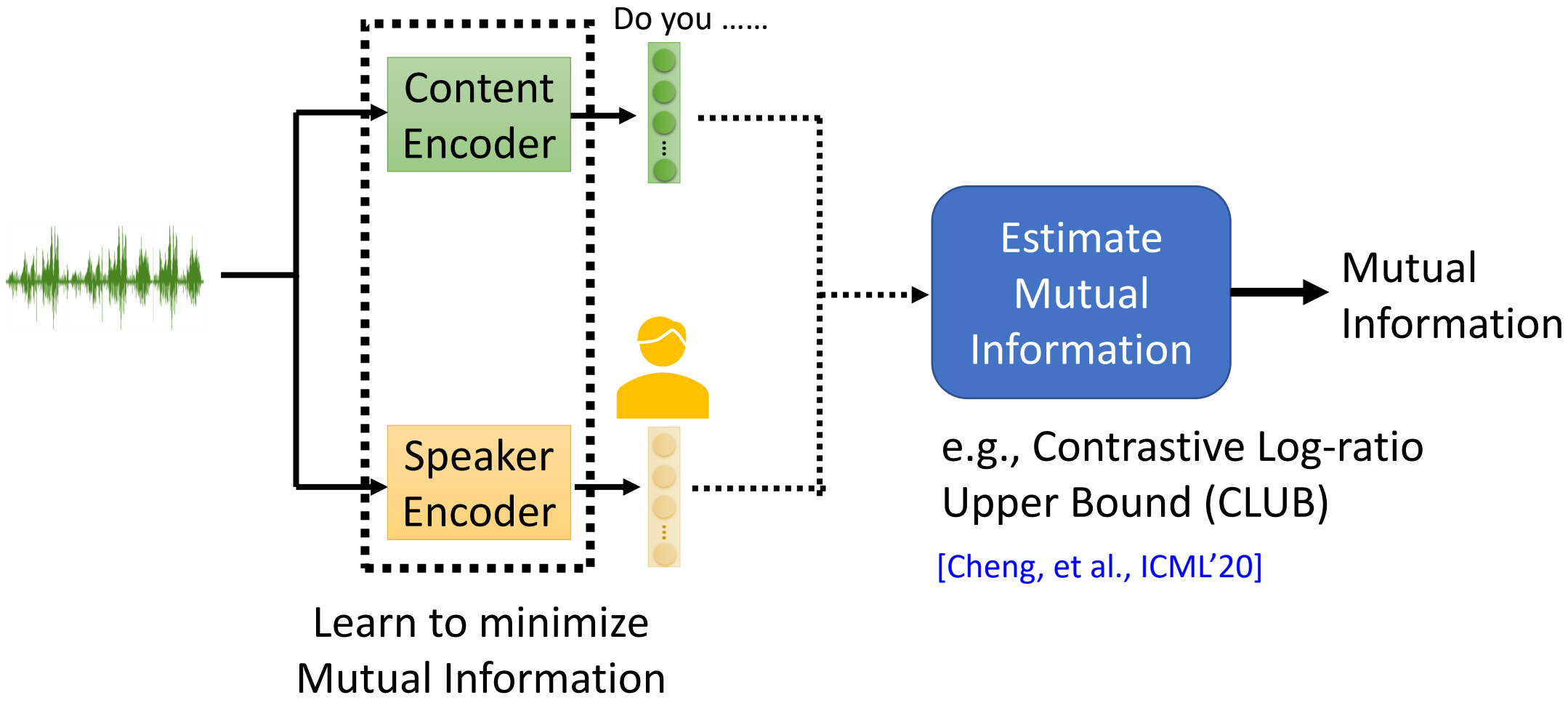
Testing on:

Dataset	Abbr.
VCTK	S
LibriTTS	LT
LibriSpeech	LS
CMU	C
THCHS-30	T

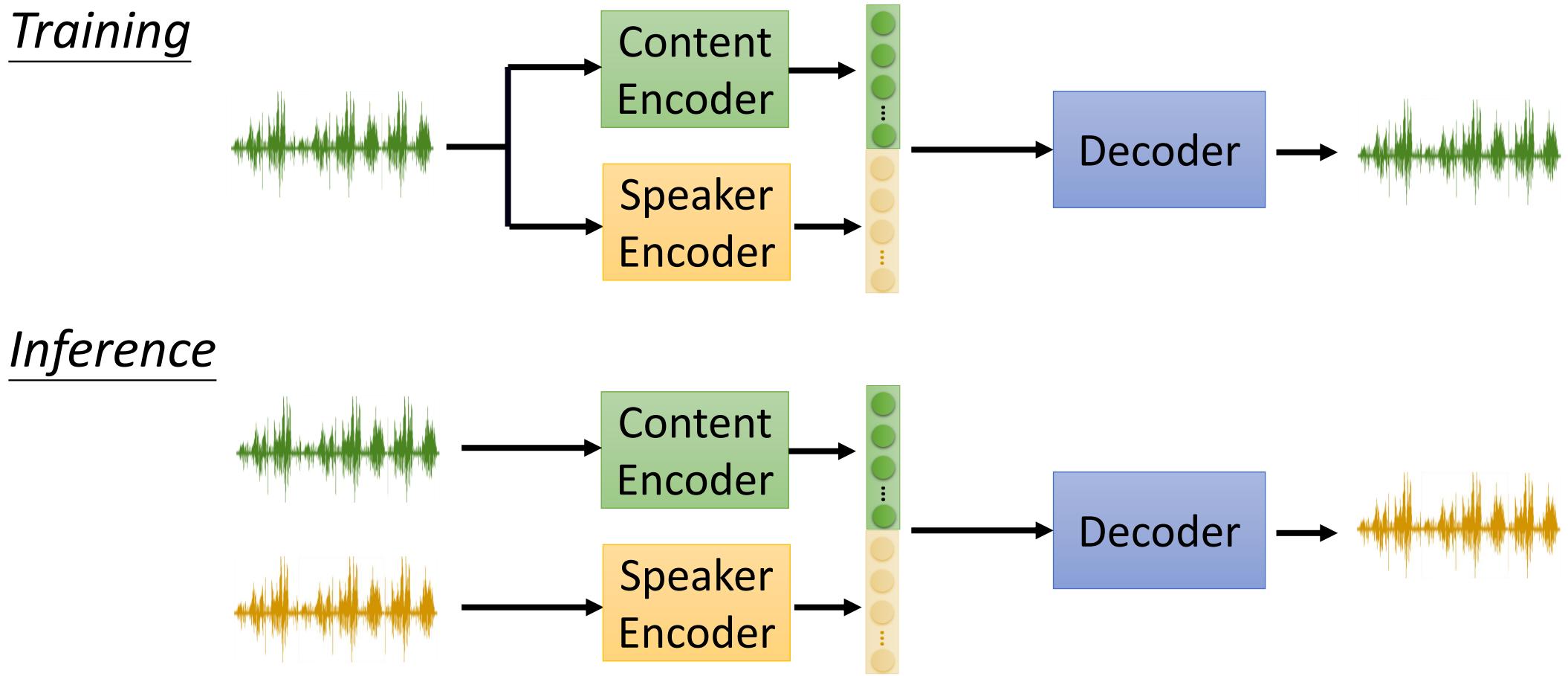


Minimize the correlation between different speech representations

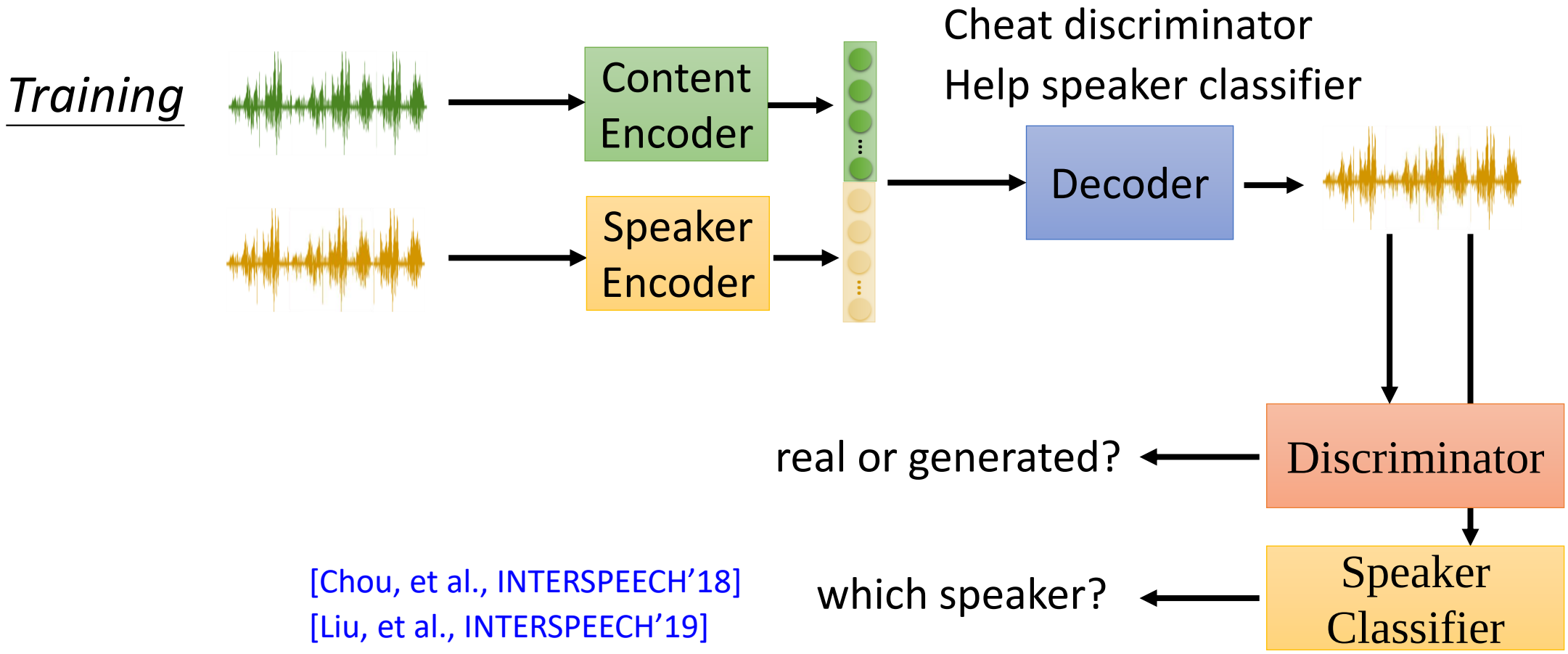
[Wang, et al., IS'21a]



Training and Inference Mismatch?

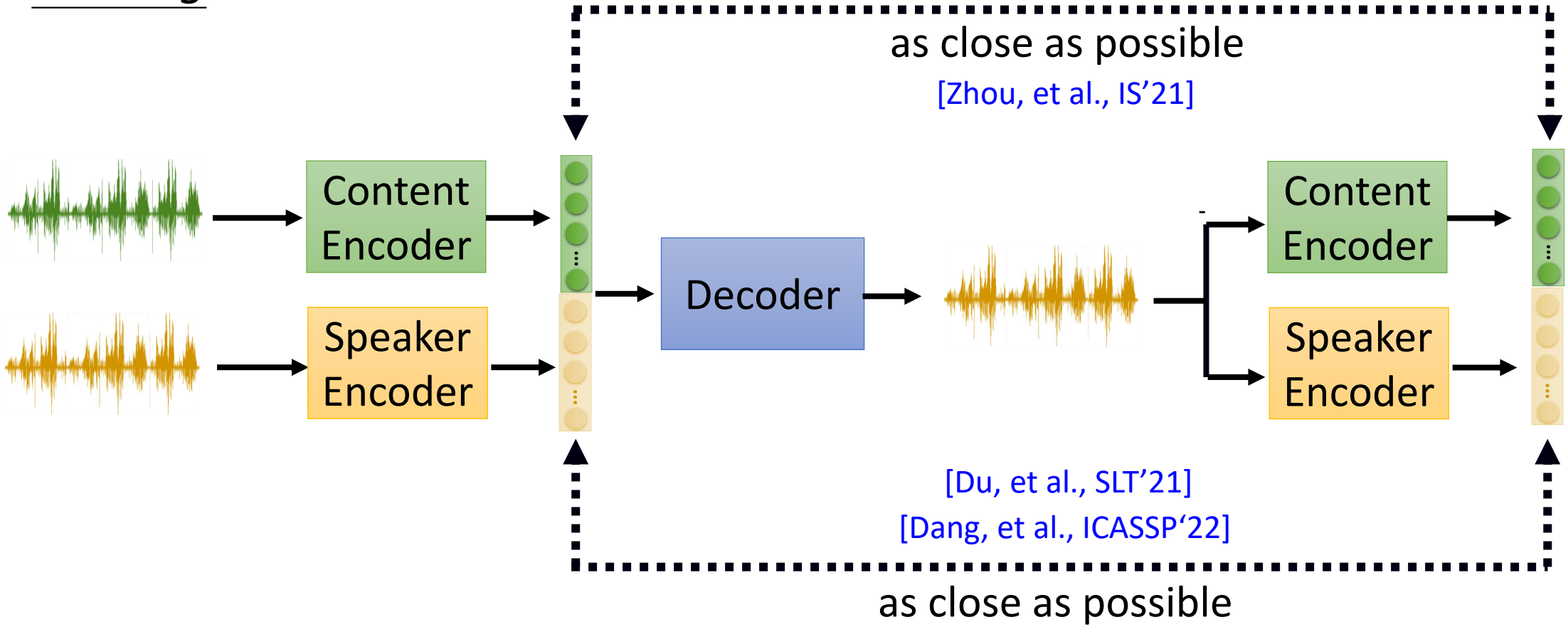


Training and Inference Mismatch?

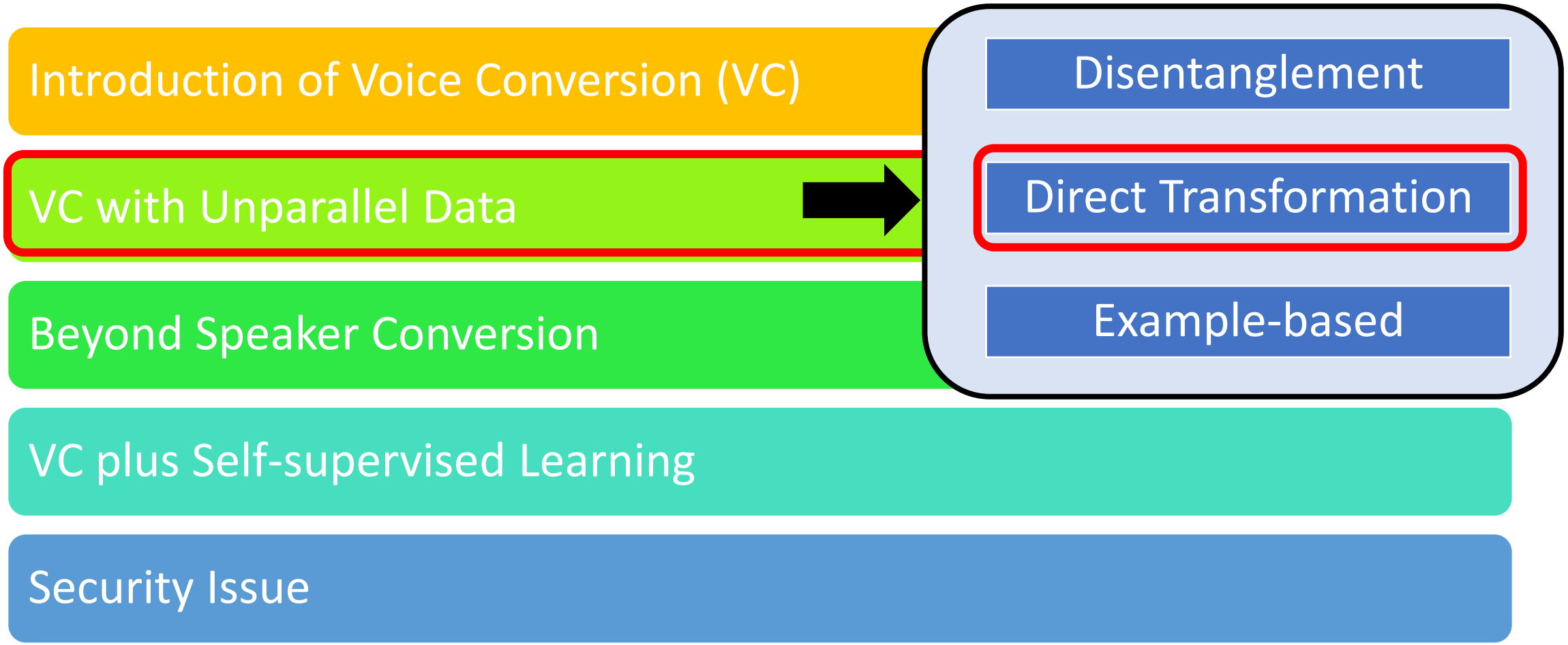


[Chou, et al., INTERSPEECH'18]
[Liu, et al., INTERSPEECH'19]

Training and Inference Mismatch?

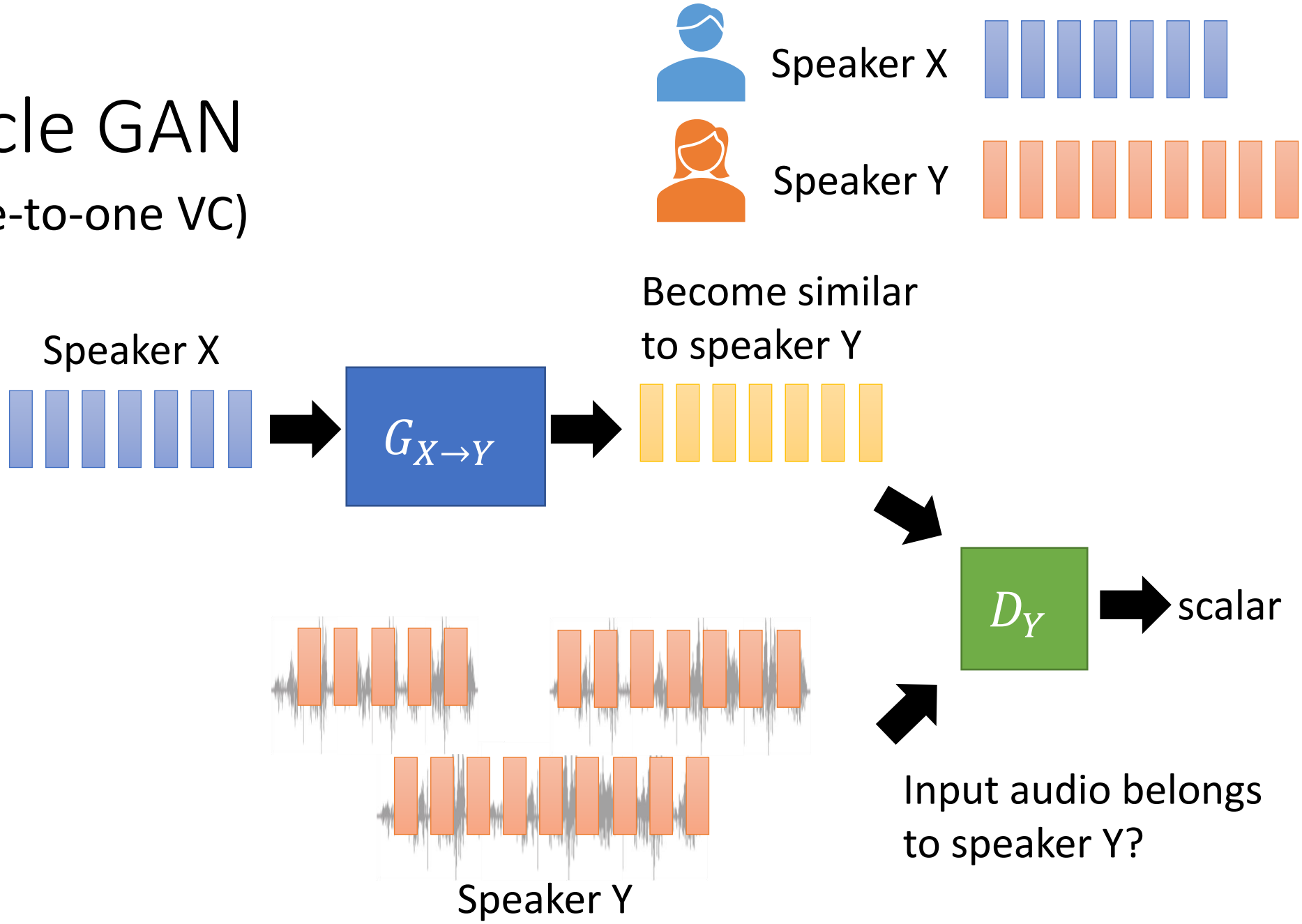


Outline


Introduction of Voice Conversion (VC)
VC with Unparallel Data
Beyond Speaker Conversion
VC plus Self-supervised Learning
Security Issue
Disentanglement
Direct Transformation
Example-based

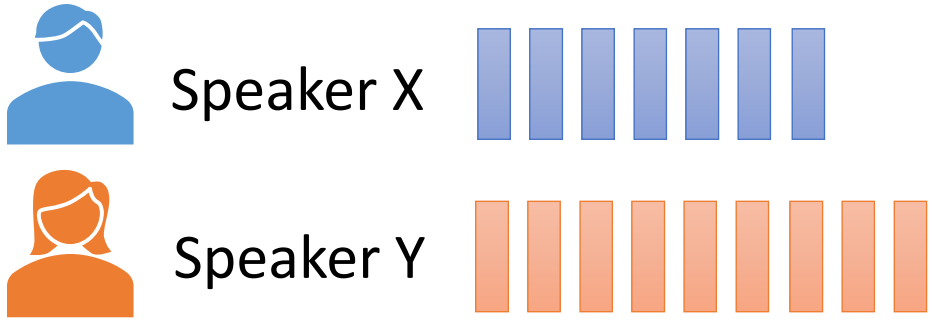
Cycle GAN

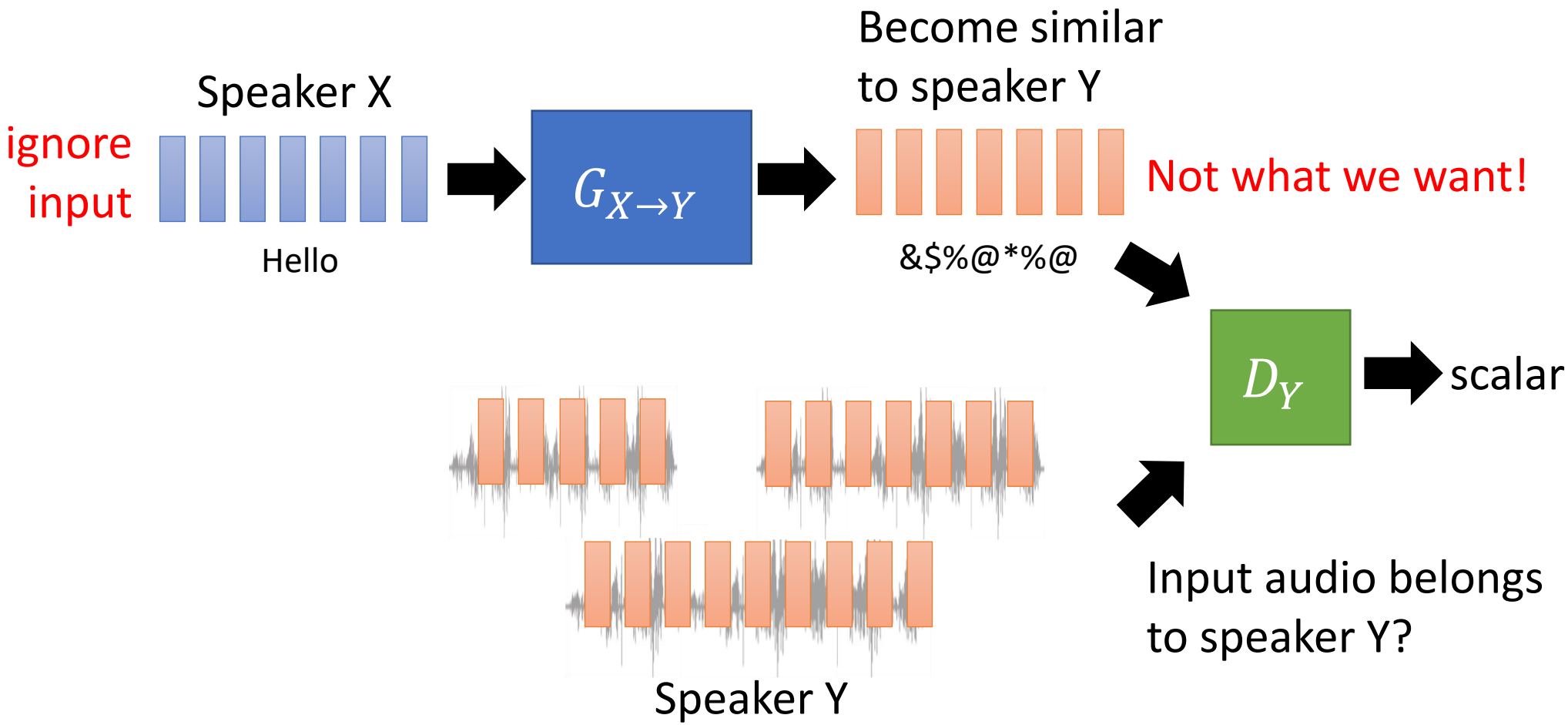
(One-to-one VC)



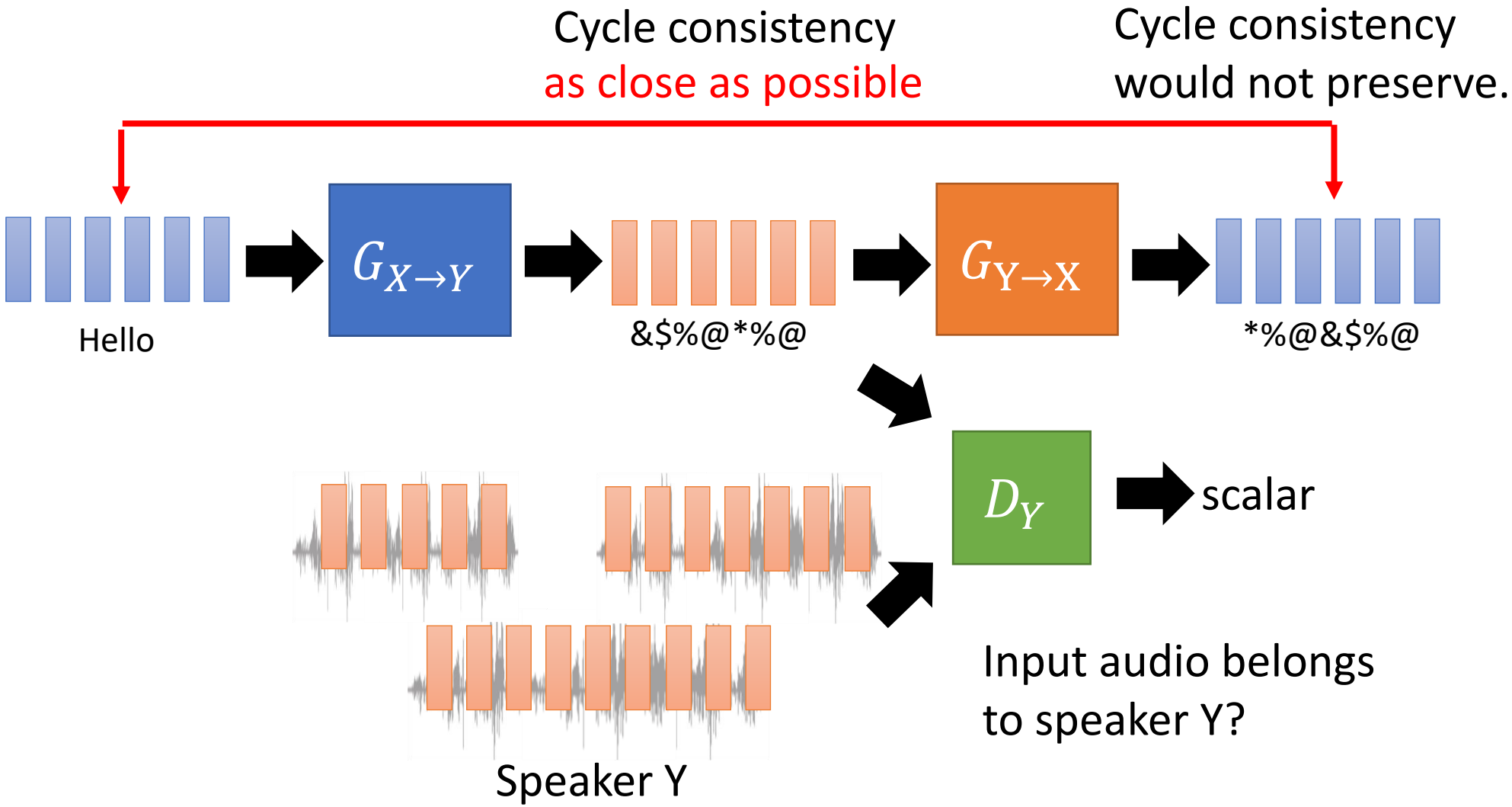
Cycle GAN

(One-to-one VC)

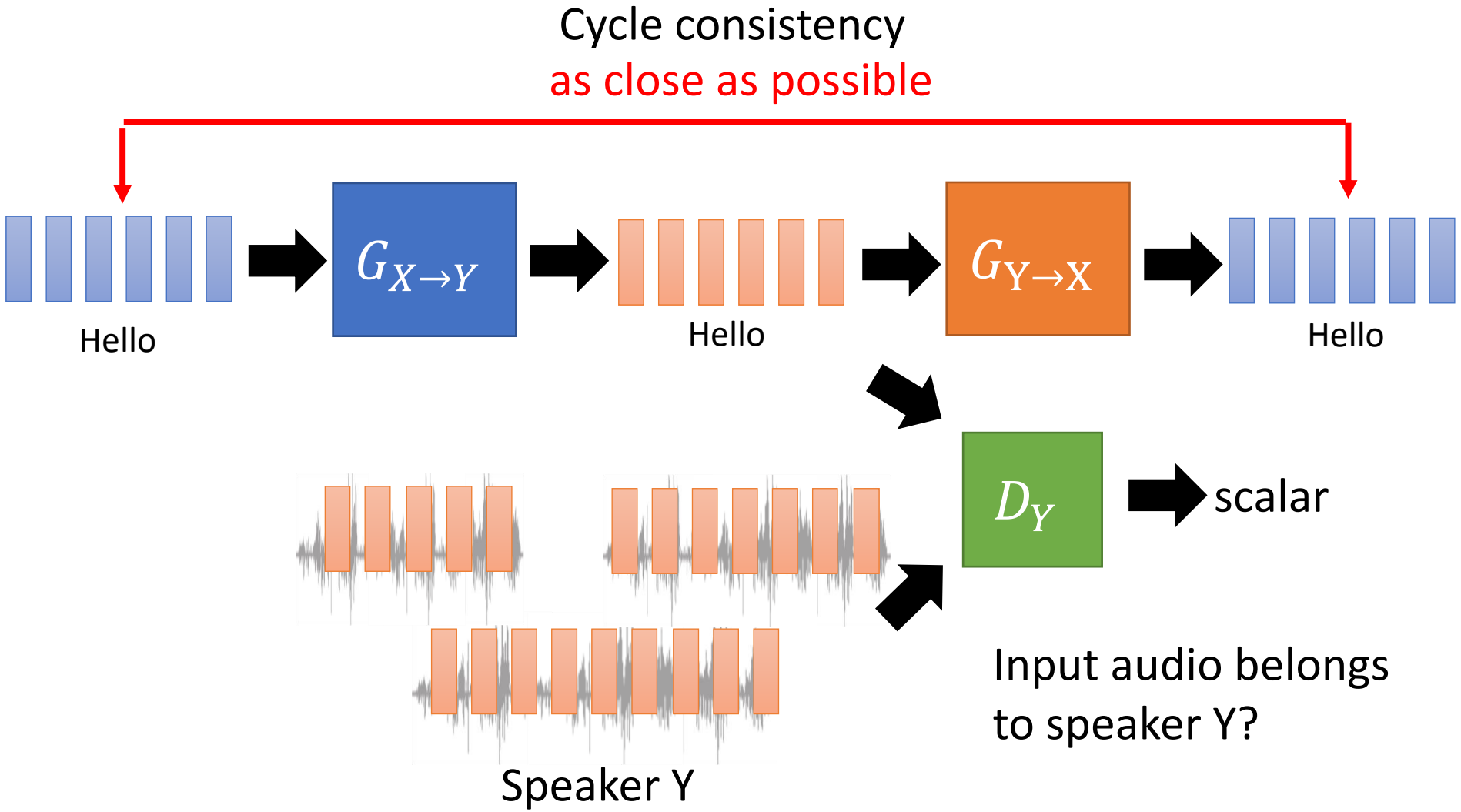




Cycle GAN


Cycle consistency
as close as possible
Cycle consistency
would not preserve.
$G_{X\to Y}$
$G_{Y\to X}$
Hello
&$%@*%@
*%@&$%@
D_Y
scalar
Input audio belongs
to speaker Y?
Speaker Y

Cycle GAN

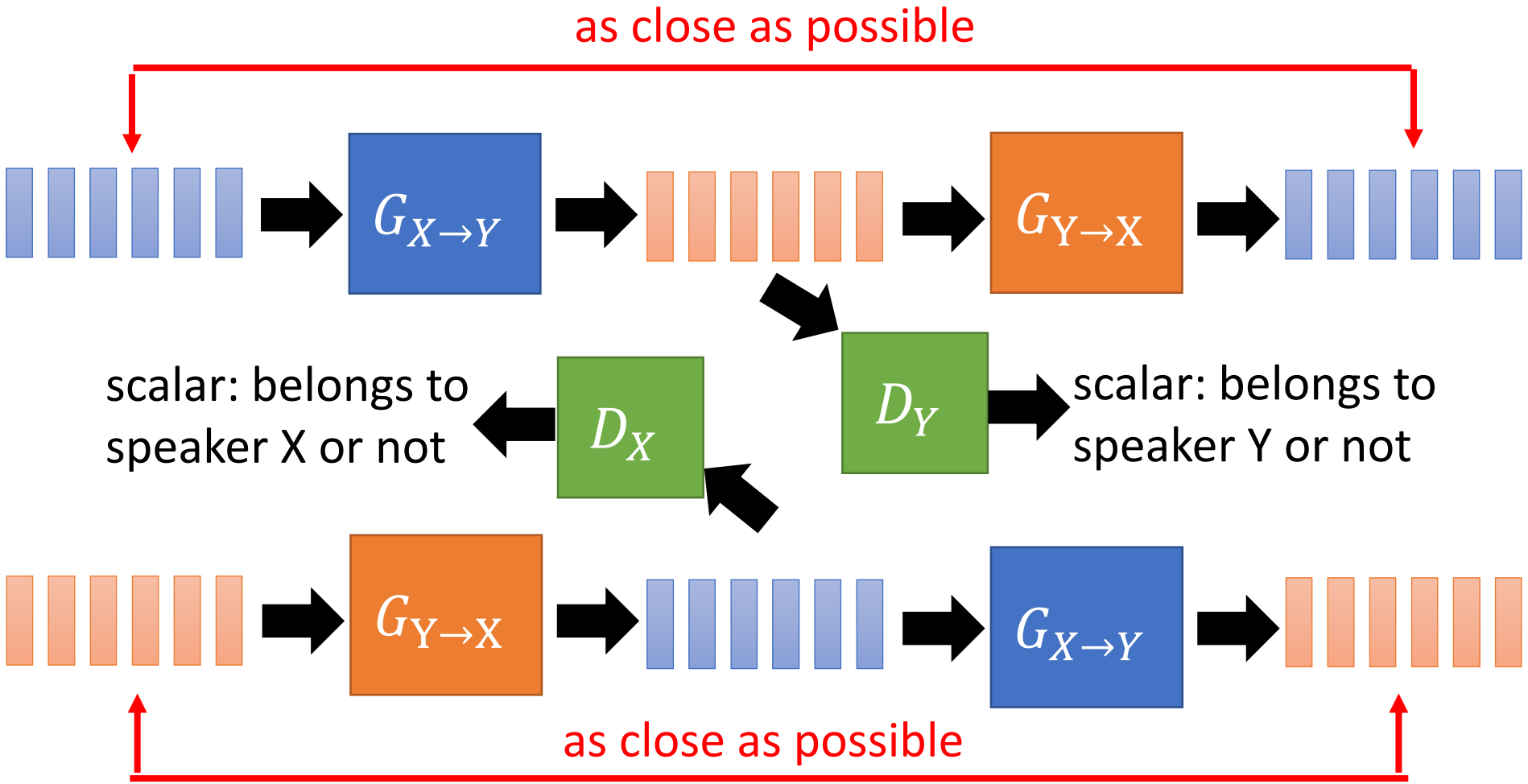

Cycle consistency
as close as possible
$G_{X\to Y}$
$G_{Y\to X}$
Hello
Hello
Hello
D_Y
scalar
Input audio belongs to speaker Y?
Speaker Y

Cycle GAN

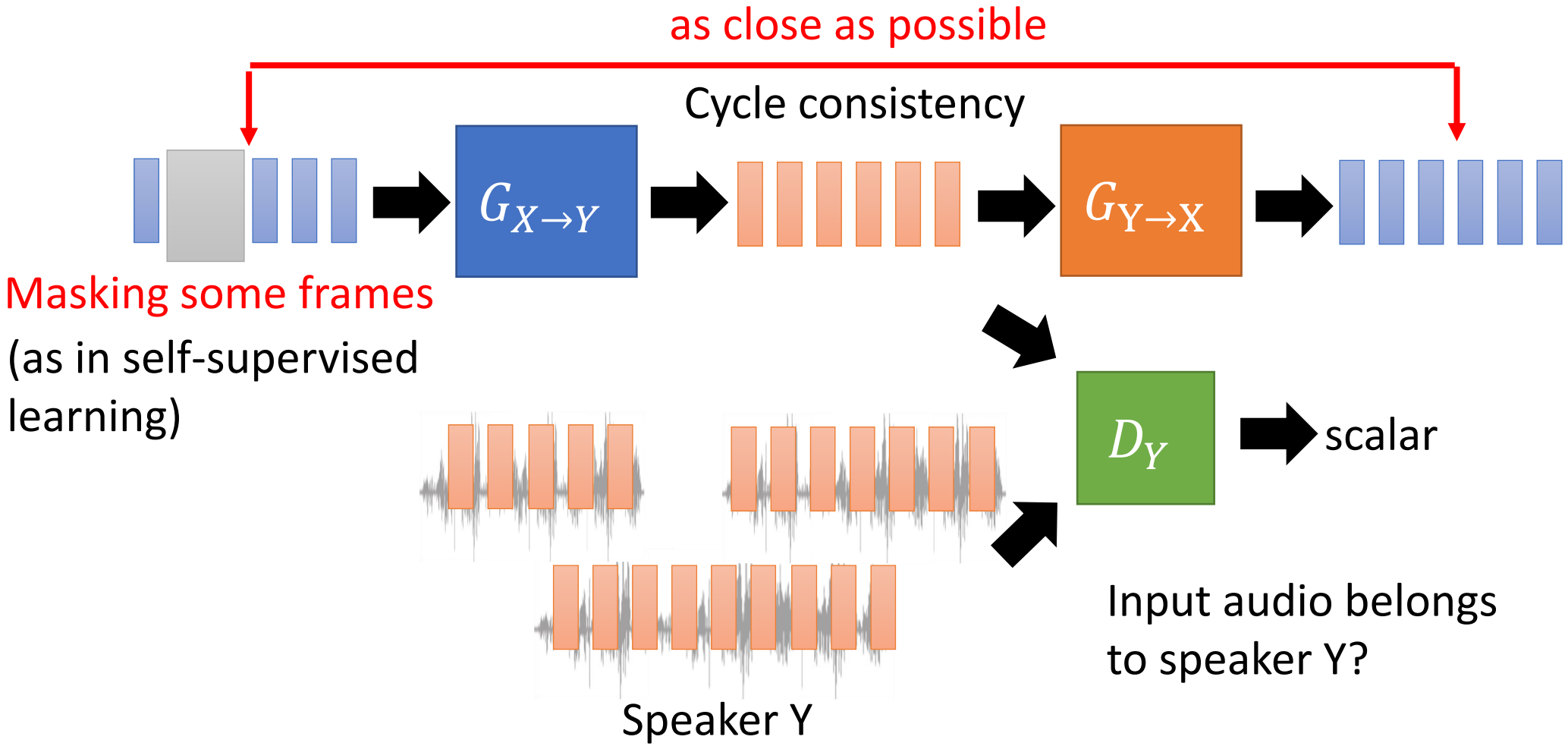
[Kaneko, et al., ICASSP'19]

CycleGAN-VC, CycleGAN-VC2, CycleGAN-VC3

[Kaneko, et al., arXiv'17] [Kaneko, et al., IS'20]

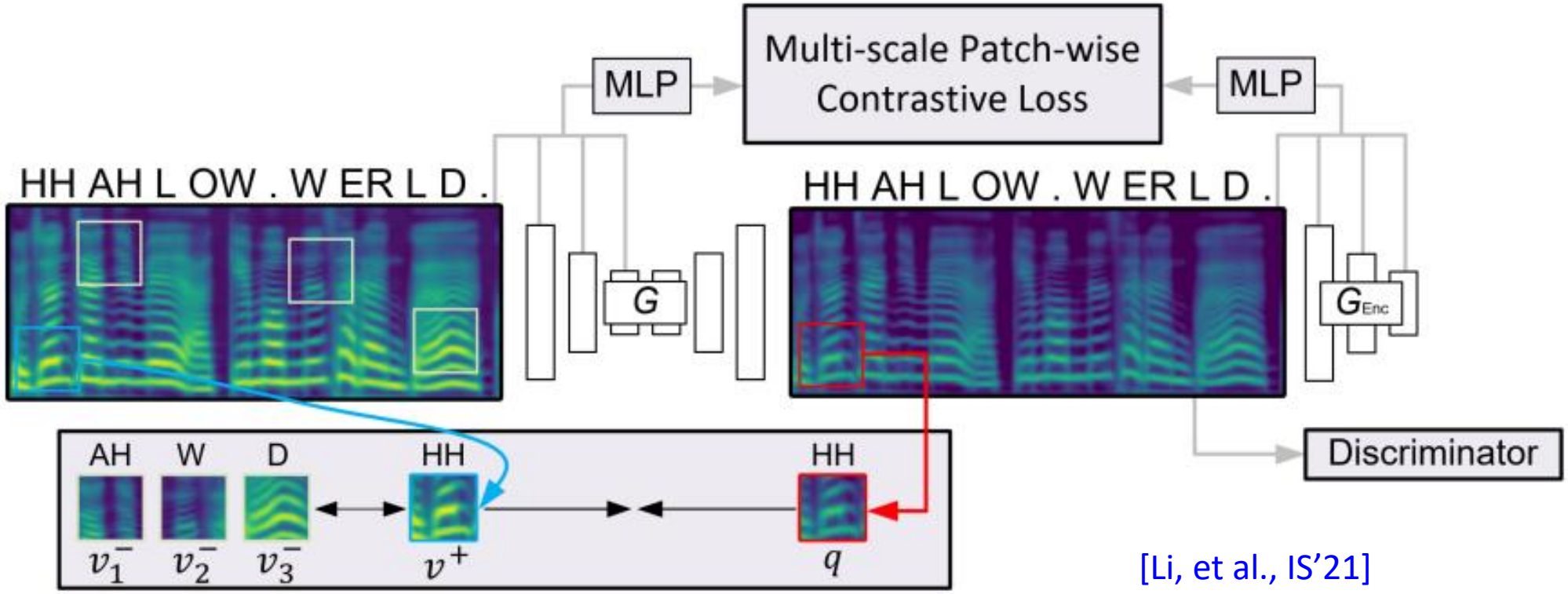


MaskCycleGAN-VC [Kaneko, et al., ICASSP'21]



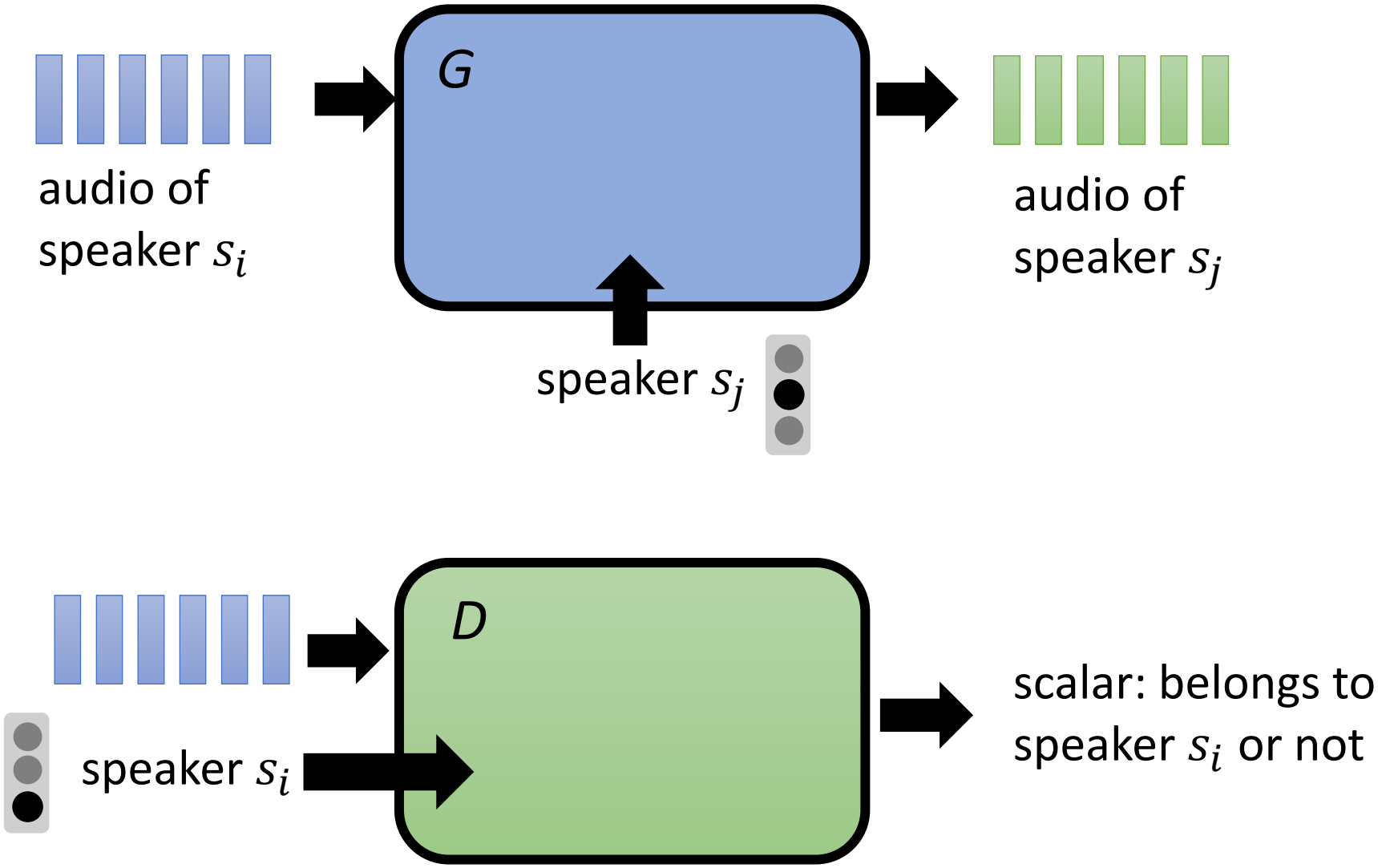
Cycle GAN

- Cycle consistency is not the only way to maintain the content



[Li, et al., IS'21]

StarGAN

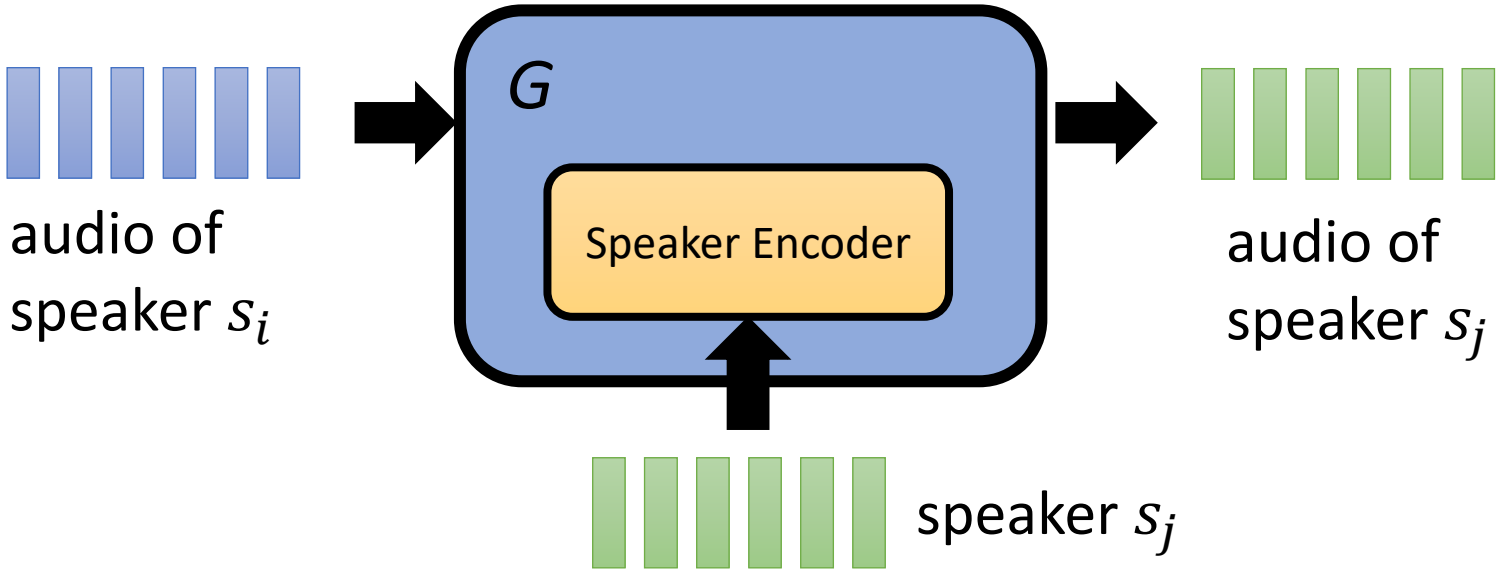


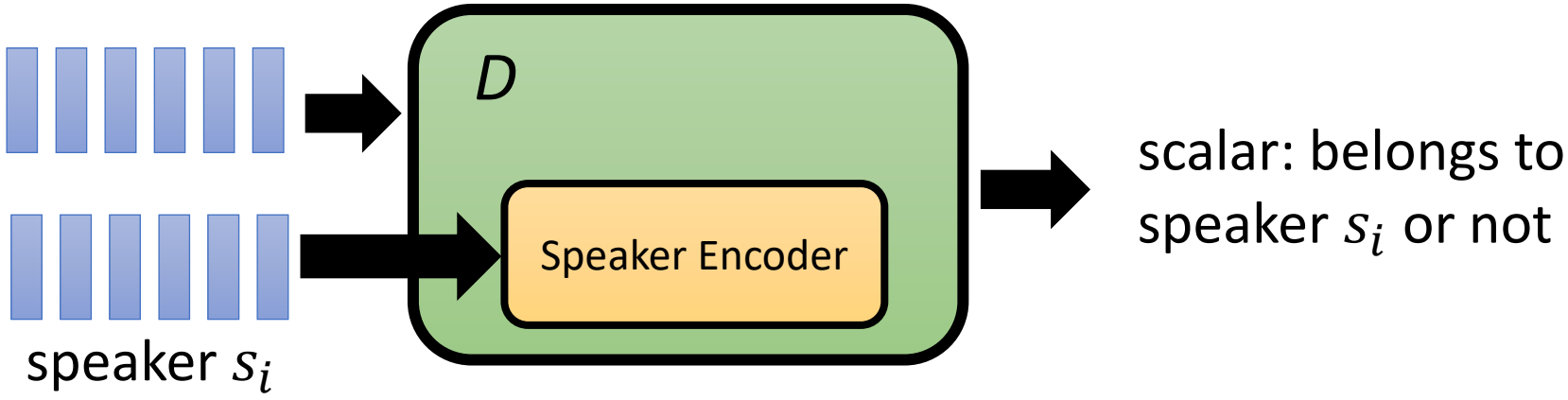
Each speaker is represented as a one-hot vector.

[Kameoka, et al., SLT'18] [Kaneko, et al., INTERSPEECH'19]

Many-to-many VC

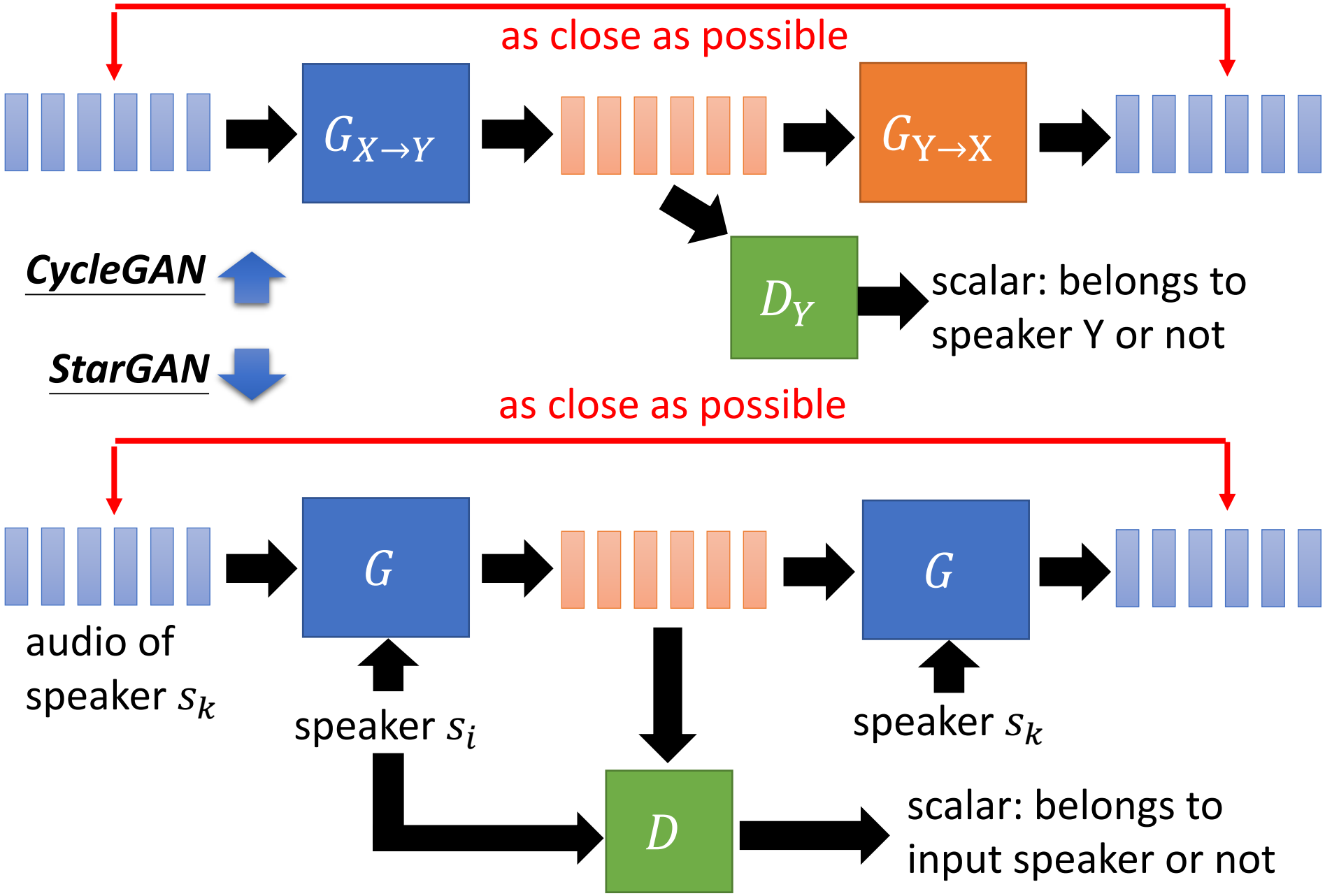
StarGAN





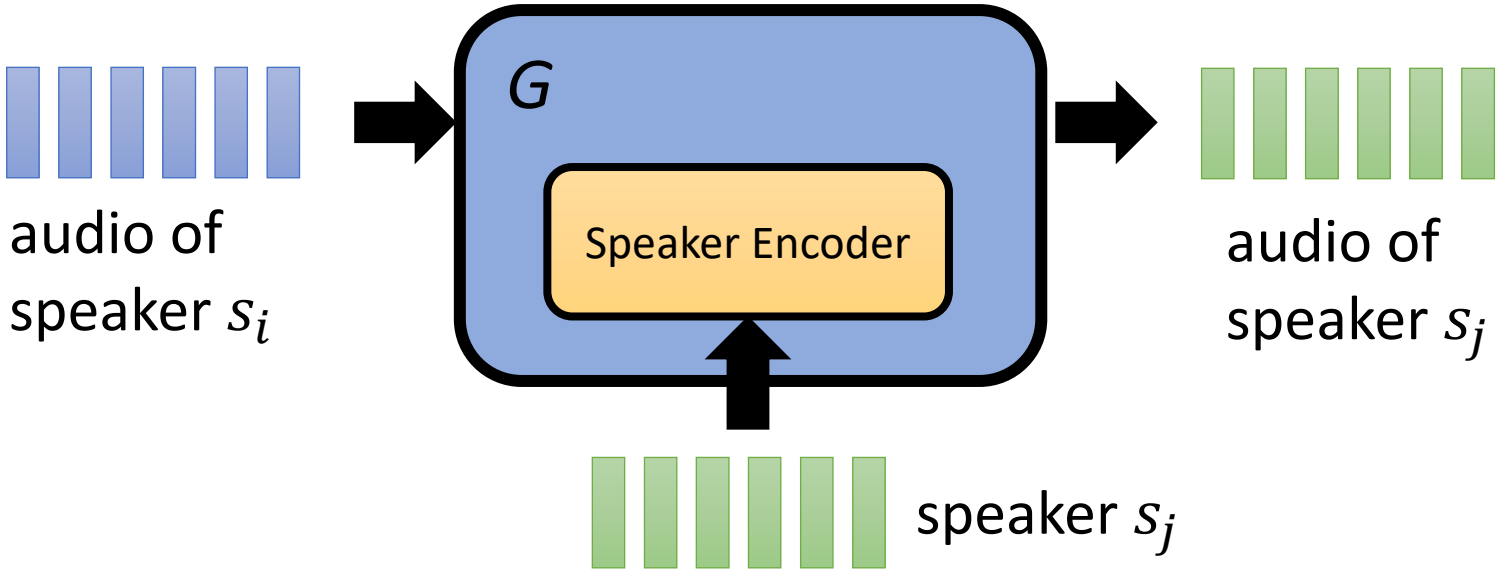
Pre-trained Speaker Encoder [Wnag, et al., ICASSP'20] [Chen, et al., ICASSP'21b]

Any-to-any VC



(The domain classifier is ignored here.)

Direct Transformation vs. Disentanglement

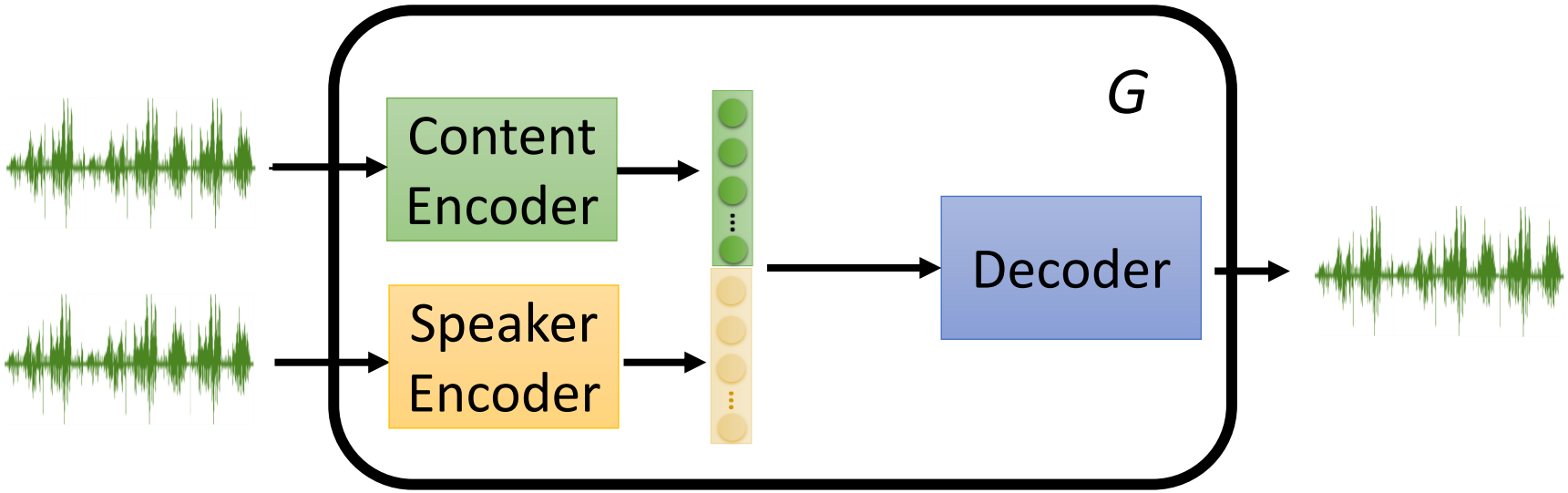


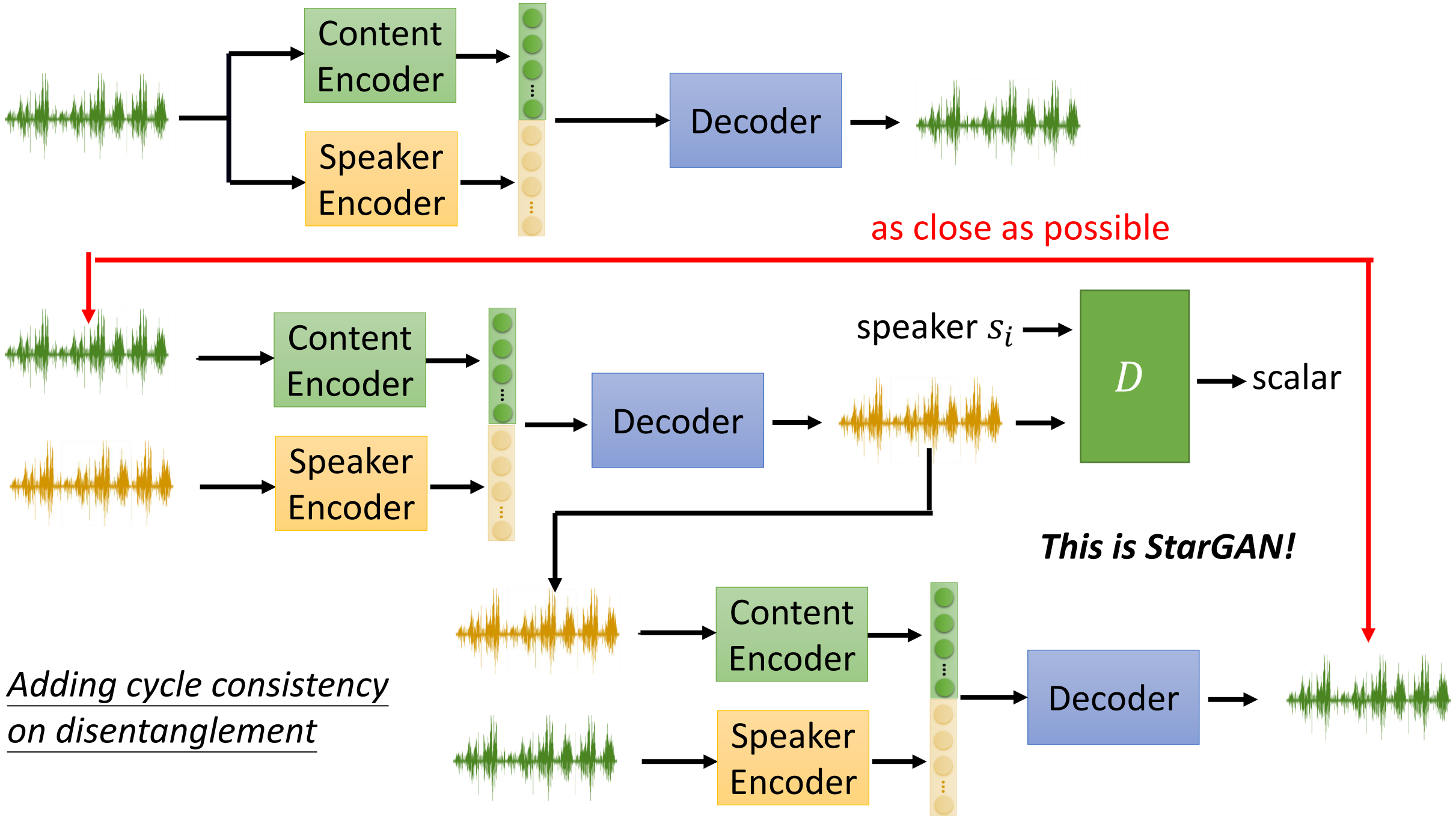
GAN-based disentangle

[Eskimez, et al., IS'21]

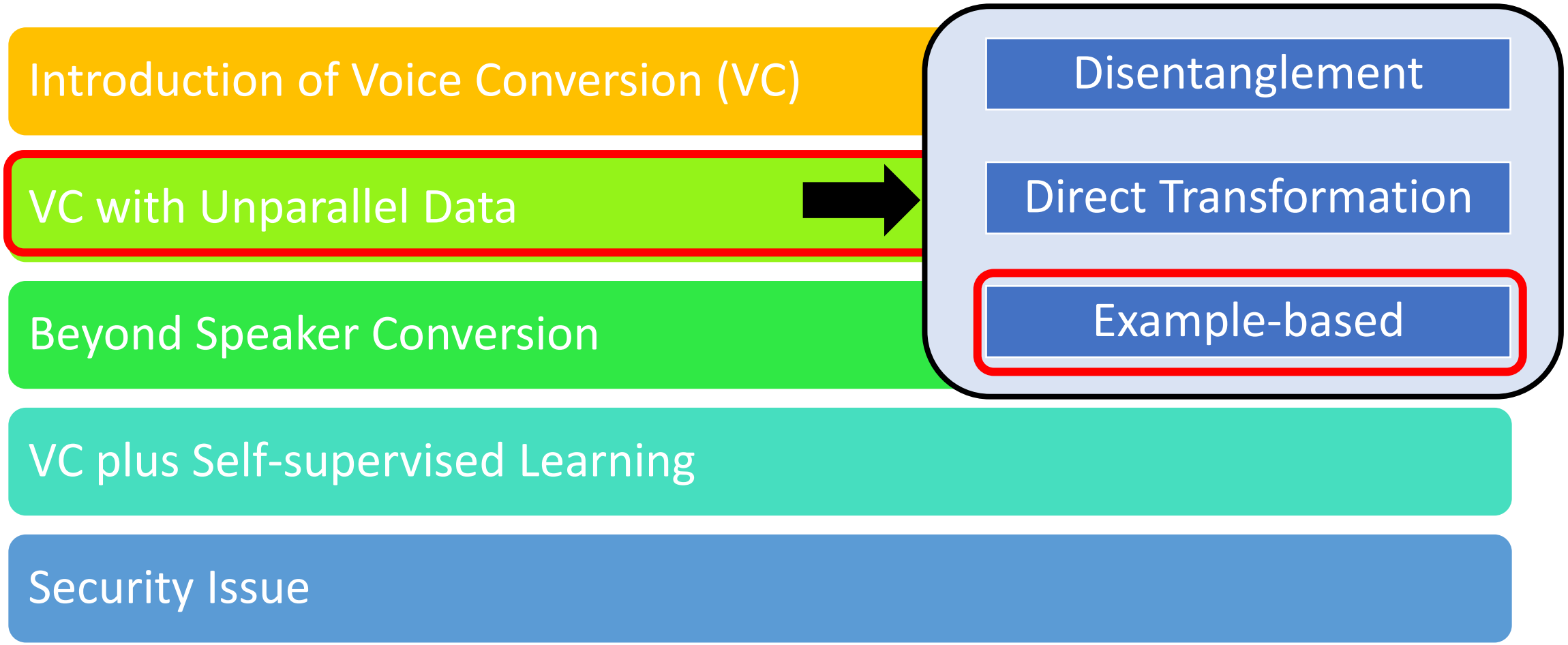
Adaptive instance normalization

[Kaneko, et al., INTERSPEECH'19]



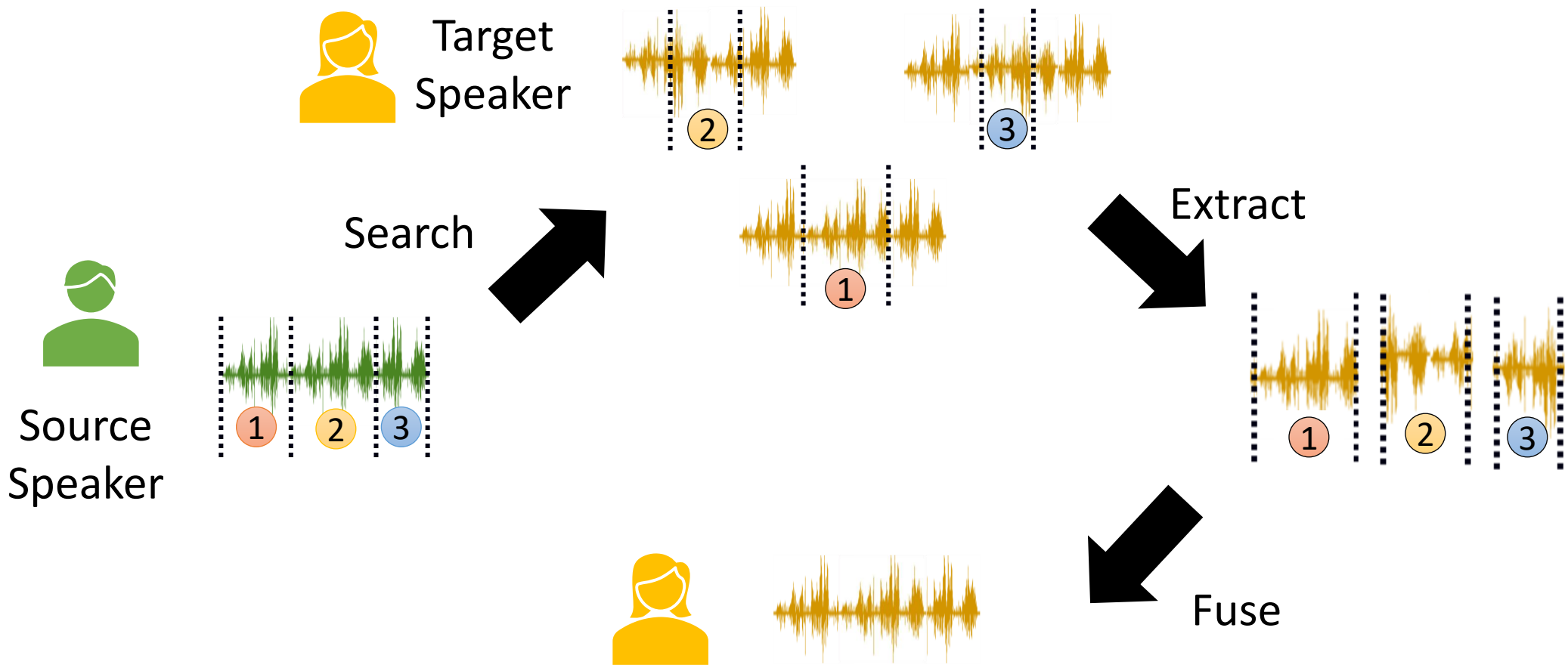

Content Encoder
Speaker Encoder
Decoder
as close as possible
Content Encoder
Speaker Encoder
Decoder
speaker s_i
D
scalar
This is StarGAN!
Content Encoder
Speaker Encoder
Decoder
Adding cycle consistency on disentanglement

Outline


Introduction of Voice Conversion (VC)
VC with Unparallel Data
Beyond Speaker Conversion
VC plus Self-supervised Learning
Security Issue
Disentanglement
Direct Transformation
Example-based

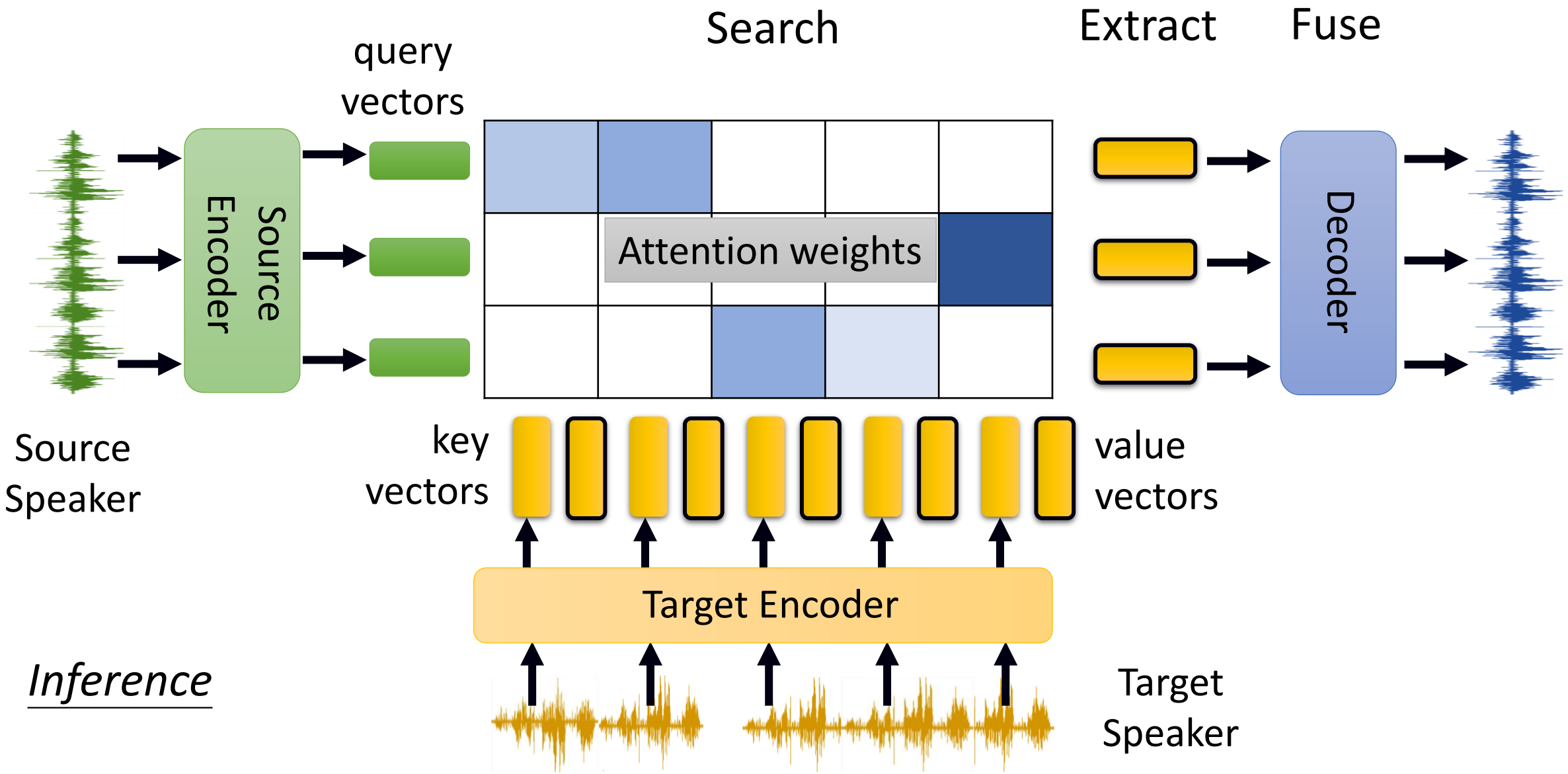
Example-based Approach

[Sundermann, et al., ICASSP'06]
[Takashima, et al., SLT'12]
[Jin, et al., ICASSP'16]



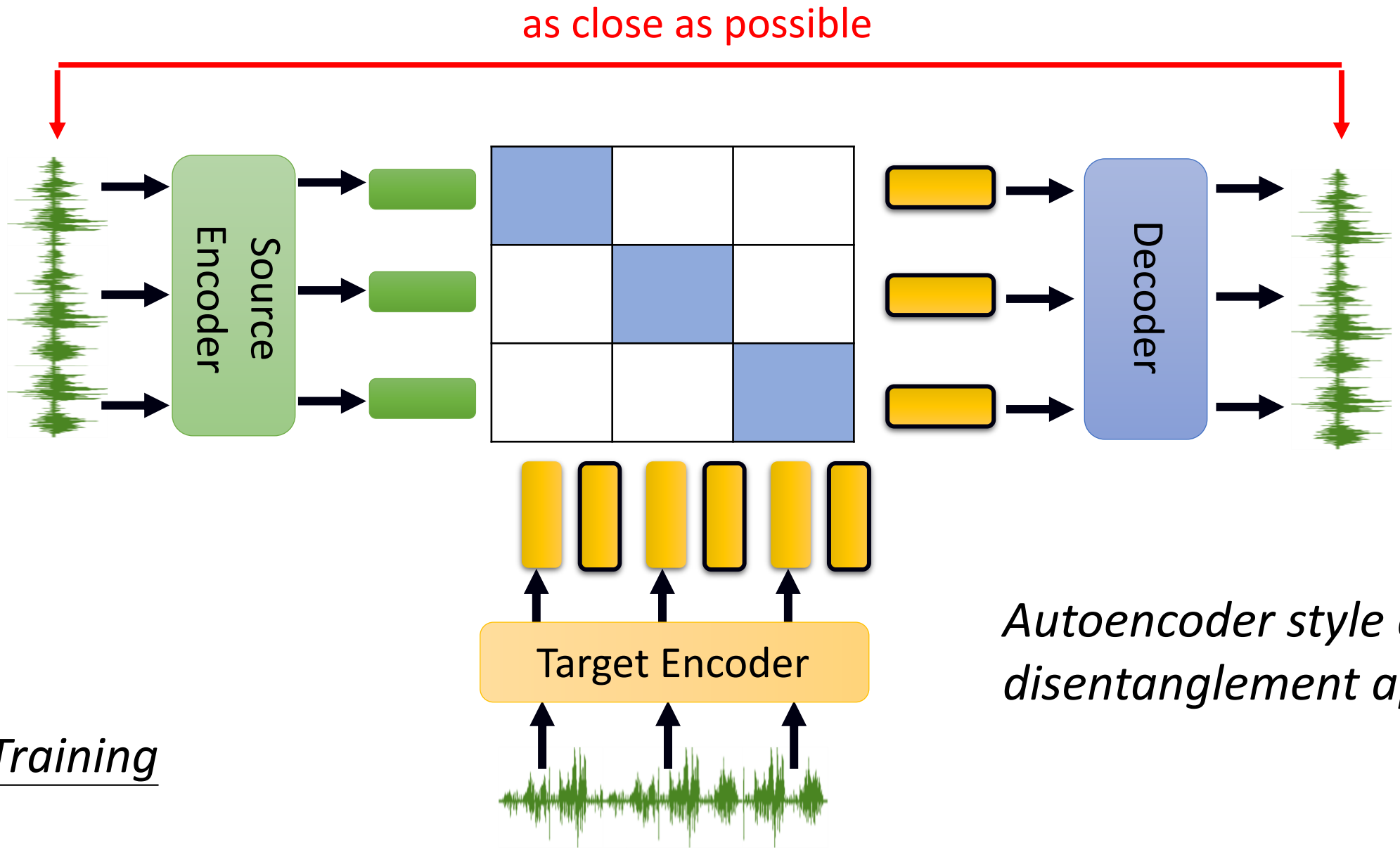
Using an end-to-end network to realize this process

Fragment VC [Lin, et al., ICASSP'21] [Lin, et al., IS'21]



Inference

Fragment VC [Lin, et al., ICASSP'21] [Lin, et al., IS'21]

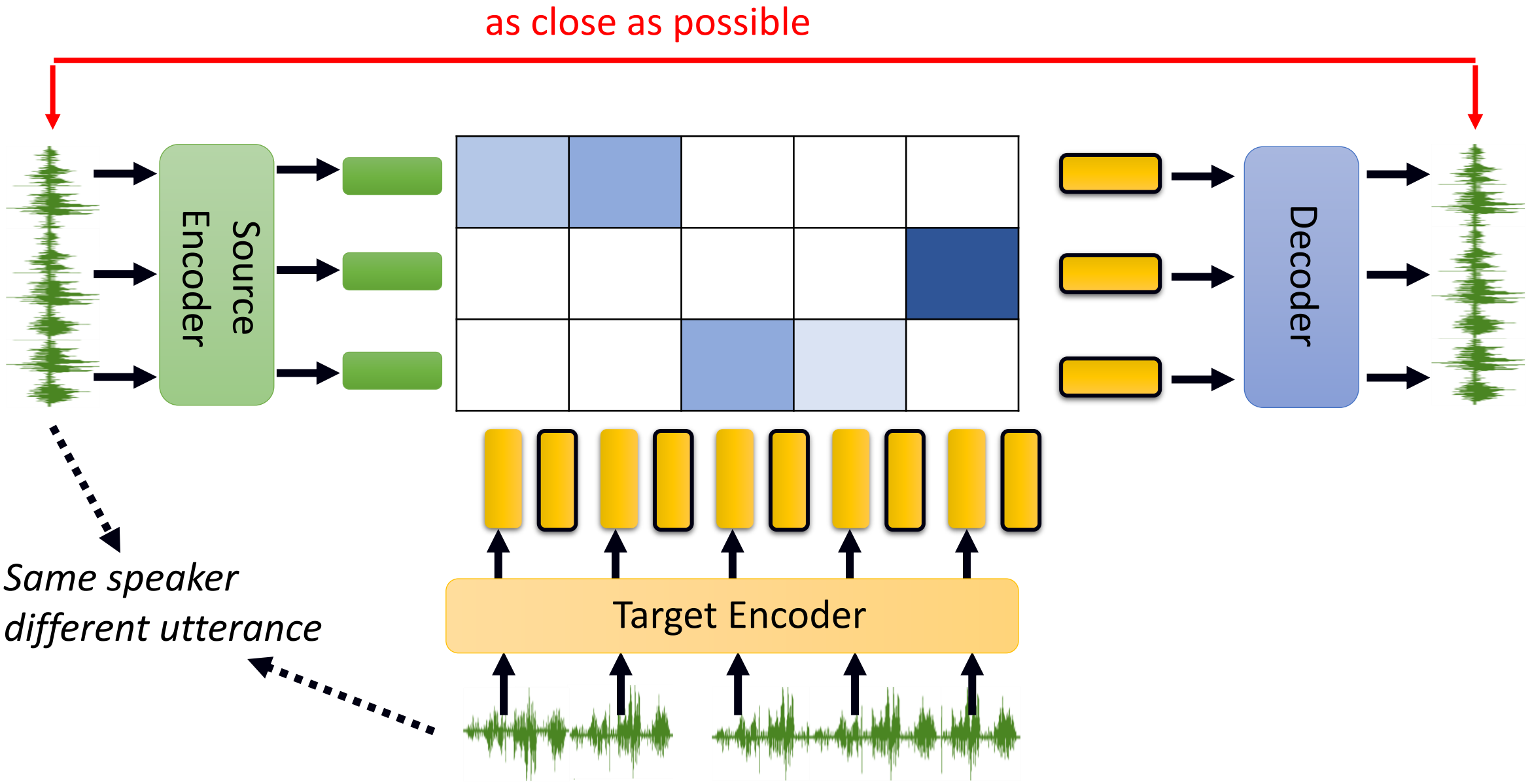


Autoencoder style as disentanglement approach

Training

Fragment VC [Lin, et al., ICASSP'21] [Lin, et al., IS'21]

Demo page:
https://yistlin.github.io/FragmentVC/



Outline

Introduction of Voice Conversion (VC)

VC with Unparallel Data

Beyond Speaker Conversion

VC plus Self-supervised Learning

Security Issue

Speech conveys rich information

Content

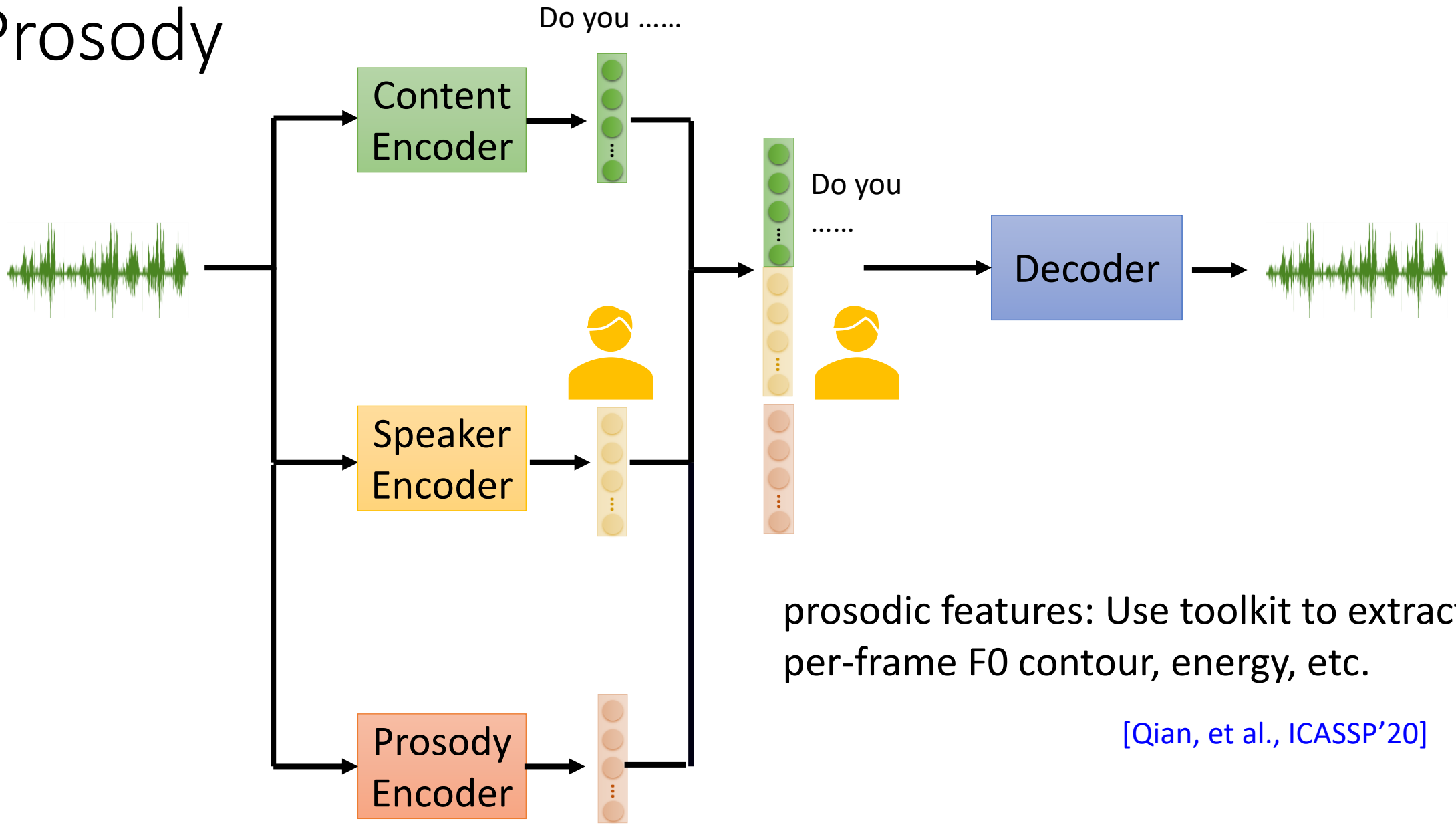
Pitch

Speaker
(timbre)

Rhythm

Environment
(e.g., background noise)

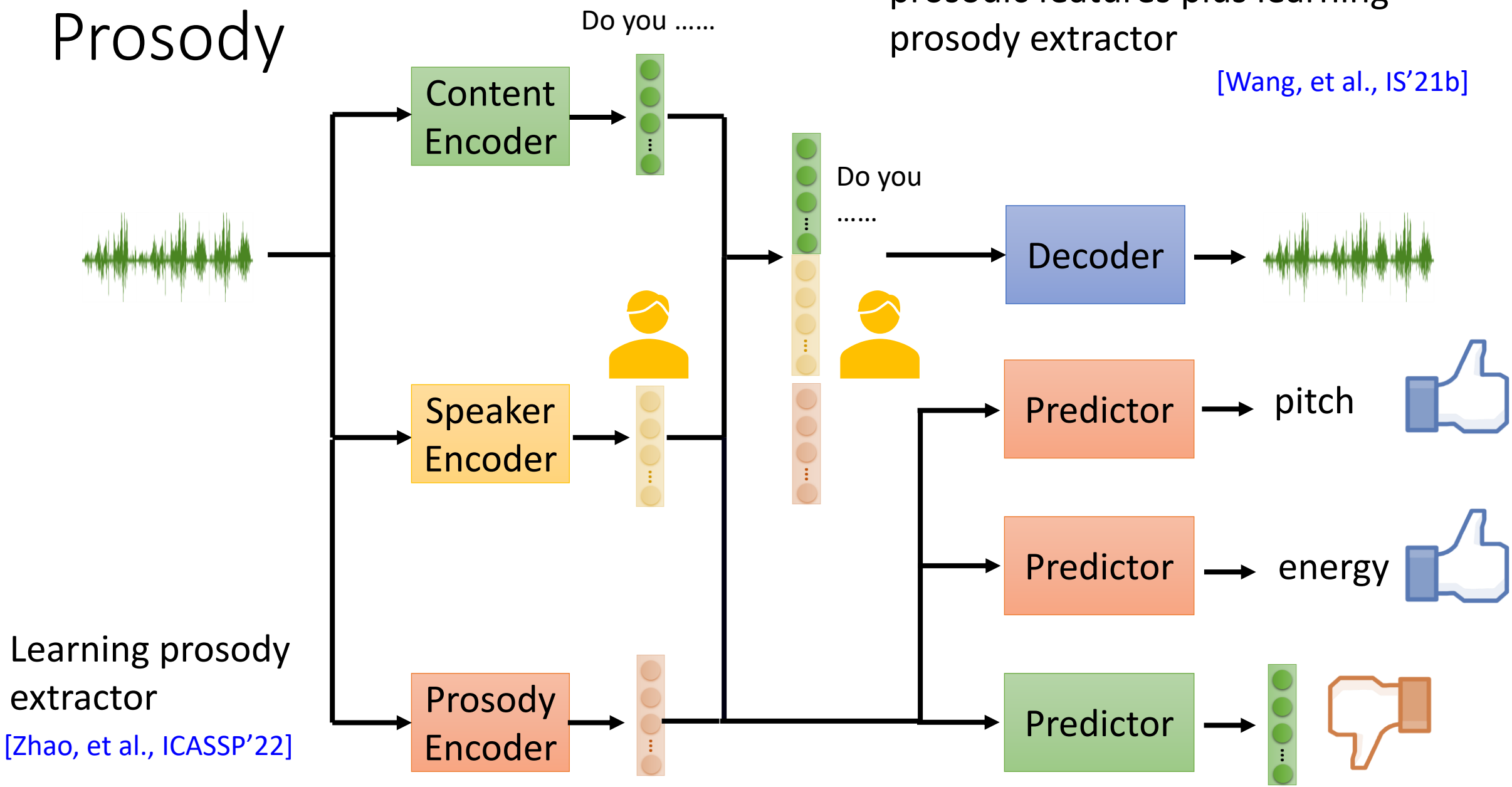
Prosody



prosodic features: Use toolkit to extract per-frame F0 contour, energy, etc.

[Qian, et al., ICASSP'20]

Prosody



prosodic features plus learning prosody extractor

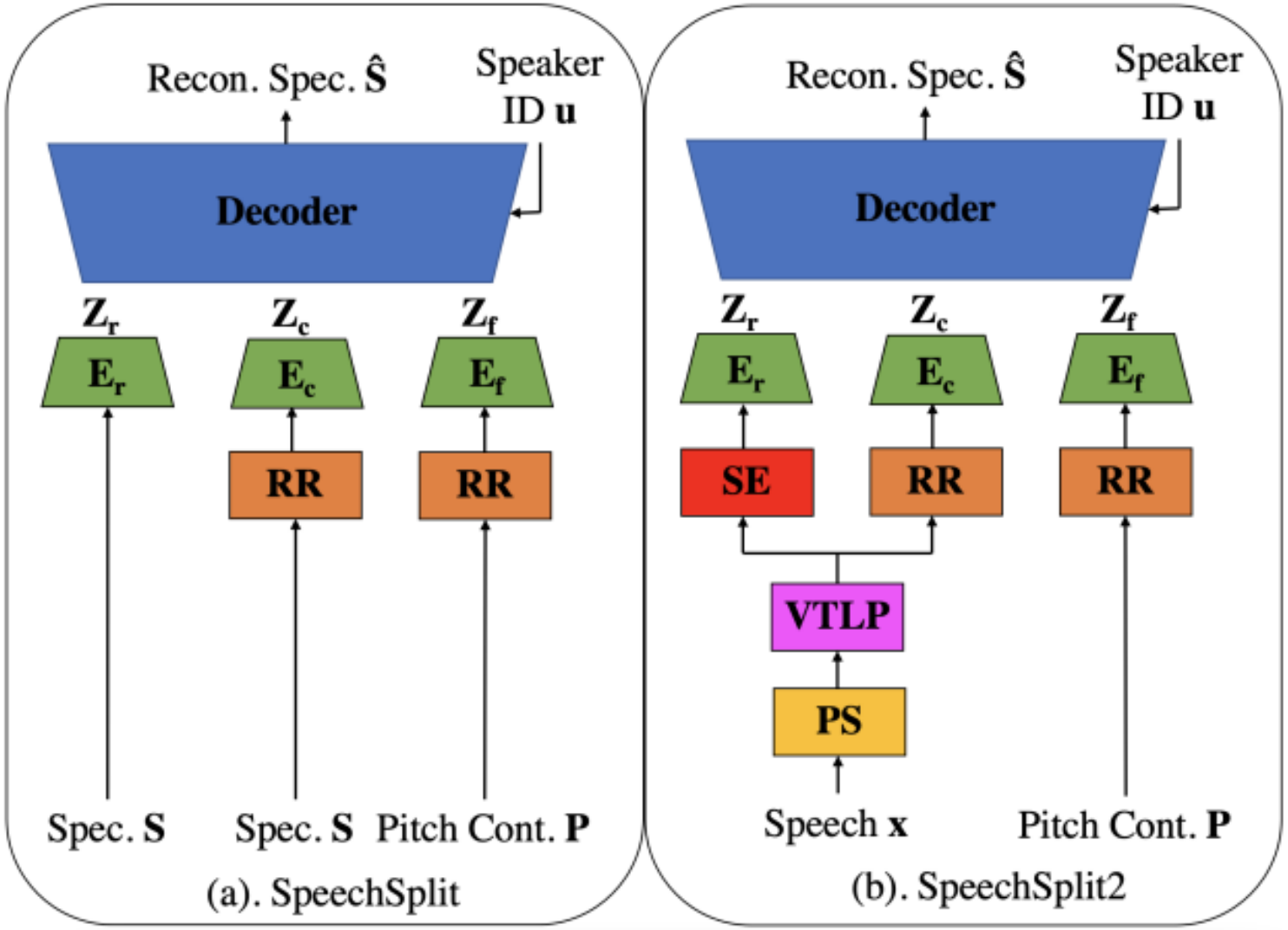
[Wang, et al., IS'21b]

Learning prosody extractor

[Zhao, et al., ICASSP'22]

Speech Split

AutoPST disentangles rhythm using similarity-based time resampling. [Qian, et al., ICML'21]

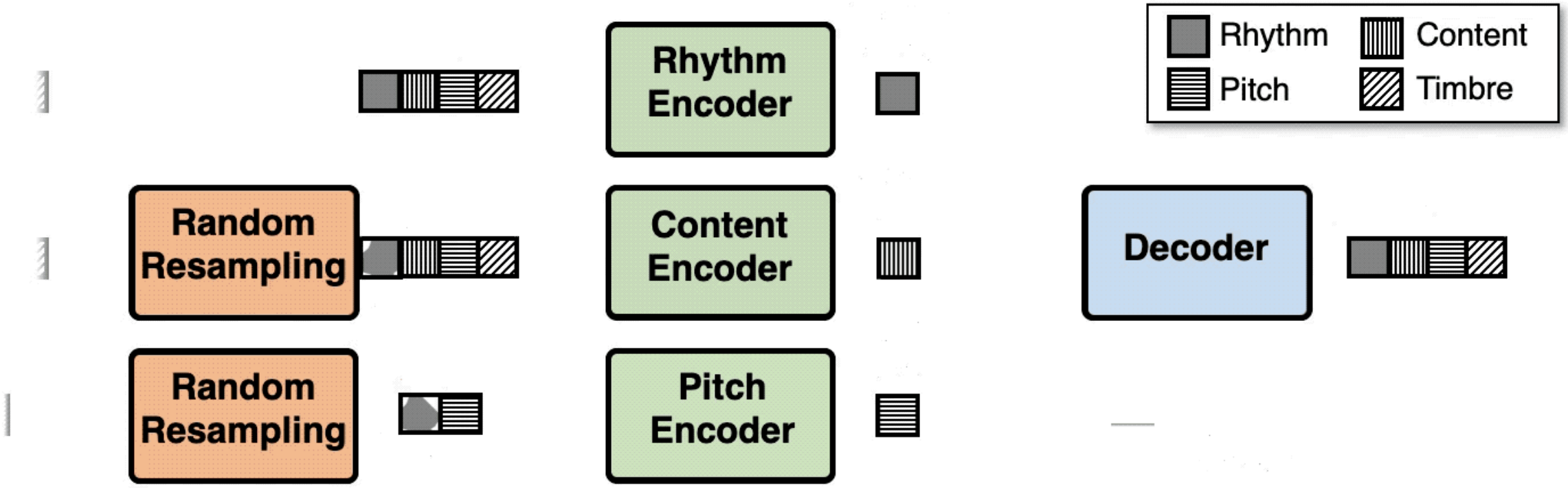


[Qian, et al., ICML'20]

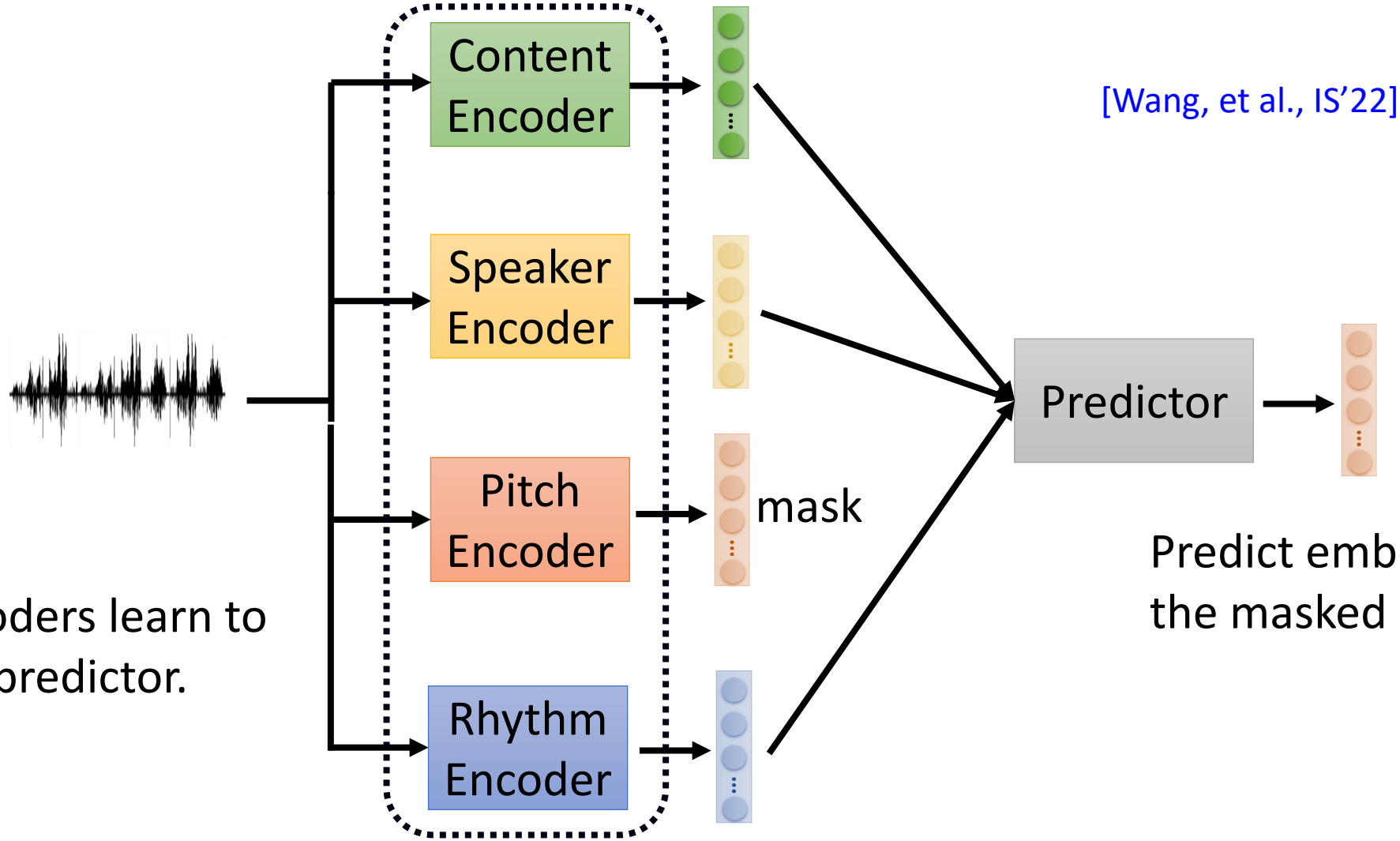
[Chan, et al., ICASSP'22]

Speech Split

- Demo: https://auspicious3000.github.io/SpeechSplit-Demo/



Adversarial Mask-And-Predict



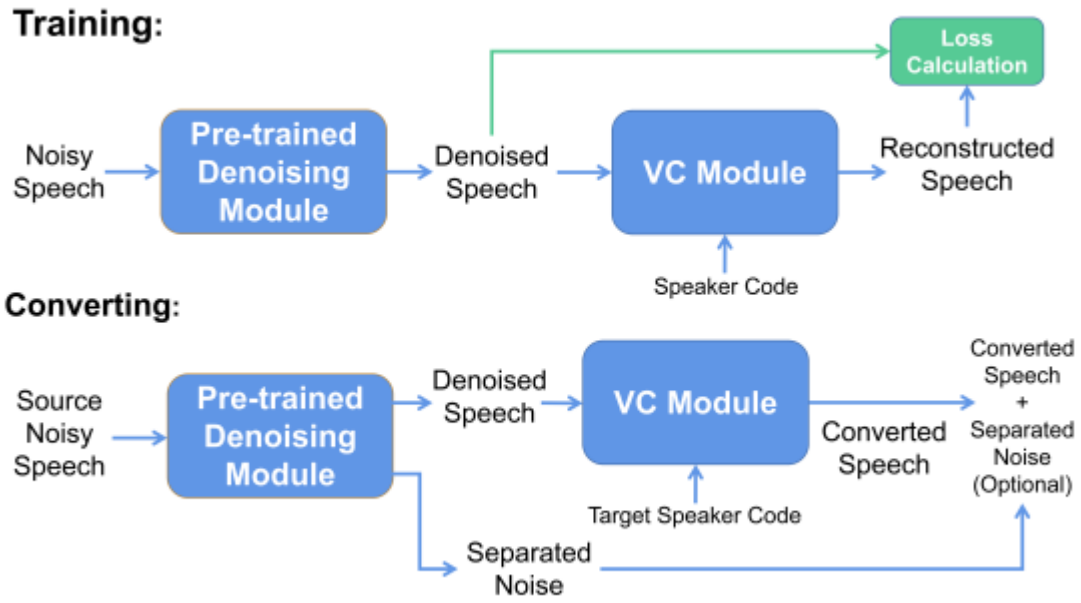
[Wang, et al., IS'22]

The encoders learn to fool the predictor.

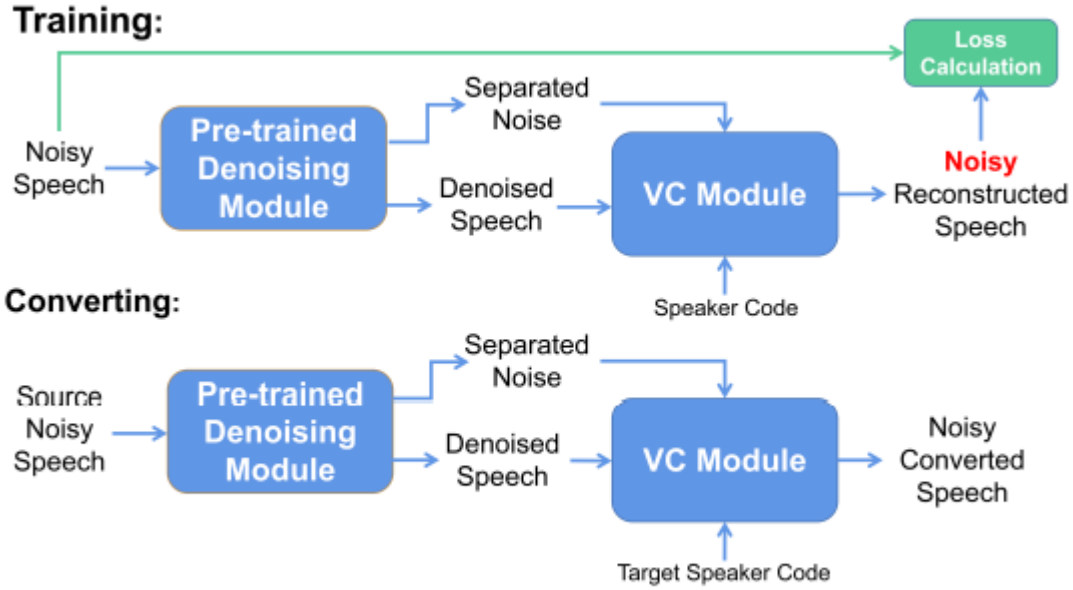
Predict embedding of the masked attribute

Background Sounds

- Background noise can harm the VC models' performance. Removal of background noise by speech enhancement before VC. [Huang, et al., ICASSP'22a]
- VC in movie/video: convert the speaker's identity while preserving the background sounds. [Xie, et al., ICASSP'22]



(a) Baseline framework



(b) Proposed framework

Outline

Introduction of Voice Conversion (VC)

VC with Unparallel Data

Beyond Speaker Conversion

VC plus Self-supervised Learning

Security Issue

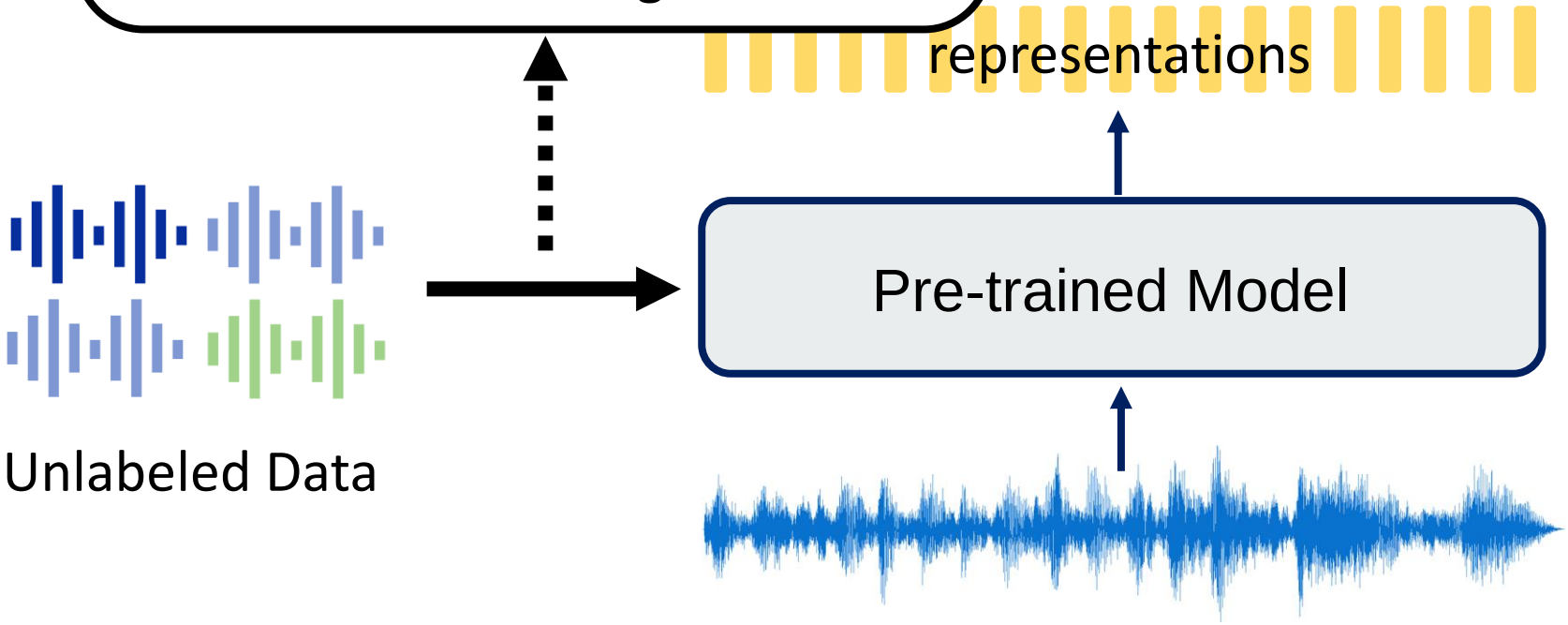
Self-supervised Learning Framework

Phase 1: Pre-train

(not complete survey)

- Mask the input signals and then reconstruct them.
- Predict the targets obtained without human efforts.
- Contrastive learning

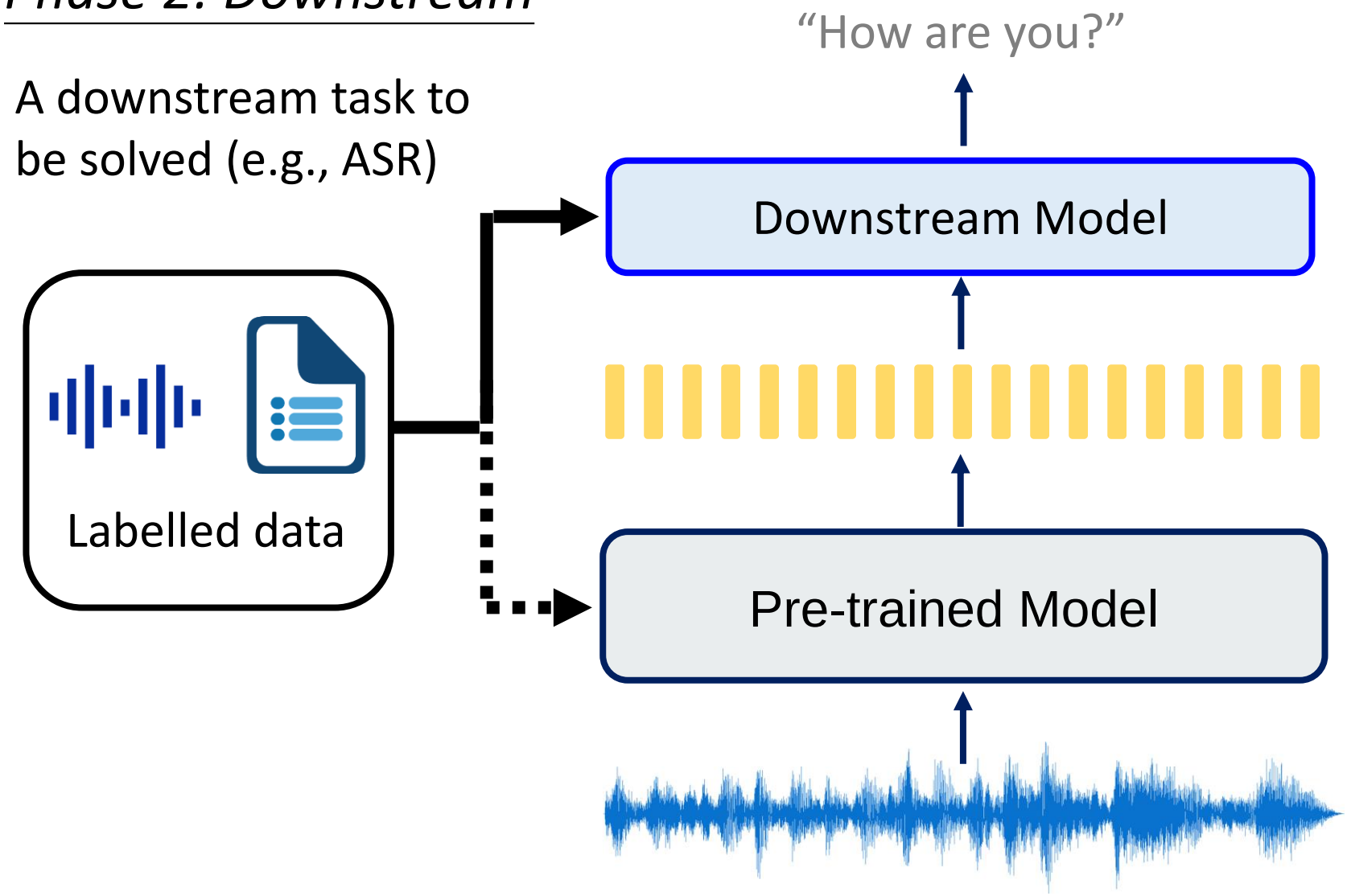
Task-agnostic



Self-supervised Learning Framework

Phase 2: Downstream

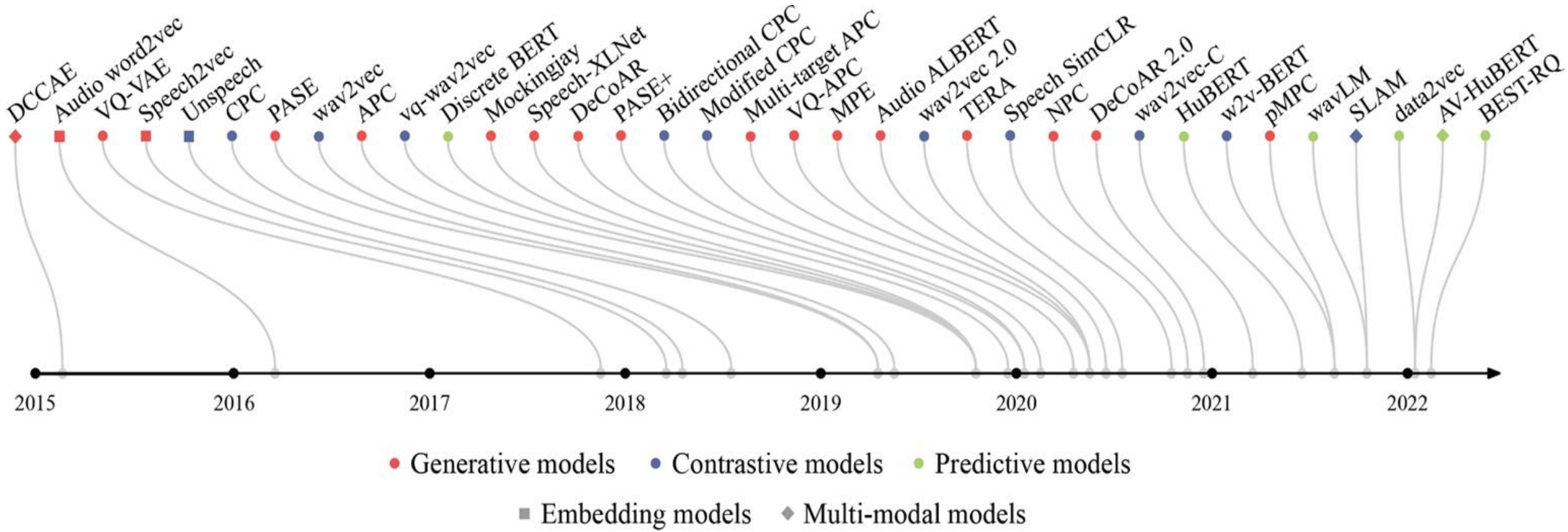
A downstream task to be solved (e.g., ASR)



Self-Supervised Speech Representation Learning: A Review

Abdelrahman Mohamed*, Hung-yi Lee*, Lasse Borgholt*, Jakob D. Havtorn*, Joakim Edin, Christian Igel
Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, Shinji Watanabe

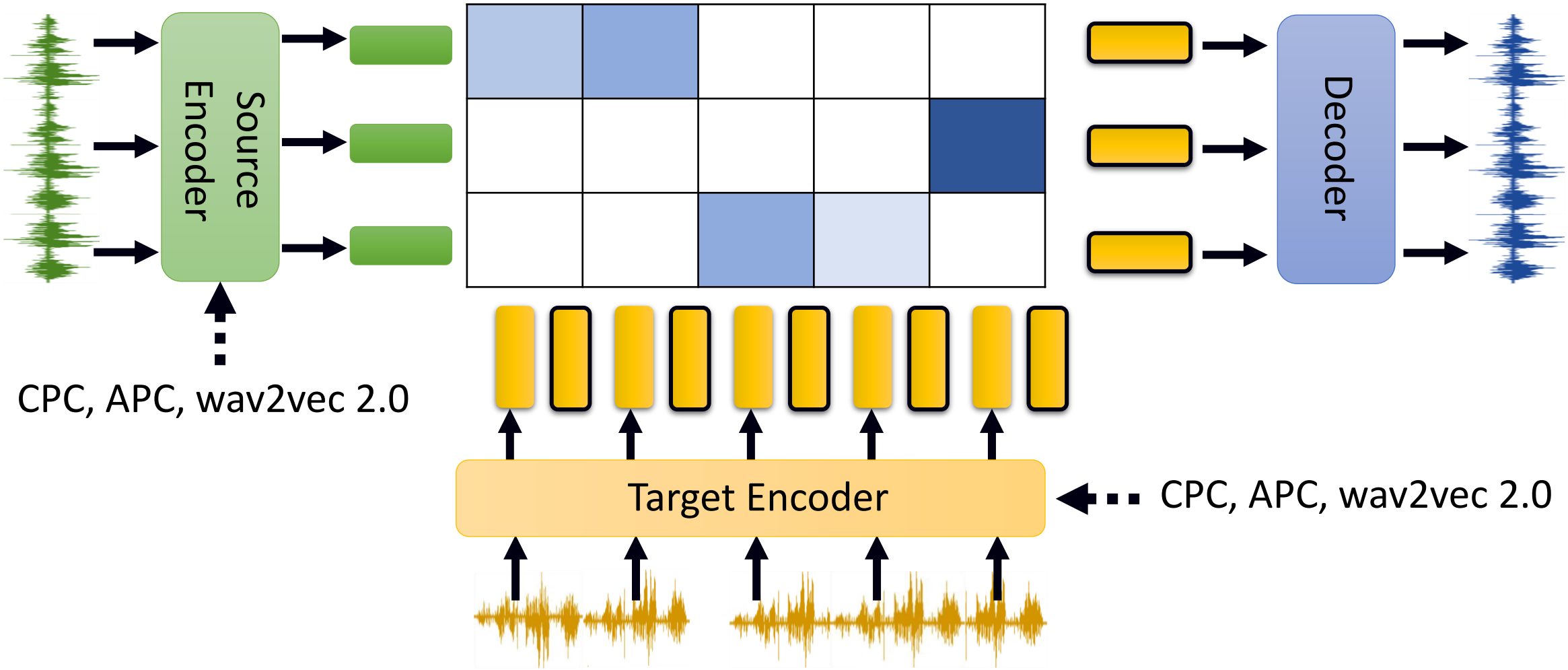
https://arxiv.org/abs/2205.10643



Fragment VC [Lin, et al., IS'21]

+ Sefl-supervised models

CPC for both source encoder and target encoder achieves the best performance.



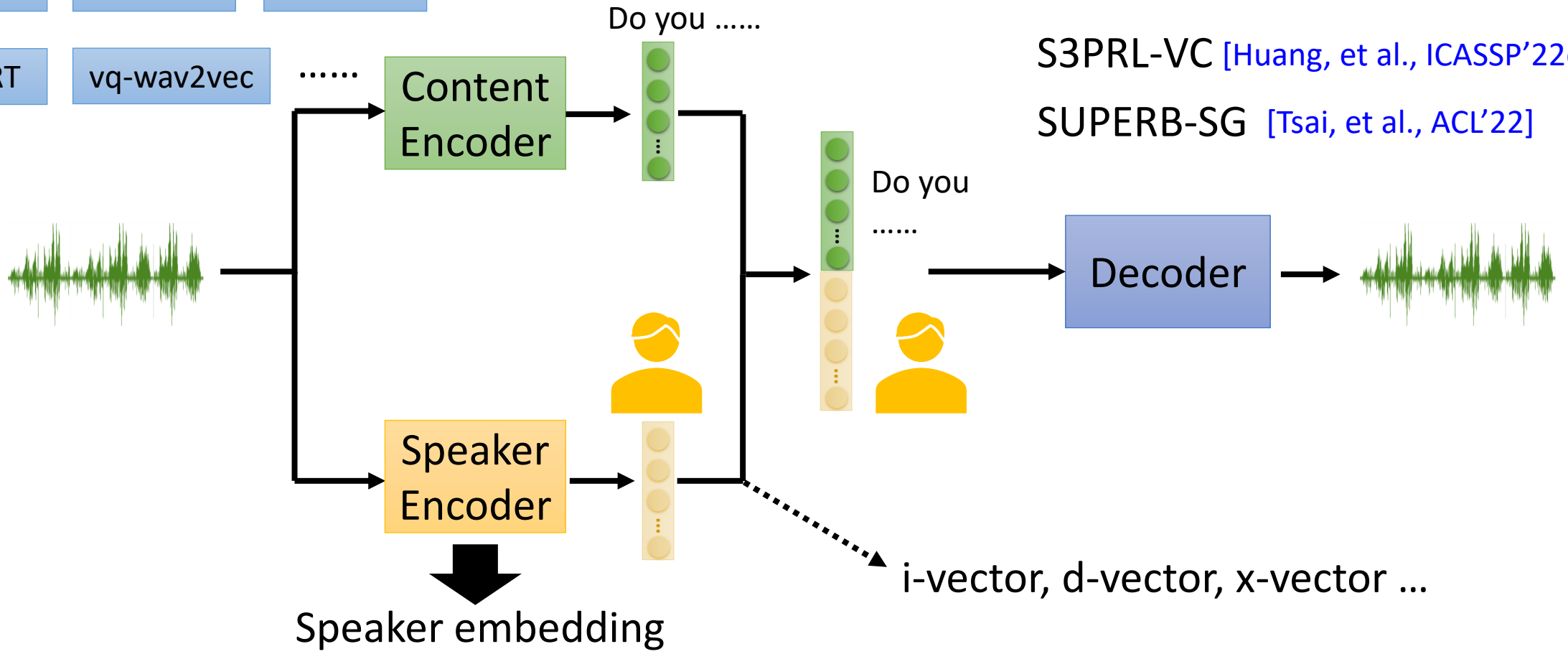
Self-supervised model as Content Encoder

- vq-wav2vec is the best
- Self-supervised models trained on a monolingual corpus can still work well in the cross-lingual setting

PASE+ | APC | NPC

Tera | DeCoAR | wav2vec

HuBERT | vq-wav2vec | .....

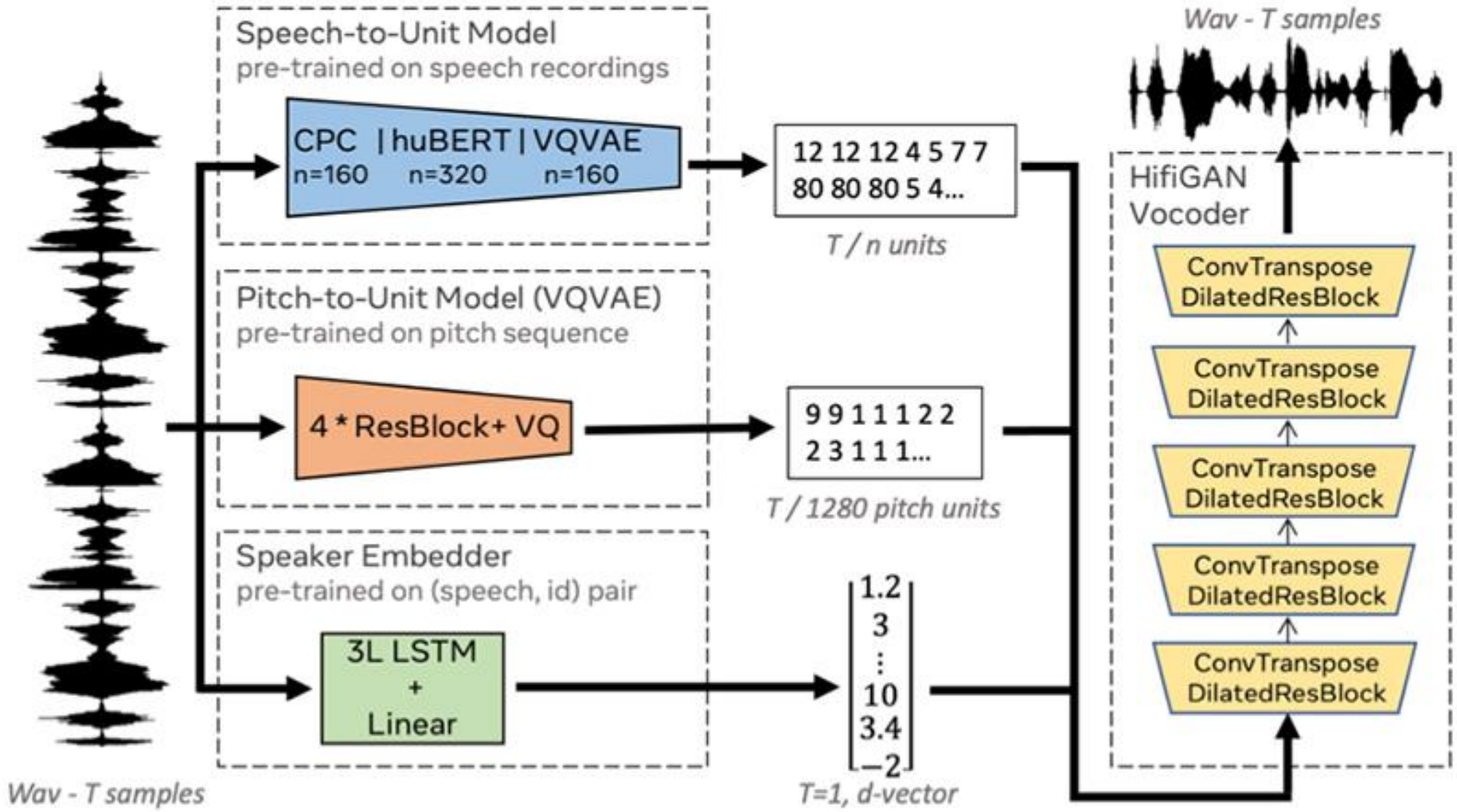


S3PRL-VC [Huang, et al., ICASSP'22c]

SUPERB-SG [Tsai, et al., ACL'22]

Self-supervised model as Content Encoder

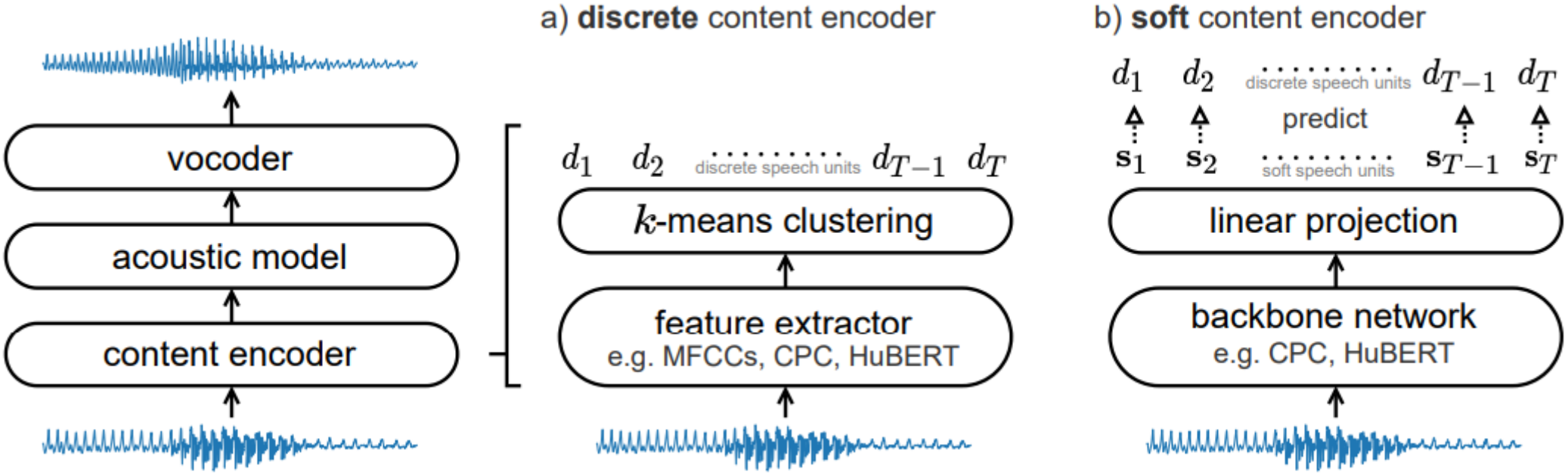
Speech Resynthesis from Discrete Disentangled Self-Supervised Representations



[Polyak, et al., IS'21]

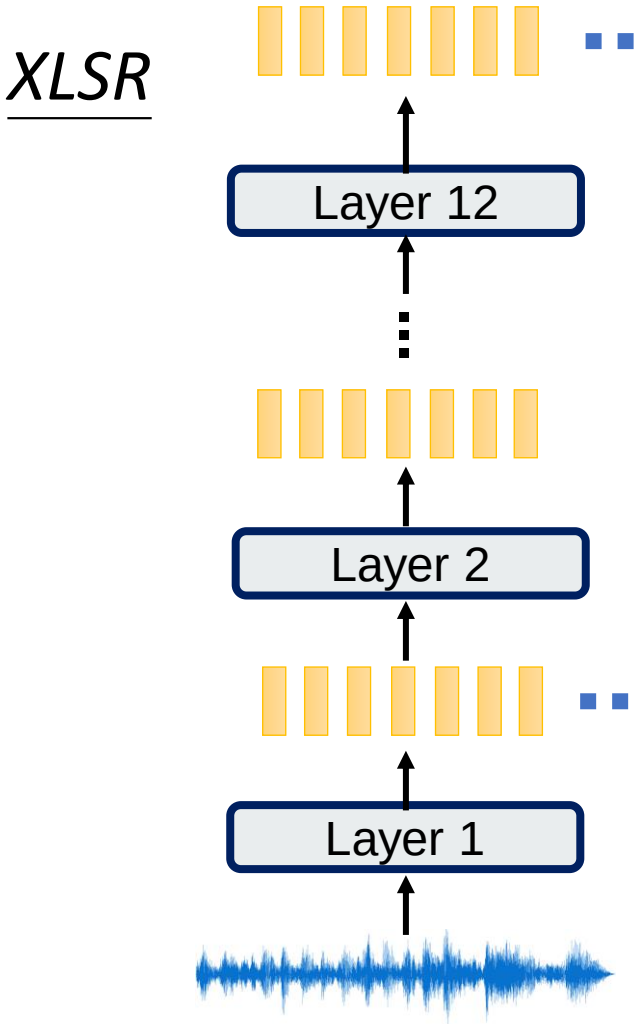
Self-supervised Model as Content Encoder

- The discrete representation effectively removes speaker information.
- But some language content is discarded resulting in mispronunciation.

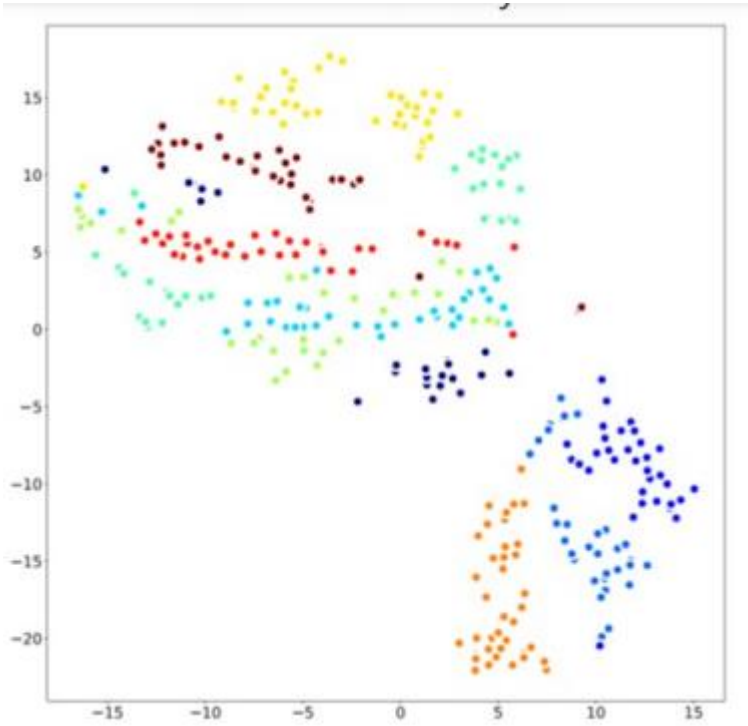


[Niekerk , et al., ICASSP'22]

Disentanglement from Self-supervised Model



Content Embedding?

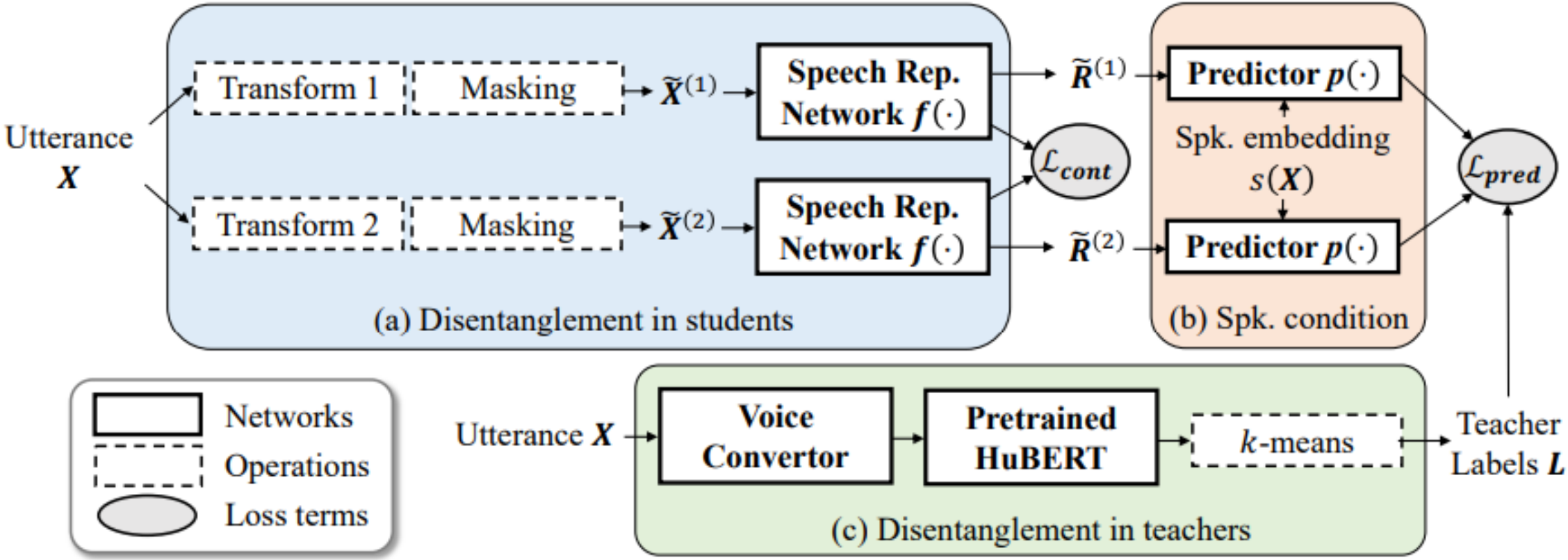


Speaker Embedding

Neural analysis and synthesis (NANSY)

[Choi, et al., NeurIPS'21]

Disentanglement from Self-supervised Model



[Qian, et al., ICML'22]

Outline

Introduction of Voice Conversion (VC)

VC with Unparallel Data

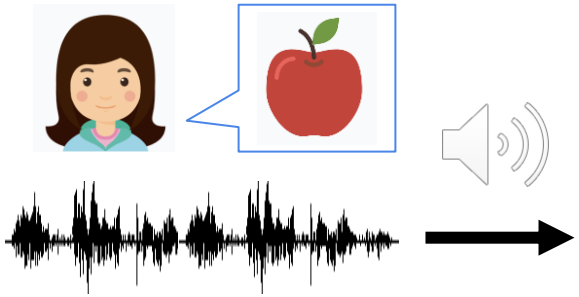
Beyond Speaker Conversion

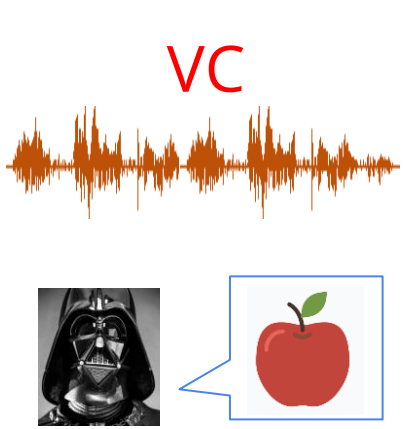
VC plus Self-supervised Learning

Security Issue

Spoofing Detection

Speech generated by voice conversion can fool both humans and speaker verification system.





VC

Anti-Spoofing Model

Other Processing

Non-Spoofing

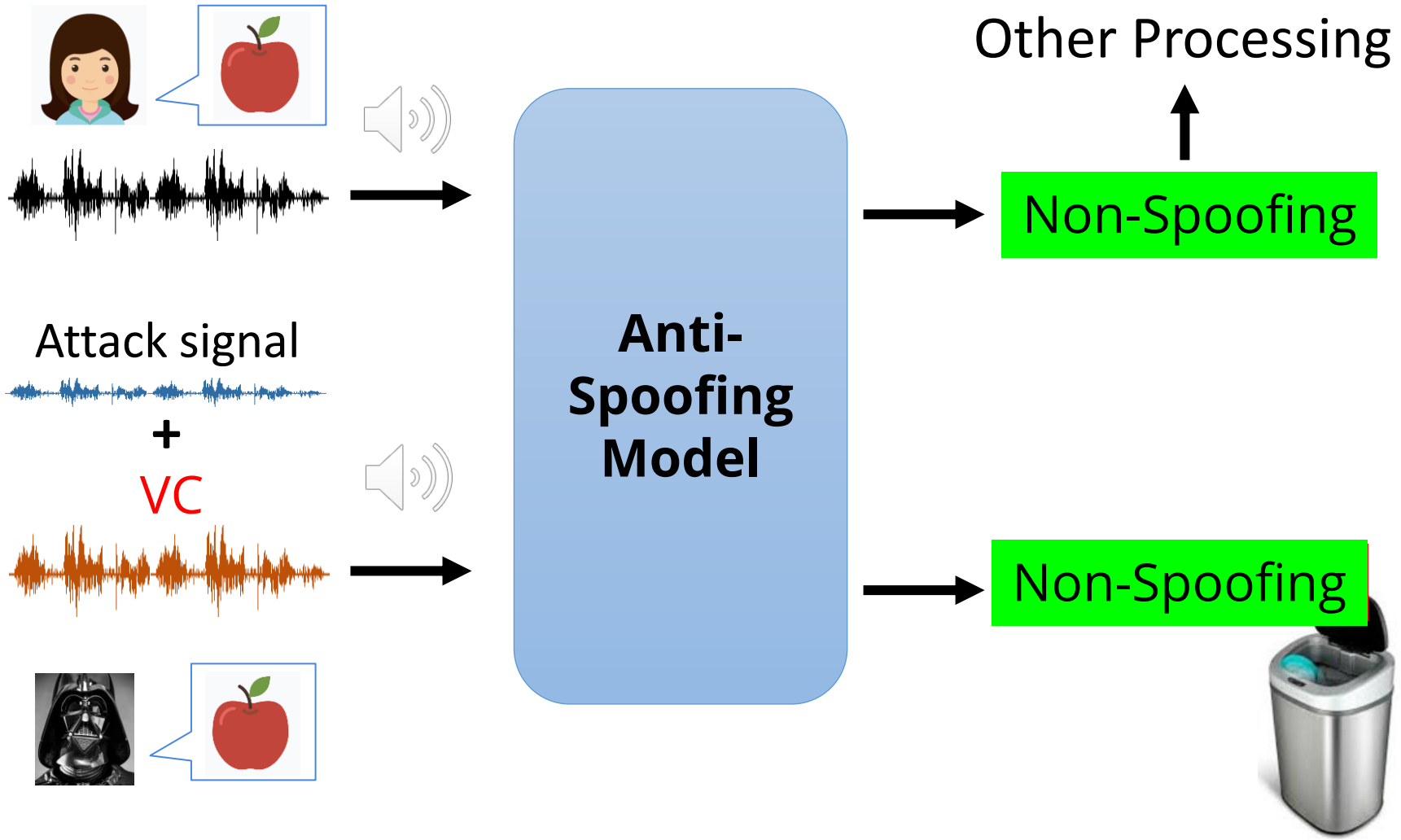
ASV spoofing challenge

https://www.asvspoof.org/

Spoofing

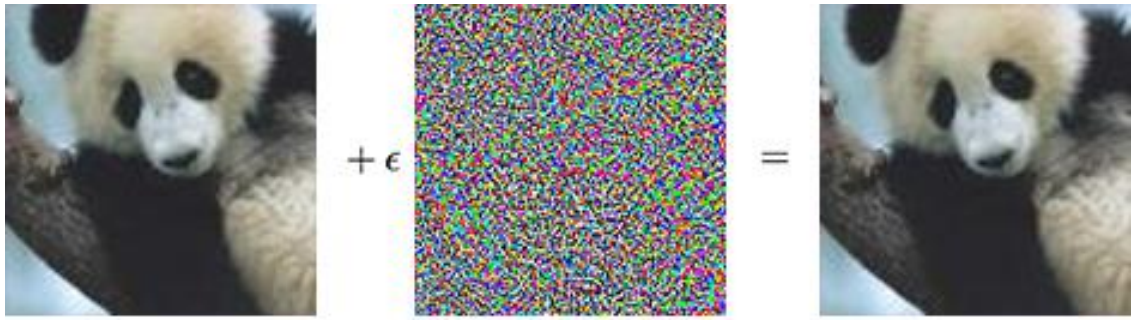
Adversarial Attack

[Ding, et al., IS'21]
[Liu, et al., ASRU'19]



Set a thief to catch a thief

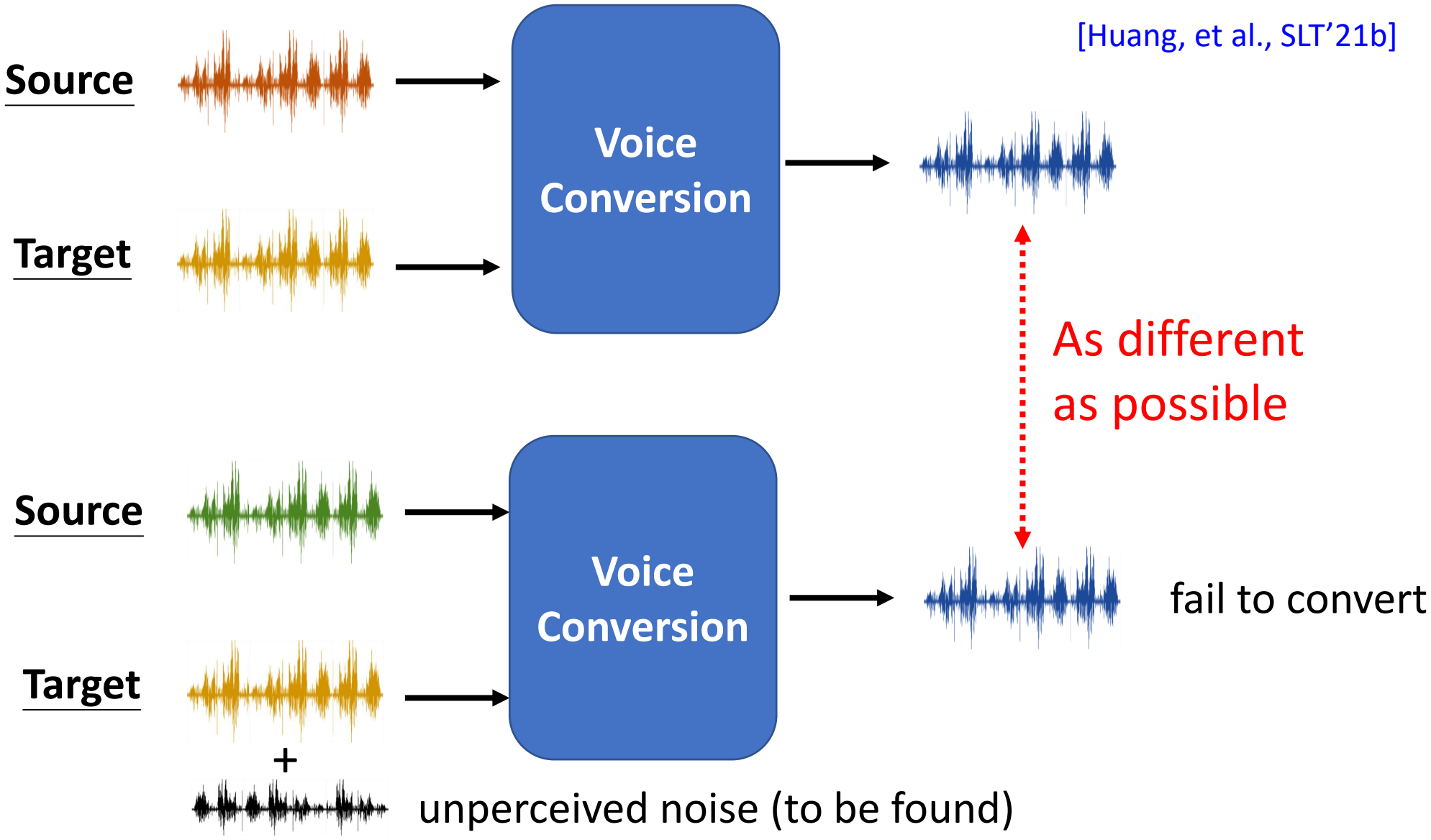
- Adversarial Attack



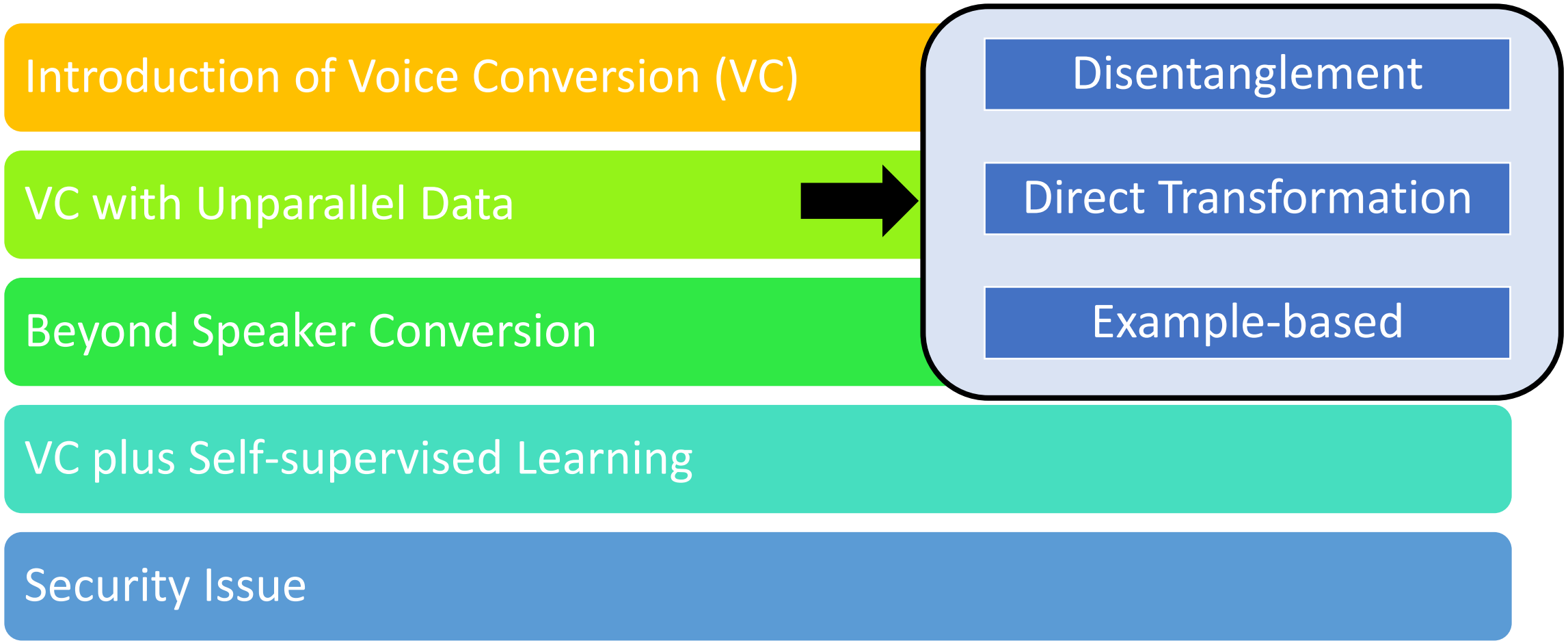
- Adversarial Attack to VC model!

Make VC model fail to convert

Adversarial Attack to VC model



Concluding Remarks



Reference

Reference

- [Biadsy, et al., INTERSPEECH'19] Fadi Biadsy, Ron J. Weiss, Pedro J. Moreno, Dimitri Kanevsky, Ye Jia, Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation, INTERSPEECH, 2019
- [Chen et al., IS'19] Li-Wei Chen, Hung-Yi Lee, Yu Tsao, Generative adversarial networks for unpaired voice transformation on impaired speech, INTERSPEECH, 2019
- [Chen, et al., IS'21] Ziyi Chen, Pengyuan Zhang, TVQVC: Transformer based Vector Quantized Variational Autoencoder with CTC loss for Voice Conversion, INTERSPEECH, 2021
- [Chen, et al., ICASSP'21a] Mingjie Chen, Yanpei Shi, Thomas Hain, Towards Low-Resource Stargan Voice Conversion Using Weight Adaptive Instance Normalization, ICASSP, 2021
- [Chen, et al., ICASSP'21b] Yen-Hao Chen, Da-Yi Wu, Tsung-Han Wu, Hung-yi Lee, AGAIN-VC: A One-shot Voice Conversion using Activation Guidance and Adaptive Instance Normalization, ICASSP 2021
- [Chan, et al., ICASSP'22] Chak Ho Chan, Kaizhi Qian, Yang Zhang, Mark Hasegawa-Johnson, SpeechSplit 2.0: Unsupervised speech disentanglement for voice conversion Without tuning autoencoder Bottlenecks, ICASSP, 2022

Reference

- [Choi, et al., NeurIPS'21] Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Hwan Lee, Hoon Heo, Kyogu Lee, Neural Analysis and Synthesis: Reconstructing Speech from Self-Supervised Representations, NeurIPS, 2021
- [Chan, et al., IS'22] David M. Chan, Shalini Ghosh, Content-context factorized representations for automated speech recognition, INTERSPEECH, 2022
- [Cheng, et al., ICML'20] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, Lawrence Carin, CLUB: A Contrastive Log-ratio Upper Bound of Mutual Information, ICML, 2020
- [Chou, et al., INTERSPEECH'18] Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee, Lin-shan Lee, "Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations", INTERSPEECH, 2018
- [Chou, et al., INTERSPEECH'19] Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee, "One-shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization", INTERSPEECH, 2019

Reference

- [Dang, et al., ICASSP‘22] Trung Dang, Dung Tran, Peter Chin, Kazuhito Koishida, Training Robust Zero-Shot Voice Conversion Models with Self-Supervised Features, ICASSP, 2022
- [Deng, et al., ICASSP’20] Chengqi Deng, Chengzhu Yu, Heng Lu, Chao Weng, Dong Yu, PitchNet: Unsupervised Singing Voice Conversion with Pitch Adversarial Network, ICASSP, 2020
- [Ding, et al., IS’21] Yi-Yang Ding, Li-Juan Liu, Yu Hu, Zhen-Hua Ling, Adversarial Voice Conversion Against Neural Spoofing Detectors, INTERSPEECH, 2021
- [Du, et al., SLT’21] Hongqiang Du, Xiaohai Tian, Lei Xie, Haizhou Li, Optimizing voice conversion network with cycle consistency loss of speaker identity, SLT, 2021
- [Eskimez, et al., IS’21] Sefik Emre Eskimez, Dimitrios Dimitriadis, Kenichi Kumatani, Robert Gmyr, One-Shot Voice Conversion with Speaker-Agnostic StarGAN, INTERSPEECH, 2021
- [Gao, et al., INTERSPEECH’19] Jian Gao, Deep Chakraborty, Hamidou Tembine, Olaitan Olaleye, Nonparallel Emotional Speech Conversion, INTERSPEECH, 2019
- [He, et al., IS’21] Xiangheng He, Junjie Chen, Georgios Rizos, Björn W. Schuller, An Improved StarGAN for Emotional Voice Conversion: Enhancing Voice Quality and Data Augmentation, INTERSPEECH, 2021

Reference

- [Huang, et al., IS'20] Wen-Chin Huang, Tomoki Hayashi, Yi-Chiao Wu,Hirokazu Kameoka,Tomoki Toda, Voice Transformer Network: Sequence-to-Sequence Voice Conversion Using Transformer with Text-to-Speech Pretraining, INTERSPEECH, 2020
- [Huang, et al., ICASSP'21] Wen-Chin Huang, Yi-Chiao Wu, Tomoki Hayashi, Tomoki Toda, Any-to-One Sequence-to-Sequence Voice Conversion using Self-Supervised Discrete Speech Representations, ICASSP, 2021
- [Huang, et al., IS'21] Wen-Chin Huang, Kazuhiro Kobayashi, Yu-Huai Peng, Ching-Feng Liu, Yu Tsao, Hsin-Min Wang, Tomoki Toda, A Preliminary Study of a Two-Stage Paradigm for Preserving Speaker Identity in Dysarthric Voice Conversion, INTERSPEECH, 2021
- [Huang, et al., ICASSP'22a] Chien-yu Huang, Kai-Wei Chang, Hung-yi Lee, Toward Degradation-Robust Voice Conversion, ICASSP, 2022
- [Huang, et al., ICASSP'22b] Wen-Chin Huang, Bence Mark Halpern, Lester Phillip Violeta, Odette Scharenborg, Tomoki Toda, Towards Identity Preserving Normal to Dysarthric Voice Conversion, ICASSP, 2022
- [Huang, et al., ICASSP'22c] Wen-Chin Huang, Shu-Wen Yang, Tomoki Hayashi, Hung-Yi Lee, Shinji Watanabe, Tomoki Toda, S3PRL-VC: Open-source Voice Conversion Framework with Self-supervised Speech Representations, ICASSP, 2022

Reference

- [Huang, et al., IS'22] Wen Chin Huang, Dejan Markovic, Alexander Richard, Israel Dejene Gebru, Anjali Menon, End-to-End Binaural Speech Synthesis, INTERSPEECH, 2022
- [Hsu, et al., APSIPA'16] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, Hsin-Min Wang, Voice Conversion from Non-parallel Corpora Using Variational Auto-encoder, APSIPA, 2016
- [Jin, et al., ICASSP'16] Zeyu Jin; Adam Finkelstein; Stephen DiVerdi; Jingwan Lu; Gautham J. Mysore, Cute: A concatenative method for voice conversion using exemplar-based unit selection, ICASSP, 2016
- [Joan, et al., NeurIPS'19] Joan Serrà, Santiago Pascual, Carlos Segura, Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion, NeurIPS, 2019
- [Kaneko, et al., arXiv'17] Takuhiro Kaneko, Hirokazu Kameoka, Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks, arXiv, 2017
- [Kaneko, et al., ICASSP'19] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, CycleGAN-VC2: Improved CycleGAN-based Non-parallel Voice Conversion, ICASSP 2019
- [Kaneko, et al., IS'19] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion, INTERSPEECH 2019

Reference

- [Kaneko, et al., IS'20] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-spectrogram Conversion, INTERSPEECH, 2020
- [Kaneko, et al., ICASSP'21] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, MaskCycleGAN-VC: Learning Non-parallel Voice Conversion with Filling in Frames, ICASSP 2021
- [Kameoka, et al., SLT'18] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, Nobukatsu Hojo, StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks, SLT, 2018
- [Keskin, et al., ICML workshop'19] Gokce Keskin, Tyler Lee, Cory Stephenson, Oguz H. Elibol, Measuring the Effectiveness of Voice Conversion on Speaker Identification and Automatic Speech Recognition Systems, ICML workshop, 2019
- [Kim, et al., ICASSP'22] Kang-wook Kim, Seung-won Park, Junhyeok Lee, Myun-chul Joe, Assem-VC: Realistic Voice Conversion by Assembling Modern Speech Synthesis Techniques, ICASSP, 2022
- [Kobayashi, et al., ICASSP'21] Kazuhiro Kobayashi, Wen-Chin Huang, Yi-Chiao Wu, Patrick Lumban Tobing, Tomoki Hayashi, Tomoki Toda, crank: An Open-Source Software for Nonparallel Voice Conversion Based on Vector-Quantized Variational Autoencoder, ICASSP, 2021

Reference

- [Li, et al., IS'20] Yanping Li, Dongxiang Xu, Yan Zhang, Yang Wang, Binbin Chen, Non-parallel Many-to-many Voice Conversion with PSR-StarGAN, INTERSPEECH, 2020
- [Li, et al., IS'21] Tingle Li, Yichen Liu, Chenxu Hu, Hang Zhao, CVC: Contrastive Learning for Non-parallel Voice Conversion, INTERSPEECH, 2021
- [Lin, et al., ICASSP'21] Yist Y. Lin, Chung-Ming Chien, Jheng-Hao Lin, Hung-yi Lee, Lin-shan Lee, FragmentVC: Any-to-Any Voice Conversion by End-to-End Extracting and Fusing Fine-Grained Voice Fragments With Attention, ICASSP, 2021
- [Lin, et al., IS'21] Jheng-hao Lin, Yist Y. Lin, Chung-Ming Chien, Hung-yi Lee, S2VC: A Framework for Any-to-Any Voice Conversion with Self-Supervised Pretrained Representations, INTERSPEECH, 2021
- [Liu, et al., INTERSPEECH'18] Songxiang Liu, Jinghua Zhong, Lifa Sun, Xixin Wu, Xunying Liu, Helen Meng, Voice Conversion Across Arbitrary Speakers Based on a Single Target-Speaker Utterance, INTERSPEECH, 2018
- [Liu, et al., IS'19] Andy T. Liu, Po-chun Hsu and Hung-yi Lee, Unsupervised End-to-End Learning of Discrete Linguistic Units for Voice Conversion, INTERSPEECH, 2019
- [Liu, et al., ASRU'19] Songxiang Liu, Haibin Wu, Hung-yi Lee, Helen Meng, Adversarial attacks on spoofing countermeasures of automatic speaker verification, ASRU, 2019

Reference

- [Luo, et al., ICASSP‘20] Yin-Jyun Luo, Chin-Chen Hsu, Kat Agres, Dorien Herremans, Singing Voice Conversion with Disentangled Representations of Singer and Vocal Technique Using Variational Autoencoders, ICASSP, 2020
- [Luong, et al., ASRU’19] Hieu-Thi Luong, Junichi Yamagishi, Bootstrapping non-parallel voice conversion from speaker-adaptive text-to-speech, ASRU, 2019
- [Mimura, et al., ASRU 2017]Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara, Cross-domain Speech Recognition Using Nonparallel Corpora with Cycle-consistent Adversarial Networks, ASRU, 2017
- [Merritt, et al., ICASSP’22] Thomas Merritt, Abdelhamid Ezzerg, Piotr Biliński, Magdalena Proszewska, Kamil Pokora, Roberto Barra-Chicote, Daniel Korzekwa, Text-free non-parallel many-to-many voice conversion using normalising flows, ICASSP, 2022
- [Nachmani, et al., INTERSPEECH’19] Eliya Nachmani, Lior Wolf, Unsupervised Singing Voice Conversion, INTERSPEECH, 2019
- [Nguyen, et al., ICASSP’22] Bac Nguyen, Fabien Cardinaux, NVC-Net: End-to-End Adversarial Voice Conversion, ICASSP, 2022

Reference

- [Niekerk , et al., ICASSP'22] Benjamin van Niekerk, Marc-André Carbonneau, Julian Zaïdi, Mathew Baas, Hugo Seuté, Herman Kamper, A Comparison of Discrete and Soft Speech Units for Improved Voice Conversion, ICASSP, 2022
- [Park, et al., IS'20] Seung-won Park, Doo-young Kim, Myun-chul Joe, Cotatron: Transcription-Guided Speech Encoder for Any-to-Many Voice Conversion without Parallel Data, INTERSPEECH, 2020
- [Patel, et al., SSW'19] Maitreya Patel, Mihir Parmar, Savan Doshi, Nirmesh Shah and Hemant A. Patil, Novel Inception-GAN for Whisper-to-Normal Speech Conversion, ISCA Speech Synthesis Workshop, 2019
- [Polyak, et al., IS'21] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, Speech Resynthesis from Discrete Disentangled Self-Supervised Representations, INTERSPEECH, 2021
- [Pucher, et al., IS'21] Michael Pucher, Thomas Woltron, Conversion of Airborne to Bone-Conducted Speech with Deep Neural Networks, INTERSPEECH, 2021

Reference

- [Qian, et al., ICML'19] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Mark Hasegawa-Johnson, AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss, ICML, 2019
- [Qian, et al., ICML'20] Kaizhi Qian, Yang Zhang, Shiyu Chang, David Cox, Mark Hasegawa-Johnson, Unsupervised Speech Decomposition via Triple Information Bottleneck, ICML, 2020
- [Qian, et al., ICASSP'20] Kaizhi Qian, Zeyu Jin, Mark Hasegawa-Johnson, Gautham J. Mysore, F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder, ICASSP, 2020
- [Qian, et al., ICML'21] Kaizhi Qian, Yang Zhang, Shiyu Chang, Jinjun Xiong, Chuang Gan, David Cox, Mark Hasegawa-Johnson, Global Rhythm Style Transfer Without Text Transcriptions, ICML, 2021
- [Qian, et al., ICML'22] Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, David Cox, Mark Hasegawa-Johnson, Shiyu Chang, ContentVec: An improved self-supervised speech representation by disentangling speakers, ICML, 2022
- [Richard, et al., ICLR'21] Alexander Richard, Dejan Markovic, Israel D. Gebru, Steven Krenn, Gladstone Alexander Butler, Fernando Torre, Yaser Sheikh, Neural Synthesis of Binaural Speech From Mono Audio, ICLR, 2021

Reference

- [Seshadri, et al., ICASSP'19] Shreyas Seshadri, Lauri Juvela, Junichi Yamagishi, Okko Räsänen, Paavo Alku,Cycle-consistent Adversarial Networks for Non-parallel Vocal Effort Based Speaking Style Conversion, *ICASSP, 2019*
- [Shi, et al., ICASSP'22] Sheng Shi, Jiahao Shao, Yifei Hao, Yangzhou Du, Jianping Fan, U-GAT-VC: Unsupervised Generative Attentional Networks for Non-Parallel Voice Conversion, ICASSP. 2022
- [Srivastava, et al., ICASSP'20] Brij Mohan Lal Srivastava, Nathalie Vauquier, Md Sahidullah, Aurélien Bellet, Marc Tommasi, Emmanuel Vincent, Evaluating Voice Conversion-based Privacy Protection against Informed Attackers, ICASSP, 2020
- [Sun, et al., ICME'16] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, Helen Meng, Phonetic posteriorgrams for many-to-one voice conversion without parallel data training, ICME, 2016
- [Sundermann, et al., ICASSP'06] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, S. Narayanan, Text-Independent Voice Conversion Based on Unit Selection, ICASSP, 2006
- [Sisman, et al., ASRU'19] Berrak Sisman, Mingyang Zhang, Minghui Dong, Haizhou Li, On the study of generative adversarial networks for cross-lingual voice conversion, ASRU, 2019

Reference

- [Takashima, et al., SLT'12] Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki, Exemplar-based voice conversion in noisy environment, SLT, 2012
- [Tsai, et al., ACL'22] Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy T. Liu, Cheng-I Jeff Lai, Jiatong Shi, Xuankai Chang, Phil Hall, Hsuan-Jui Chen, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, Hung-yi Lee, SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities, ACL, 2022
- [Valle, et al., ICASSP'20] Rafael Valle, Jason Li, Ryan Prenger, Bryan Catanzaro, Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens, ICASSP, 2020
- [Wang, et al., ICASSP'20] Ruobai Wang, Yu Ding, Lincheng Li, Changjie Fan, One-Shot Voice Conversion Using Star-Gan, ICASSP, 2020
- [Wang, et al., ICASSP'22] Disong Wang, Songxiang Liu, Xixin Wu, Hui Lu, Lifa Sun, Xunying Liu, Helen Meng, Speaker Identity Preservation in Dysarthric Speech Reconstruction by Adversarial Speaker Adaptation, ICASSP, 2022
- [Wang, et al., IS'21a] Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, Helen Meng, VQMIVC: Vector Quantization and Mutual Information-Based Unsupervised Speech Representation Disentanglement for One-shot Voice Conversion, INTERSPEECH, 2021

Reference

- [Wang, et al., IS'21b] Zhichao Wang, Xinyong Zhou, Fengyu Yang, Tao Li, Hongqiang Du, Lei Xie, Wendong Gan, Haitao Chen, Hai Li, Enriching Source Style Transfer in Recognition-Synthesis based Non-Parallel Voice Conversion, INTERSPEECH, 2021
- [Wang, et al., IS'22] Jie Wang, Jingbei Li, Xintao Zhao, Zhiyong Wu, Shiyin Kang, Helen Meng, Adversarially learning disentangled speech representations for robust multi-factor voice conversion, INTERSPEECH, 2022
- [Wu, et al., ICASSP'20] Da-Yi Wu, Hung-yi Lee, ONE-SHOT VOICE CONVERSION BY VECTOR QUANTIZATION, ICASSP, 2020
- [Wu, et al., IS'20] Da-Yi Wu, Yen-Hao Chen, Hung-Yi Lee, VQVC+: One-Shot Voice Conversion by Vector Quantization and U-Net architecture, INTERSPEECH, 2020
- [Xie, et al., ICASSP'22] Chao Xie, Yi-Chiao Wu, Patrick Lumban Tobing, Wen-Chin Huang, Tomoki Toda, Direct Noisy Speech Modeling for Noisy-to-Noisy Voice Conversion, ICASSP, 2022
- [Yeh, et al., SLT'18] Cheng-chieh Yeh, Po-chun Hsu, Ju-chieh Chou, Hung-Yi Lee, Lin-shan Lee, Rhythm-Flexible Voice Conversion without Parallel Data Using Cycle-GAN over Phoneme Posteriorgram Sequences, SLT 2018

Reference

- [Zhang, et al., ICASSP'22] Haozhe Zhang, Zexin Cai, Xiaoyi Qin, Ming Li, SIG-VC: A Speaker Information Guided Zero-Shot Voice Conversion System for Both Human Beings and Machines, ICASSP, 2022
- [Zhao, et al., IS'19] Guanlong Zhao, Shaojin Ding, Ricardo Gutierrez-Osuna, Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams, INTERSPEECH, 2019
- [Zhao, et al., ICASSP'21] Shengkui Zhao; Hao Wang; Trung Hieu Nguyen; Bin Ma, Towards Natural and Controllable Cross-Lingual Voice Conversion Based on Neural TTS Model and Phonetic Posteriorgram, ICASSP, 2021
- [Zhao, et al., ICASSP'22] Xintao Zhao, Feng Liu, Changhe Song, Zhiyong Wu, Shiyin Kang, Deyi Tuo, Helen Meng, Disentangleing Content and Fine-grained Prosody Information via Hybrid ASR Bottleneck Features for Voice Conversion, ICASSP, 2022
- [Zhou, et al., ASRU'19] Yi Zhou, Xiaohai Tian, Emre Yilmaz, Rohan Kumar Das, Haizhou Li, A modularized neural network with language-specific output layers for cross-lingual voice conversion, ASRU, 2019
- [Zhou, et al., IS'21] Yi Zhou, Xiaohai Tian, Zhizheng Wu, Haizhou Li, Cross-Lingual Voice Conversion with a Cycle Consistency Loss on Linguistic Representation, INTERSPEECH, 2021
- [Zhou, et al., ICASSP'21] Yi Zhou, Xiaohai Tian, Haizhou Li, Multi-Task WaveRNN With an Integrated Architecture for Cross-Lingual Voice Conversion, ICASSP, 2021