

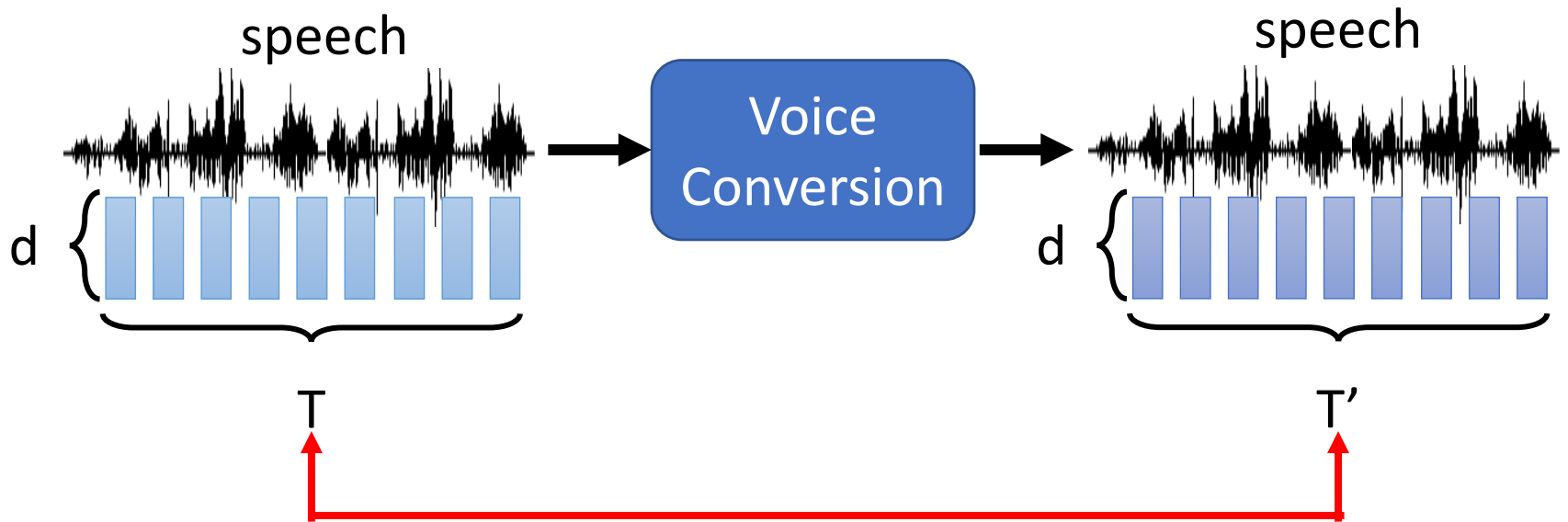
# VOICE CONVERSION

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/](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)



# 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/](https://shreyas253.github.io/SpStyleConv_CycleGAN/)

‘lip thrill’ (彈唇) or ‘vibrato’ (顫音)

# Improving Communication

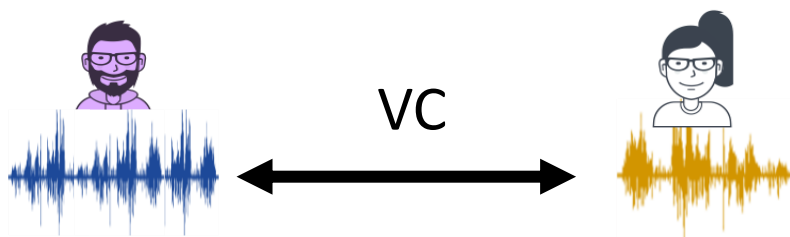
- Improving the speech intelligibility
  - surgical patients who have had parts of their articulators removed

[Biadys, 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

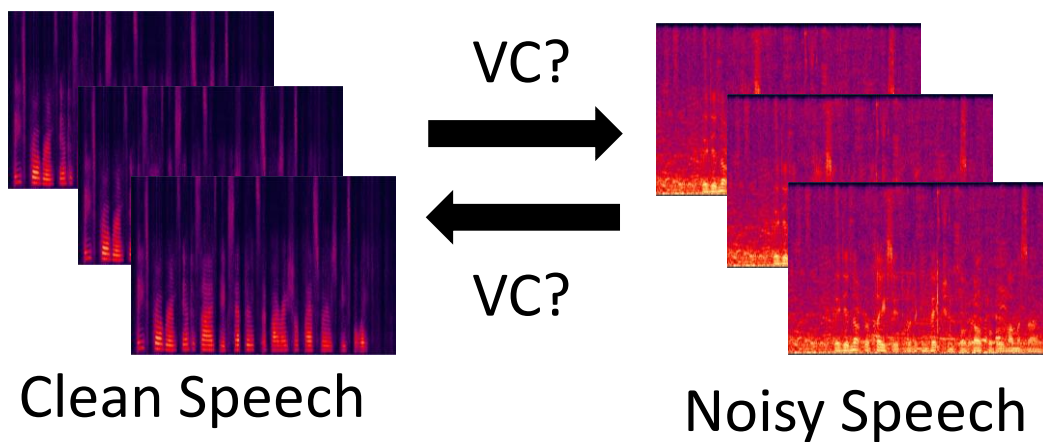
[Zhao, et al., INTERSPEECH'19]

# Data Augmentation



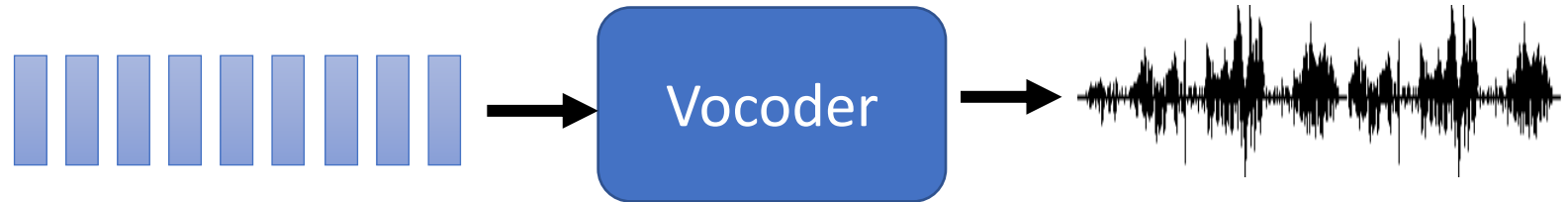
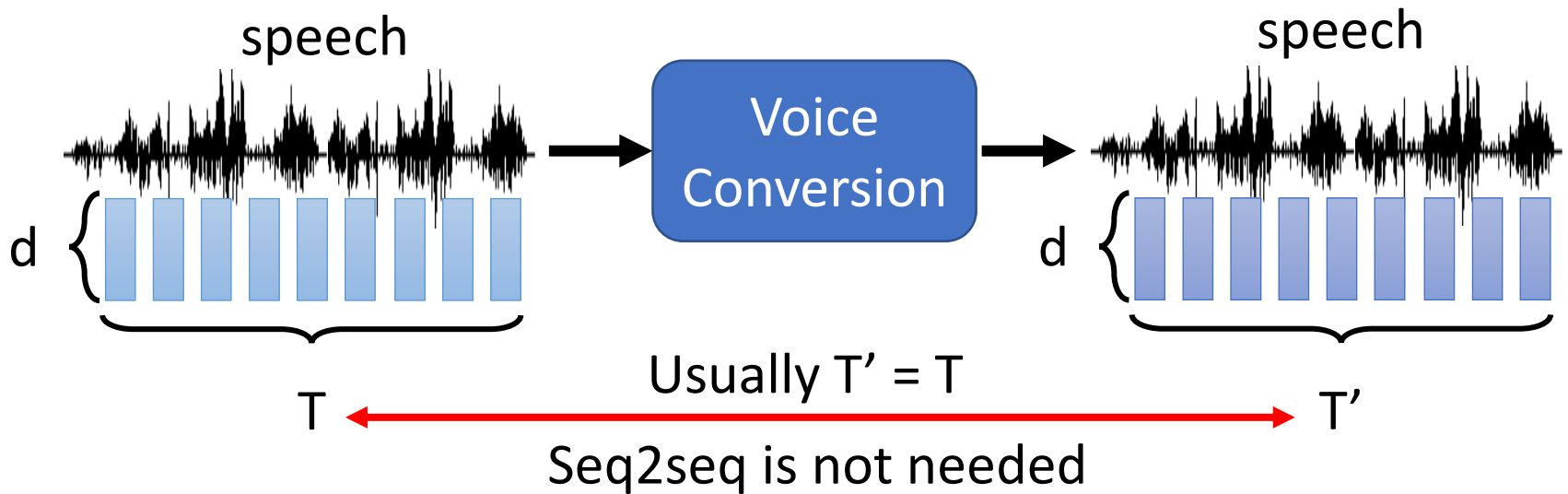
Training  
Data x 2

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



[Mimura, et al.,  
ASRU 2017]

# In real implementation ...



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

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



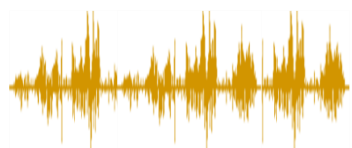
# Categories

Lack of training data:

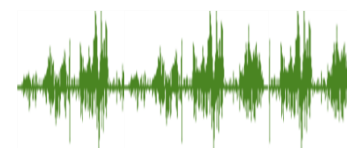
- Model Pre-training [Huang, et al., arXiv'19]
- Synthesized data!

[Biadsy, et al., INTERSPEECH'19]

Parallel Data



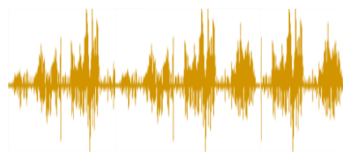
How are you?



How are you?

Unparallel Data

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



天氣真好

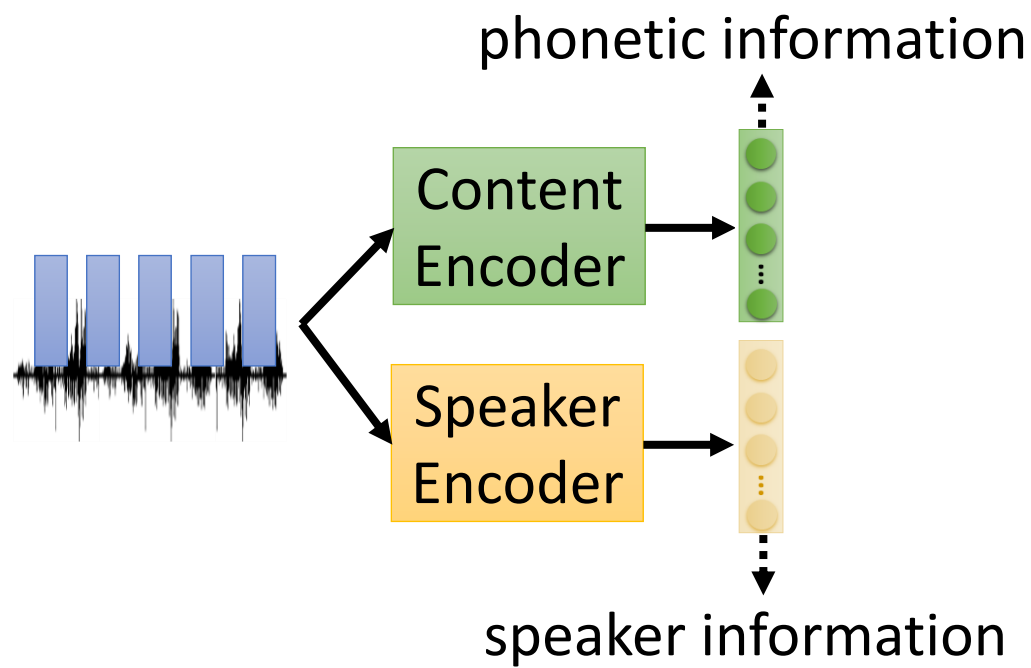


How are you?

# Categories

Parallel Data

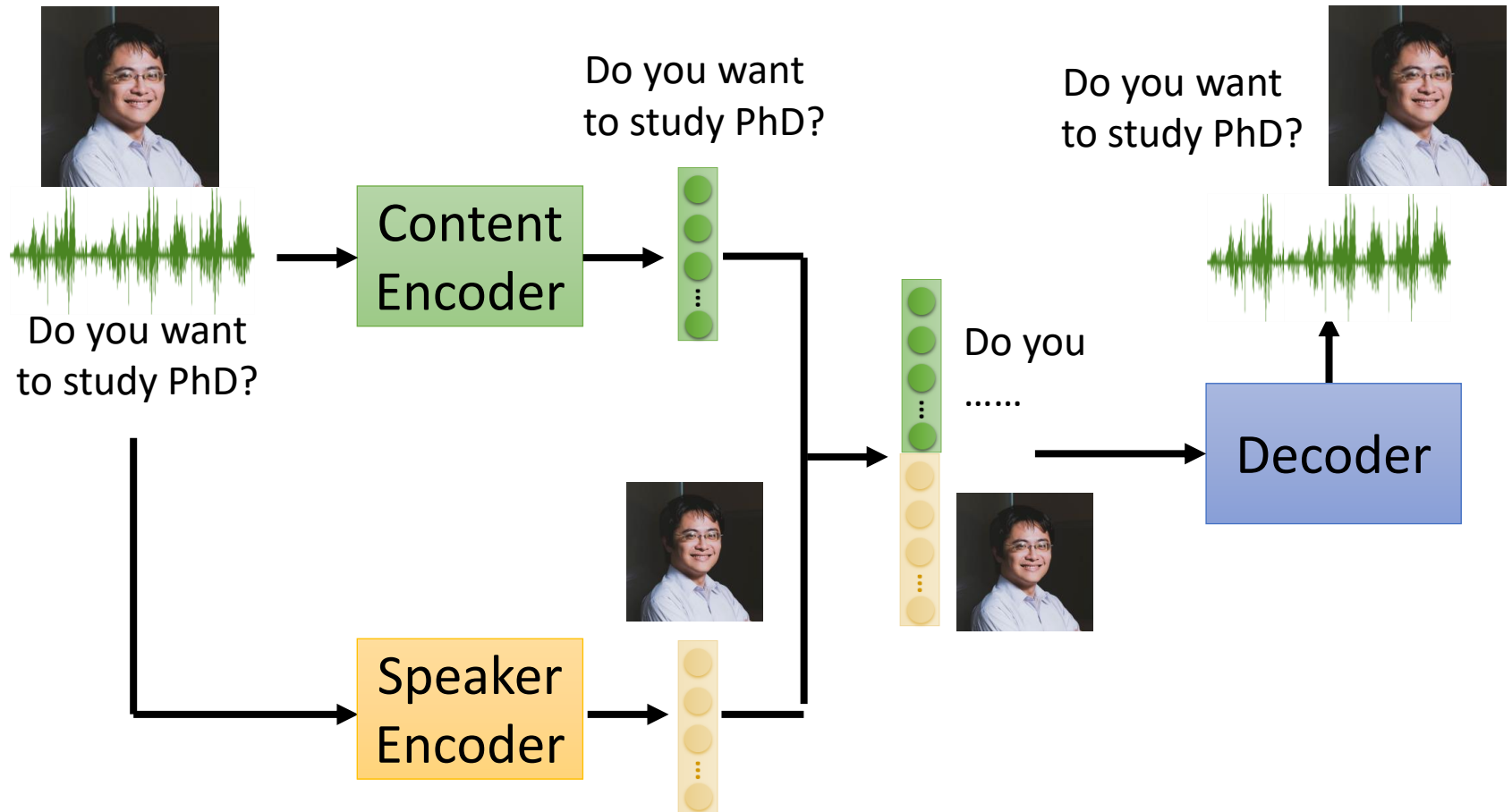
Unparallel Data



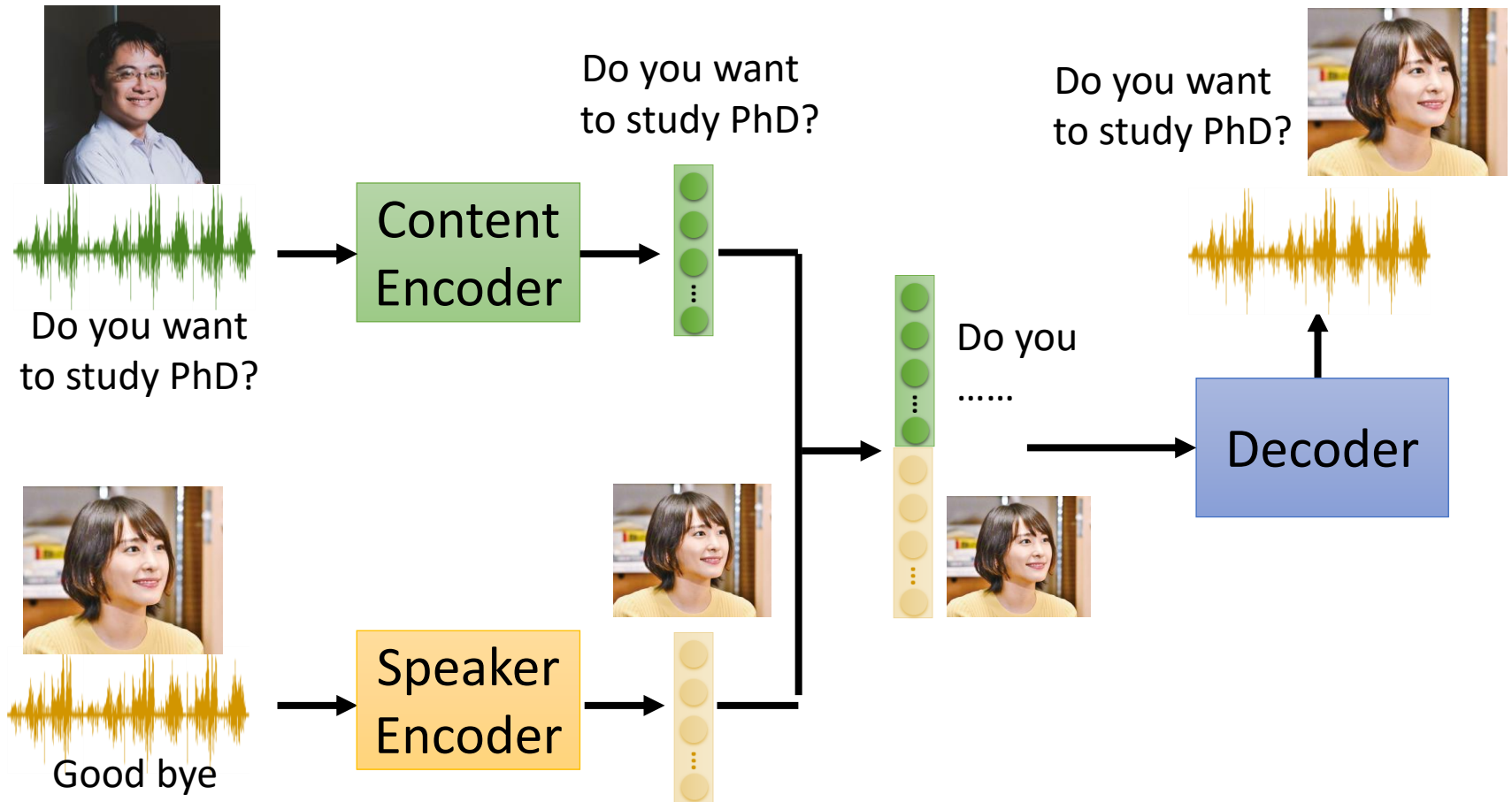
Feature Disentangle

Direct Transformation

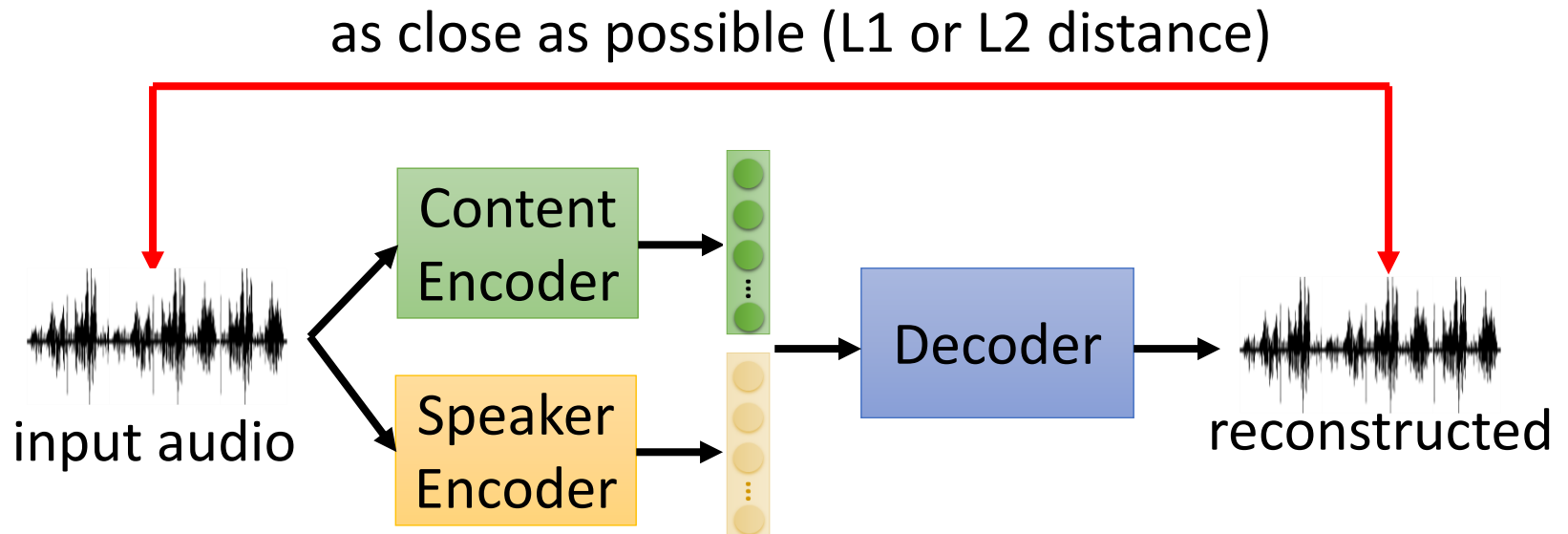
# Feature Disentangle



# Feature Disentangle



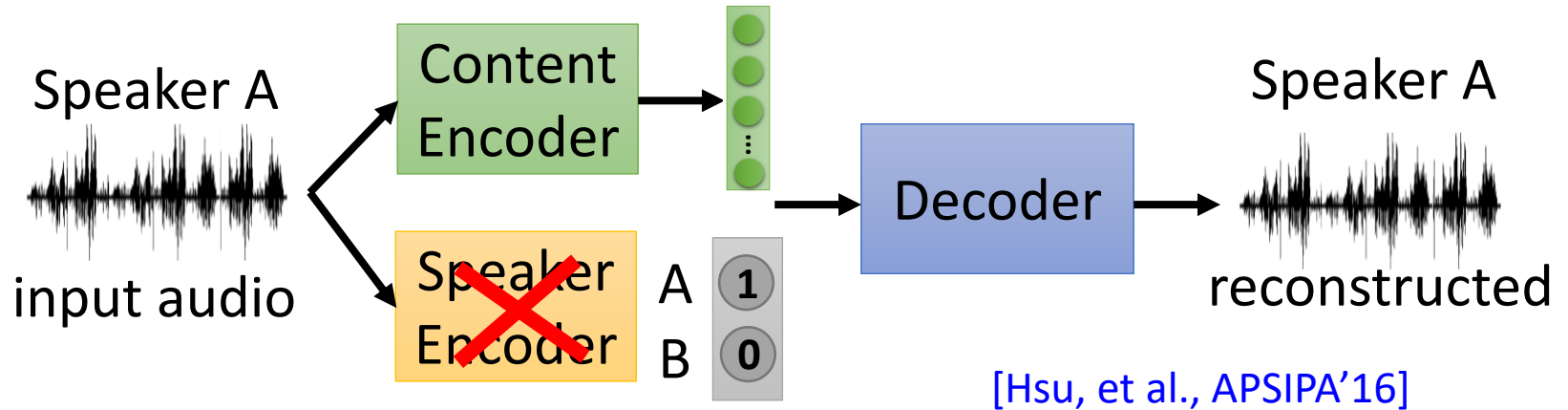
# Feature Disentangle



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

# Using Speaker Information

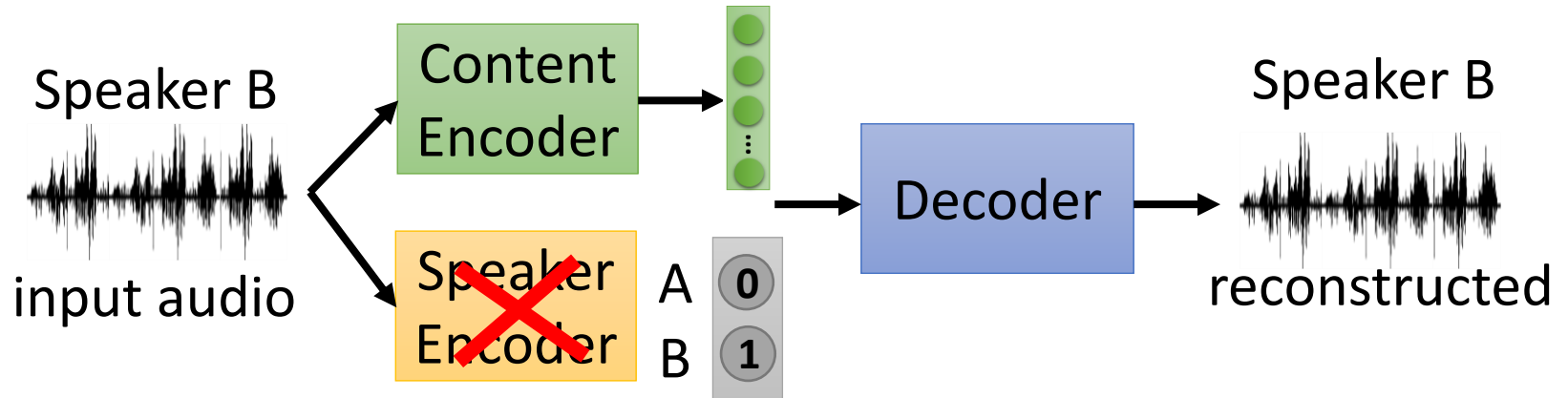
Assume we know the speakers of training utterances



- One-hot vector for each speaker

# Using Speaker Information

Assume we know the speakers of training utterances

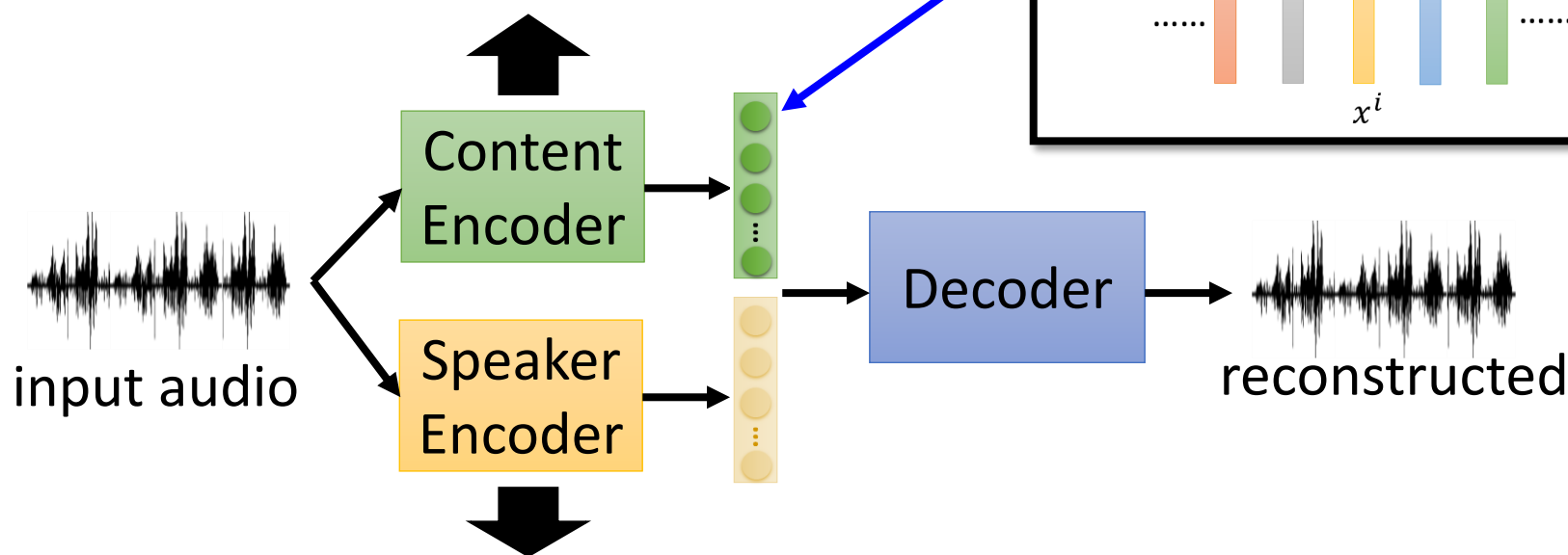


- One-hot vector for each speaker

# 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

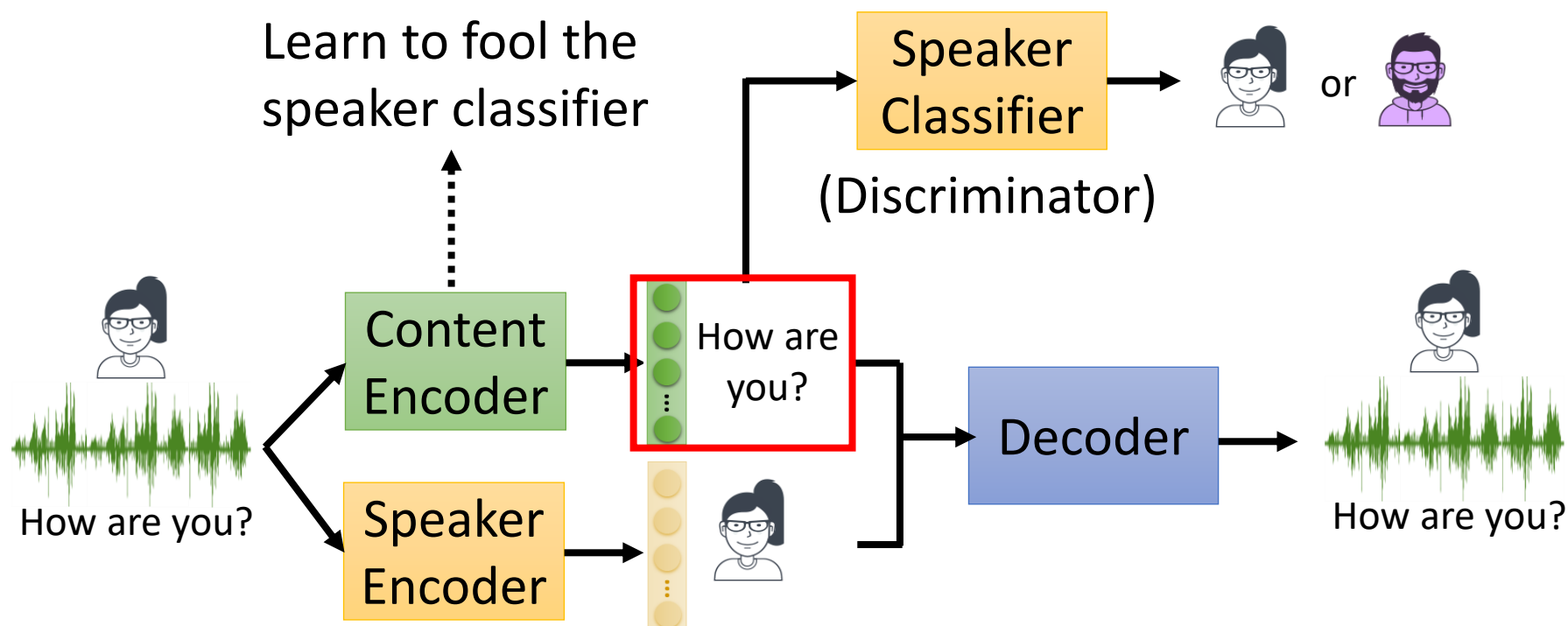
- Speaker embedding (i-vector, d-vector, x-vector ... )

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



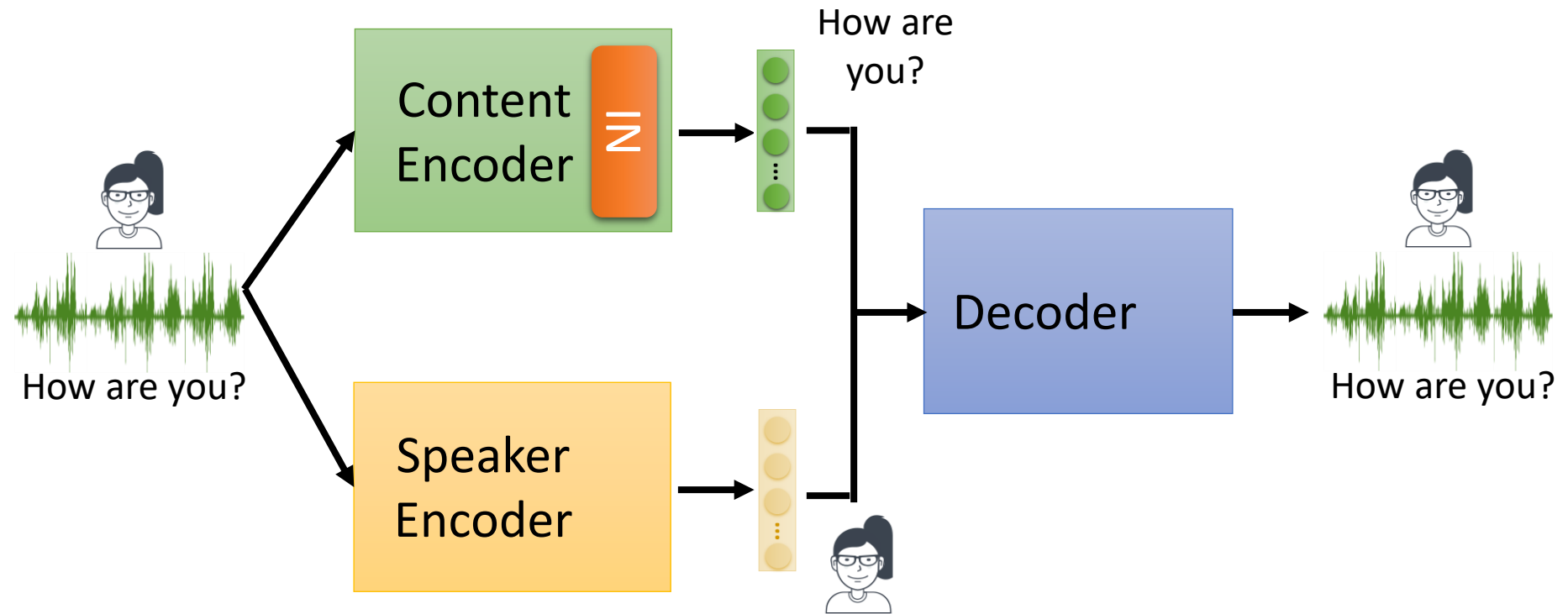
# Adversarial Training

[Chou, et al., INTERSPEECH'18]



Speaker classifier and encoder are learned iteratively

# Designing network architecture



IN

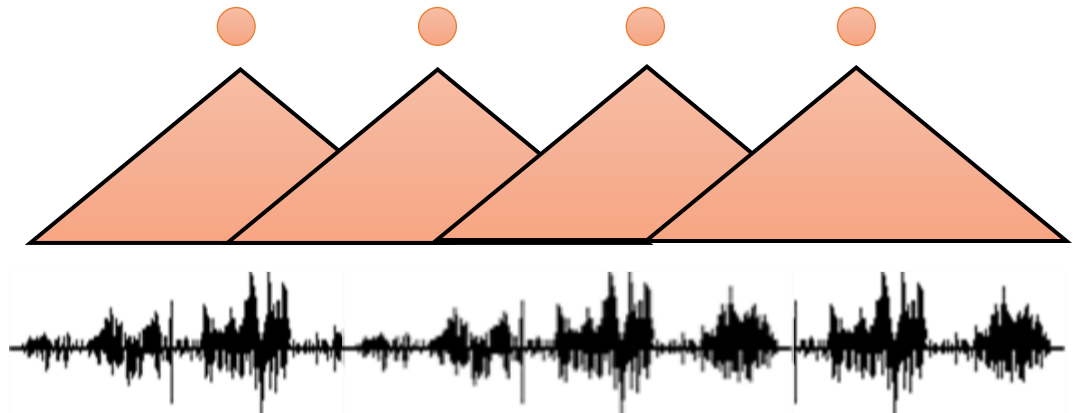
= instance normalization (remove global information)

# Designing network architecture

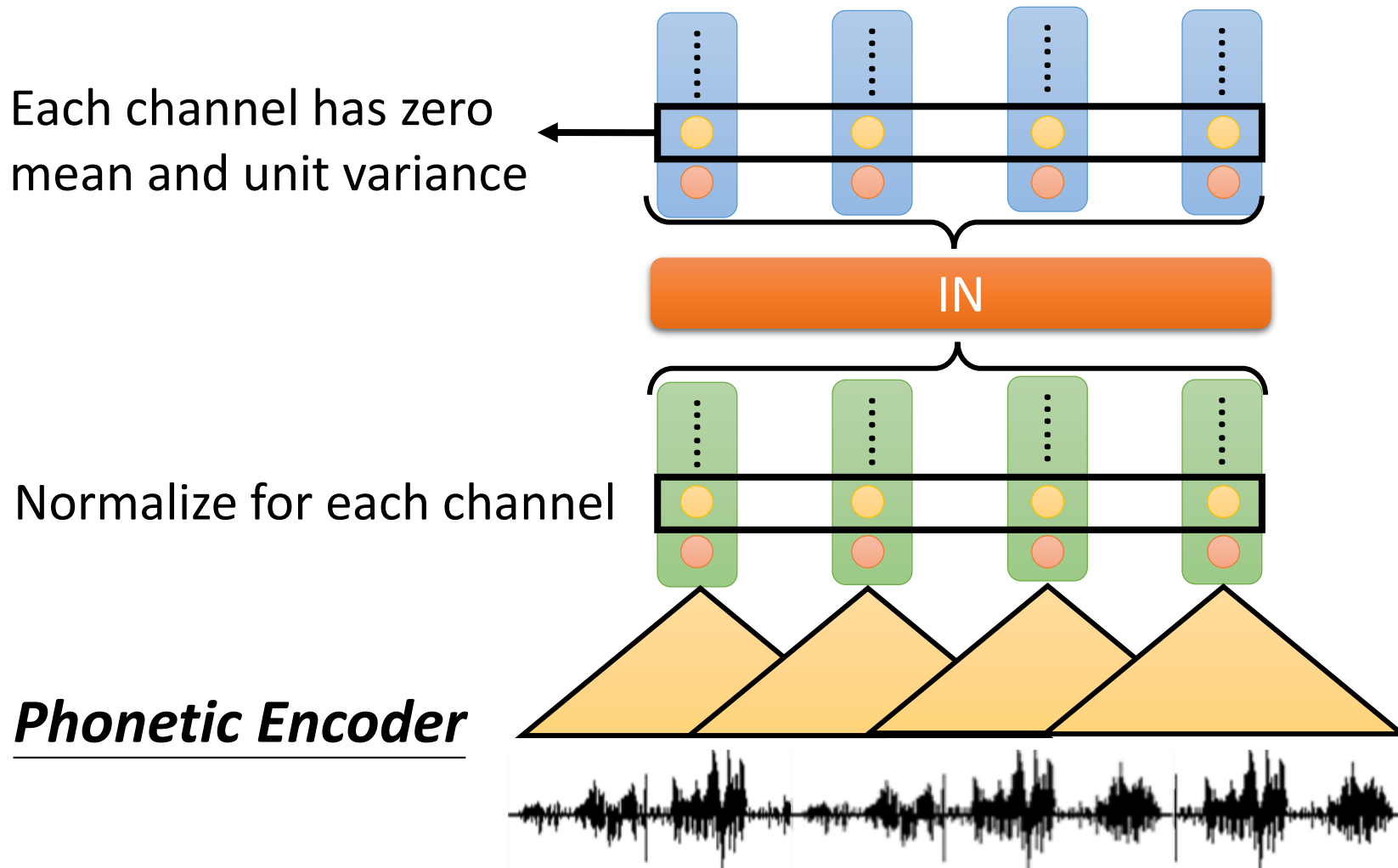
IN

= instance normalization (remove global information)

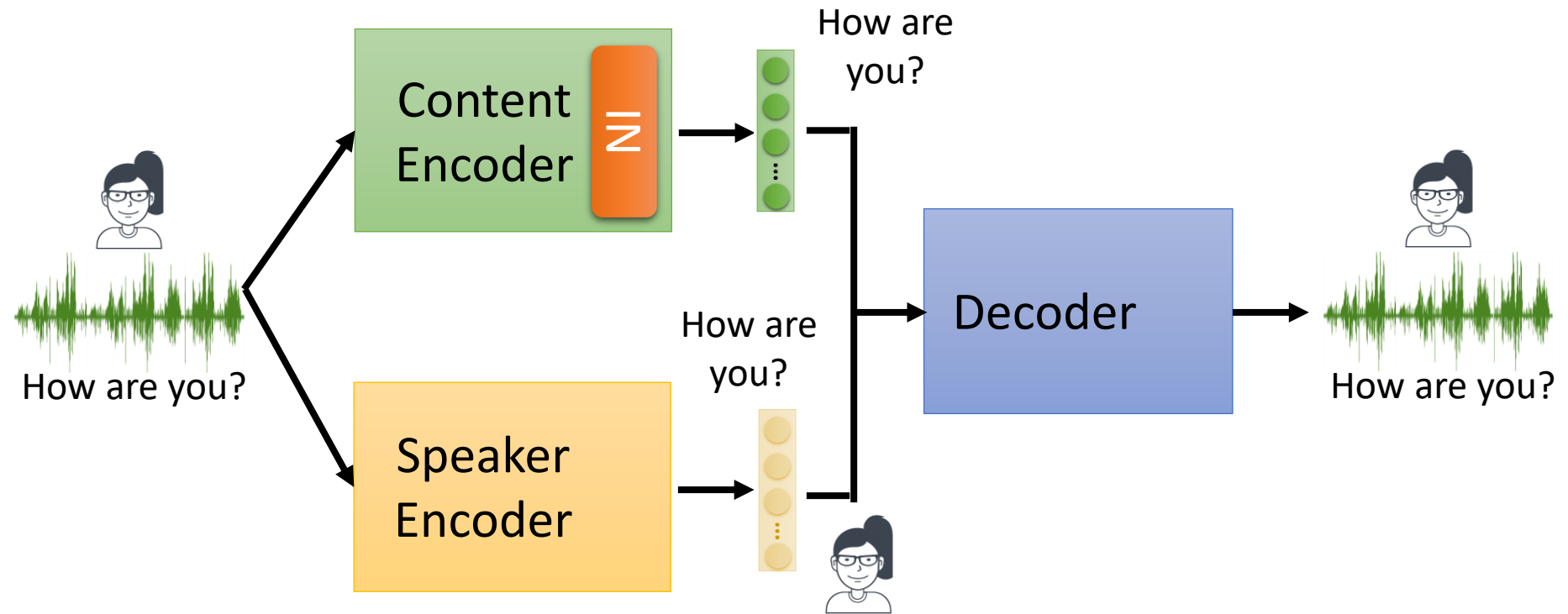
*Phonetic Encoder*



# Designing network architecture



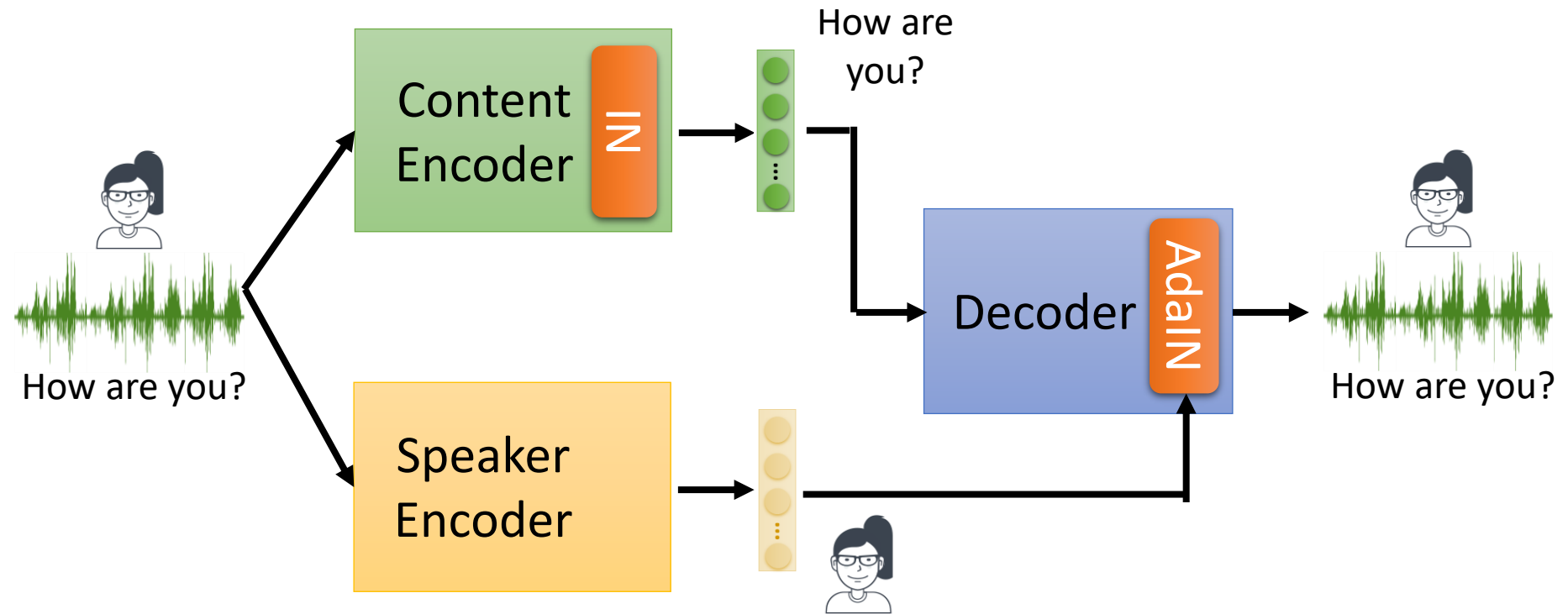
# Designing network architecture



IN

= instance normalization (remove global information)

# Designing network architecture



**IN** = instance normalization (remove global information)

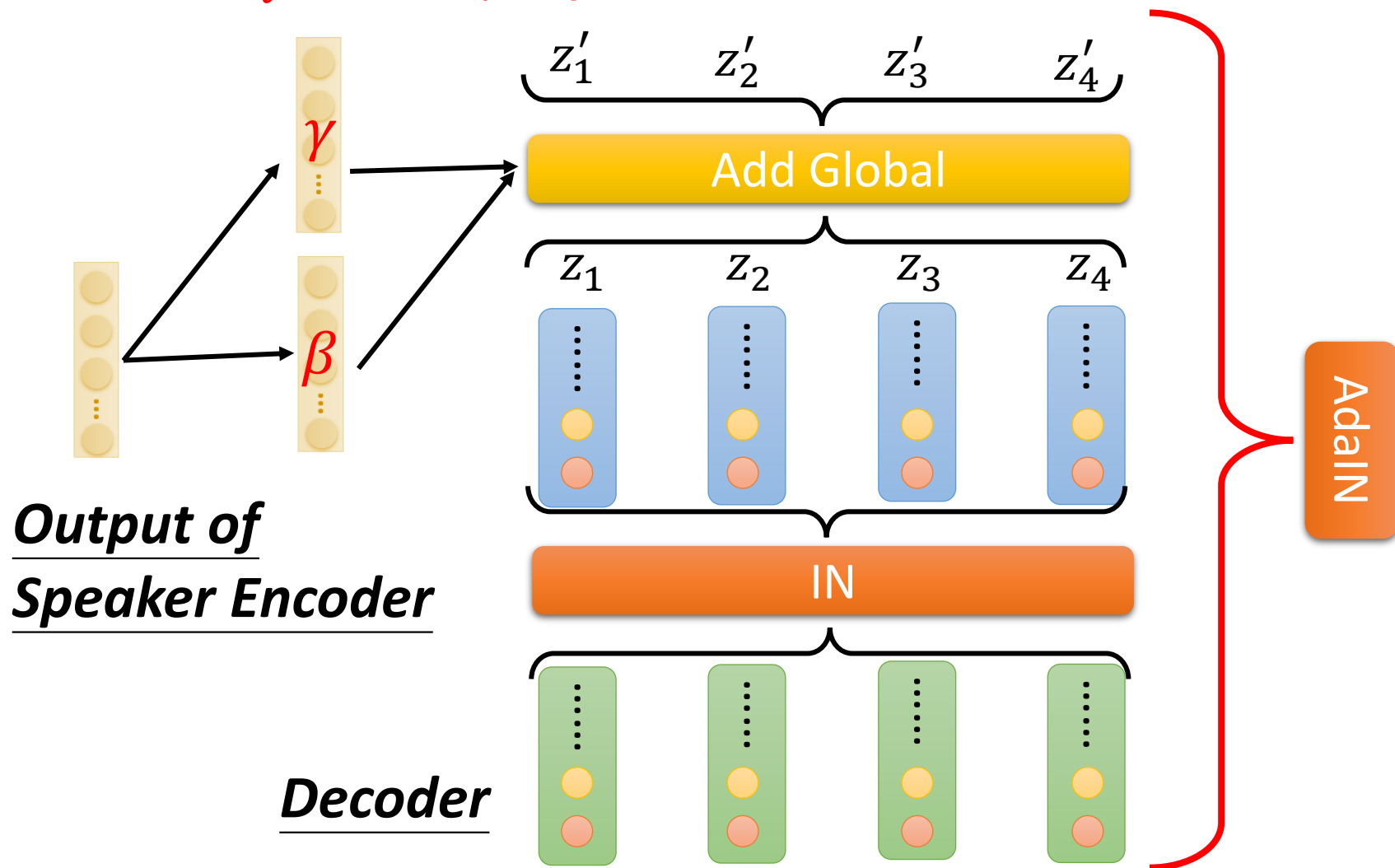
**AdaIN** = adaptive instance normalization  
(only influence global information)

AdaIN

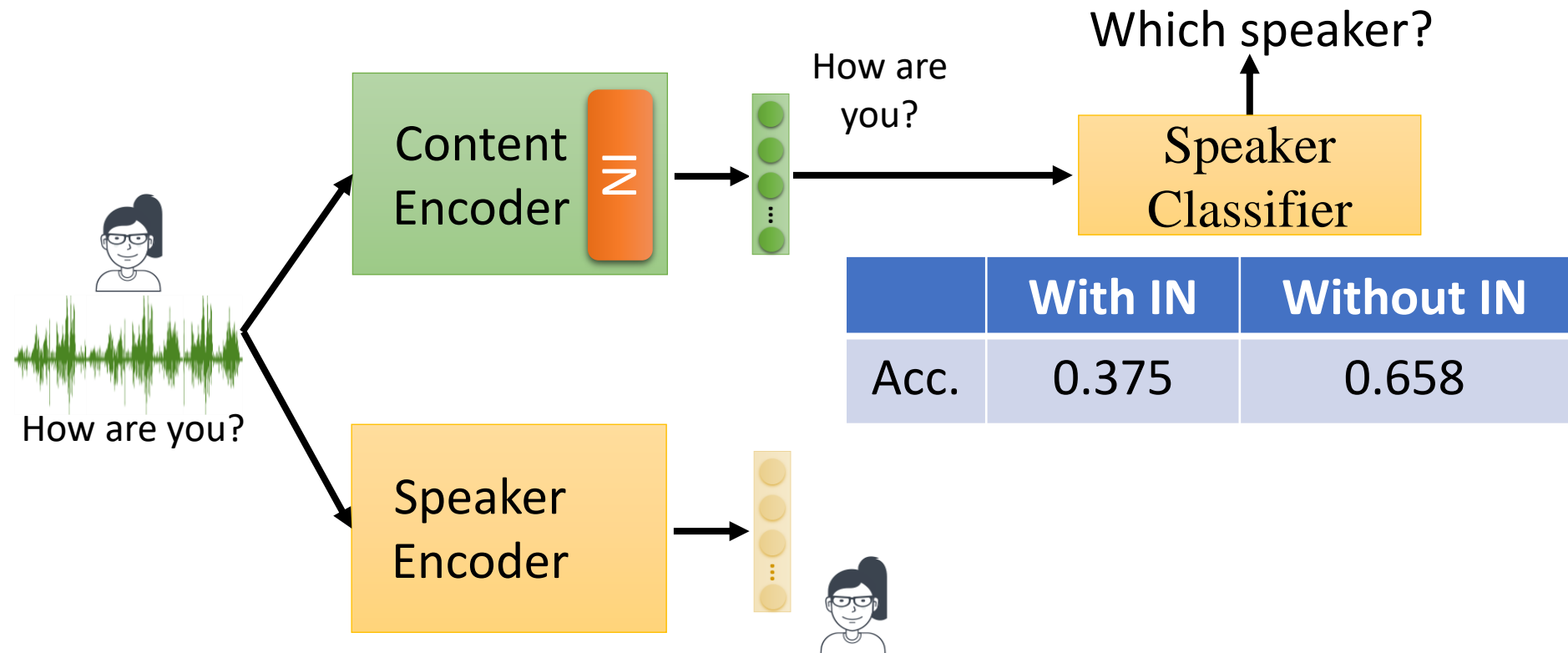
= adaptive instance normalization

(only influence global information)

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



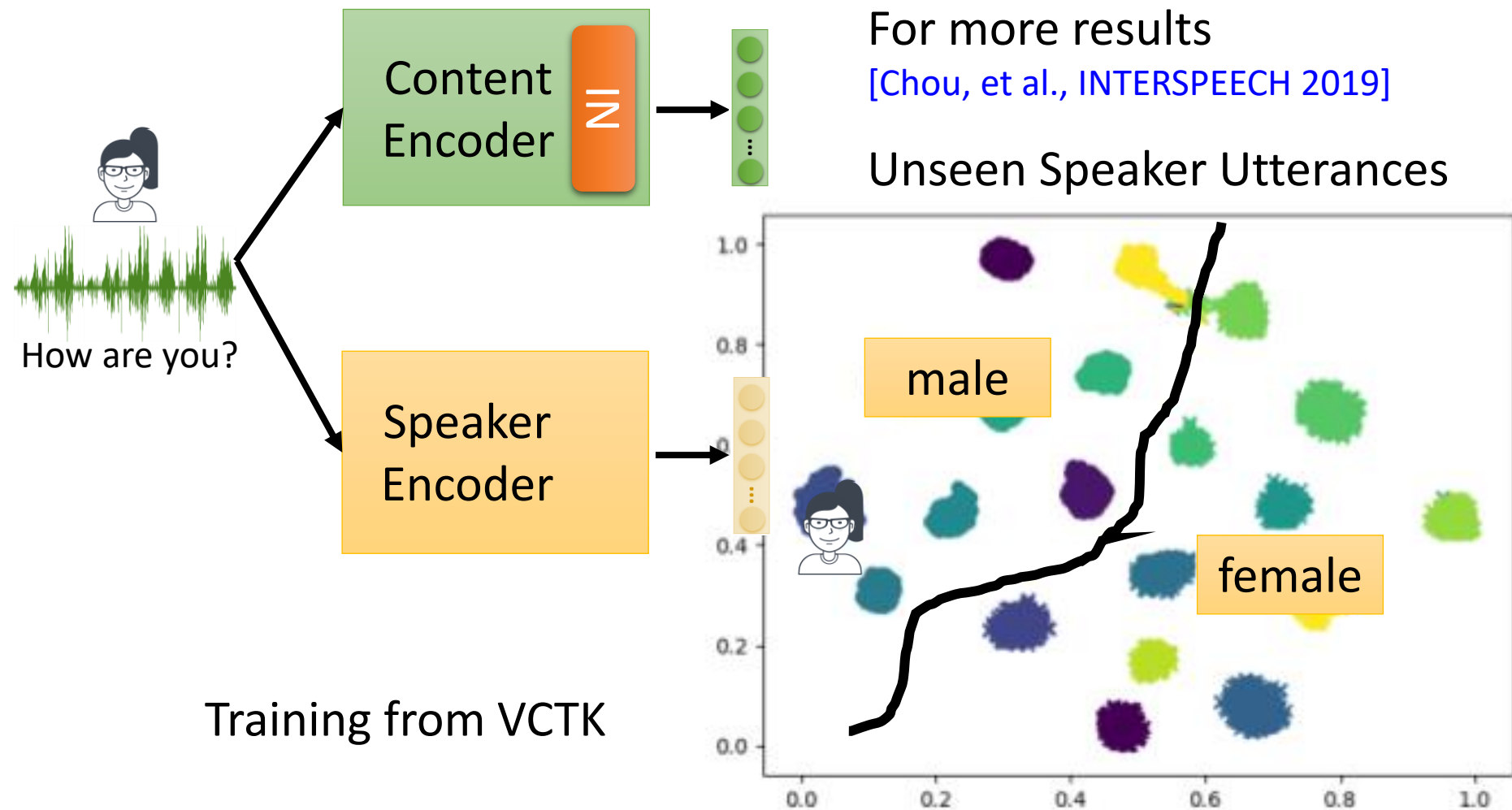
# Designing network architecture



Training from VCTK

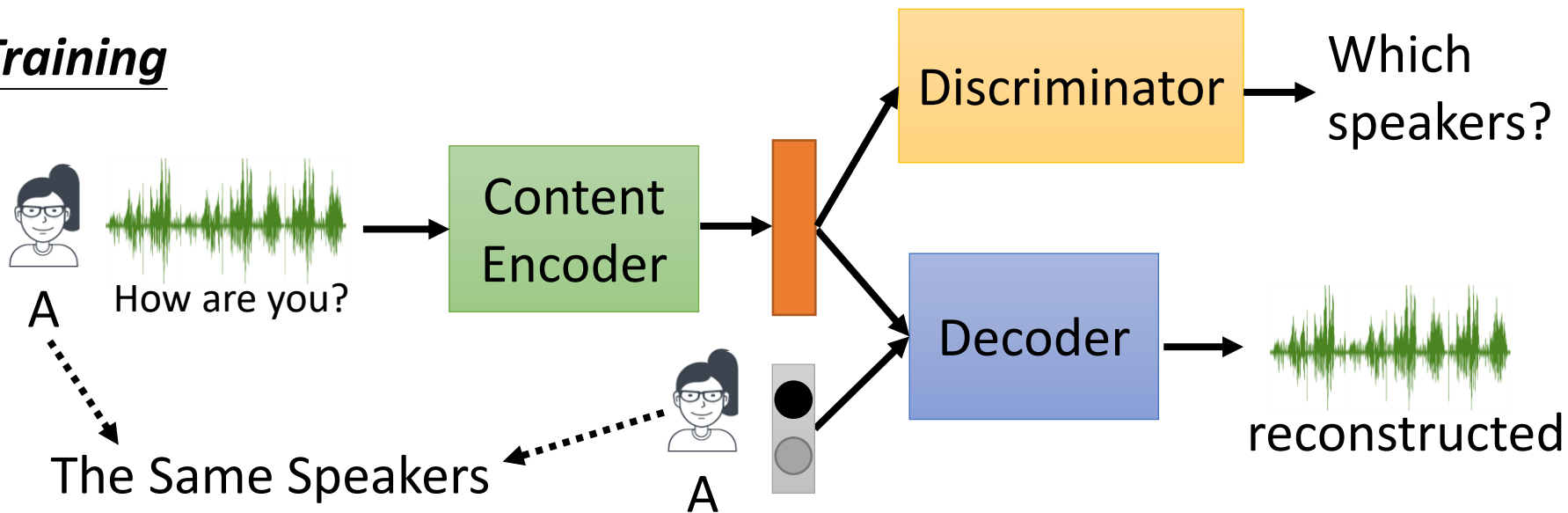


# Designing network architecture

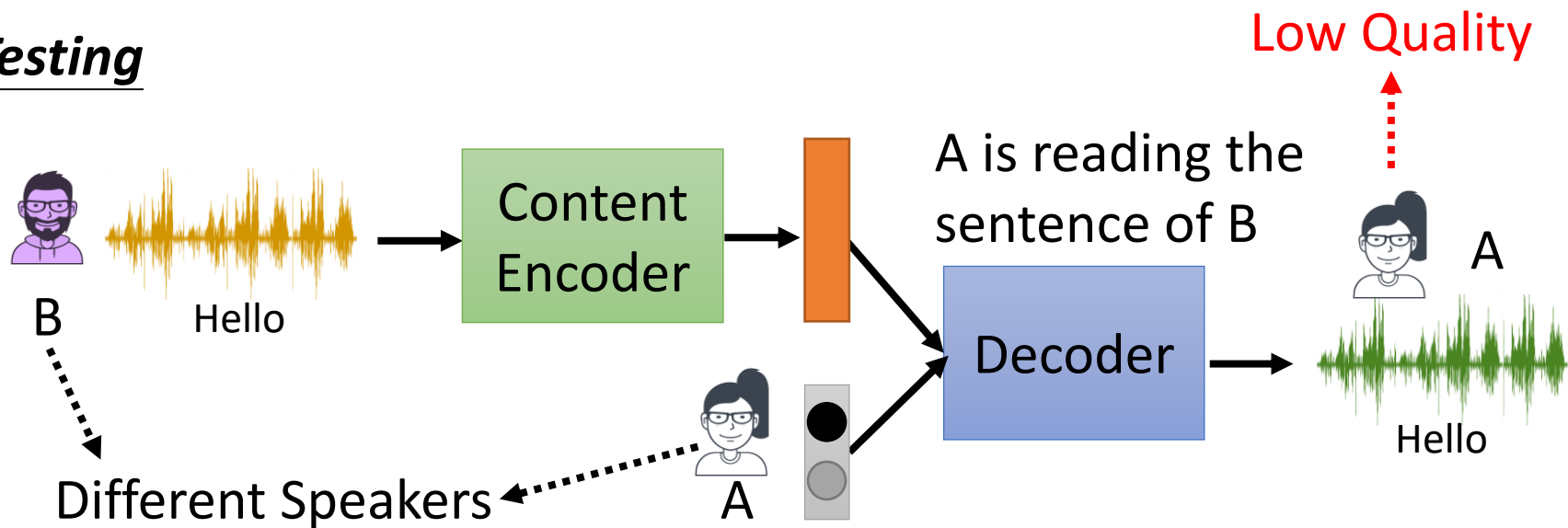


# Issues

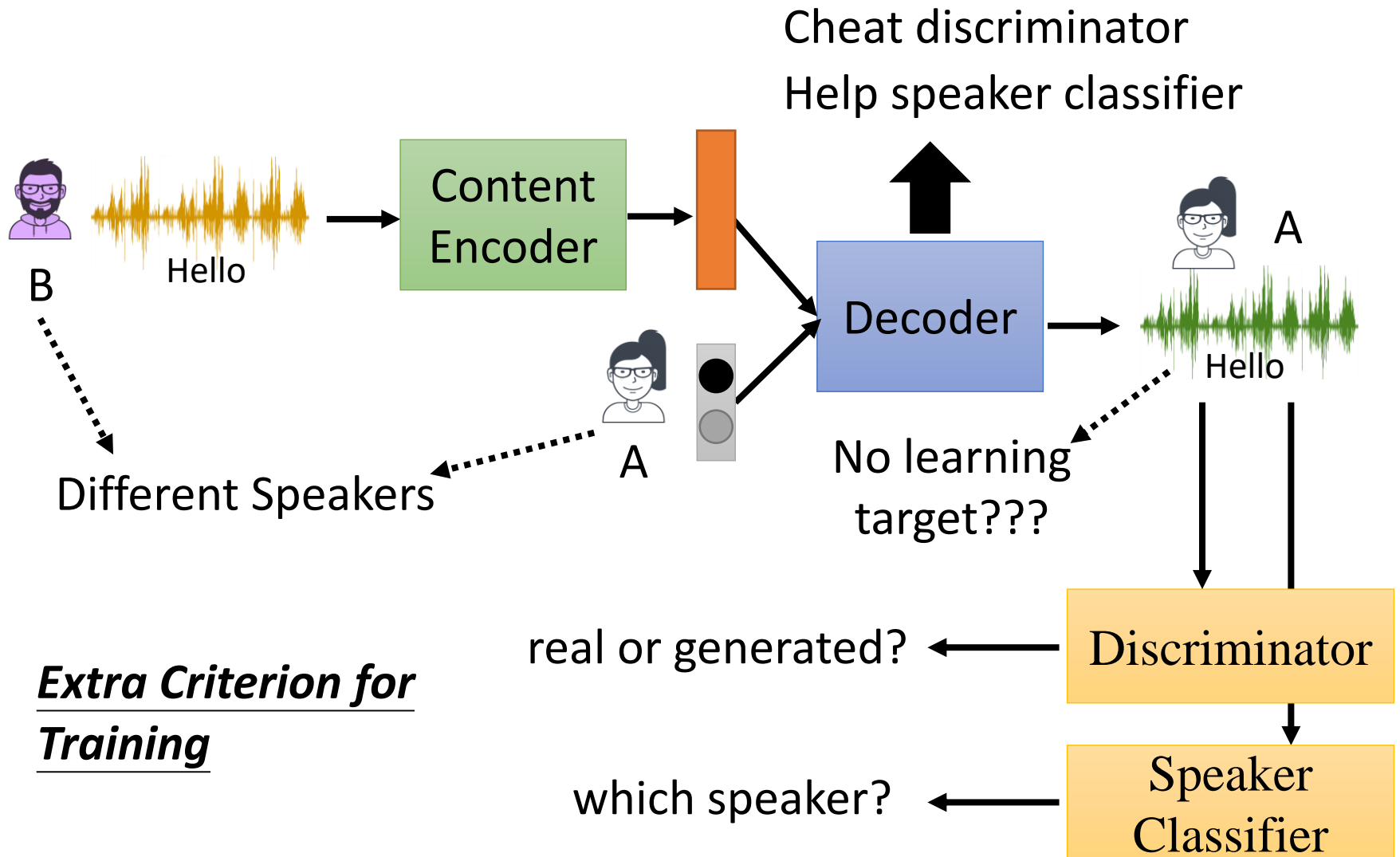
## Training



## Testing

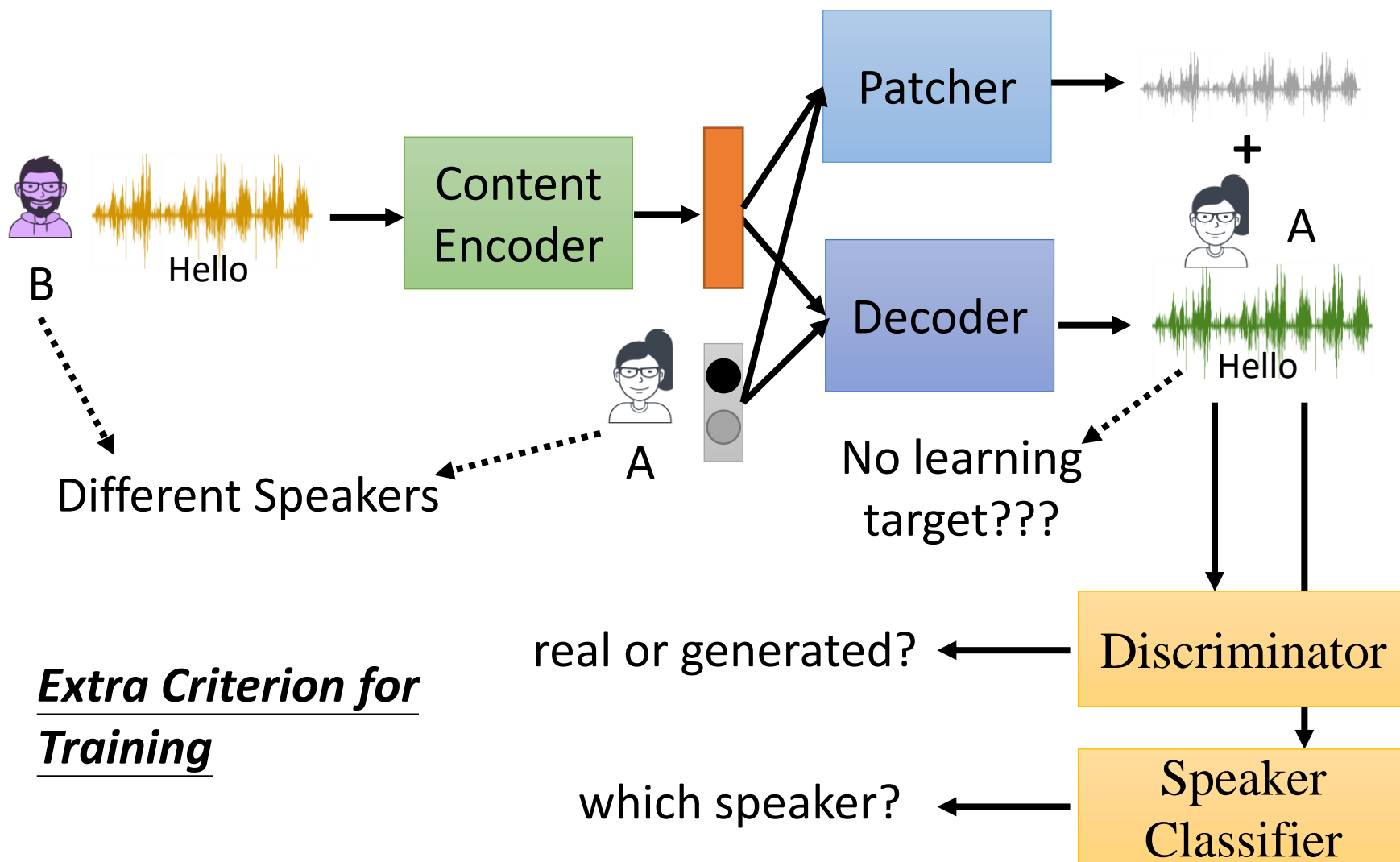


# 2nd Stage Training



# 2nd Stage Training

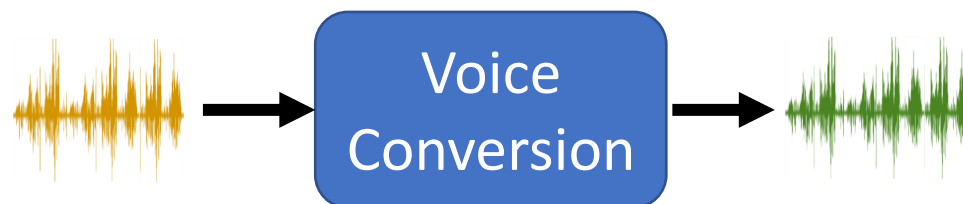
Only learn the patcher  
in the 2<sup>nd</sup> stage



# Categories

- Training without parallel data
- Using CycleGAN

Parallel Data



Unparallel Data

Feature Disentangle

Direct Transformation

# Cycle GAN

[Kaneko, et al., ICASSP'19]



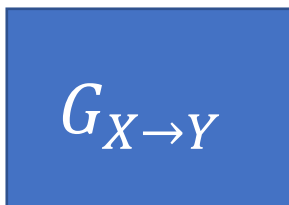
Speaker X



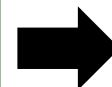
Speaker Y



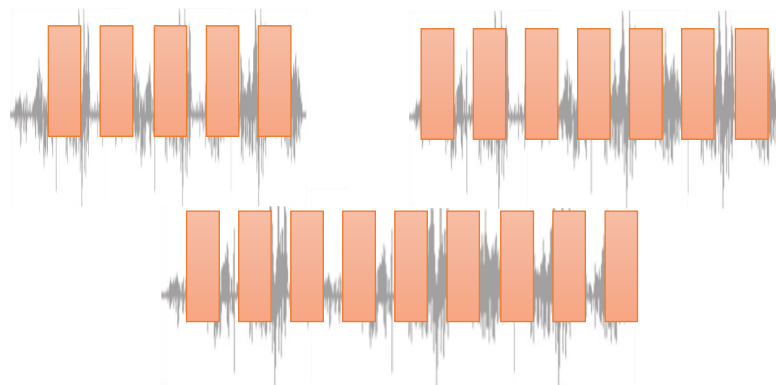
Speaker X



Become similar  
to speaker Y



scalar



Speaker Y



Input audio belongs  
to speaker Y?

# Cycle GAN

[Kaneko, et al., ICASSP'19]



Speaker X



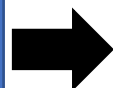
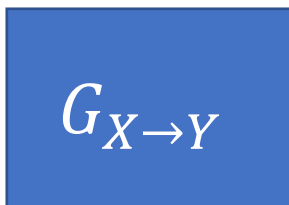
Speaker Y



Speaker X



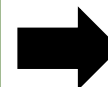
ignore input



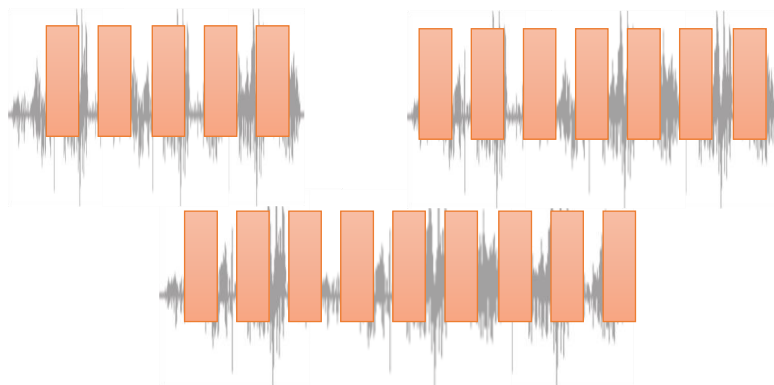
Become similar  
to speaker Y



Not what we want!



scalar

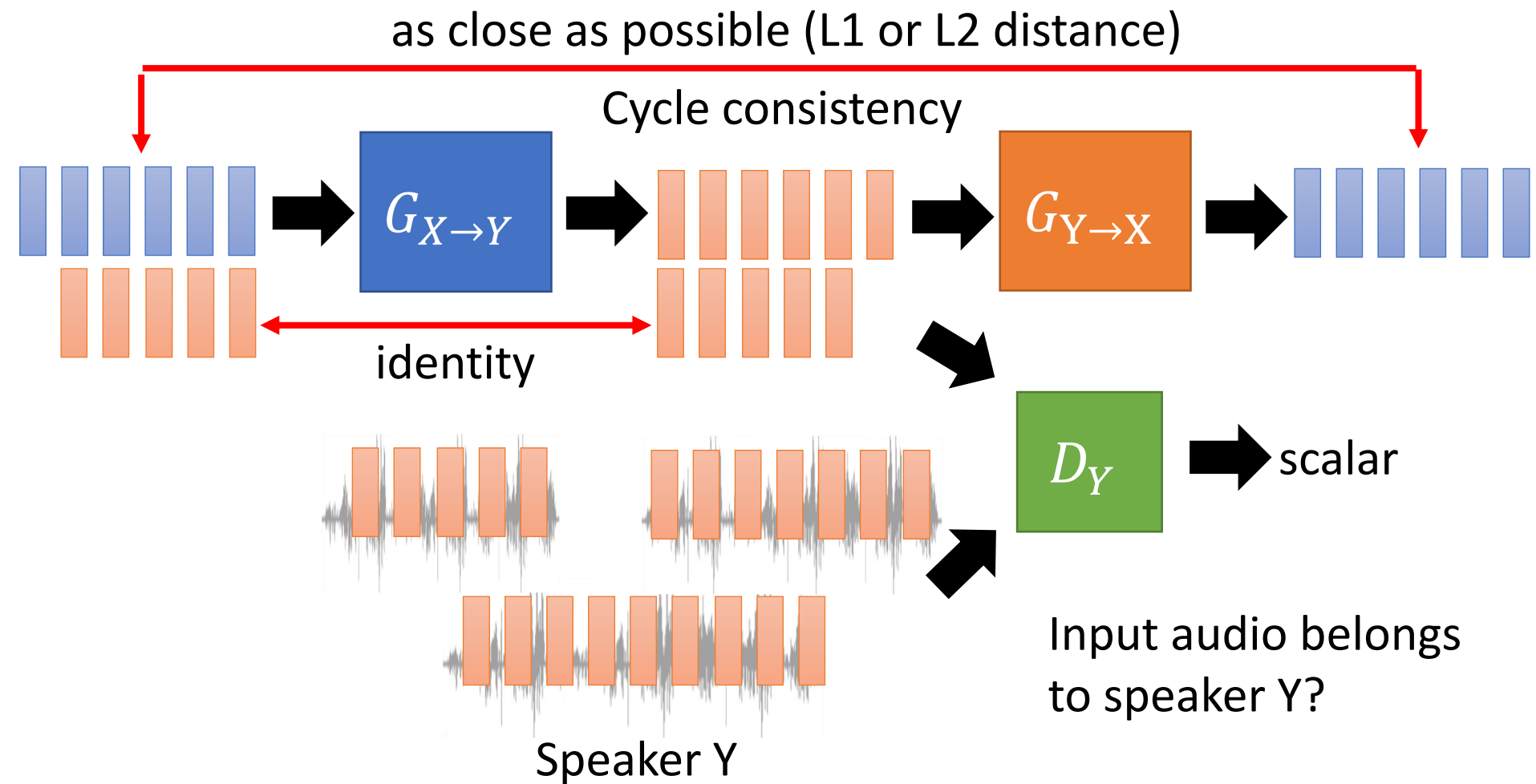


Speaker Y



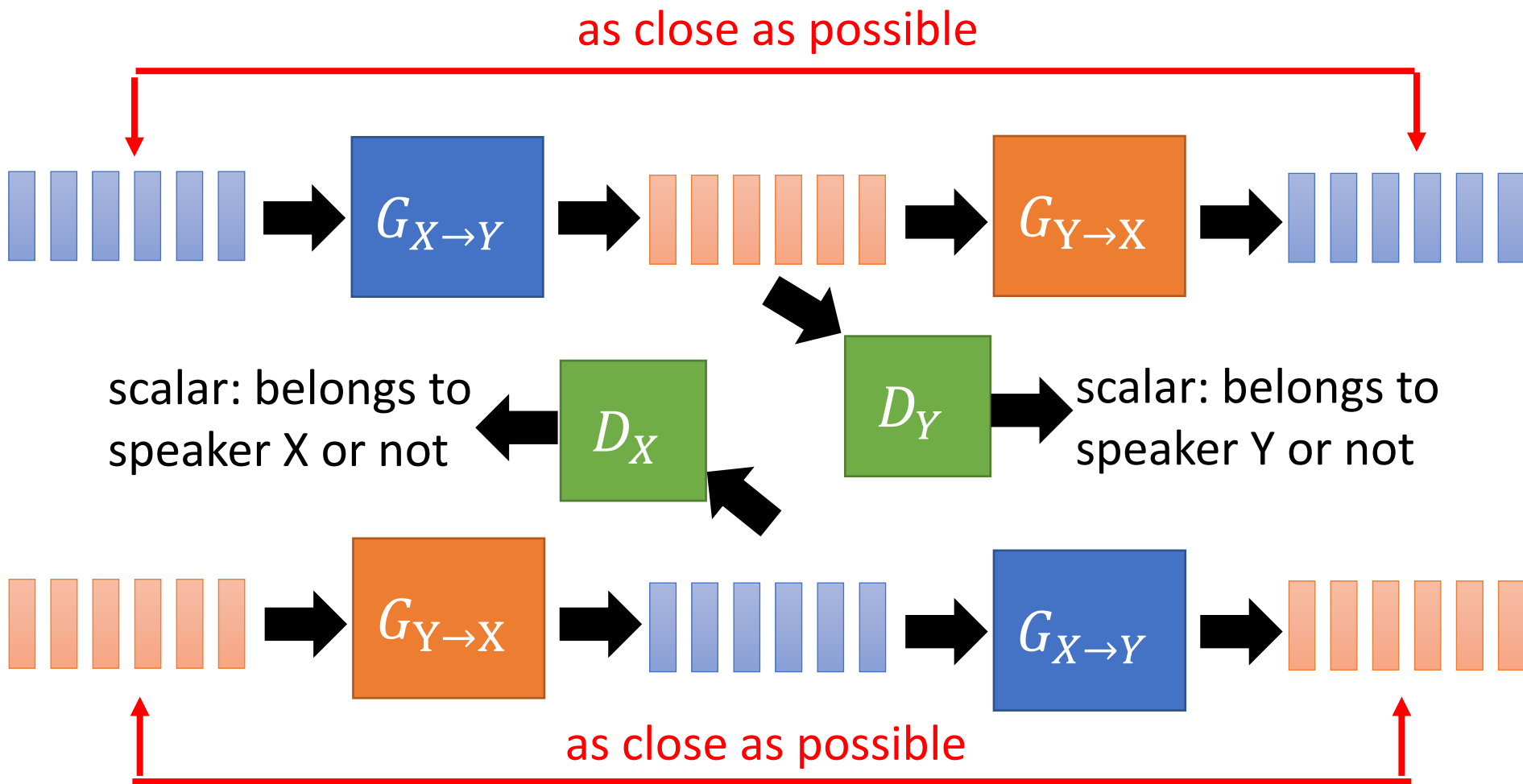
Input audio belongs  
to speaker Y?

# Cycle GAN





# Cycle GAN

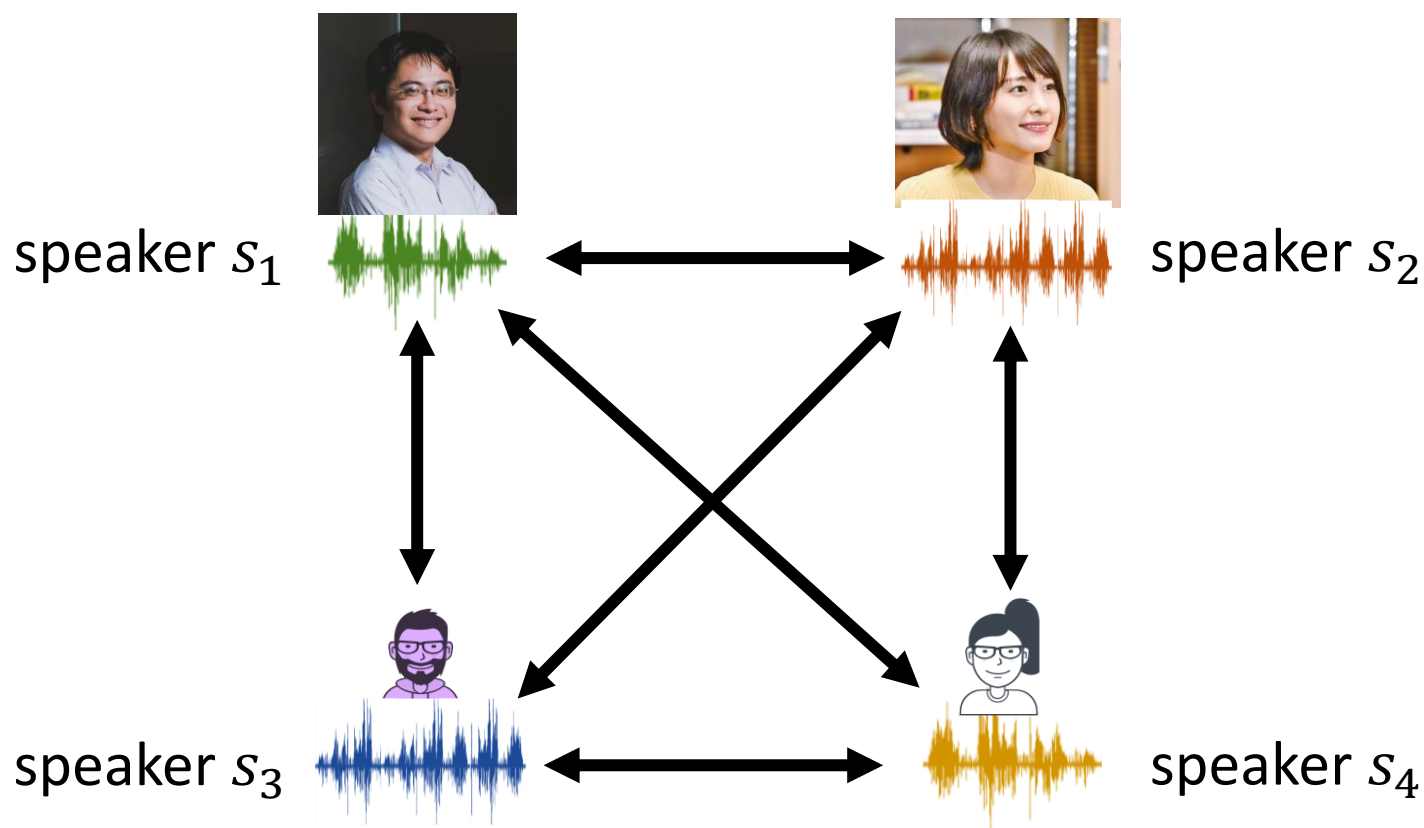


# StarGAN

[Kaneko, et al., INTERSPEECH'19]

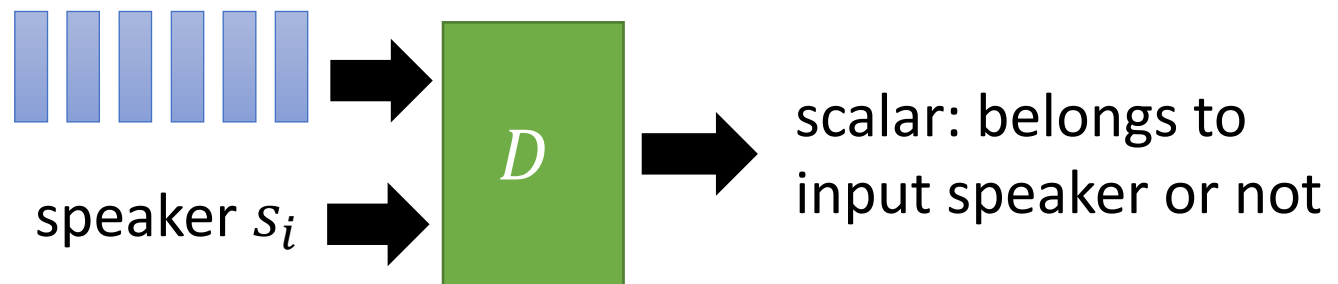
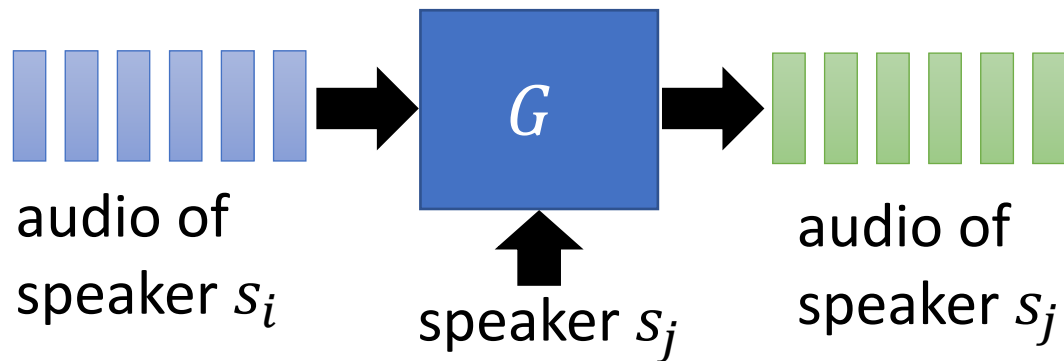
For CycleGAN:

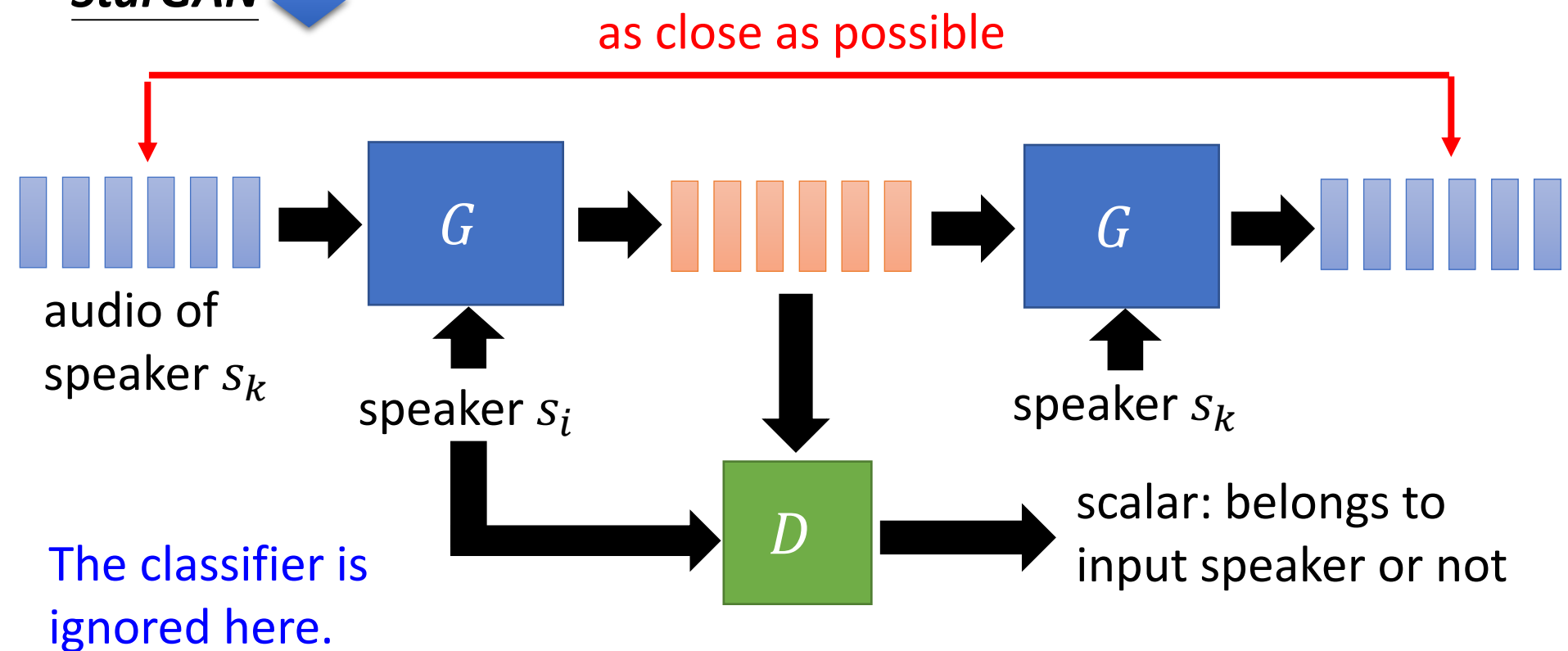
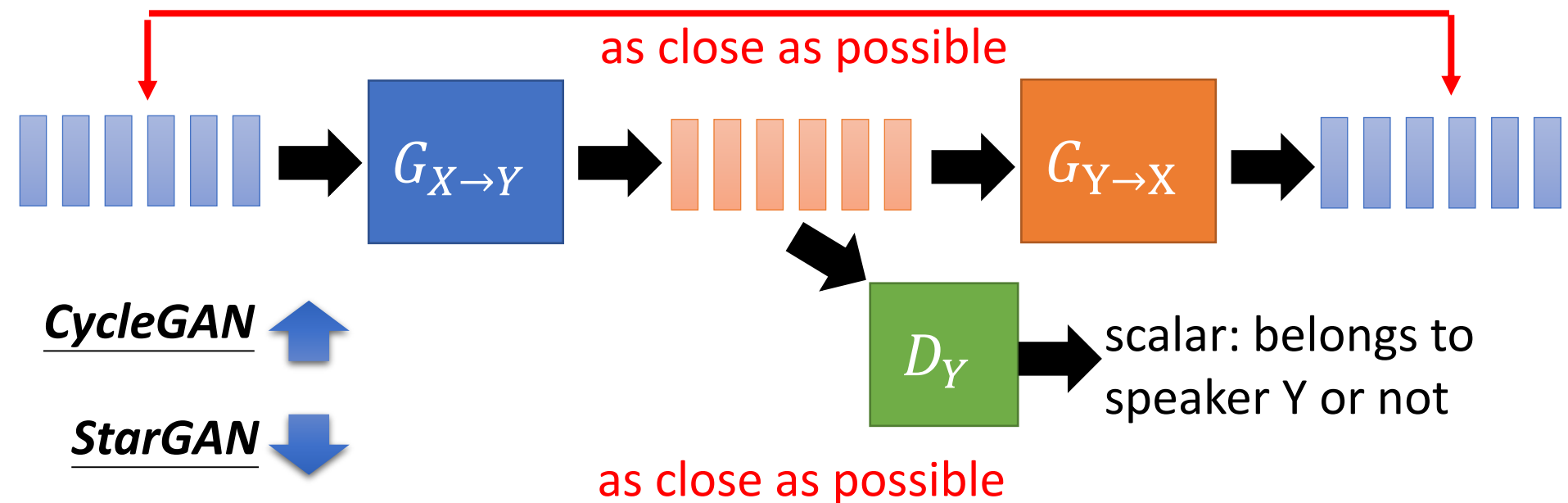
If there are  $N$  speakers, you need  $N \times (N-1)$  generators.



# StarGAN

Each speaker is represented as a vector.





# Blow

[Joan, et al., NeurIPS'19]

## Flow-based model for VC

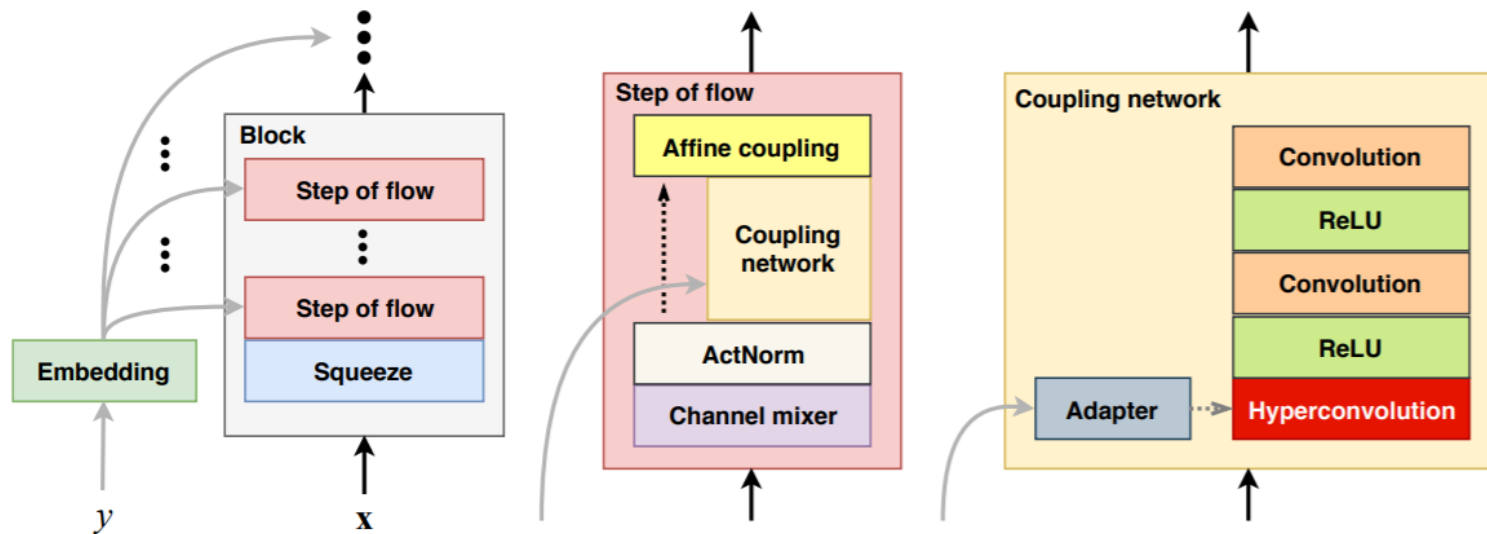
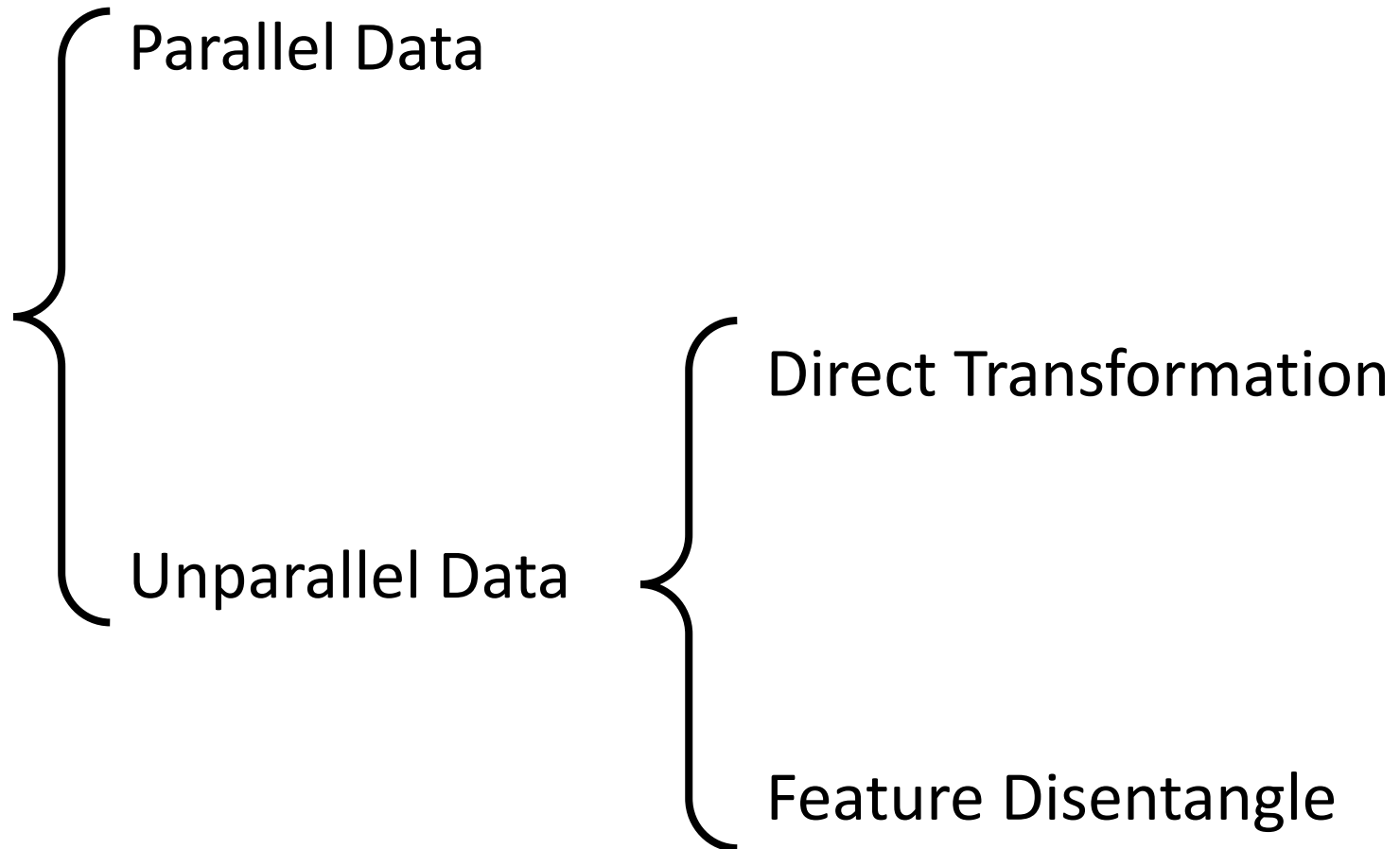


Figure 1: Blow schema featuring its block structure (left), steps of flow (center), and coupling network with hyperconvolution module (right).

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

# Concluding Remarks



# 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