Lecture 17:

Voice Conversion

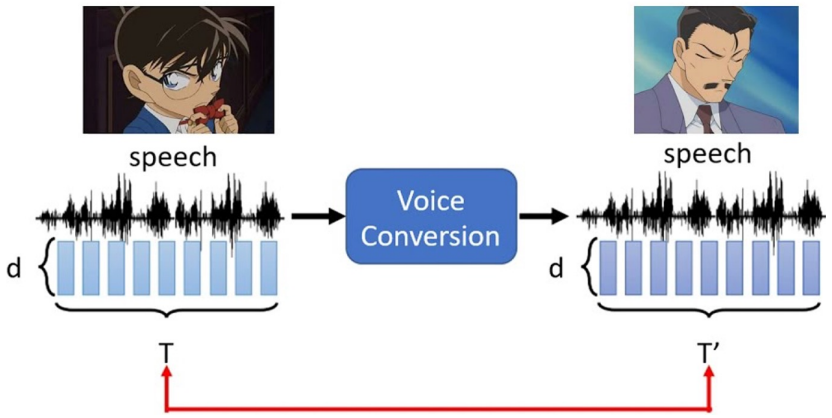
Shuai Wang

Outline

- Introduction
- Basics & Methods
- Beyond common voice conversion
- Appendix
- Q & A

Definition

A common definition:



Voice conversion (VC) is a task that

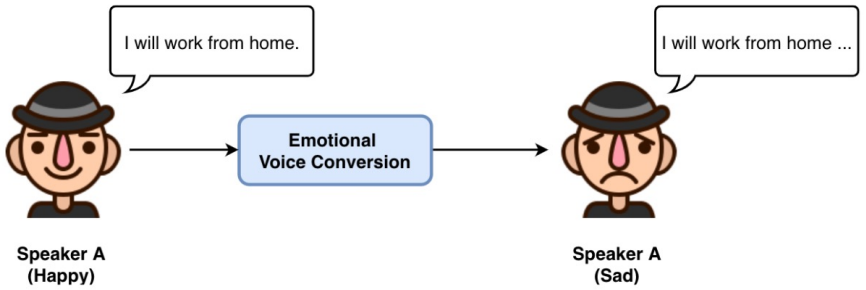
- transforms a speaker's voice into that of another speaker

without altering

- the linguistic content
- prosody and other paralinguistic information

Definition

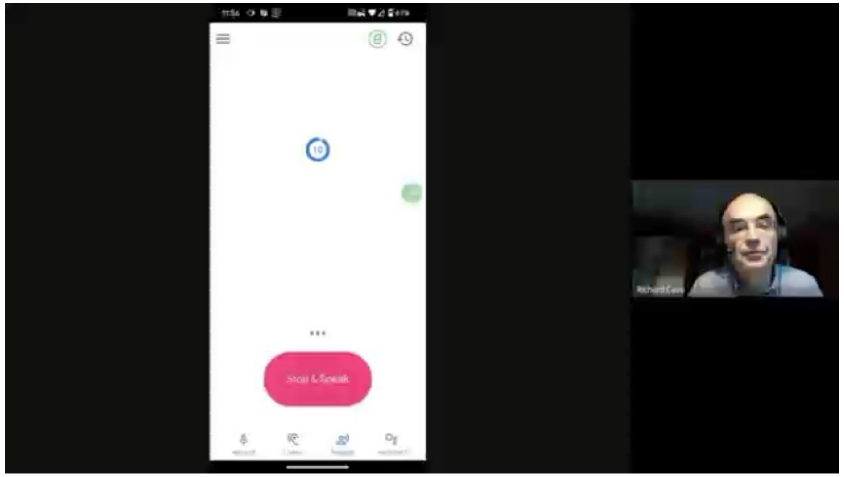
Broader definition:



Picture from https://hltsingapore.github.io/ESD/index.html

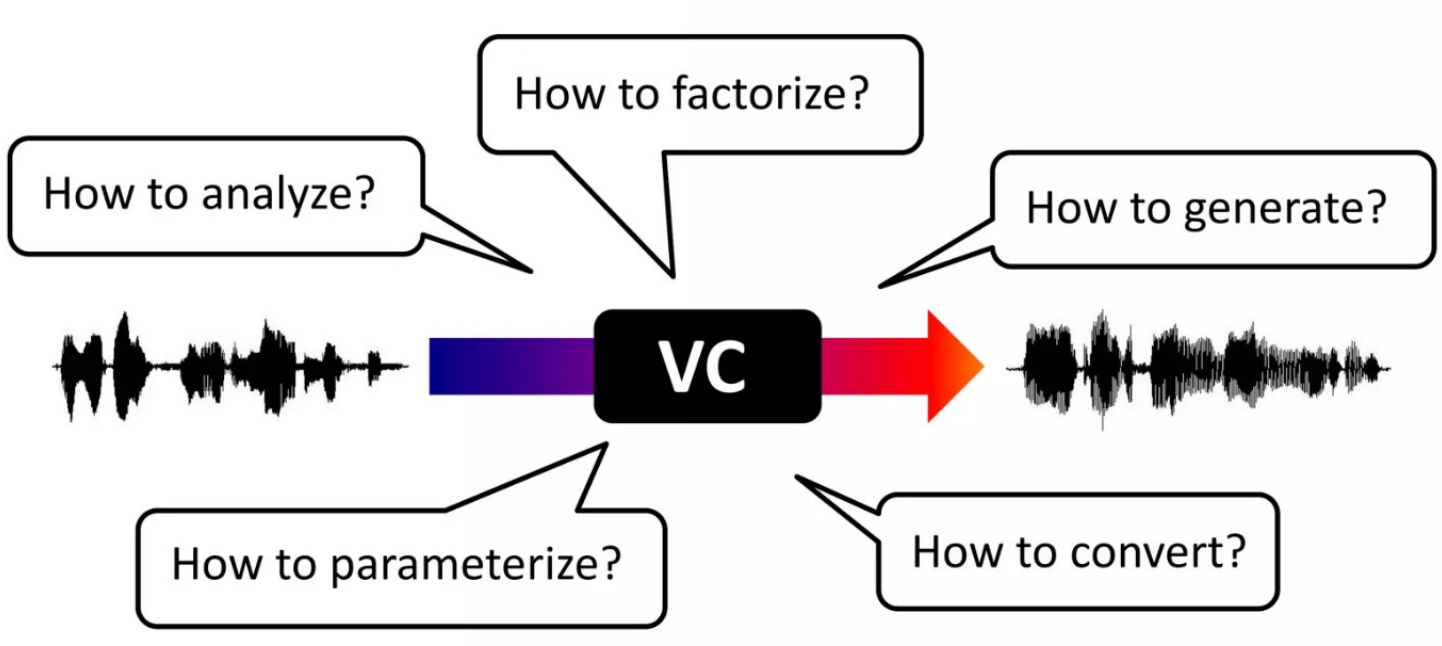
Besides the timbre conversion, can be extended to,

- Emotion conversion
- Dysarthria-to-normal



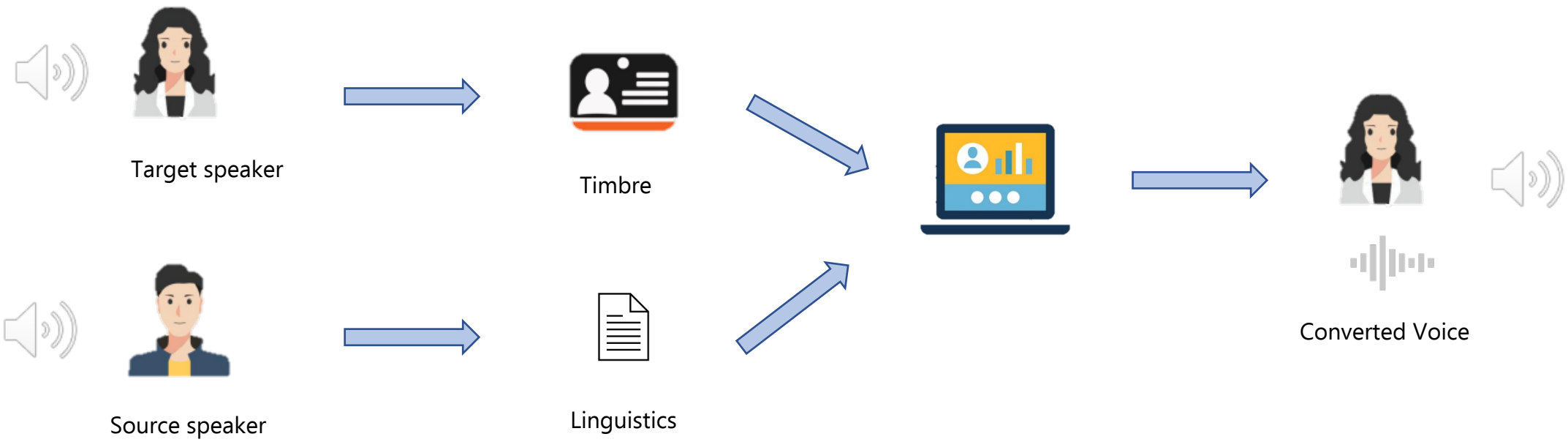
Overall Definition

Voice conversion is a technique to modify the speech waveform to convert non-/para- linguistic information while preserving linguistic information



Tomoki Toda: Recent progress on voice conversion: what is next

Example


Target speaker
Timbre
Source speaker
Linguistics
Converted Voice

Applications

- Voice over for movies
- Livestreaming using the target voice
- Speaker anonymization

Dubbing / voice over

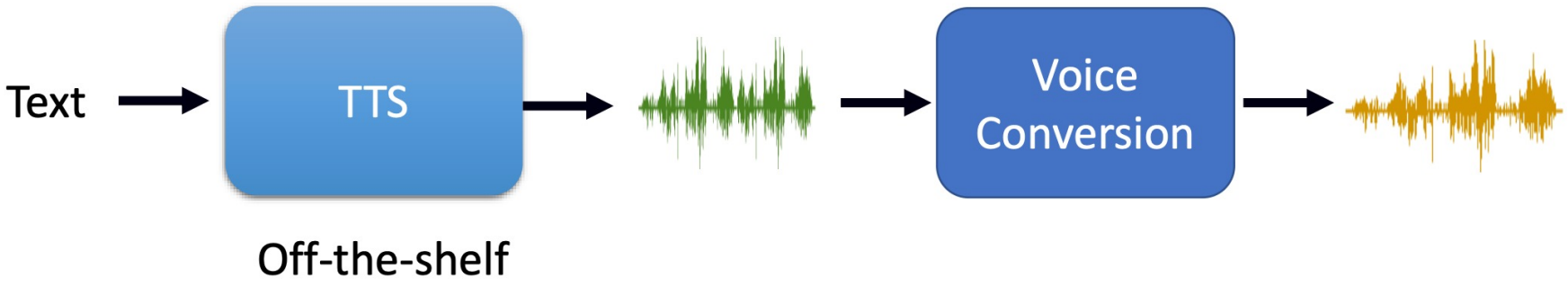
virtual idol

privacy protection

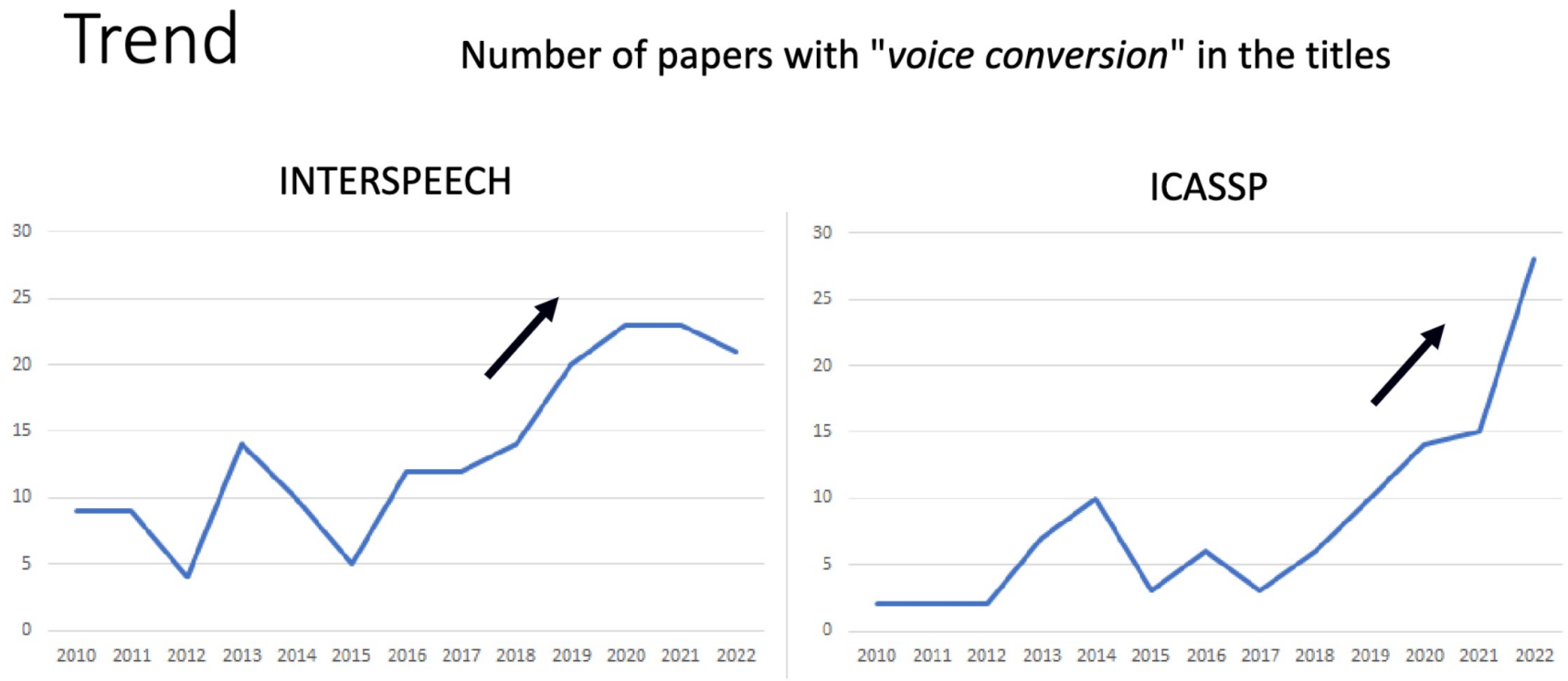
Applications

Adaptive TTS:

Leverage the existing TTS system and change the speaker information



Thriving research interest



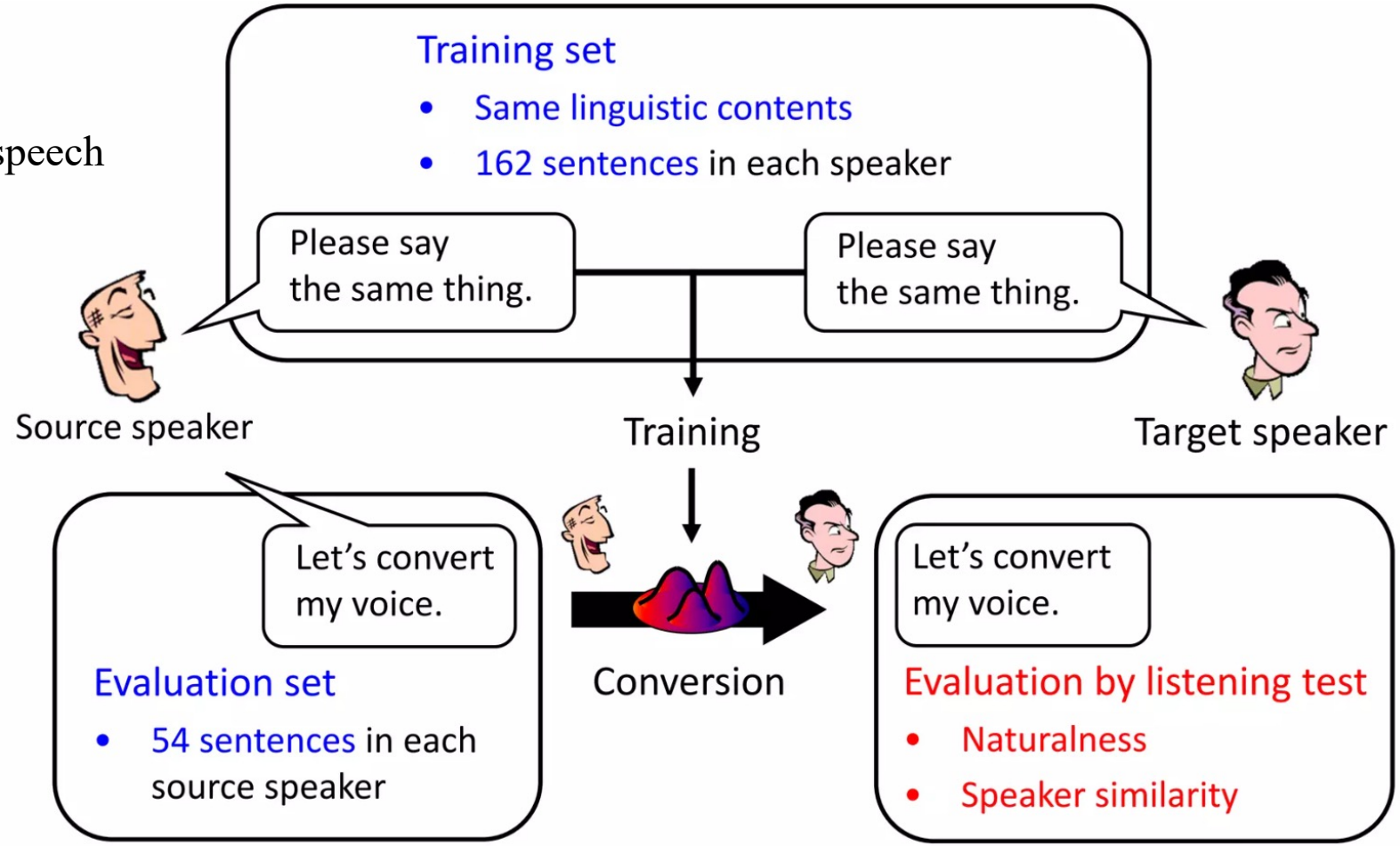
Picture from the interspeech2022 voice conversion tutorial given by Hung-yi Lee
https://github.com/tts-tutorial/interspeech2022/blob/main/INTERSPEECH_Tutorial_VC.pdf

Basics

Data available: Parallel data

VCC 2016

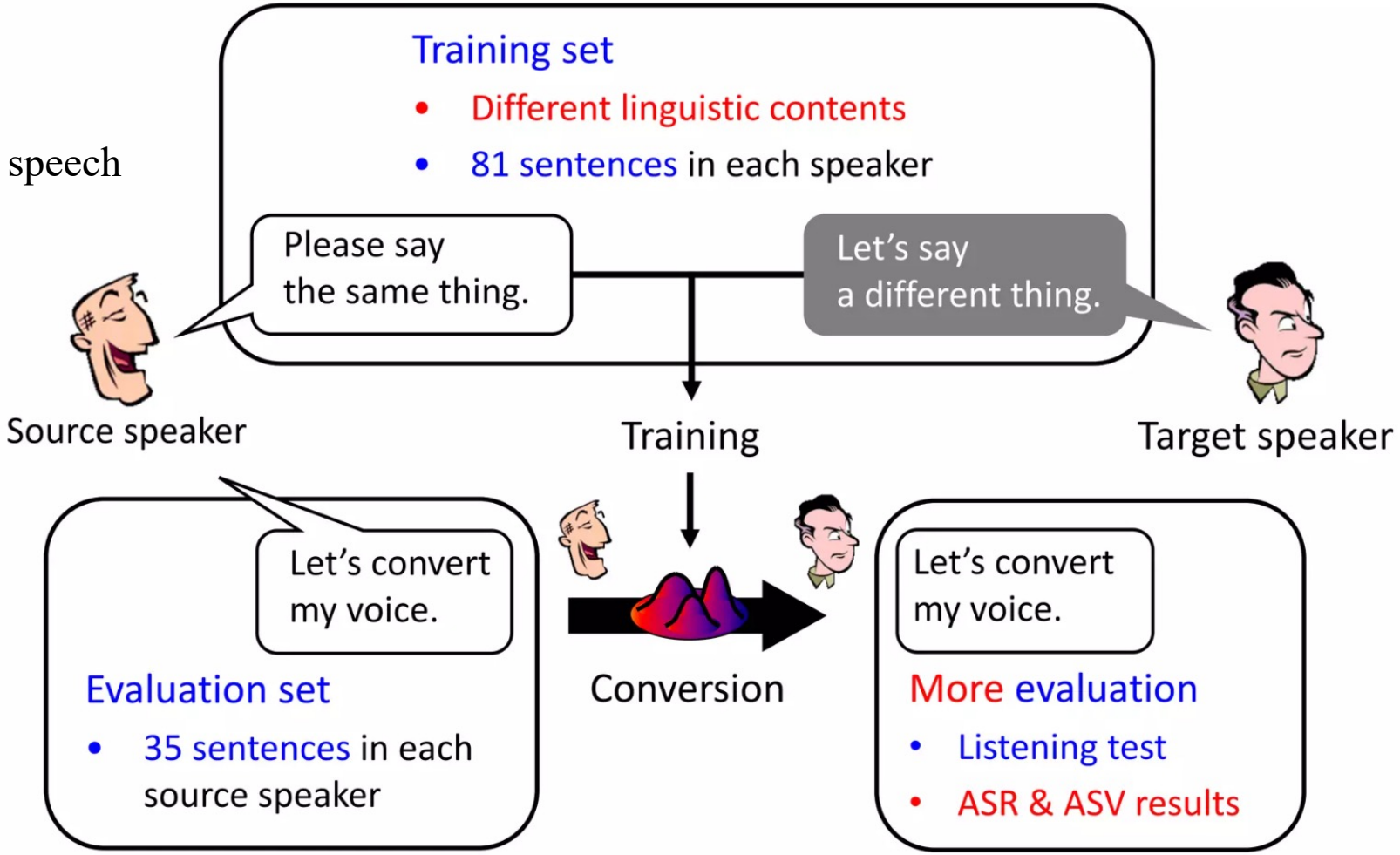
Utterance pairs of source-target speech



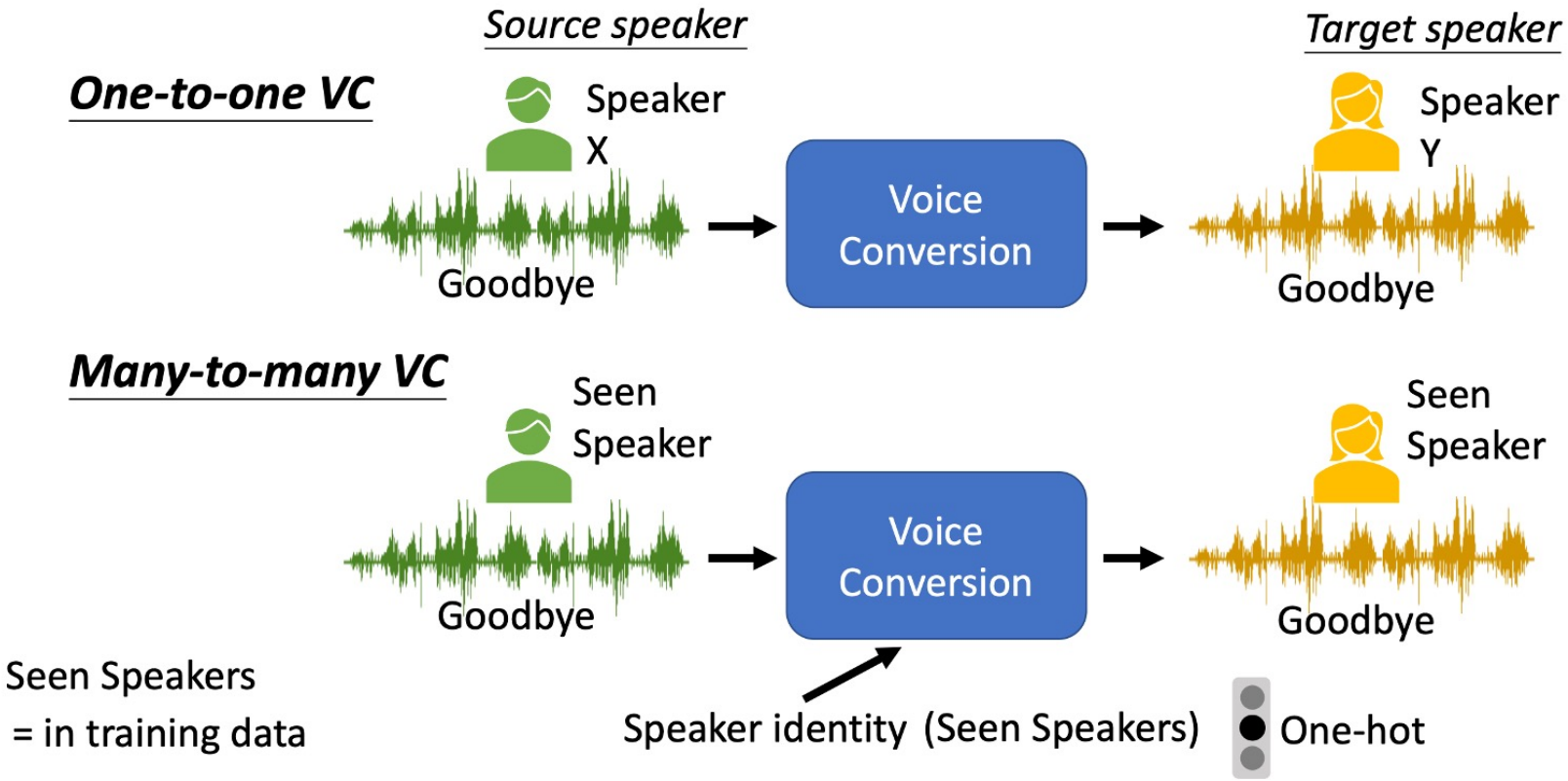
Data available: Unparallel data

VCC 2018

Arbitrary utterances of source/target speech

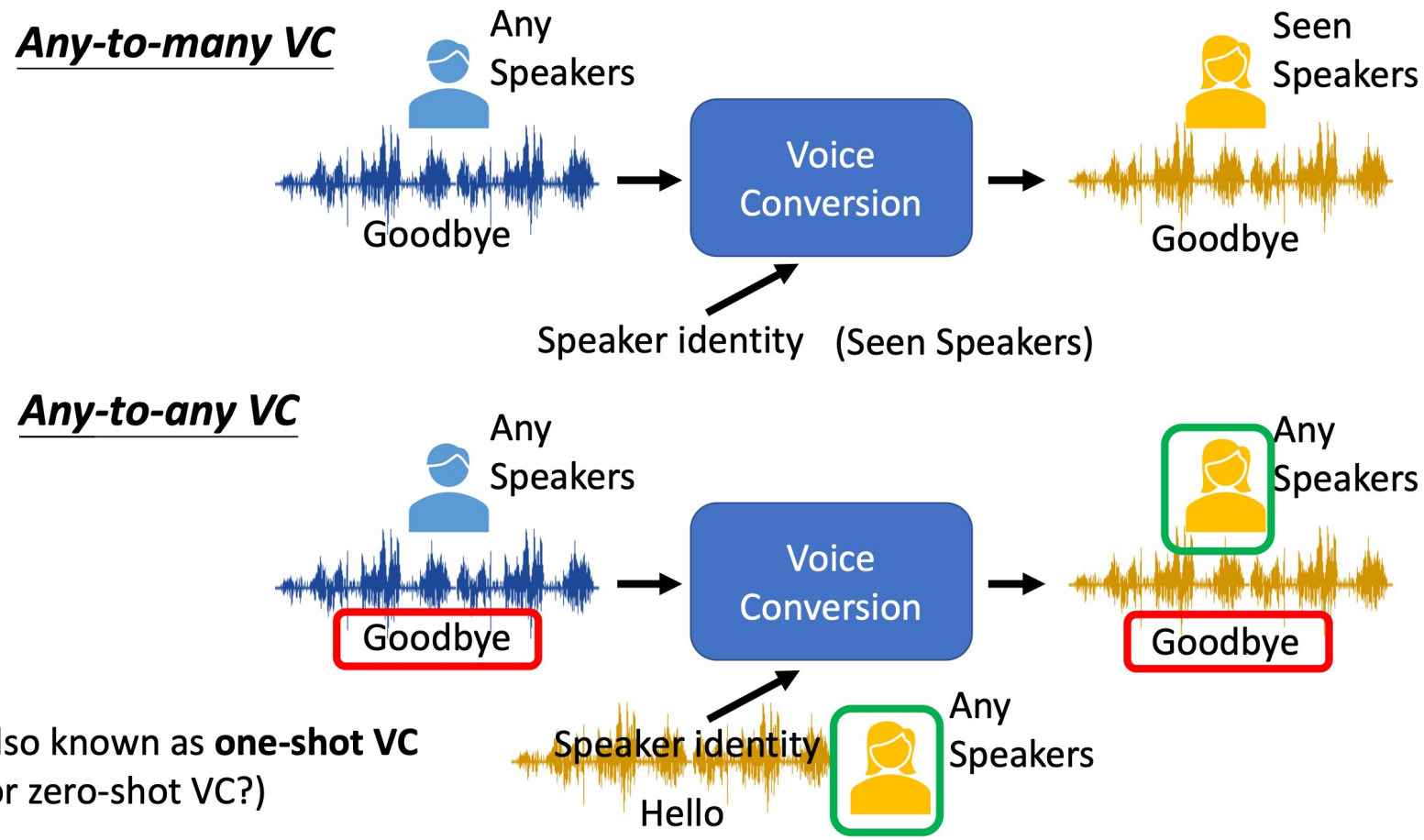


Capabilities: Input vs. Output



Picture from the interspeech2022 voice conversion tutorial given by Hung-yi Lee
https://github.com/tts-tutorial/interspeech2022/blob/main/INTERSPEECH_Tutorial_VC.pdf

Capabilities: Input vs. Output



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

Picture from the interspeech2022 voice conversion tutorial given by Hung-yi Lee
https://github.com/tts-tutorial/interspeech2022/blob/main/INTERSPEECH_Tutorial_VC.pdf

Evaluation metrics

Objective metrics

Mel-cepstral distortion (MCD)

$$D_{\mathrm{MCD}} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\frac{2}{M}\sum_{m=1}^{M}\left(\log_e(c_{1,n,m}) - \log_e(c_{2,n,m})\right)^2}$$

Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(F0_k^c - F0_k^t)^2}$$

Subjective metrics

Mean Opinion Score (MOS)

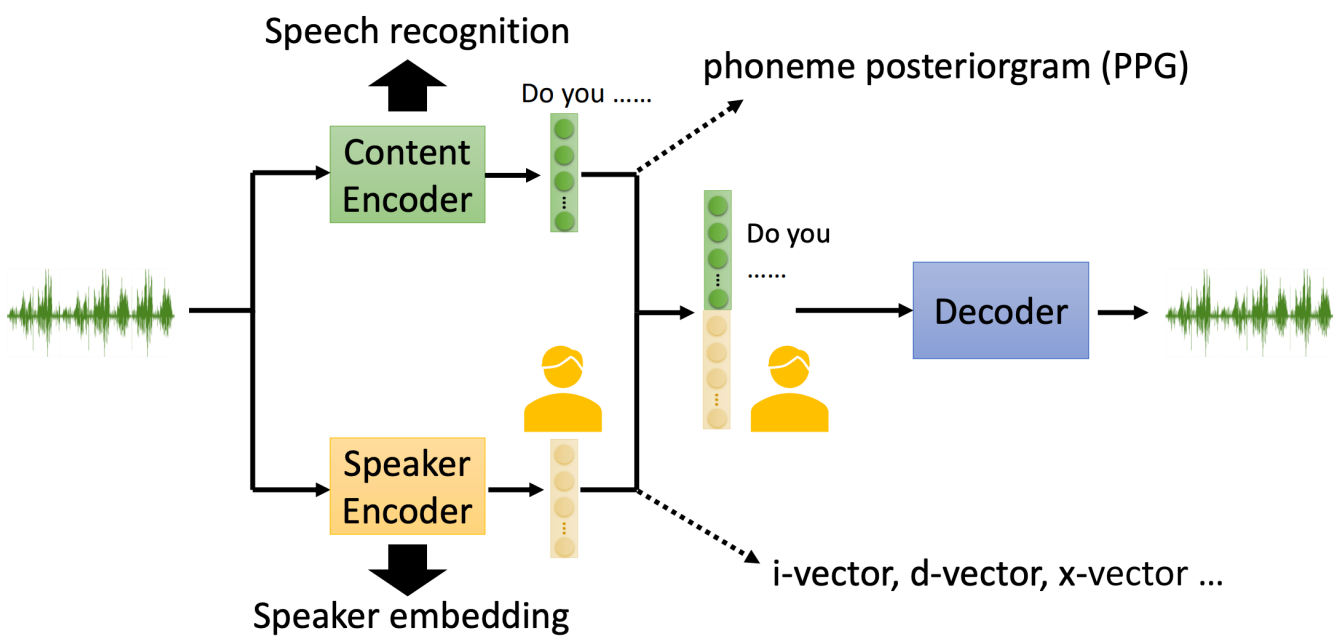
MOS Score	Description
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

ABX Test: Which one do you prefer

Pretrained model based methods

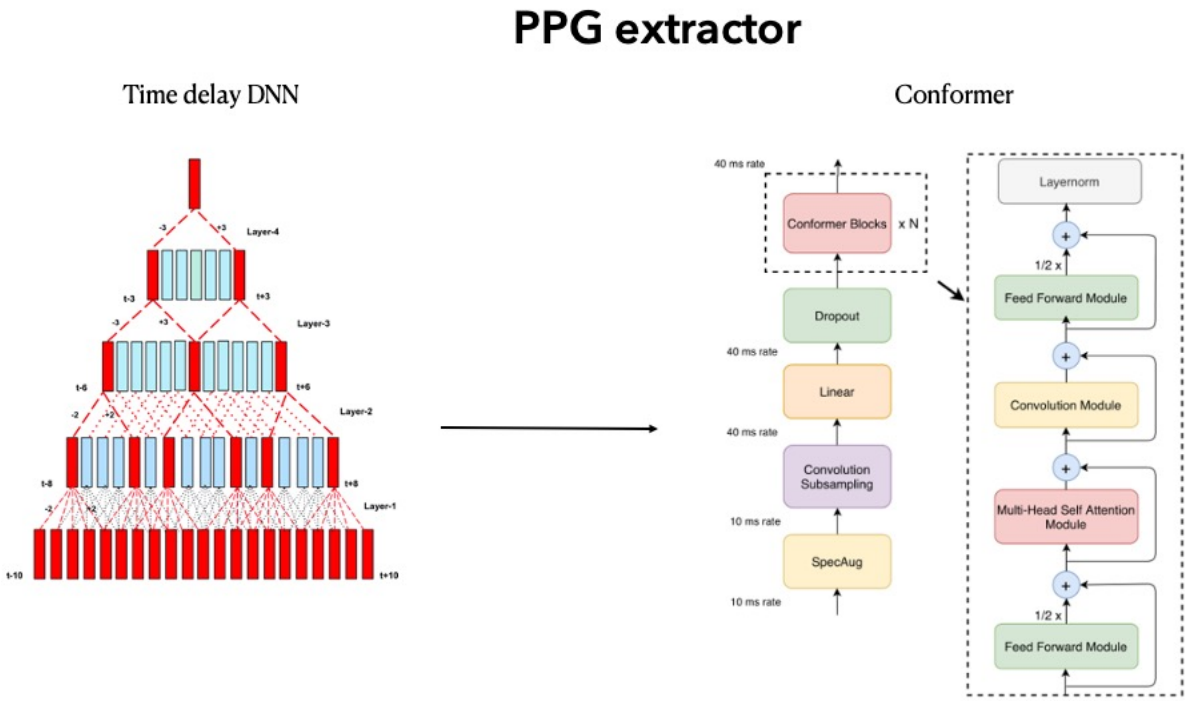
Voice conversion pipeline

- ASR based VC systems
- Frame-to-frame conversion
- Modeling
  - PPG/Bottleneck feature extraction
  - Speaker embedding extraction
  - Decoder
    - VC AM
    - Vocoder



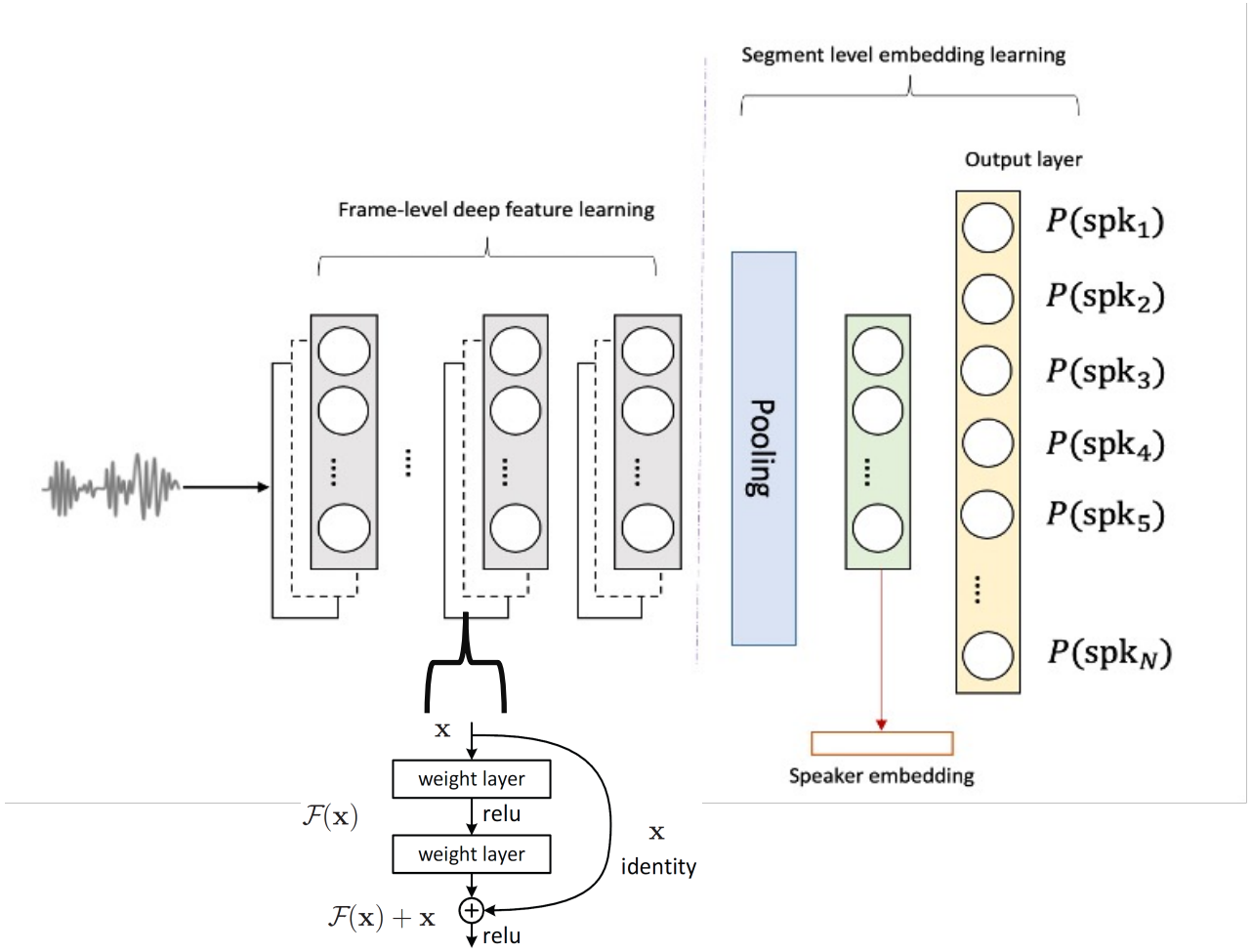
Content Embedding

- Extract content embeddings from pretrained ASR models
- Recall:
  - ASR aims to transcribe the input audio, and is expected to be robust against
    - Speaker identities
    - Environment
    - Channels
    - …
  - Perfect for content representation learning

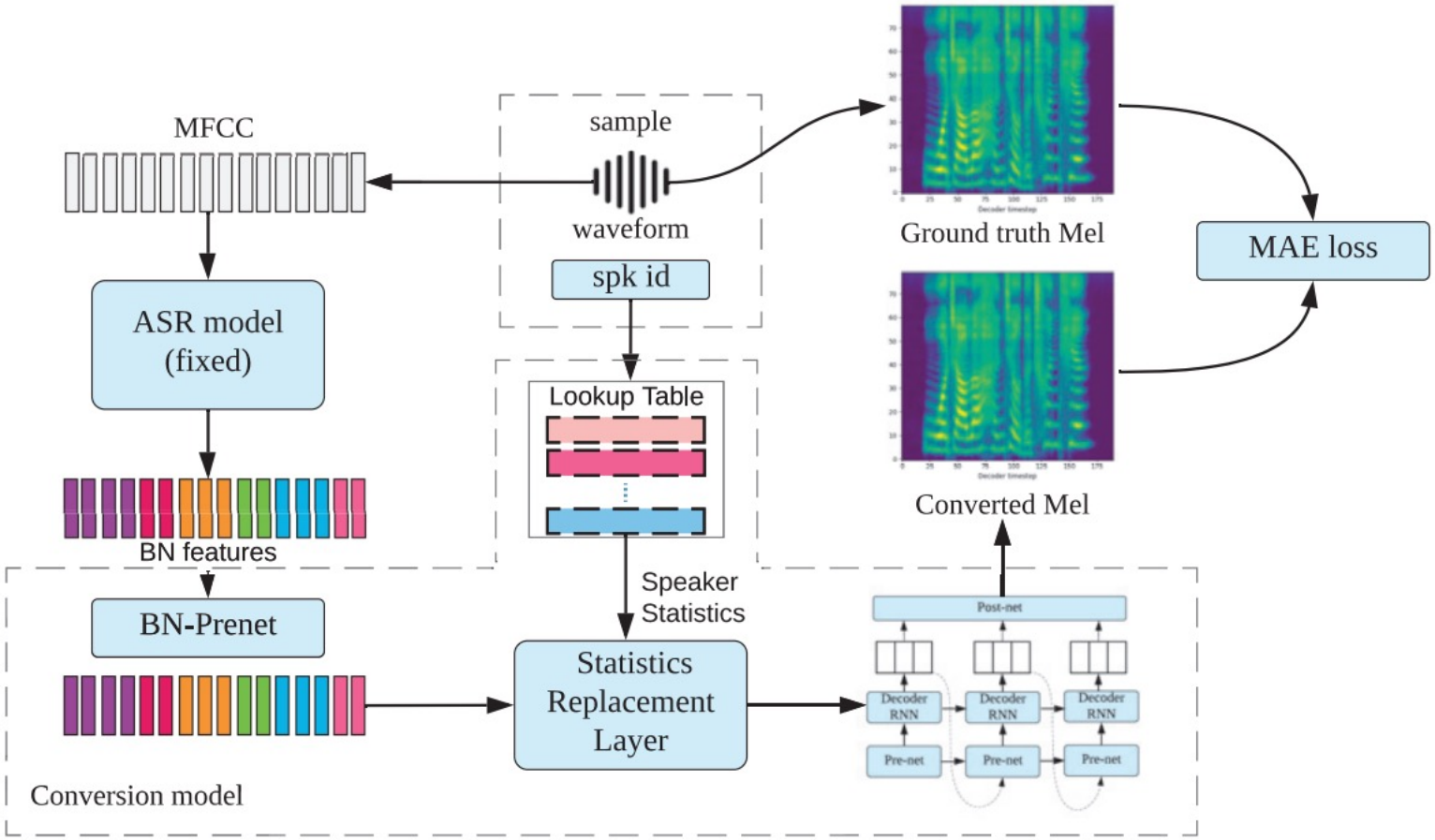


Speaker Embedding

- Extract content embeddings from pretrained speaker classification models
- Can be pretrained on a large-scale speaker classification dataset
- Segment-level representation
- Frame layers + pooling + segment layers + loss function



Acoustic Model (Optional)



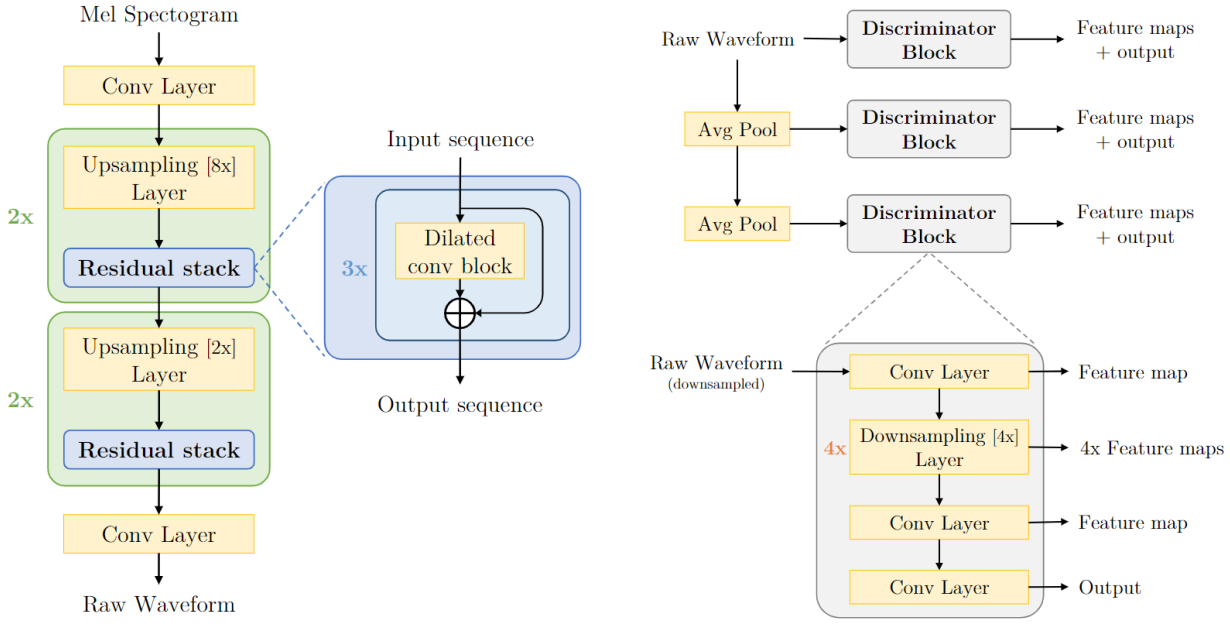
- In Text-to-speech, acoustic model performs the text-to-mel alignment and conversion
- For voice conversion:
  - No need for alignment (frame-to-frame)
  - AM aims to enhance the modeling capabilities.
  - Mapping PPG to Mel

Yufei Liu, et al. Non-Parallel Any-to-Many Voice Conversion by Replacing Speaker Statistics

Vocoder

A neural vocoder takes an acoustic feature such as mel spectrogram as input and outputs a waveform using deep learning networks

- Can be pretrained on a large dataset (only audio data is needed)
- Current dominating approach:
  - GAN based vocoder
  - Fast adaptation
  - High-quality
  - Fast inference: Non-autoregressive



(a) Generator (b) Discriminator

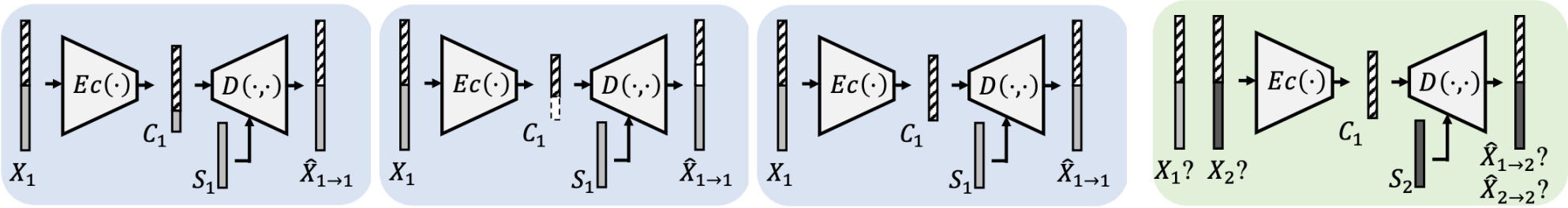
Kumar, Kundan, et al. "Melgan: Generative adversarial networks for conditional waveform synthesis."

End-to-end methods (self-disentangle)

End-to-end systems

Learn the disentanglement

AutoVC: Carefully design the bottleneck



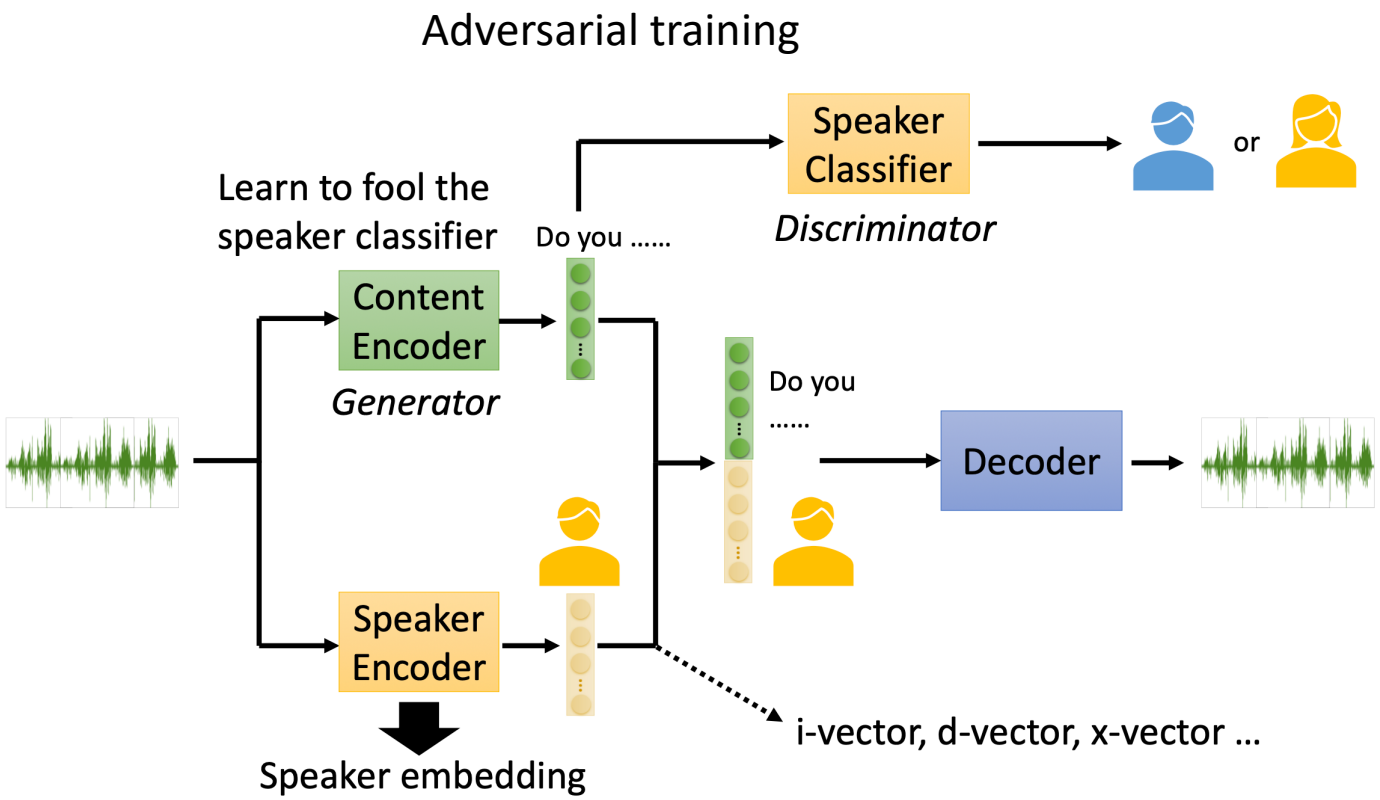
(a) Bottleneck too wide (b) Bottleneck too narrow (c) Bottleneck just right (d) Conversion

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

Qian, Kaizhi, et al. "Autovc: Zero-shot voice style transfer with only autoencoder loss." ICML 2019

End-to-end systems

Learn the disentanglement

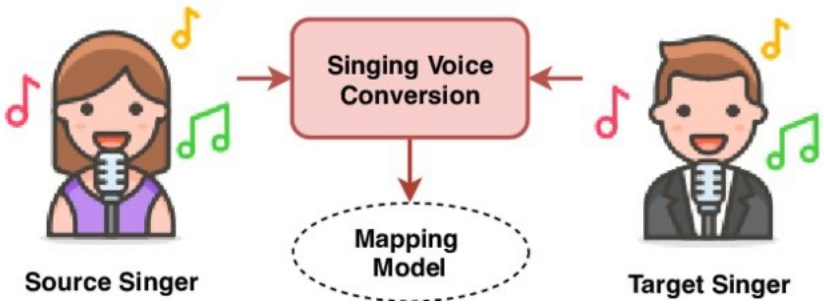


Qian, Kaizhi, et al. "Autovc: Zero-shot voice style transfer with only autoencoder loss." ICML 2019
Chin-Cheng Hsu, et al. "Voice Conversion from Non-parallel Corpora Using Variational Auto-encoder." APSIPA, 2016

Beyond common voice conversion

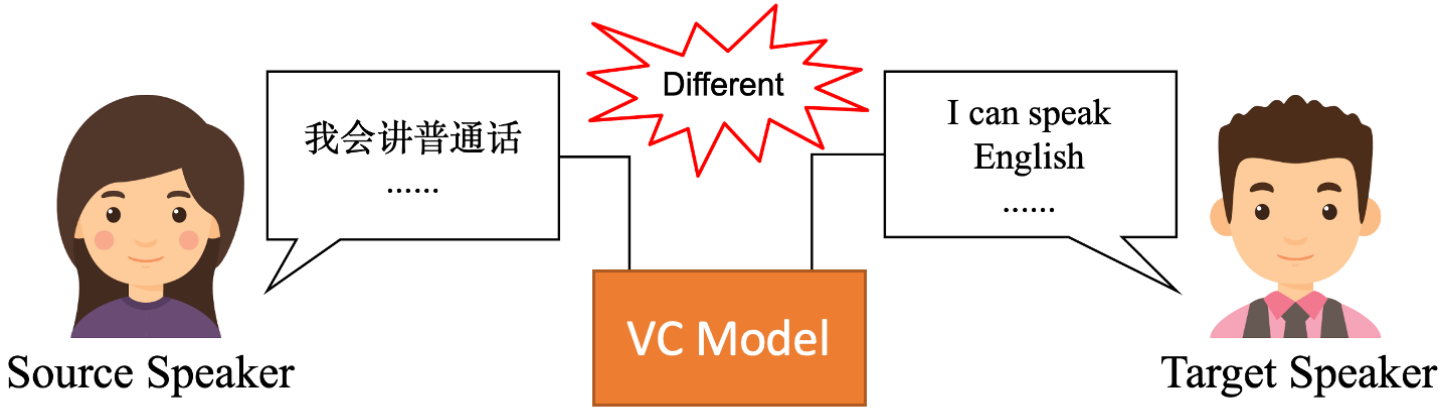
Singing voice conversion

- Prosody needs explicit modelling
- Vocoder needs improvement for singing voice modeling
  - Usually accepts pitch as extra information
- The problem of cross-gender conversion (large pitch shift)



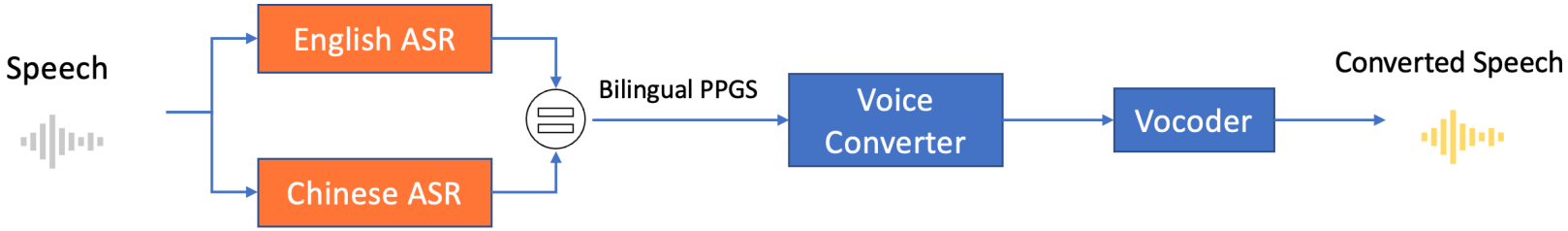
Cross-lingual voice conversion

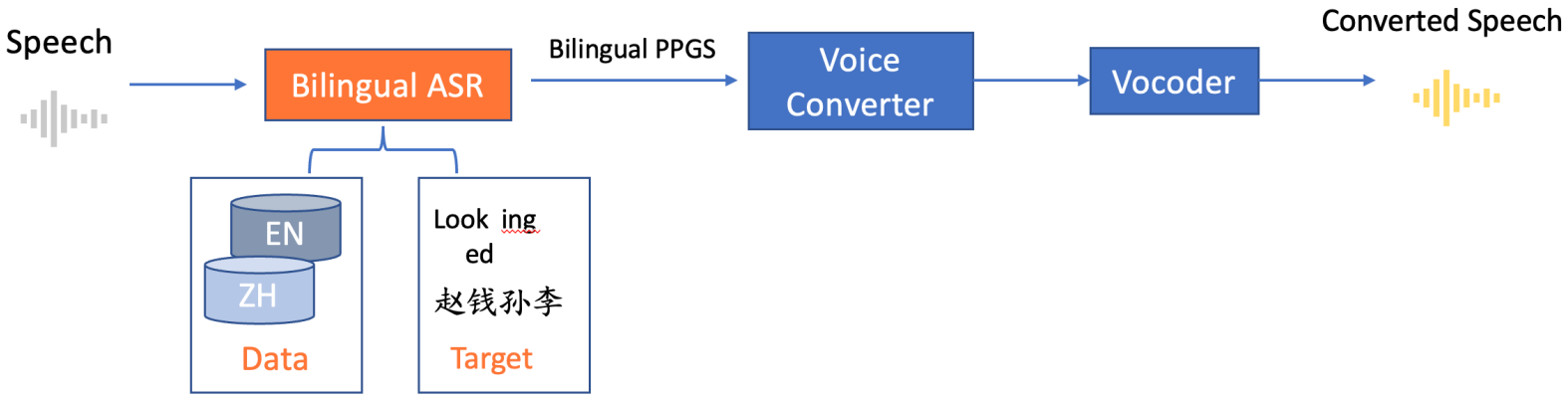
- Problem: Accent
- Solution: Multi-lingual content modeling
  - Multi-lingual ASR



Cross-lingual voice conversion

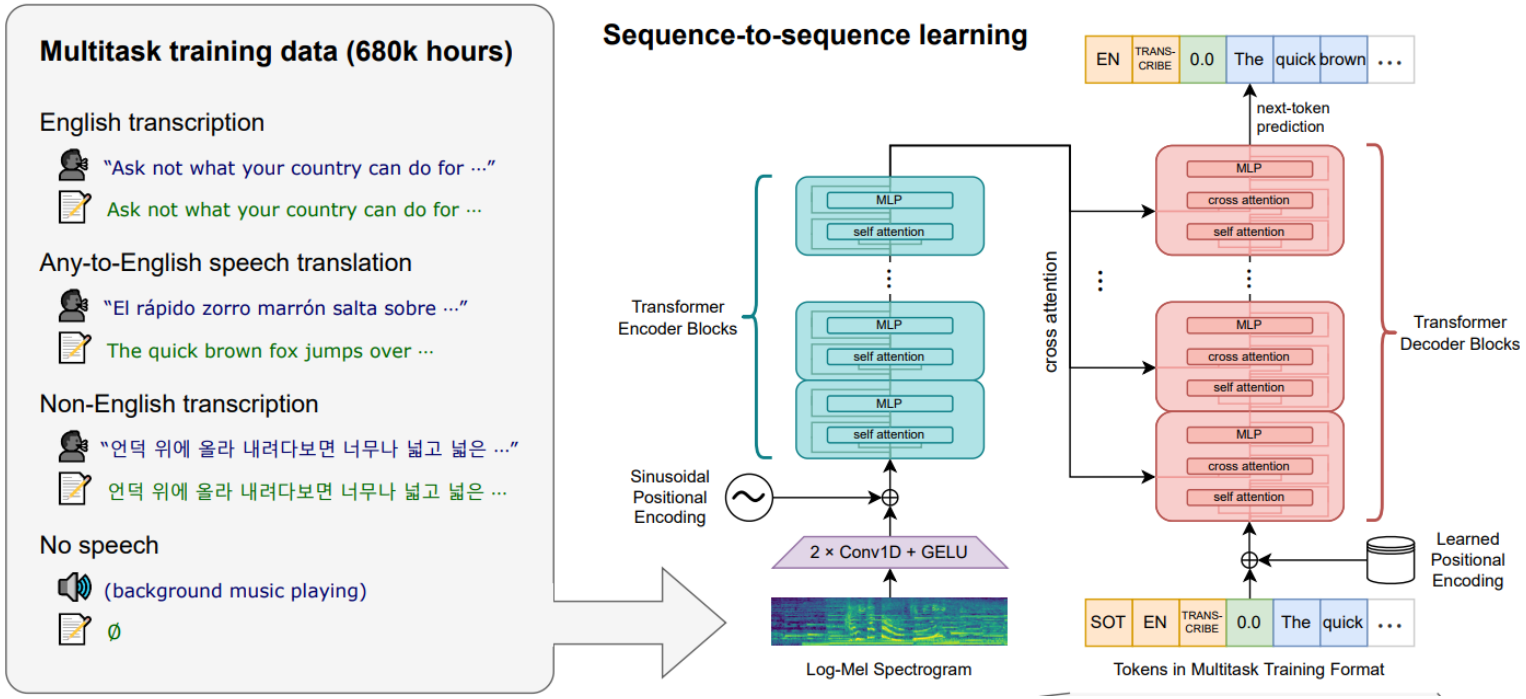
Cross-lingual Voice Conversion





Cross-lingual voice conversion

Whisper from OpenAI



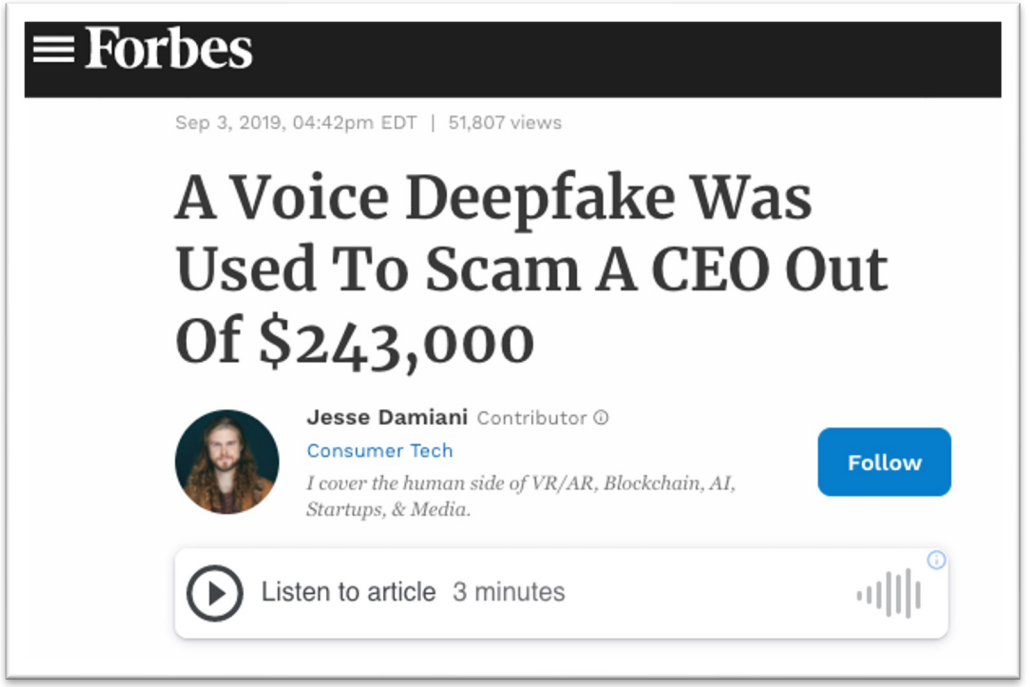
Real-time Streaming Voice Conversion

- Live broadcasting
- Real-time communication (RTC)

- Challenges:
  - Extreme low latency
  - Streaming mode leads to inaccurate modeling (short context, no future information)

Risk of Voice Conversion

- The possibility of the misusage for spoofing
  - VC makes it possible for someone to speaker in your voice
- What can we do?
  - Anti-spoofing!
  - Attracting growing interest along with the development of speech generation techniques


Forbes

Sep 3, 2019, 04:42pm EDT | 51,807 views

A Voice Deepfake Was Used To Scam A CEO Out Of $243,000

Jesse Damiani Contributor

Consumer Tech

I cover the human side of VR/AR, Blockchain, AI, Startups, & Media.

Follow

Listen to article 3 minutes


THE WALL STREET JOURNAL.

PRO CYBER NEWS

Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case

Scams using artificial intelligence are a new challenge for companies

Appendix

VCC: Voice Conversion Challenge

http://www.vc-challenge.org/

2016 — 1st VCC (VCC2016)
- Parallel training

2017

2018 — 2nd VCC (VCC2018)
- Parallel training
- Non-parallel training

2019

2020 — 3rd VCC (VCC2020)
- Semi-parallel training
- Non-parallel training across different languages (cross-lingual VC)

SINGING VOICE CONVERSION CHALLENGE

Singing Voice Conversion Challenge 2023

VCC2023: SVC

ASVspoof: Detecting the synthesized speech

https://www.asvspoof.org/

ASVspoof

Automatic Speaker Verification and Spoofing Countermeasures Challenge

ASVspoof5

We Need You!

Call For Spoofed/Speech DeepFake Data Contributors

if you are interested in becoming a contributor, send an email to info@asvspoof.org

Focus on VC and TTS DeepFake detection

- x formal definition
- x benchmark
- x database

- ✓ Task definition
- ✓ Metrics
- ✓ Database

- ✓ More unseen attacks
- ✓ Channel mismatch (codec & trans.)

2015 — ASVspoof 2015

2017

2019 — ASVspoof 2019

2021 — ASVspoof 2021

Practice: Build a VC system

- PPG：
  - Wenet: https://github.com/wenet-e2e/wenet
  - Whisper: https://github.com/openai/whisper
- Speaker embedding
  - Wespeaker: https://github.com/wenet-e2e/wespeaker
- Vocoder
  - Hifi-gan: https://github.com/jik876/hifi-gan

Q & A