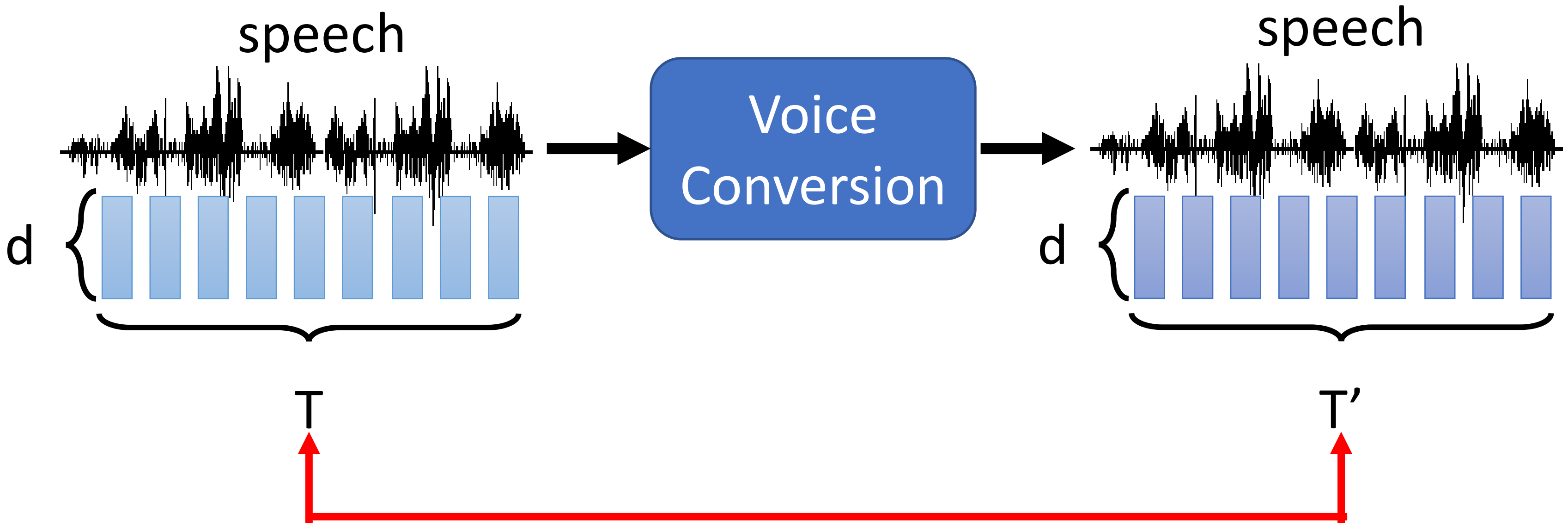
VOICE CONVERSION

Hung-yi Lee

李宏毅

What is Voice Conversion (VC)?



What is preserved? Content

What is changed? Many different aspects …

Speaker

- The same sentence said by different people has different effect.
- Deep Fake: Fool humans / speaker verification system
- One simple way to achieve personalized TTS
- Singing

[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/

Speaker

- Privacy Preserving
  [Srivastava, et al., arXiv’19]

(詳見獵人第八卷)

拷貝的帶子也壞了。
有必要連拷貝帶都毀掉嗎？
可能是不想留下線索吧。
因爲只從聲音也可以得到很多訊息。
像身高、體重、性別、年齡、臉型、健康狀況，還能推測對方的心理狀態。
另外，常常可以從背景的雜音，得知錄音的地點。

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]

🔊 Normal 🔊 Lombard

Source of audio: https://shreyas253.github.io/SpStyleConv_CycleGAN/

‘lip thrill’ (彈唇) or ‘vibrato’ (顫音)

Improving Intelligibility

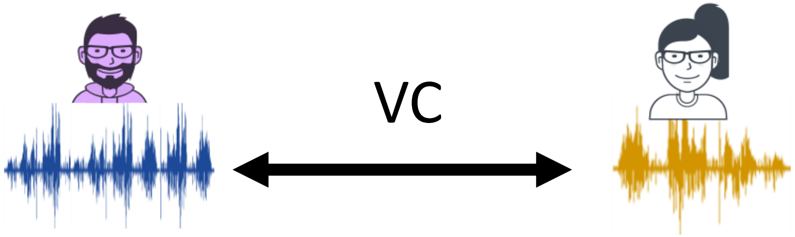
- Improving the speech intelligibility
  - surgical patients who have had parts of their articulators removed

[Biadsy, et al., INTERSPEECH’19][Chen et al., INTERSPEECH’19]

- Accent conversion
  - voice quality of a non-native speaker and the pronunciation patterns of a native speaker
  - Can be used in language learning

[Zhao, et al., INTERSPEECH’19]

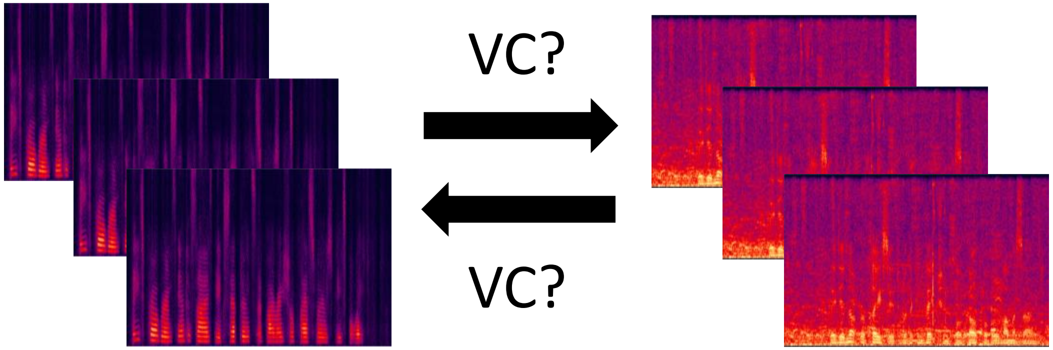
Data Augmentation



VC

Training Data x 2

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



VC?

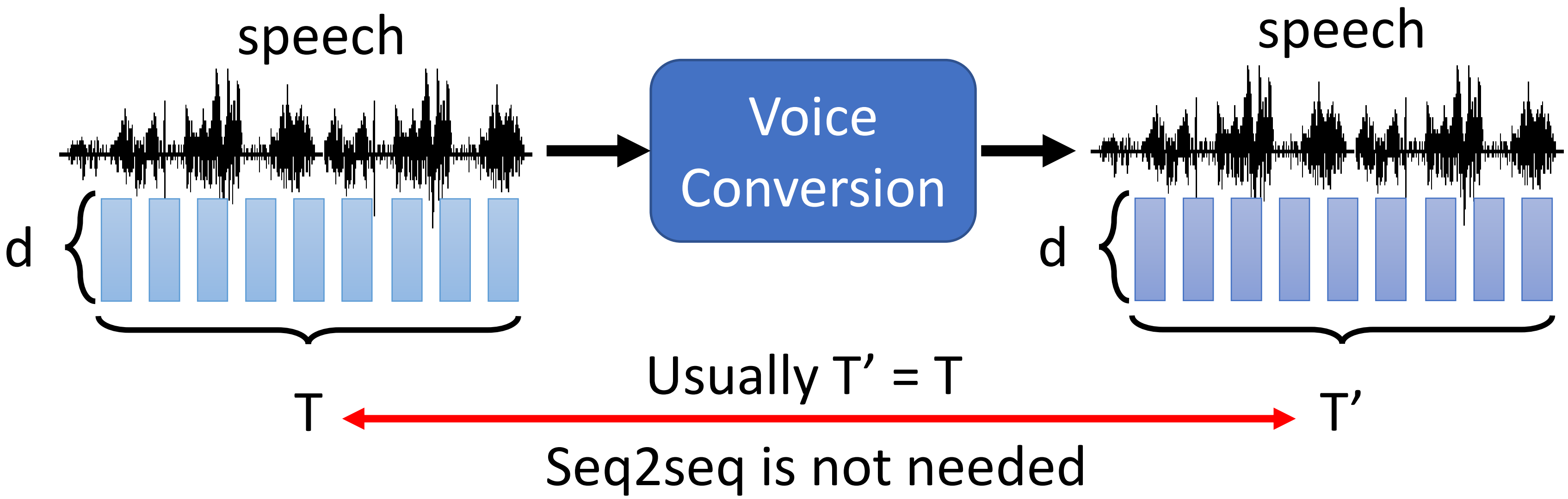
VC?

Clean Speech

Noisy Speech

[Mimura, et al., ASRU 2017]

In real implementation …



speech

Voice Conversion

speech

d

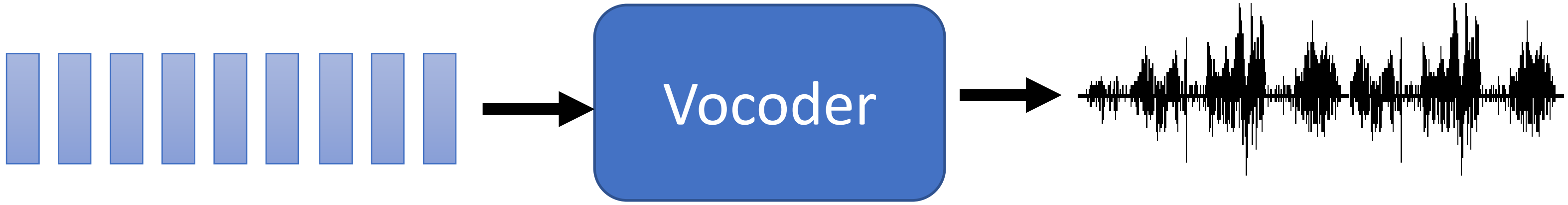
d

T

Usually T’ = T

T’

Seq2seq is not needed



Vocoder

- Rule-based: Griffin-Lim algorithm
- Deep Learning: WaveNet

Used in VC, TTS, Speech Separation, etc. (not today)

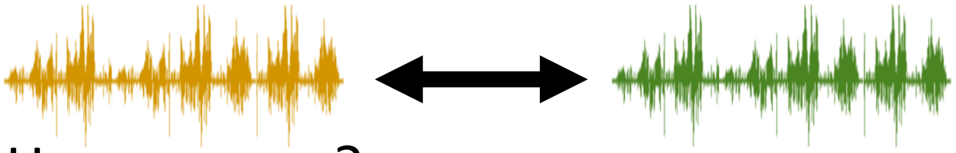
Categories

Lack of training data:

- Model Pre-training [Huang, et al., arXiv’19]
- Synthesized data!

[Biadsy, et al., INTERSPEECH’19]

Parallel Data



How are you?

How are you?

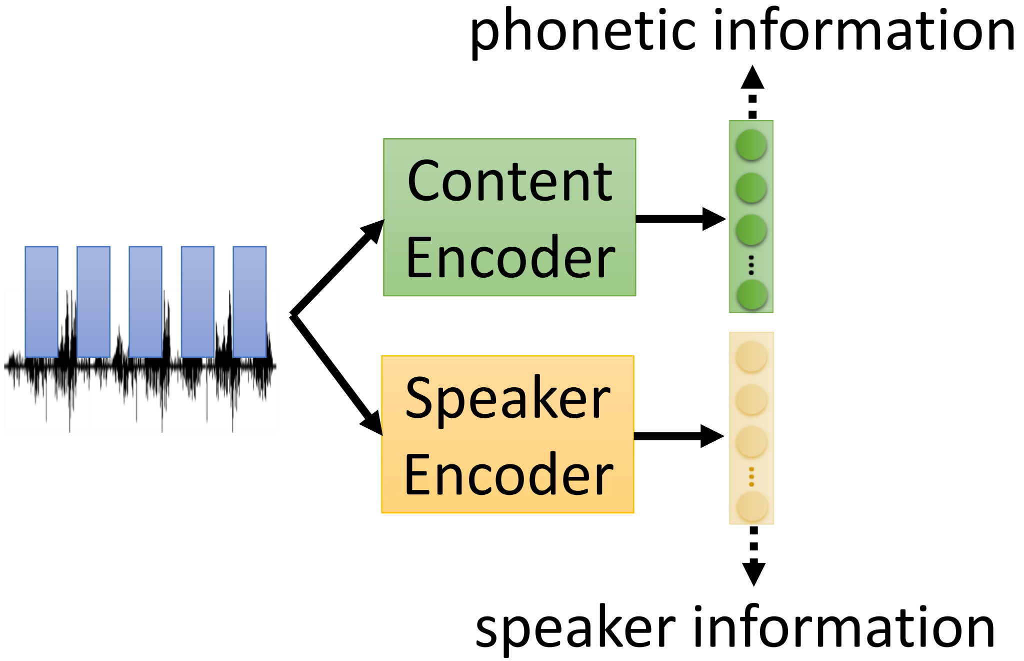
- This is “*audio style transfer*”
- Borrowing techniques from image style transfer

Unparallel Data

天氣真好

How are you?

Categories

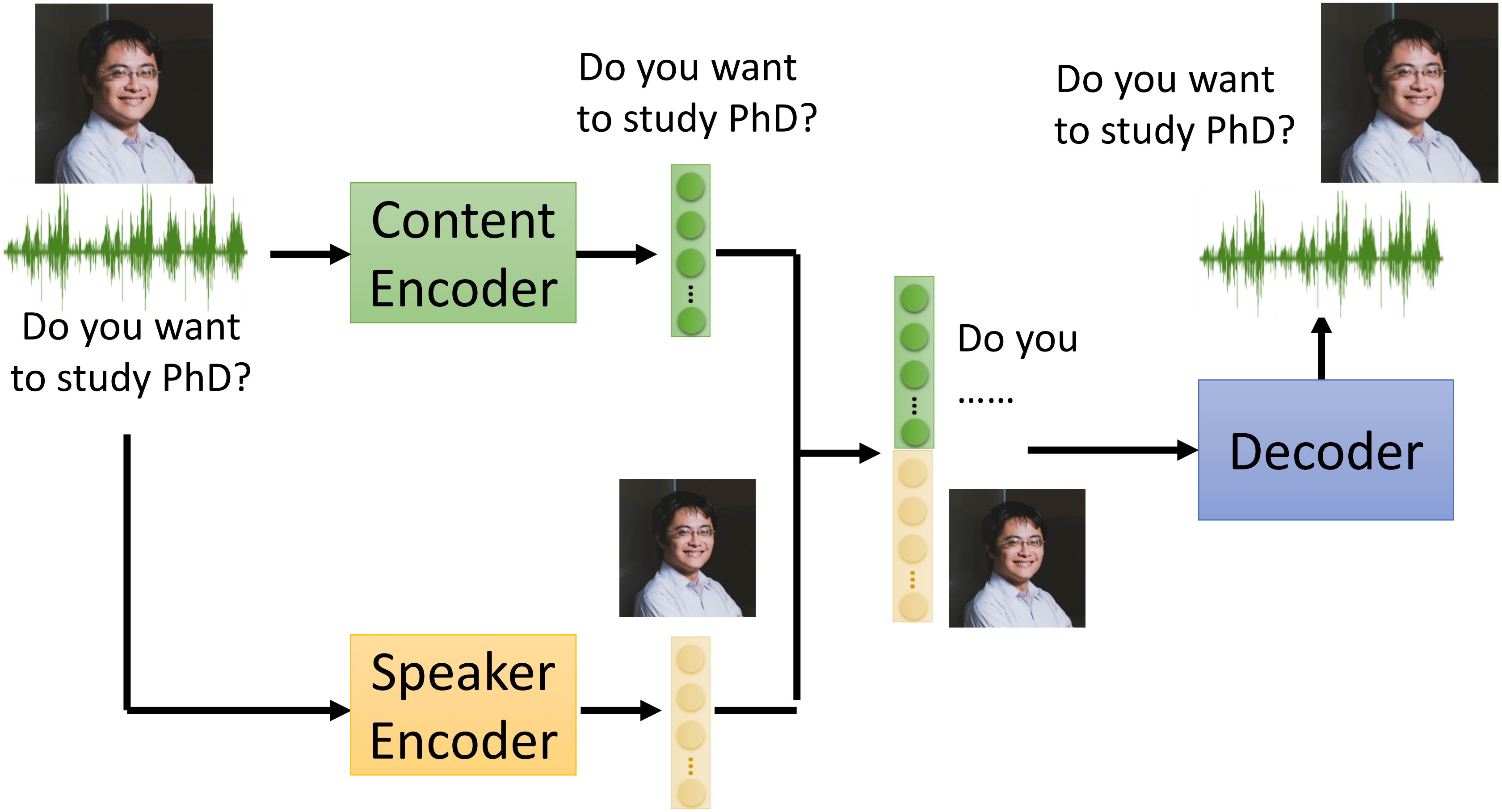


Parallel Data

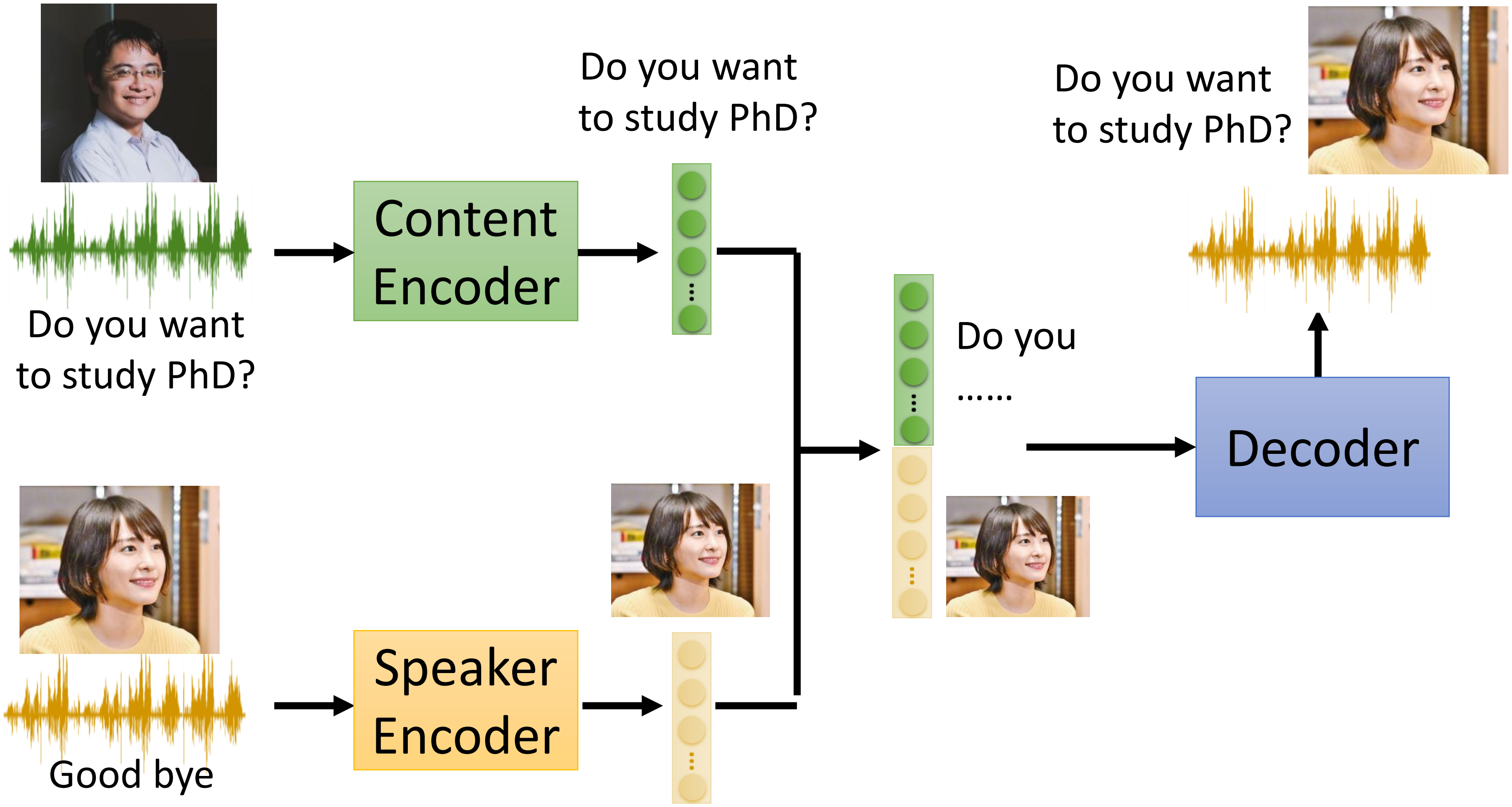
Unparallel Data

- Feature Disentangle
- Direct Transformation

Feature Disentangle



Feature Disentangle



Do you want to study PhD?

Content Encoder

Do you want to study PhD?

Good bye

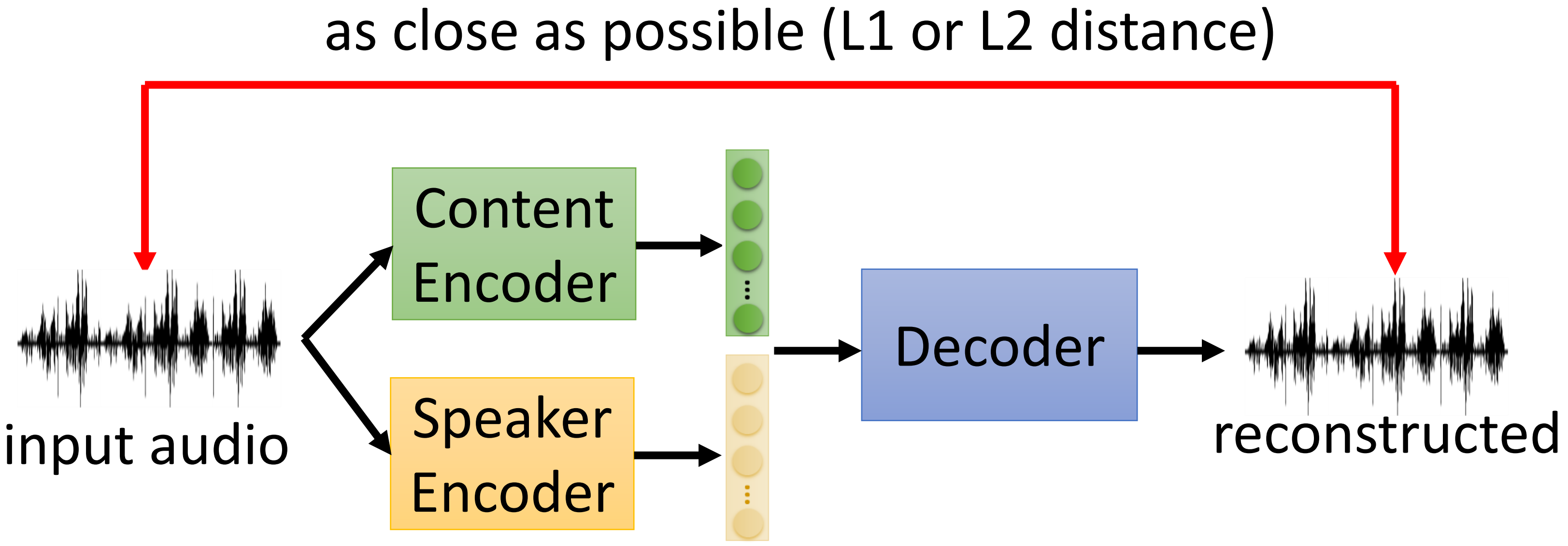
Speaker Encoder

Do you ……

Decoder

Do you want to study PhD?

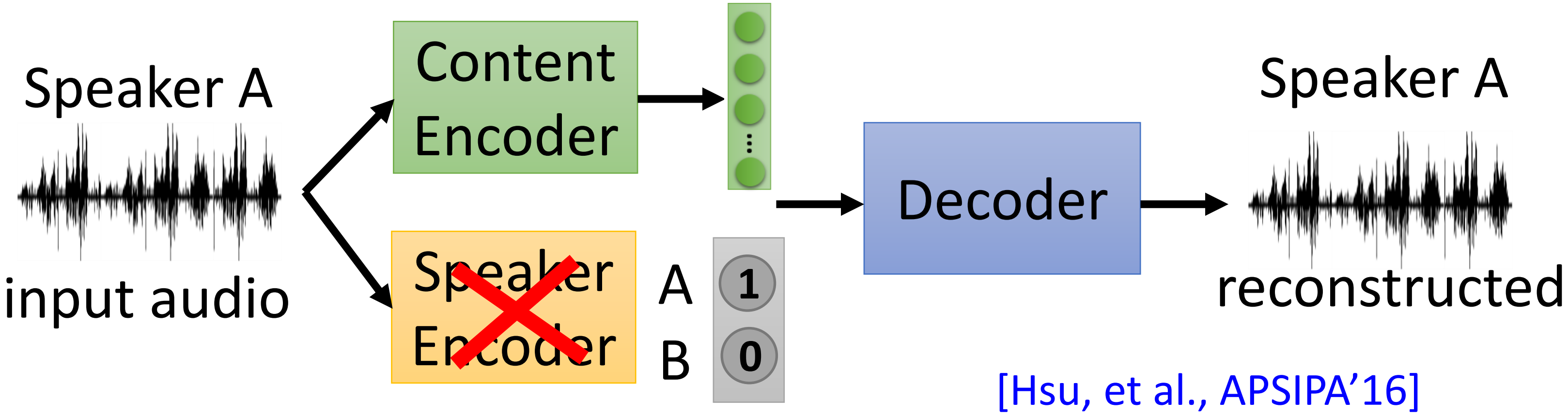
Feature Disentangle



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

Using Speaker Information

Assume we know the speakers of training utterances

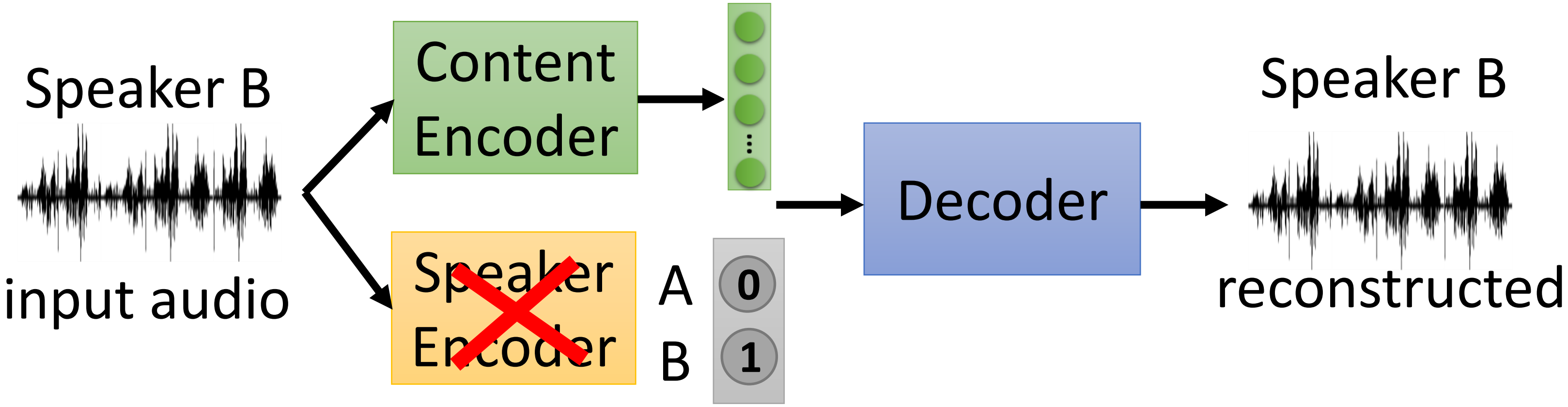


[Hsu, et al., APSIPA'16]

- One-hot vector for each speaker

Using Speaker Information

Assume we know the speakers of training utterances

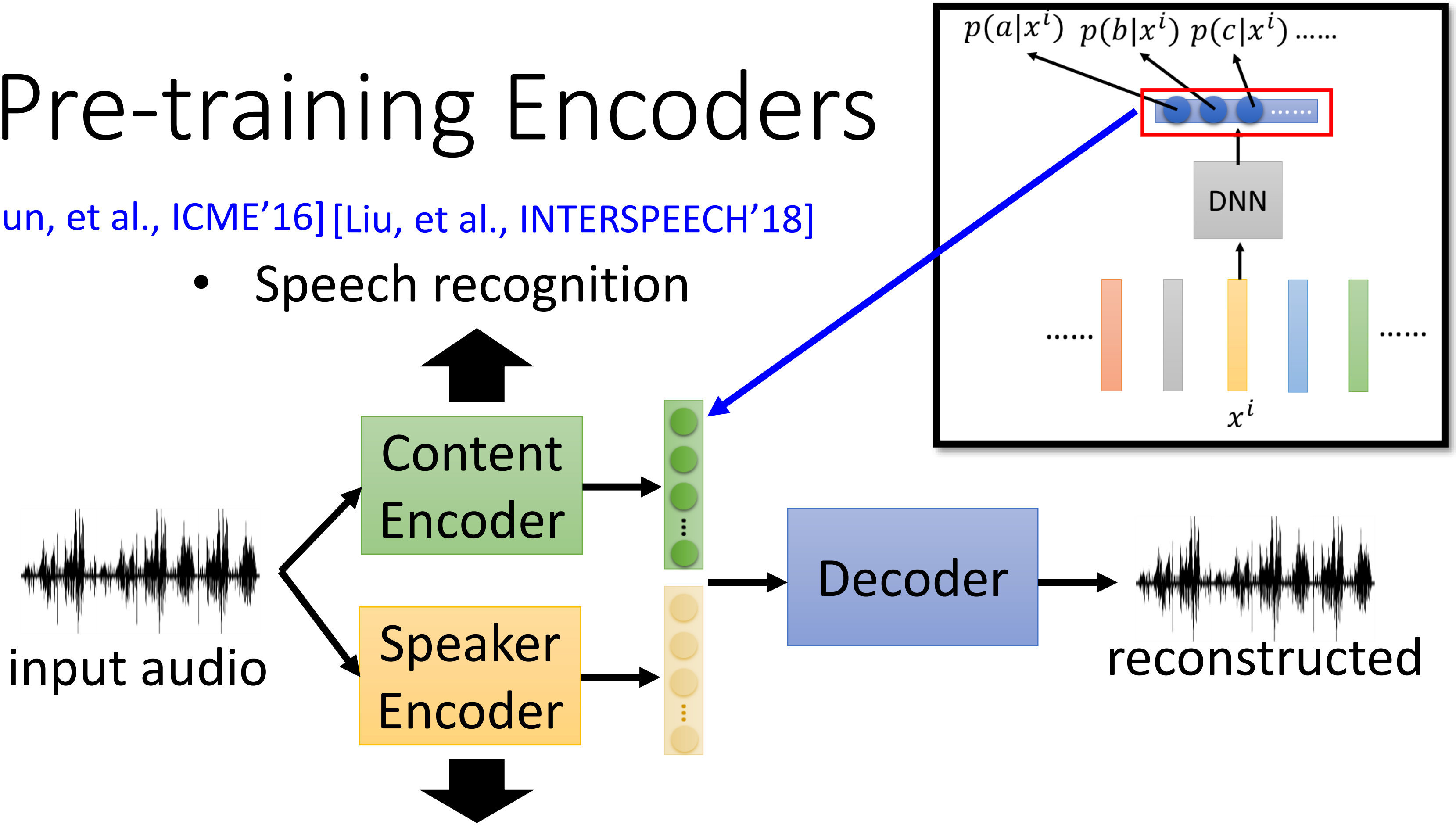


- One-hot vector for each speaker

Pre-training Encoders

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

- Speech recognition



- One-hot vector for each speaker

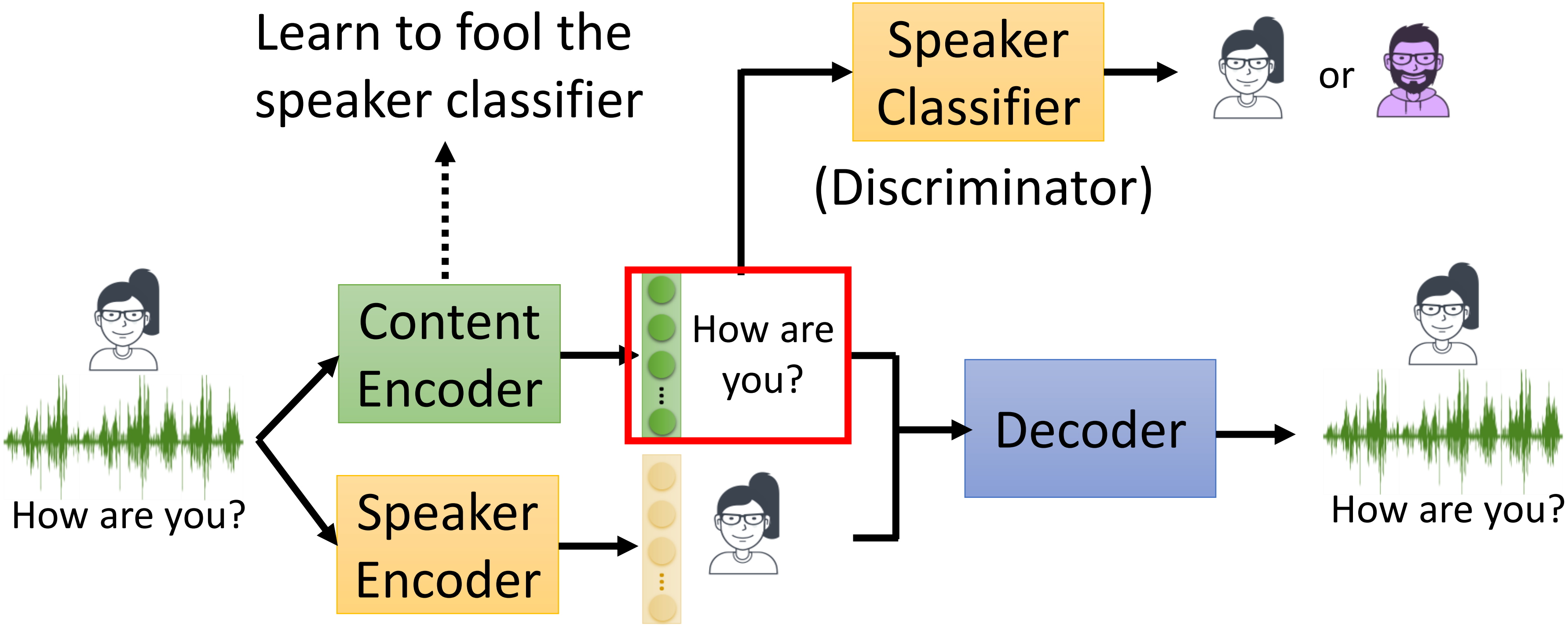
Issue: difficult to consider new speakers

- Speaker embedding (i-vector, d-vector, x-vector …)

[Qian, et al., ICML’19][Liu, et al., INTERSPEECH’18]

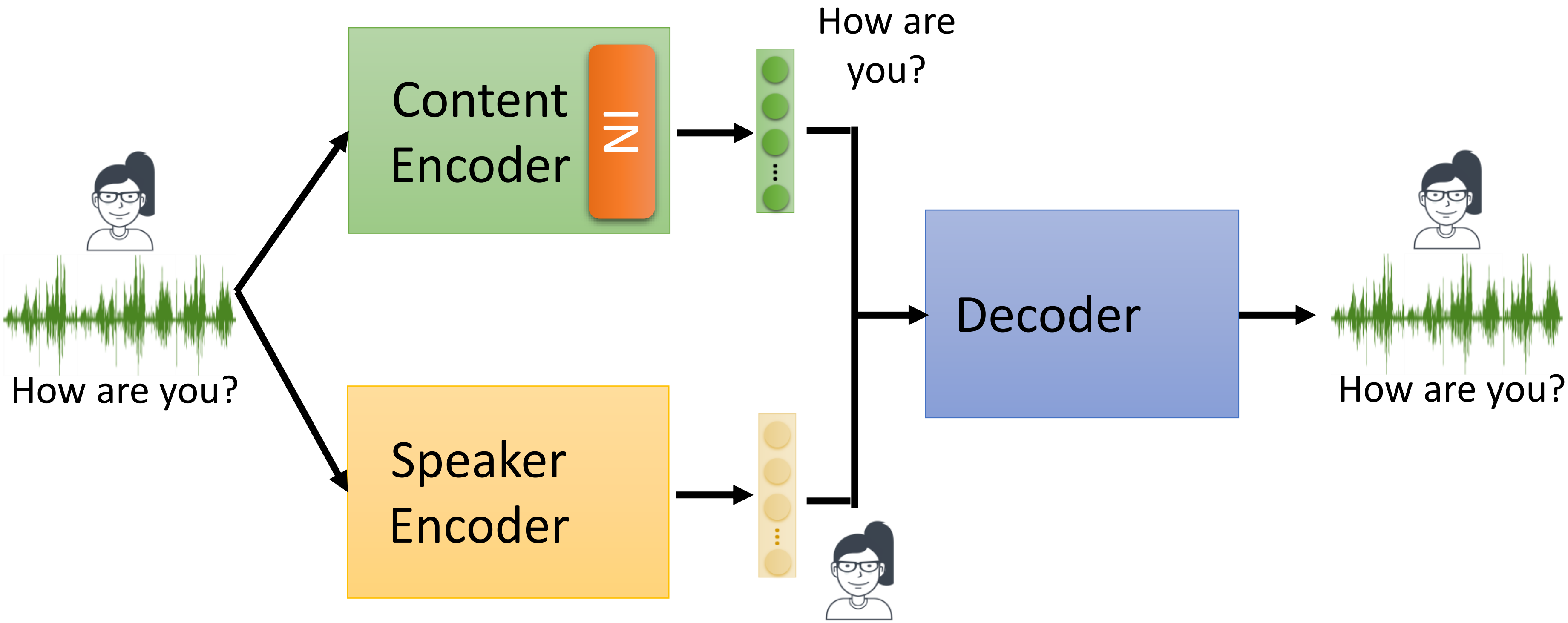
Adversarial Training

[Chou, et al., INTERSPEECH’18]



Speaker classifier and encoder are learned iteratively

Designing network architecture

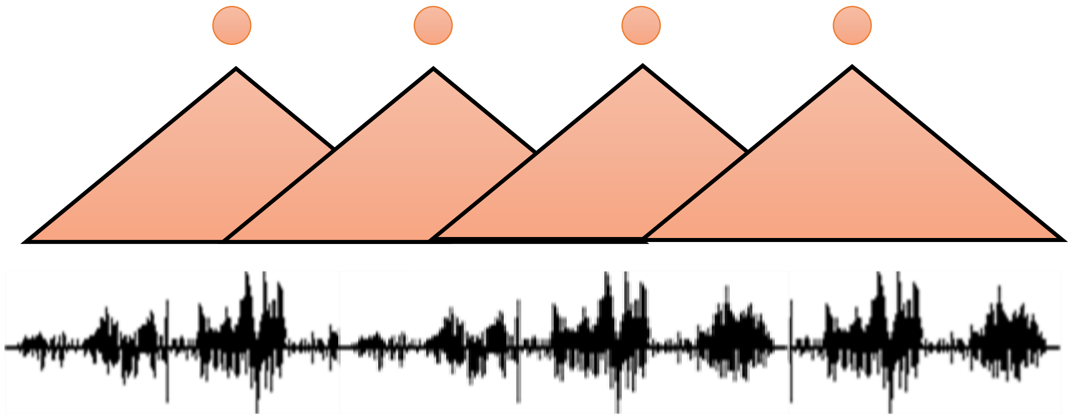


IN = instance normalization (remove speaker information)

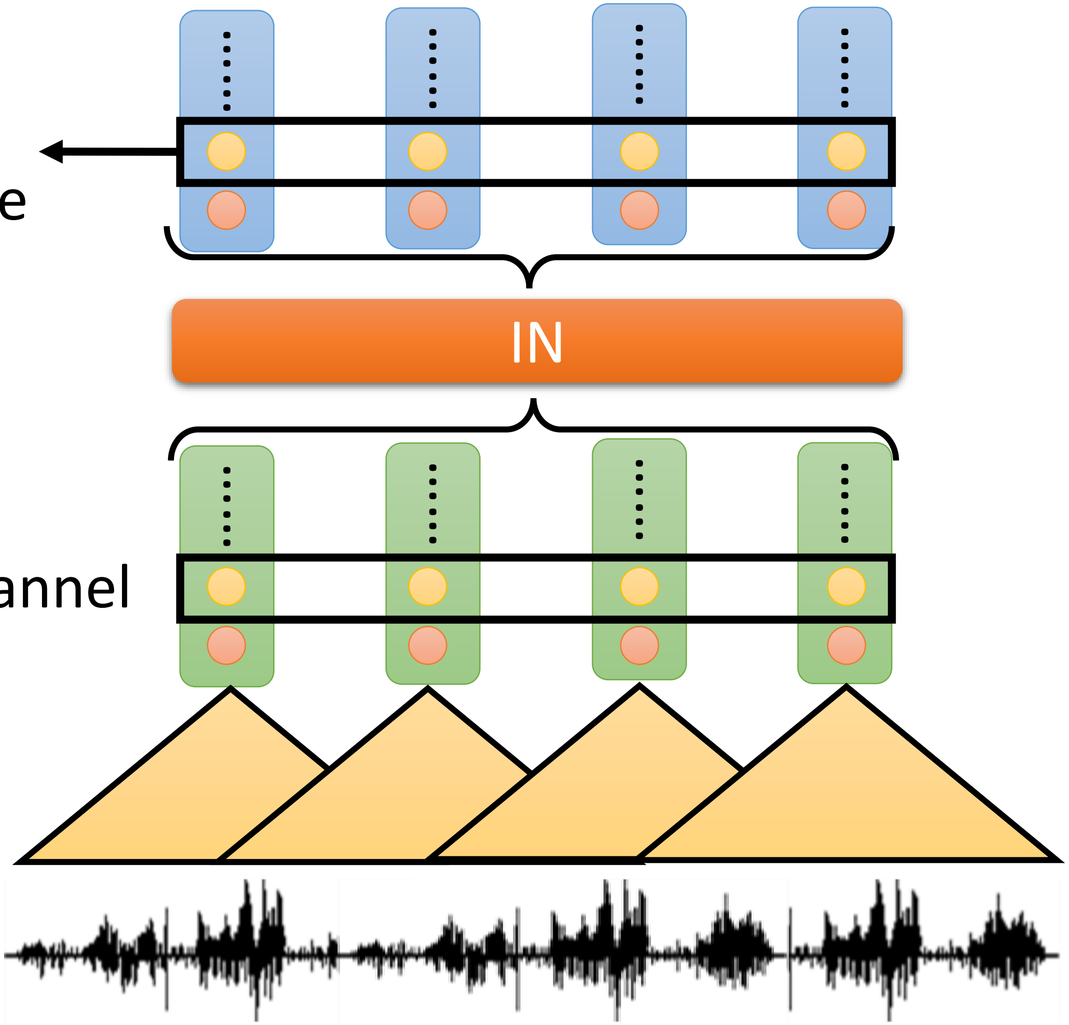
Designing network architecture

IN = instance normalization (remove speaker information)

Content Encoder



Designing network architecture

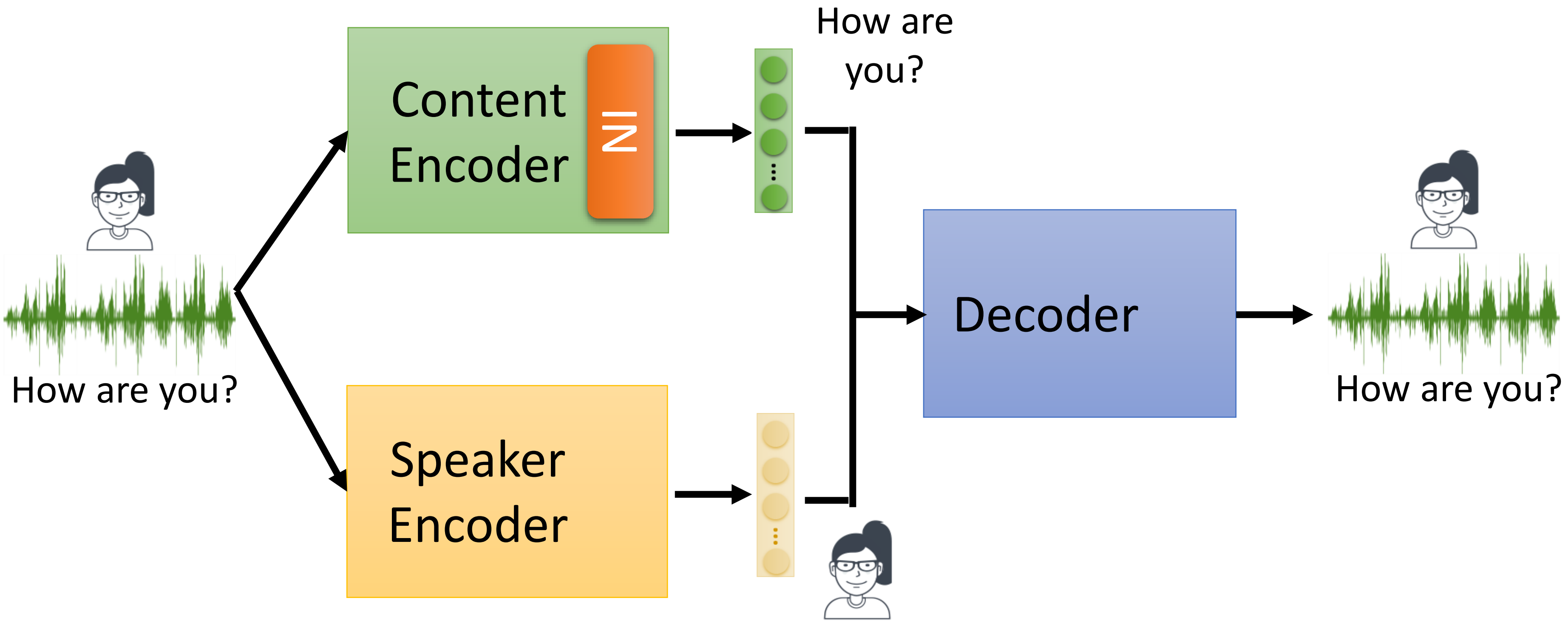


Each channel has zero mean and unit variance

Normalize for each channel

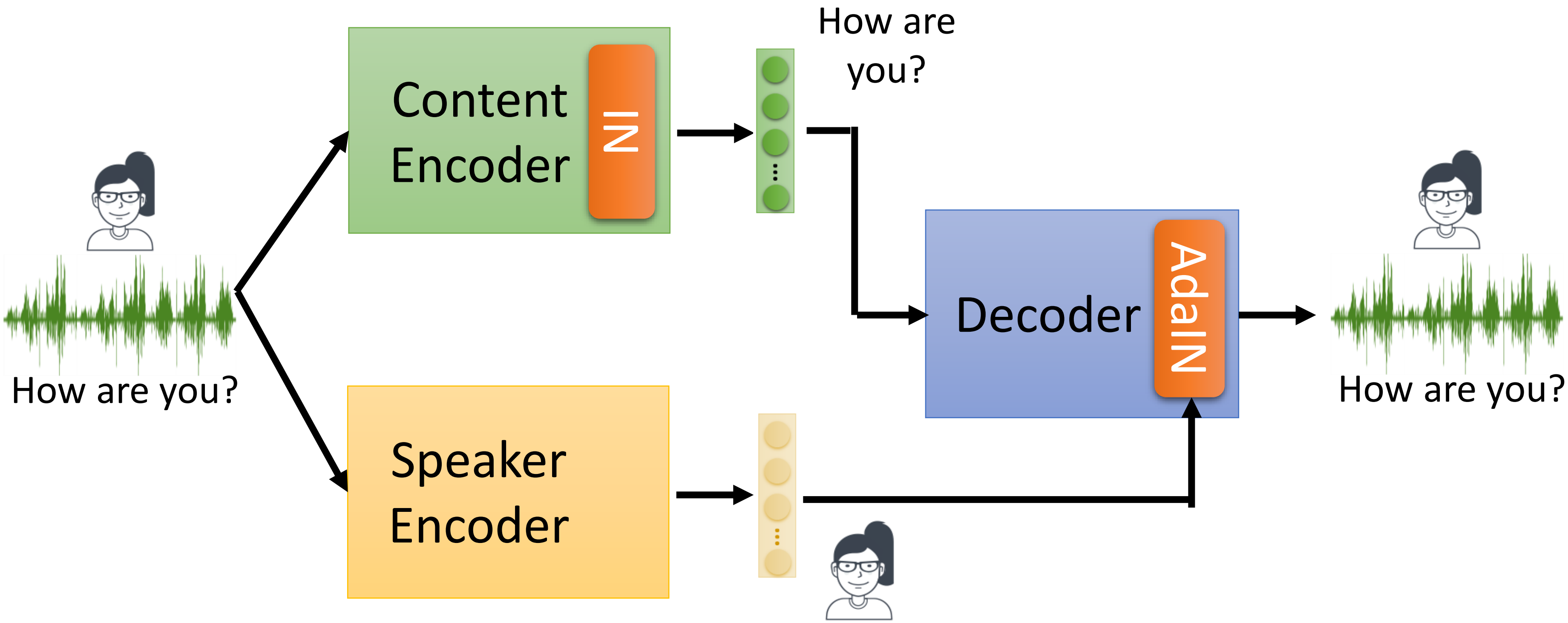
Content Encoder

Designing network architecture



IN = instance normalization (remove speaker information)

Designing network architecture



IN = instance normalization (remove speaker information)

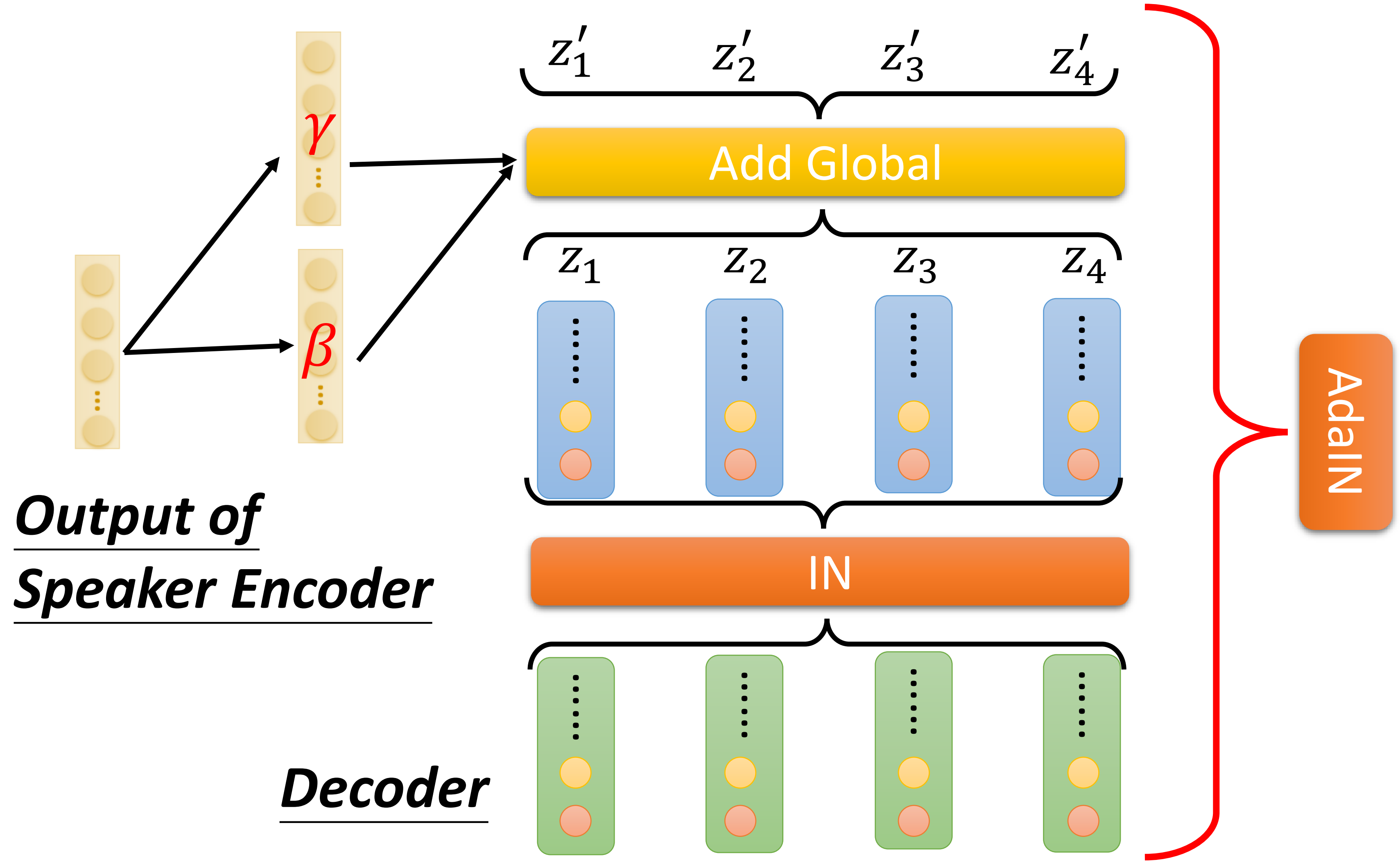
AdaIN = adaptive instance normalization

(only influence speaker information)

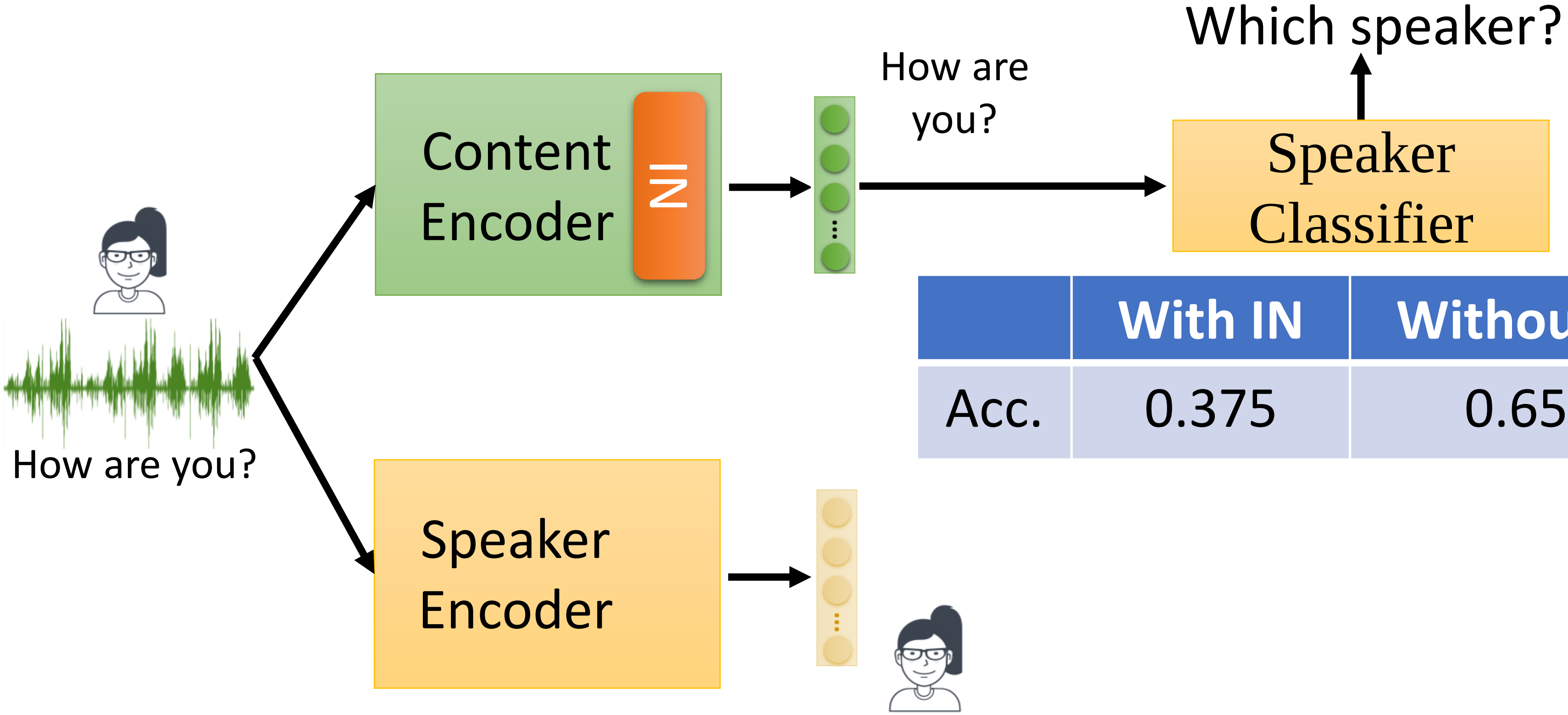
AdaIN = adaptive instance normalization

(only influence speaker information)

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



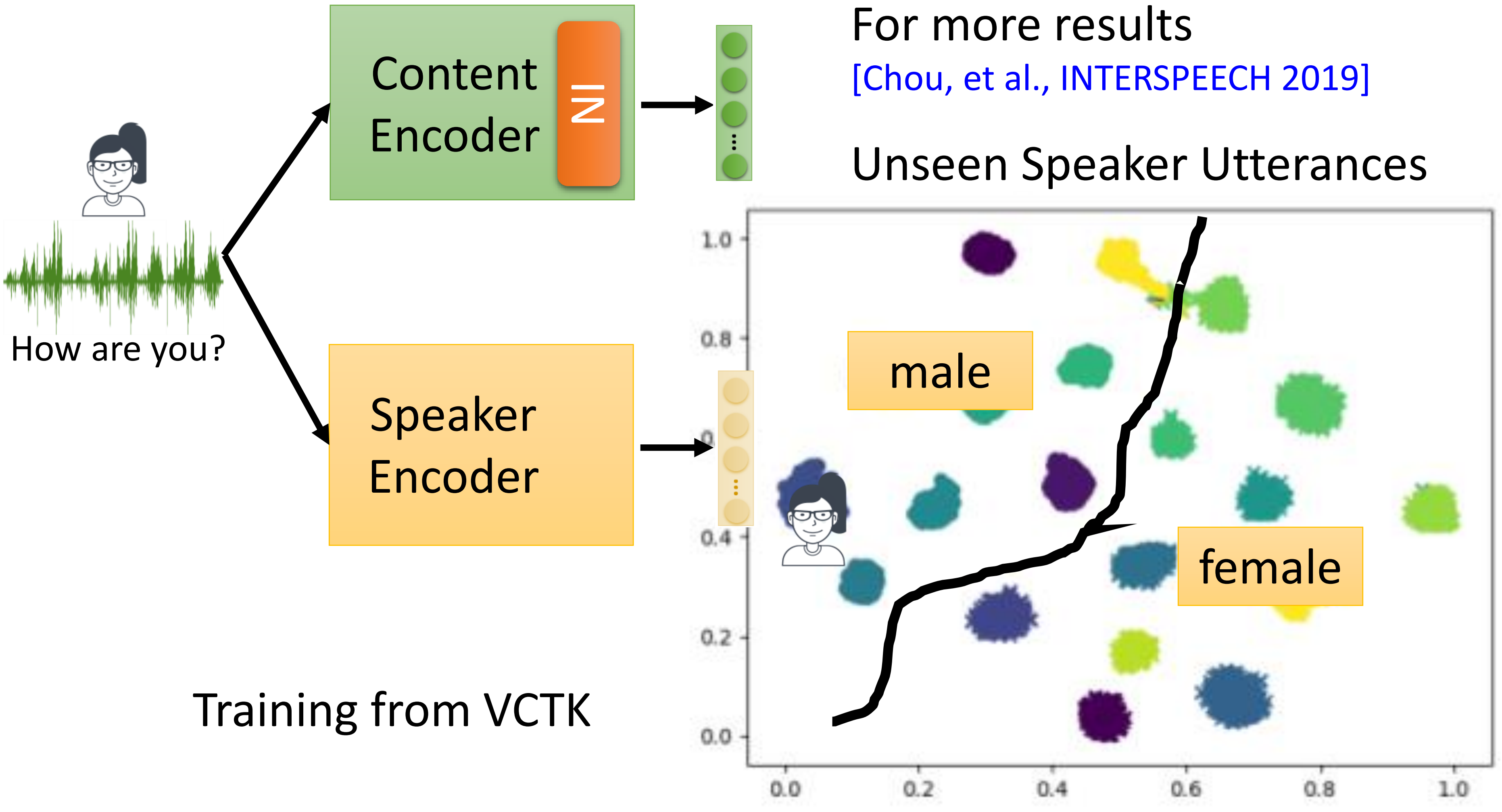
Designing network architecture



	With IN	Without IN
Acc.	0.375	0.658

Training from VCTK

Designing network architecture



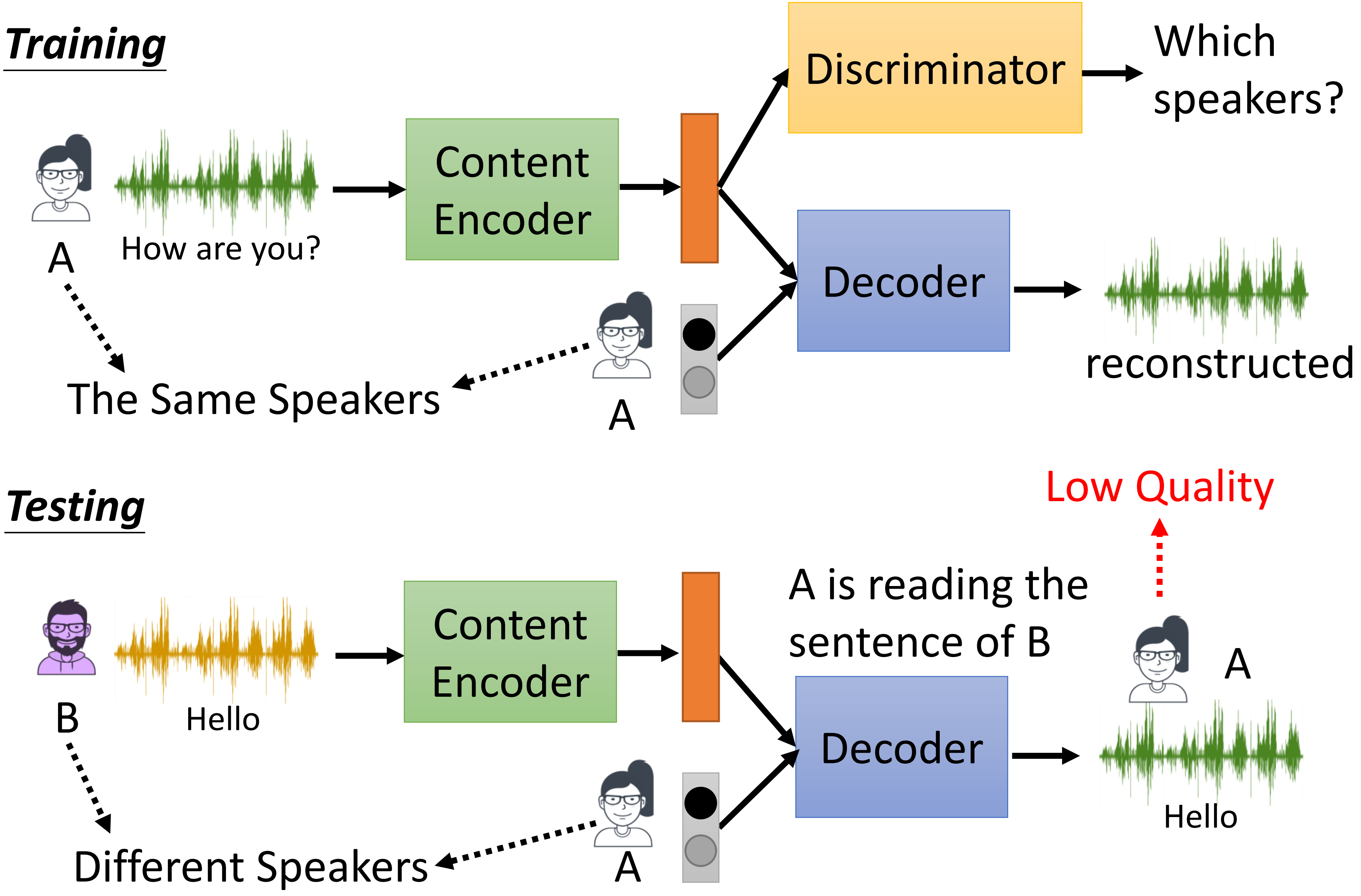
For more results

[Chou, et al., INTERSPEECH 2019]

Unseen Speaker Utterances

Training from VCTK

Issues



Training

A

How are you?

Content Encoder

Discriminator

Which speakers?

Decoder

reconstructed

A

The Same Speakers

Testing

B

Hello

Content Encoder

A is reading the sentence of B

Low Quality

Decoder

A

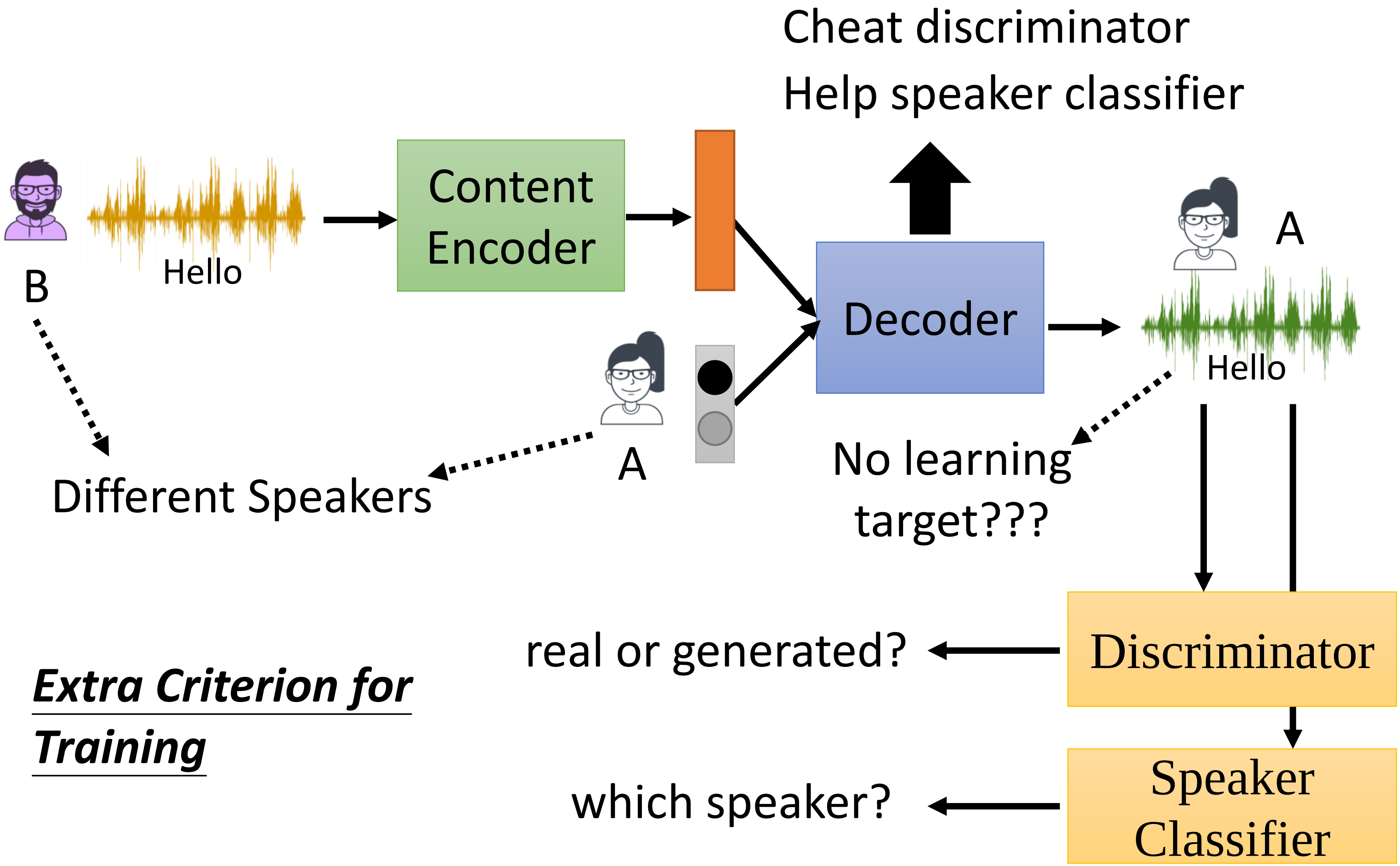
Hello

A

Different Speakers

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

2nd Stage Training



Cheat discriminator

Help speaker classifier

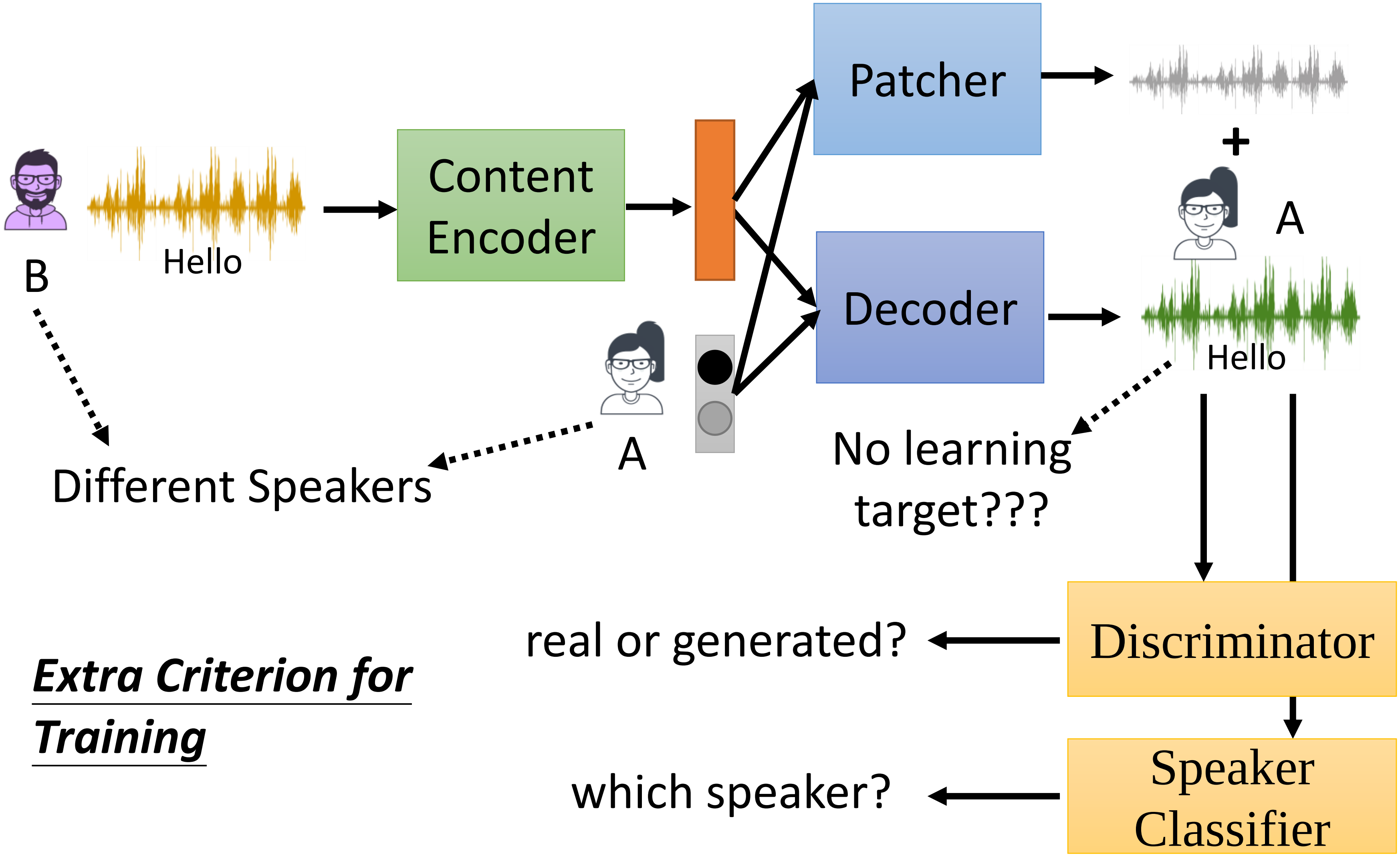
real or generated?

which speaker?

Extra Criterion for Training

2nd Stage Training

Only learn the patcher in the 2nd stage



Extra Criterion for Training

Categories

Parallel Data

Unparallel Data

- Training without parallel data
- Using CycleGAN

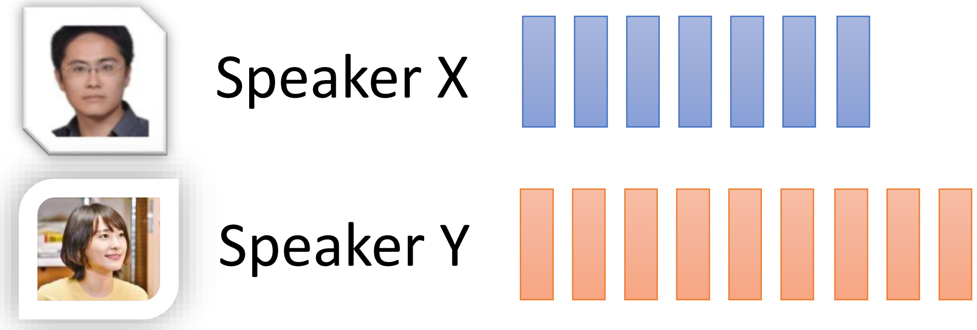
Voice Conversion

Feature Disentangle

Direct Transformation

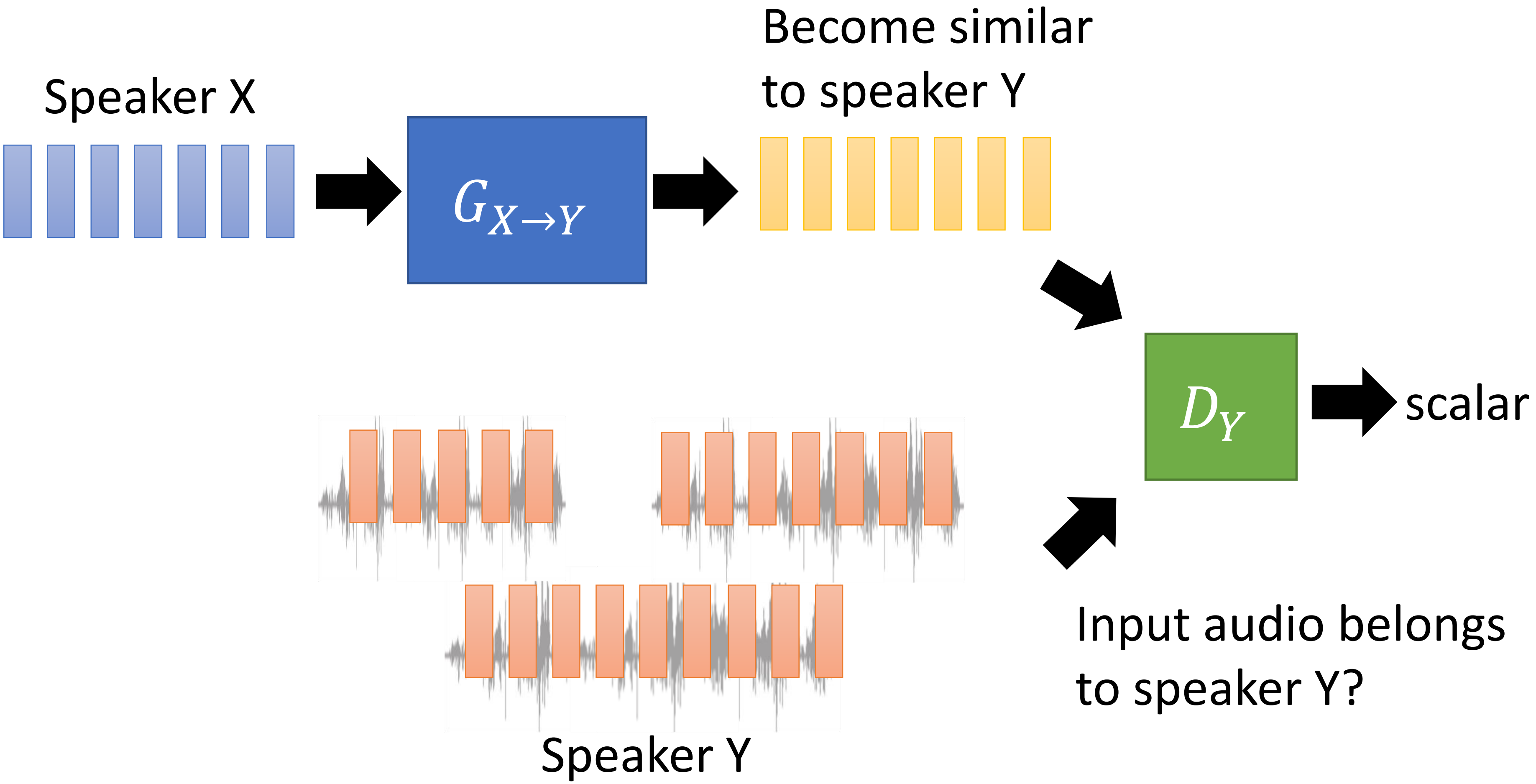
Cycle GAN

[Kaneko, et al., ICASSP’19]



Speaker X

Speaker Y



Speaker X

Become similar to speaker Y

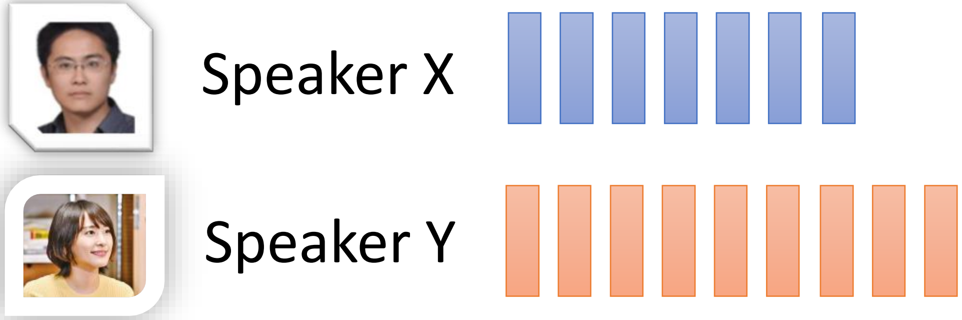
scalar

Speaker Y

Input audio belongs to speaker Y?

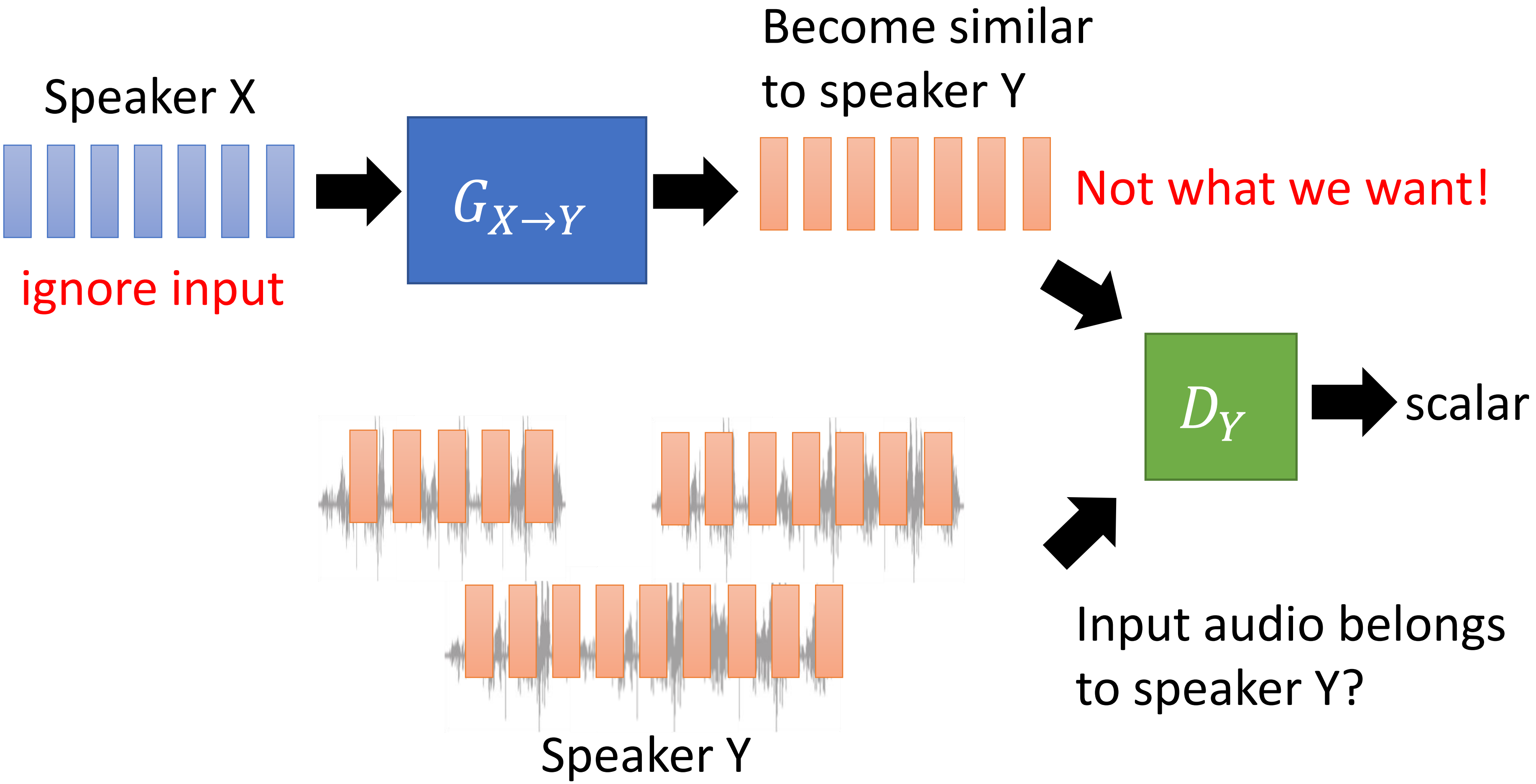
Cycle GAN

[Kaneko, et al., ICASSP'19]



Speaker X

Speaker Y



Speaker X

ignore input

$G_{X \to Y}$

Become similar to speaker Y

Not what we want!

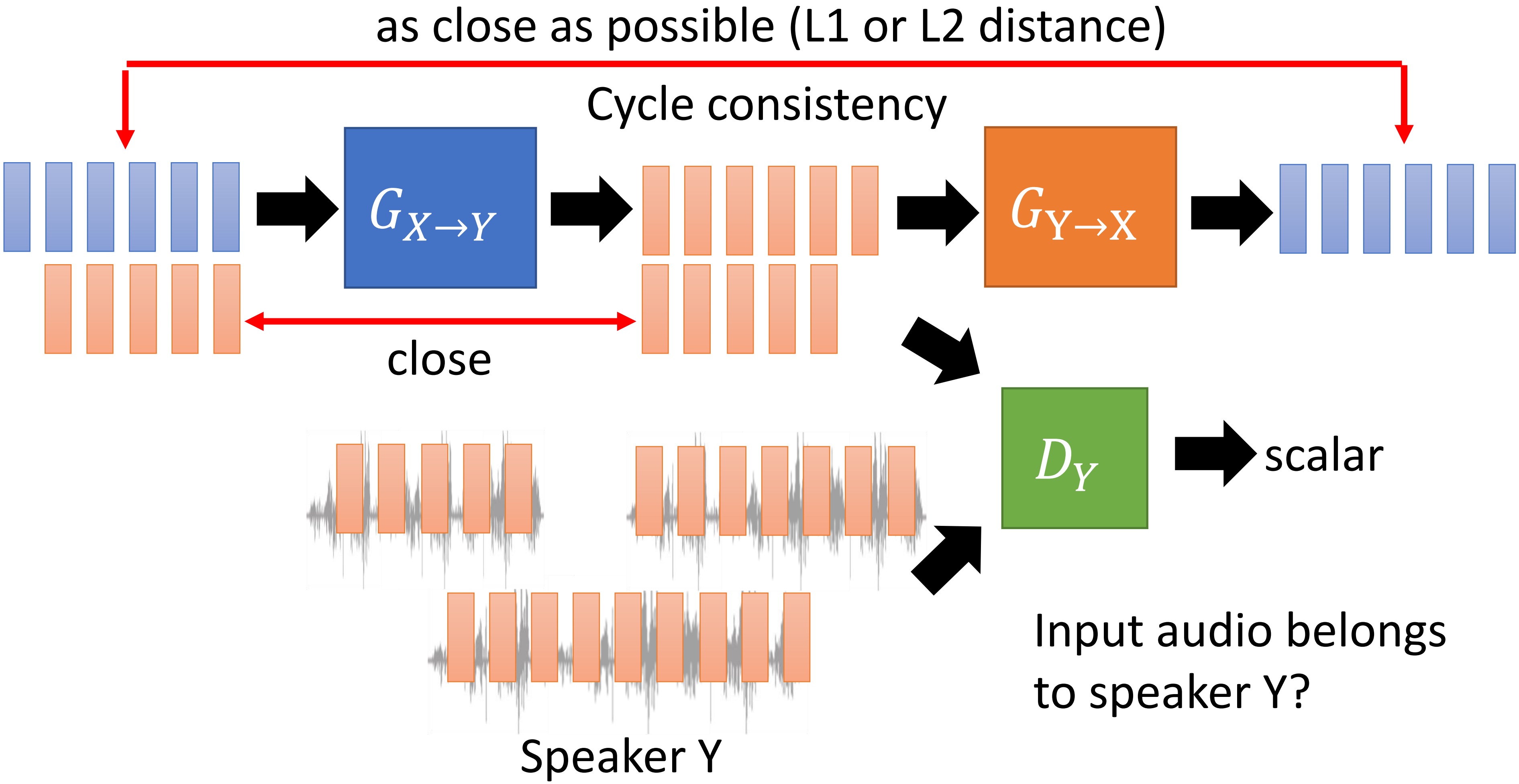
D_Y

scalar

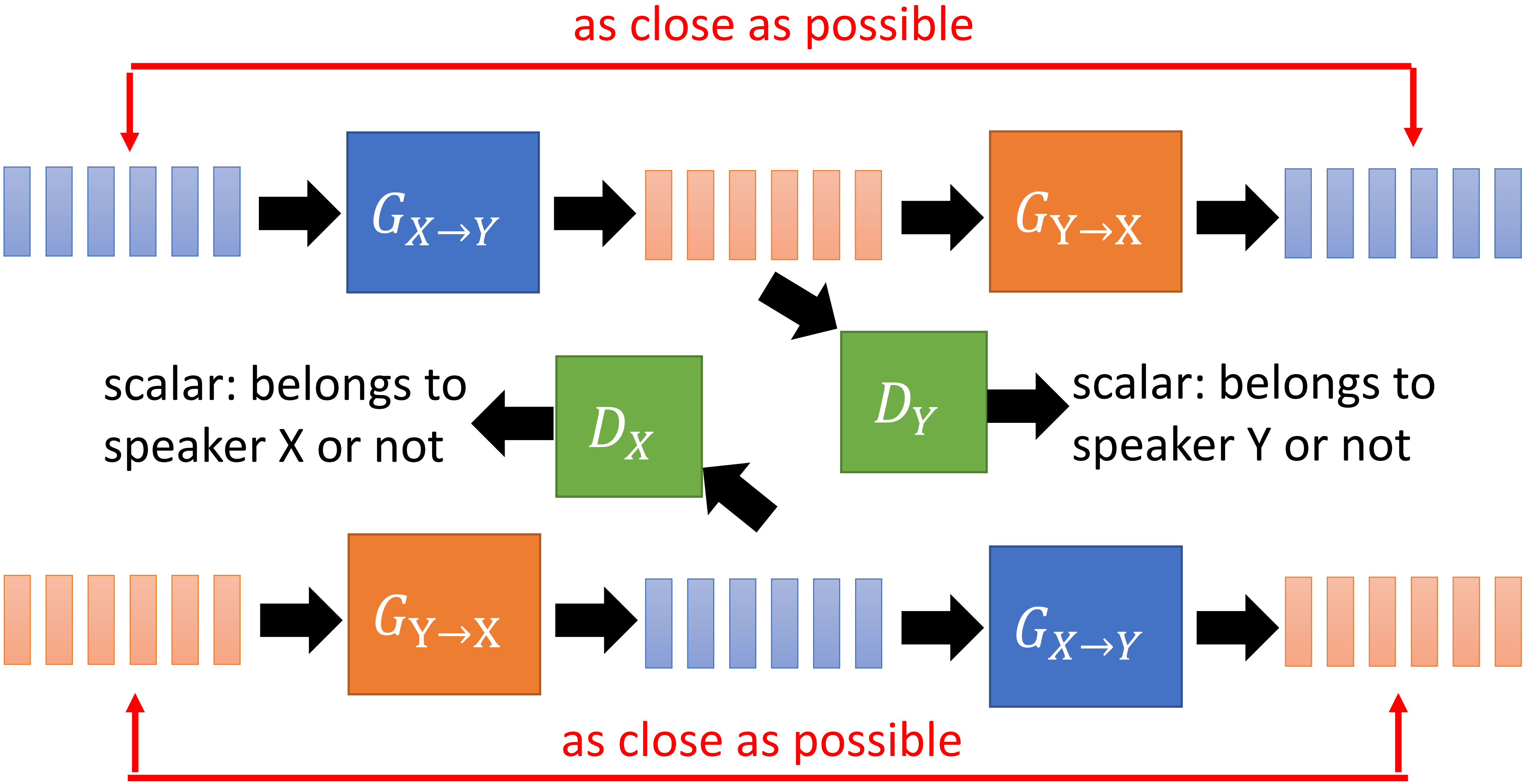
Speaker Y

Input audio belongs to speaker Y?

Cycle GAN


as close as possible (L1 or L2 distance)
Cycle consistency
$G_{X \to Y}$
$G_{Y \to X}$
close
D_Y
scalar
Input audio belongs to speaker Y?
Speaker Y

Cycle GAN



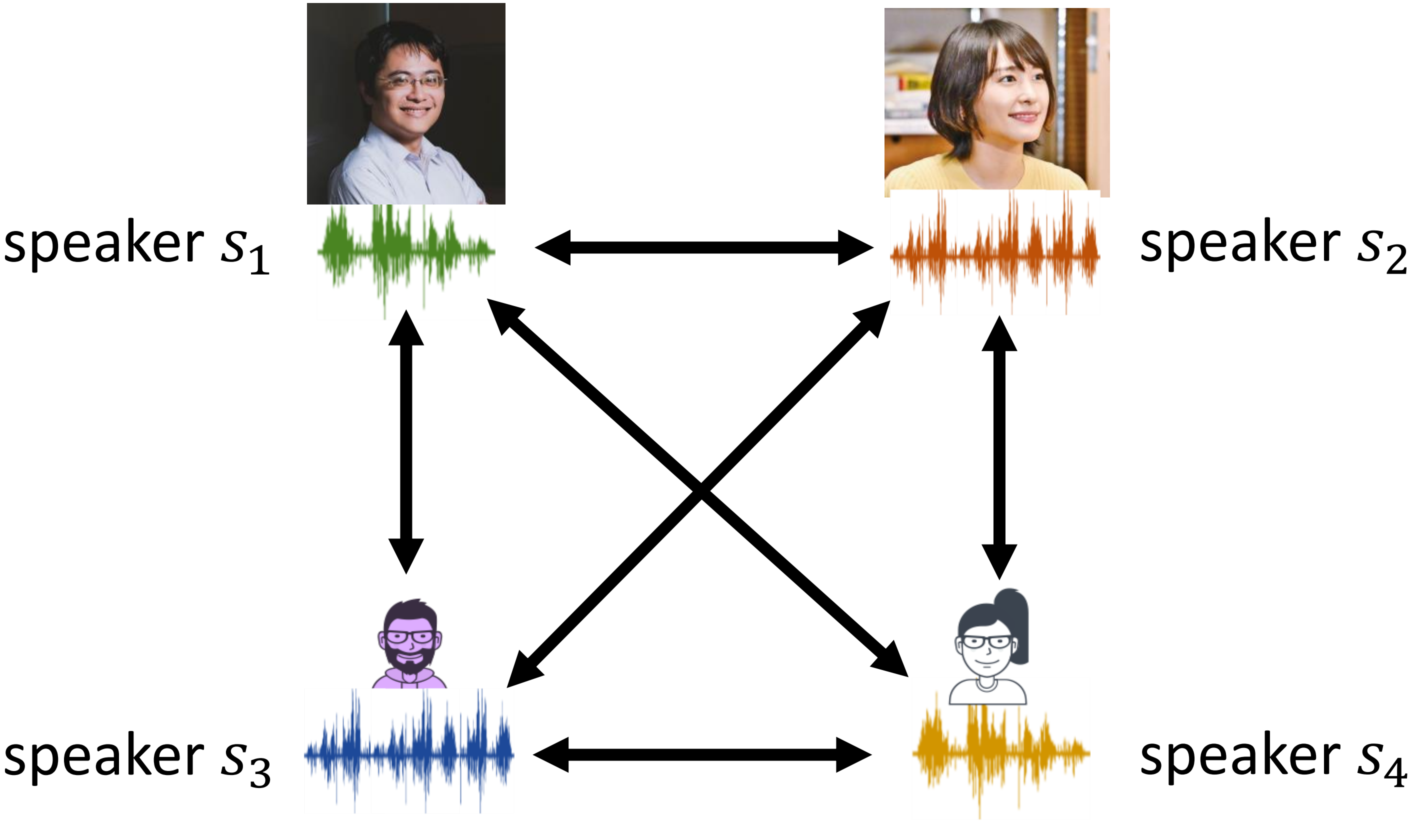
scalar: belongs to speaker X or not

scalar: belongs to speaker Y or not

StarGAN

[Kaneko, et al., INTERSPEECH’19]

For CycleGAN:
If there are N speakers, you need
N x (N-1) generators.



speaker s_1

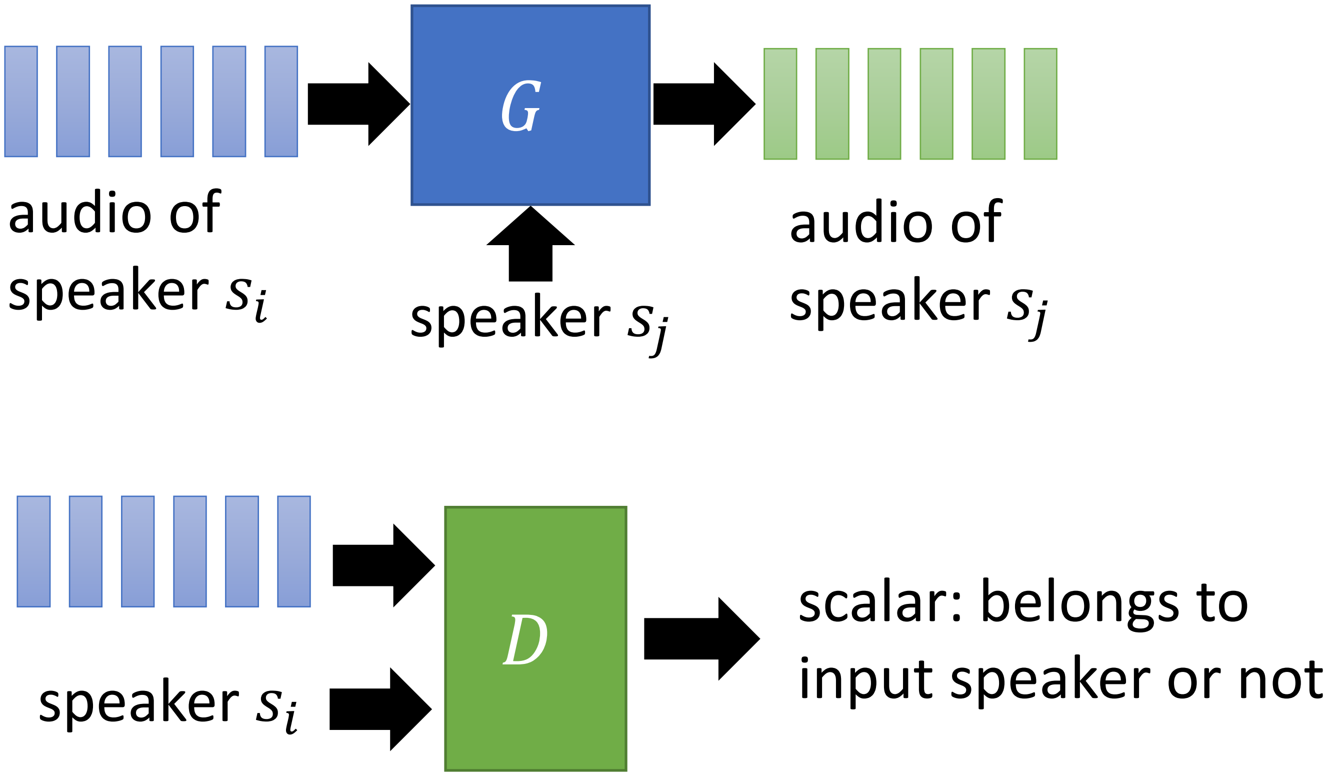
speaker s_2

speaker s_3

speaker s_4

StarGAN

Each speaker is represented as a vector.



audio of speaker s_i

G

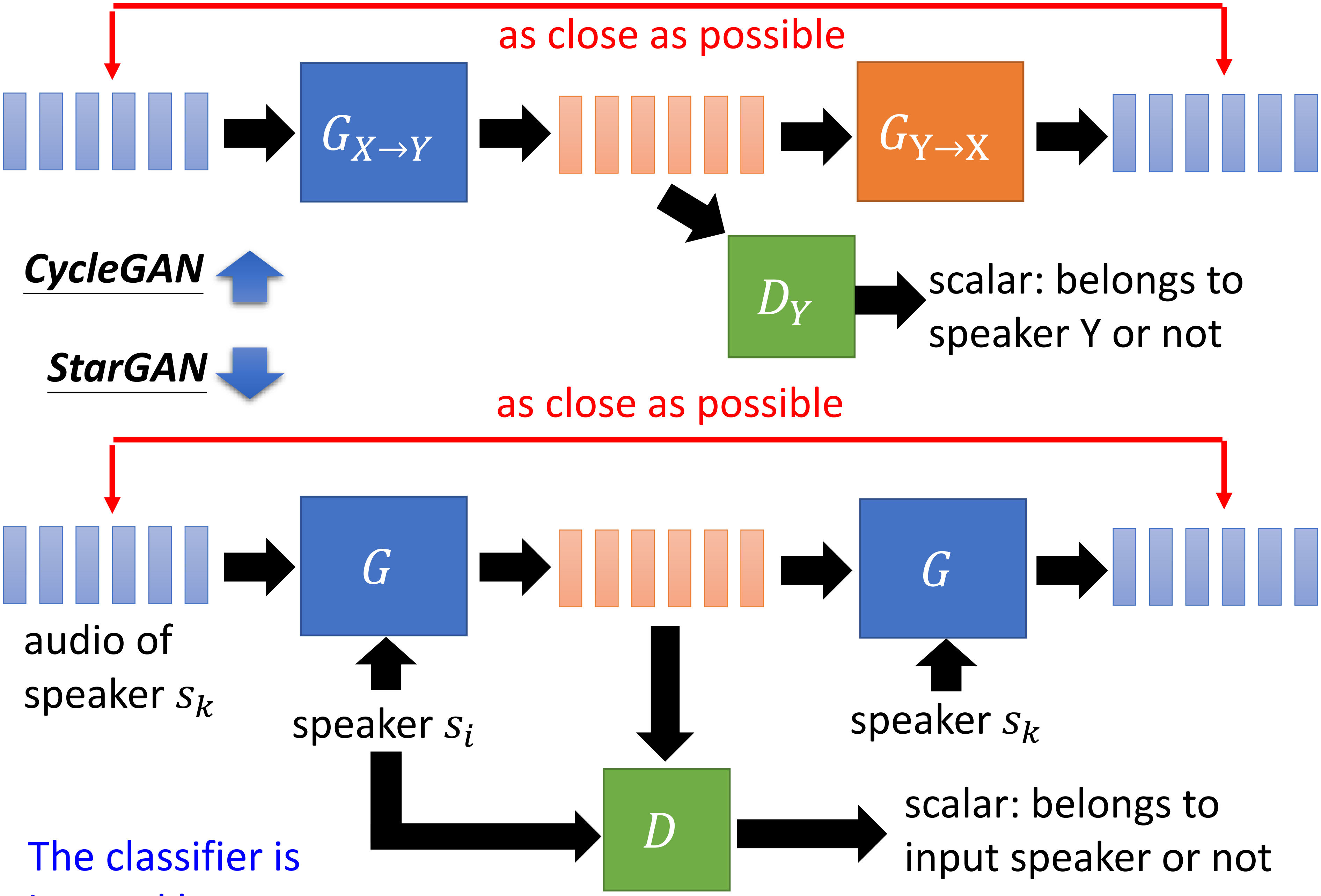
speaker s_j

audio of speaker s_j

speaker s_i

D

scalar: belongs to input speaker or not



as close as possible

$G_{X\rightarrow Y}$

$G_{Y\rightarrow X}$

D_Y

scalar: belongs to speaker Y or not

CycleGAN

StarGAN

as close as possible

G

G

audio of speaker s_k

speaker s_i

speaker s_k

D

scalar: belongs to input speaker or not

The classifier is ignored here.

Blow

[Joan, et al., NeurIPS'19]

Flow-based model for VC

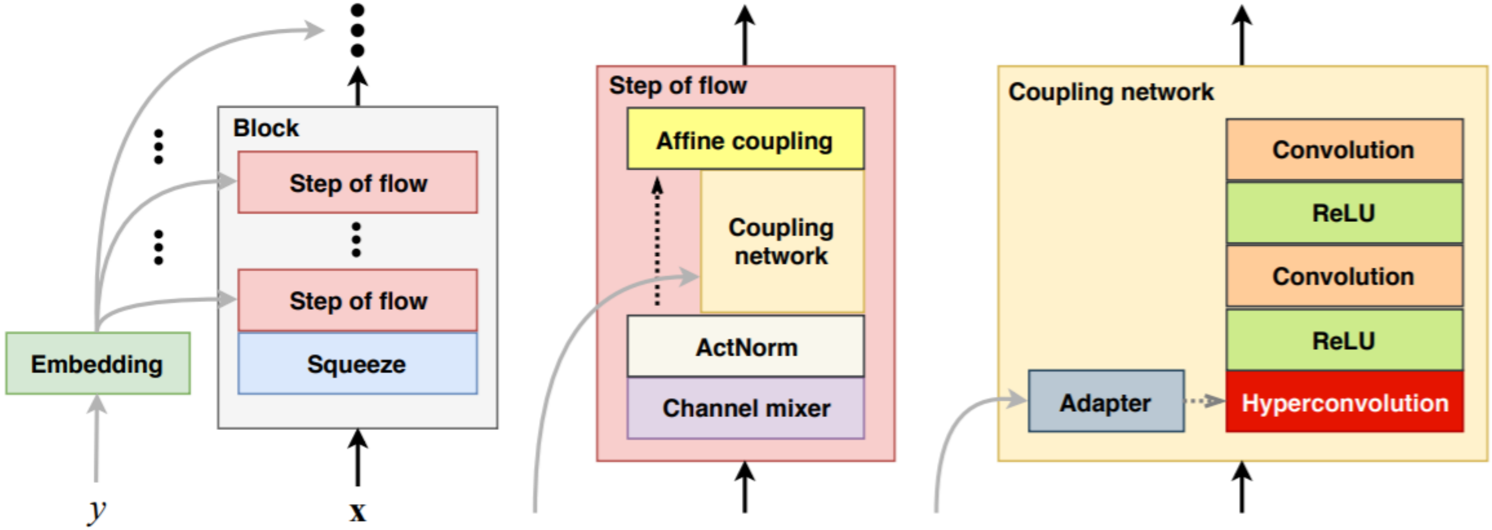
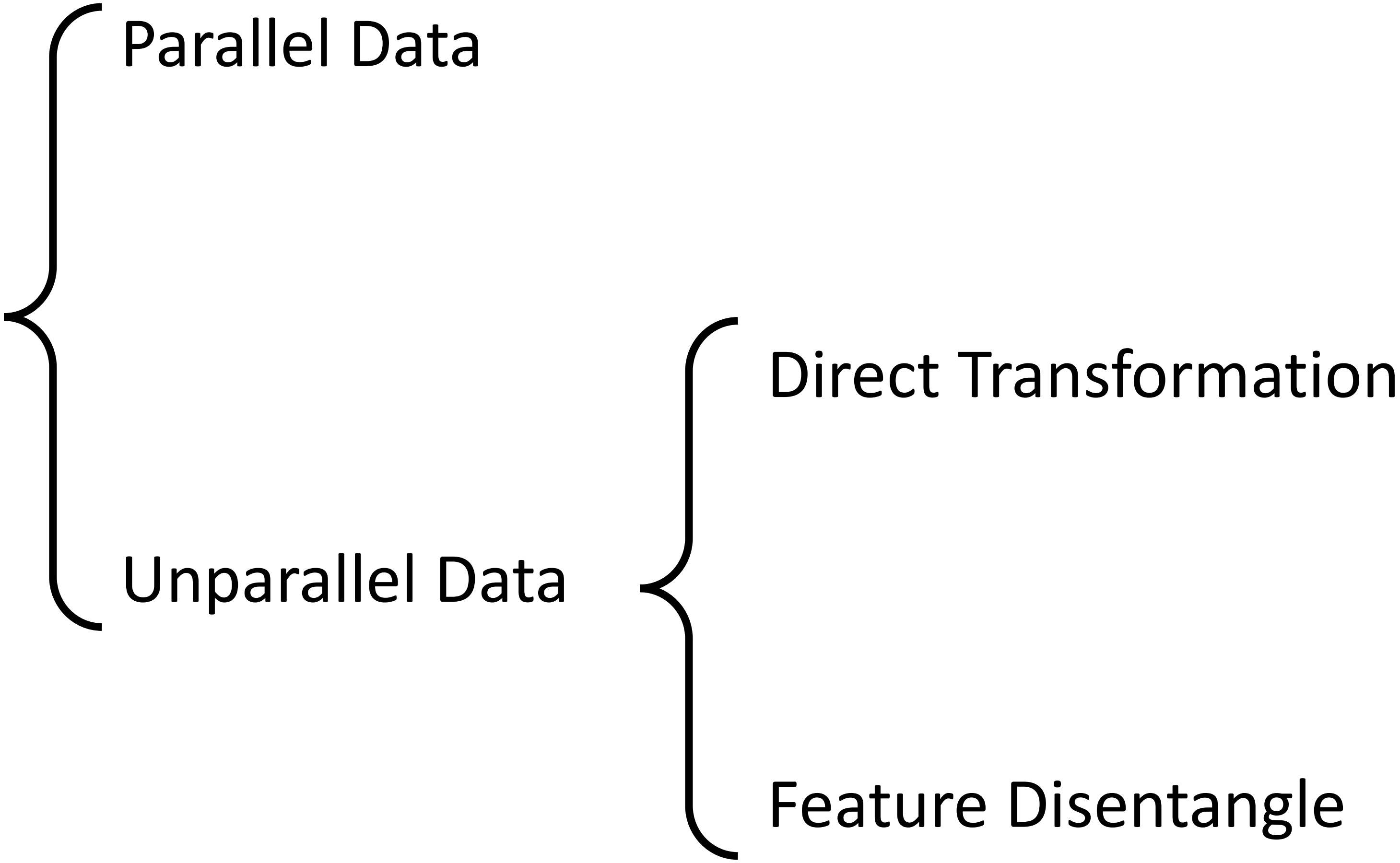


Figure 1: Blow schema featuring its block structure (left), steps of flow (center), and coupling network with hyperconvolution module (right).

Ref for flow-based model: https://youtu.be/uXY18nzdSsM

Concluding Remarks



Parallel Data

Unparallel Data

Direct Transformation

Feature Disentangle

Reference

- [Huang, et al., arXiv’19] 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, arXiv, 2019
- [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
- [Nachmani, et al., INTERSPEECH’19] Eliya Nachmani, Lior Wolf, Unsupervised Singing Voice Conversion, INTERSPEECH, 2019
- [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*

Reference

- [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
- [Gao, et al., INTERSPEECH’19] Jian Gao, Deep Chakraborty, Hamidou Tembine, Olaitan Olaleye, Nonparallel Emotional Speech Conversion, INTERSPEECH, 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
- [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., INTERSPEECH’19] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, StarGAN-VC2: Rethinking Conditional Methods for StarGAN-Based Voice Conversion, INTERSPEECH *2019*

Reference

- [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
- [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
- [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
- [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

Reference

- [Chen et al., INTERSPEECH’19] Li-Wei Chen, Hung-Yi Lee, Yu Tsao, Generative adversarial networks for unpaired voice transformation on impaired speech, INTERSPEECH, 2019
- [Zhao, et al., INTERSPEECH’19] Guanlong Zhao, Shaojin Ding, Ricardo Gutierrez-Osuna, Foreign Accent Conversion by Synthesizing Speech from Phonetic Posteriorgrams, INTERSPEECH, 2019
- [Srivastava, et al., arXiv’19] Brij Mohan Lal Srivastava, Nathalie Vauquier, Md Sahidullah, Aurélien Bellet, Marc Tommasi, Emmanuel Vincent, Evaluating Voice Conversion-based Privacy Protection against Informed Attackers, arXiv, 2019
- [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
- [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

Reference

- [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
- [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
- [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
- [Liu, et al., INTERSPEECH’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