Setting - 2 Task Setting - 1

Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations

Method

Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee, Lin-shan Lee

Best student paper award nominated in Interspeech 2018.

Speech Processing Laboratory,
National Taiwan University

Outline

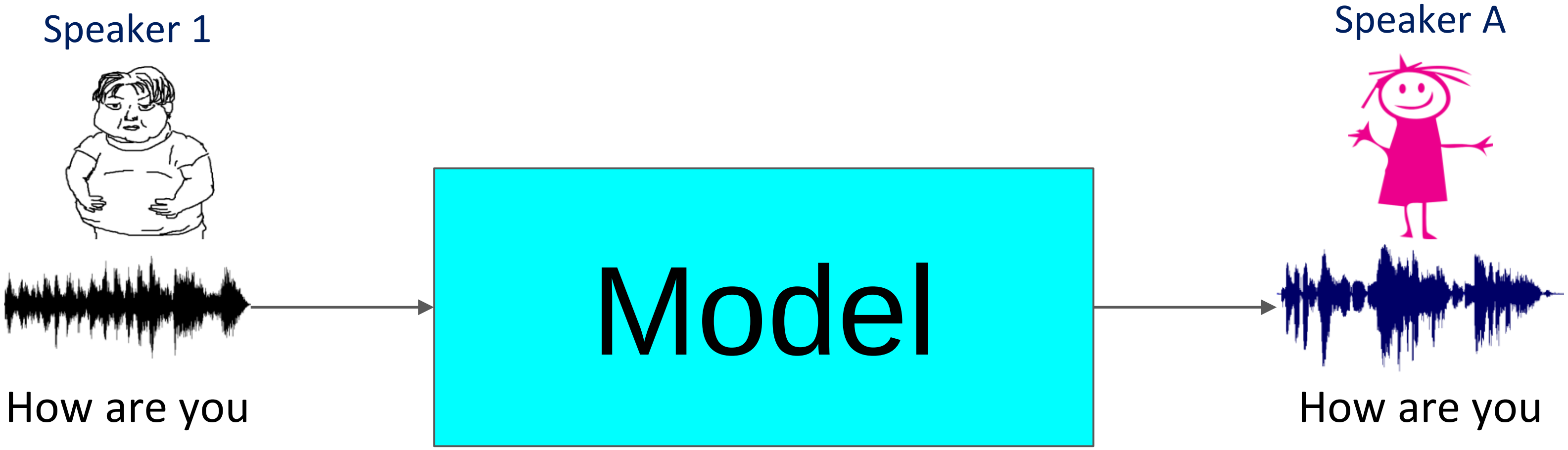
- Introduction
  - Convertional: supervised with paired data
  - This work: unsupervised with non-parallel data
  - This work: multi-target with non-parallel data
- Multi-target scenario (our contribution)
  - Model
  - Experiments

Outline

- Introduction
  - Convertional: supervised with paired data
  - This work: unsupervised with non-parallel data
  - This work: multi-target with non-parallel data
- Multi-target scenario (our contribution)
  - Model
  - Experiments

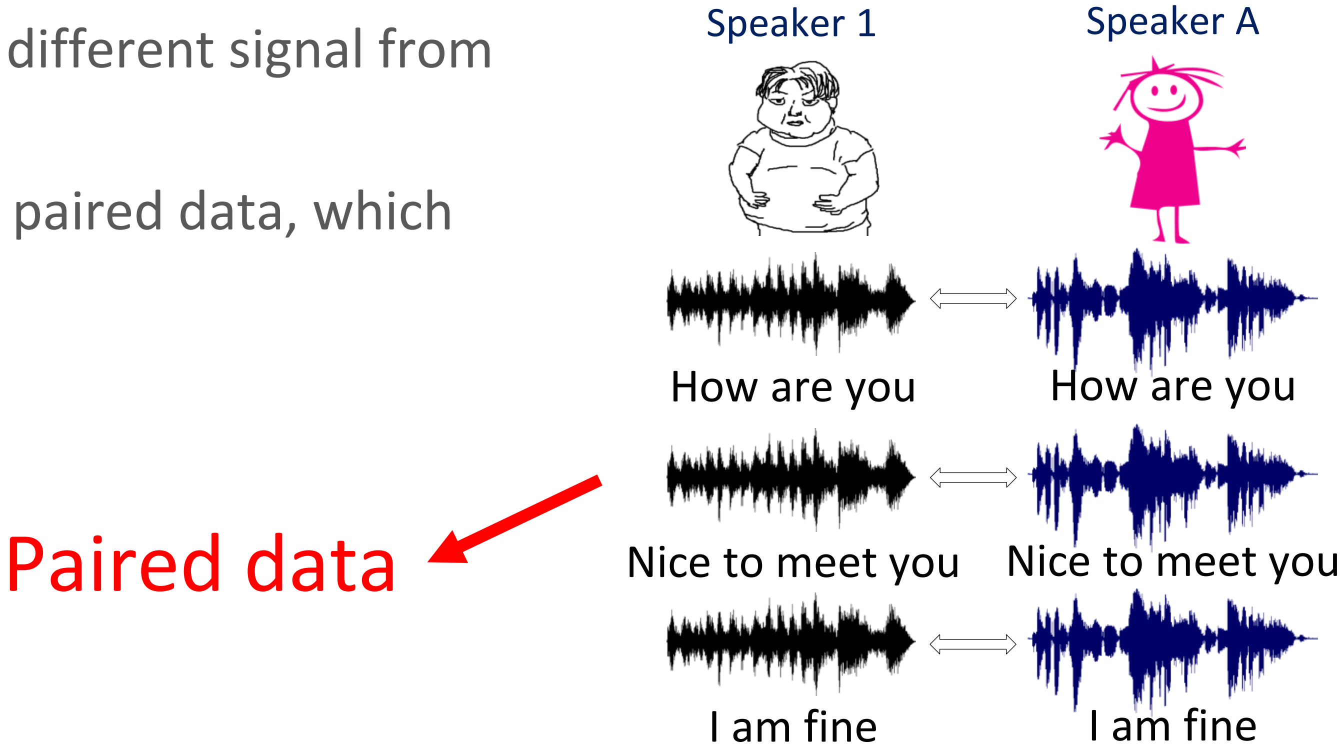
Voice conversion

- Change the **characteristic** of an utterance while maintaining the linguistic content the same.
- Characteristic: accent, speaker identity, emotion...
- This work: focus on speaker identity conversion.



Conventional: supervised with paired data

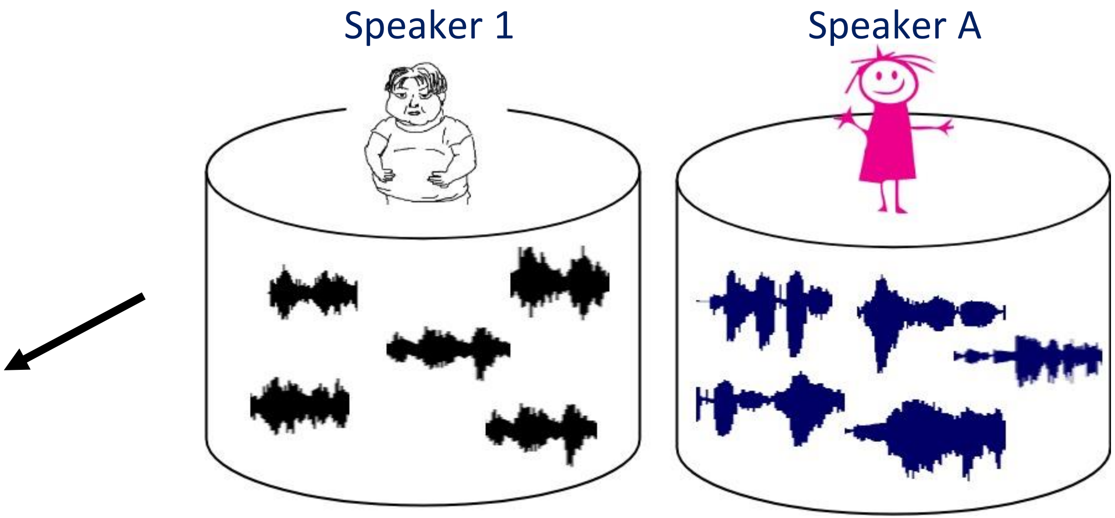
- Same sentences, different signal from 2 speakers.
- Problem: require paired data, which is hard to collect.



This work: unsupervised with non-parallel data

- Trained on non-parallel corpus, which is more attainable.
- Actively investigated.
- Prior work: utilize deep generative model, ex. VAE, GAN, cycleGAN [1].

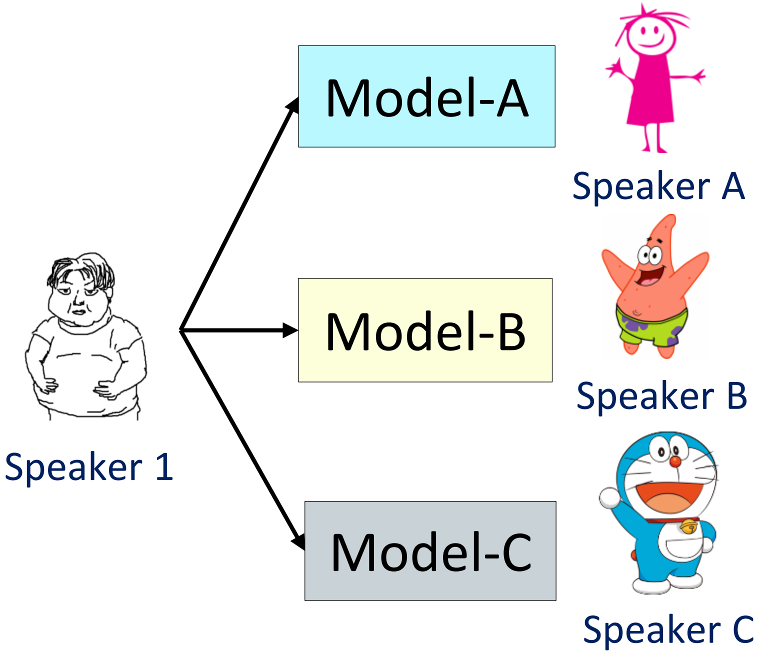
Don't have to speak same sentences.



CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks. Kaneko et.al. EUSIPCO 2018 [1]

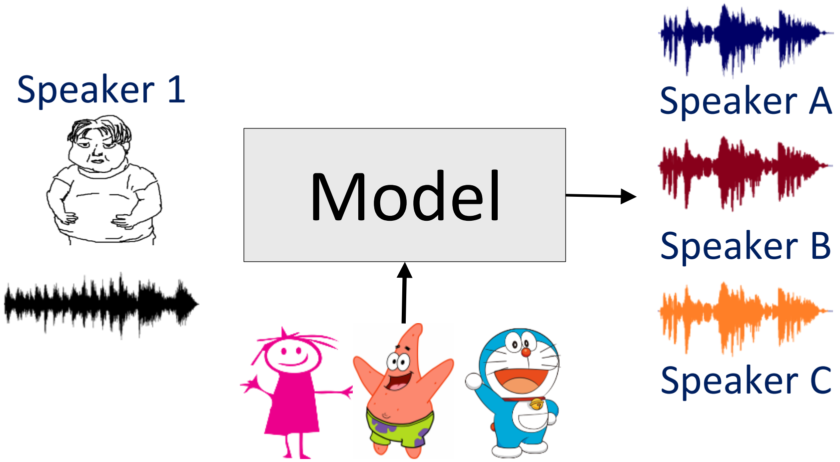
This work: multi-target unsupervised with non-parallel data

3 models are needed for 3 target speakers.



N^2 models for N speakers.

Only one model is needed.

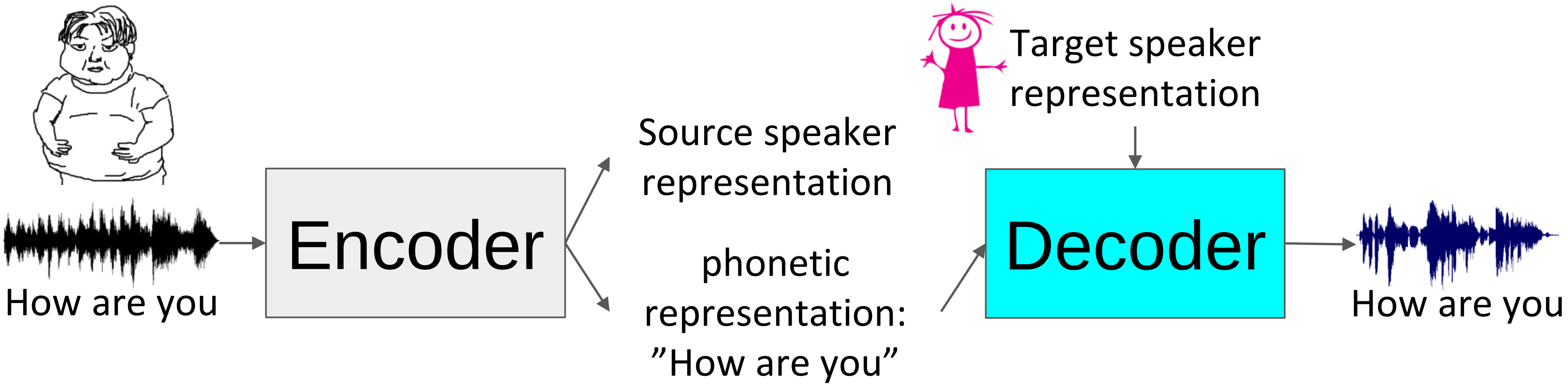


Outline

- Introduction
  - Convertional: supervised with paired data
  - This: unsupervised with non-parallel data
  - This: multi-target with non-parallel data
- Multi-target scenario (our contribution)
  - Model
  - Experiments

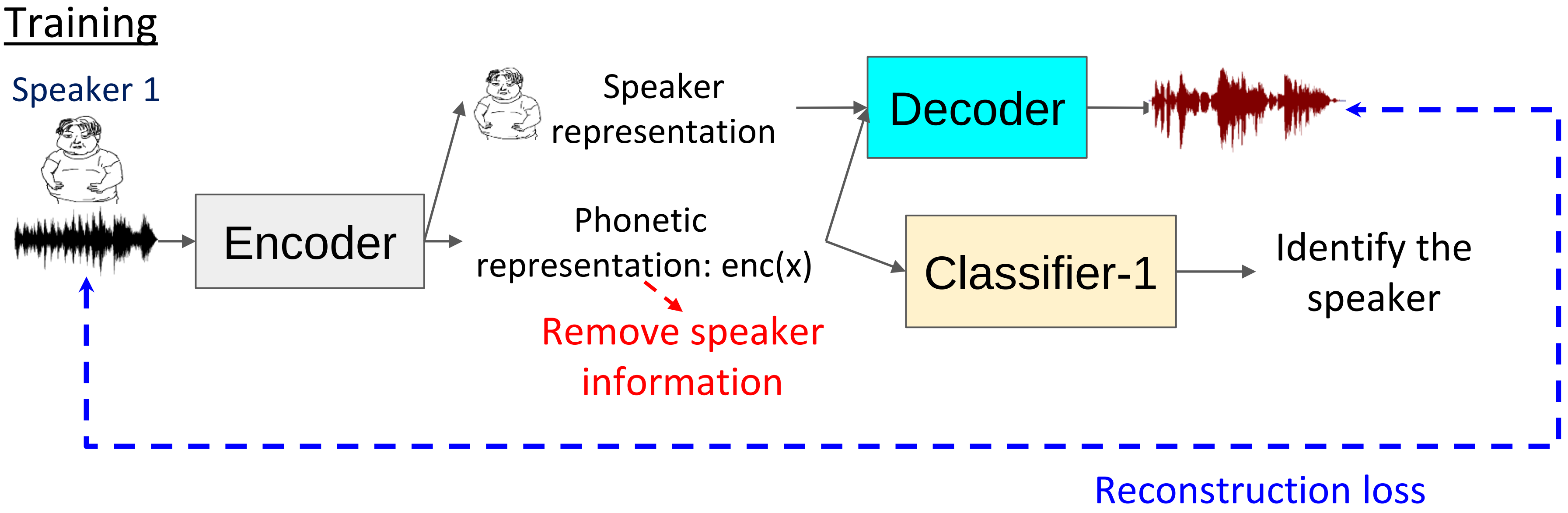
Multi-target Scenario (main contribution)

- Intuition: speech signals inherently carry both phonetic and speaker information.
- Learn the phonetic/speaker representation separately.
- Synthesize the target voice by combining the source phonetic representation and target speaker representation.



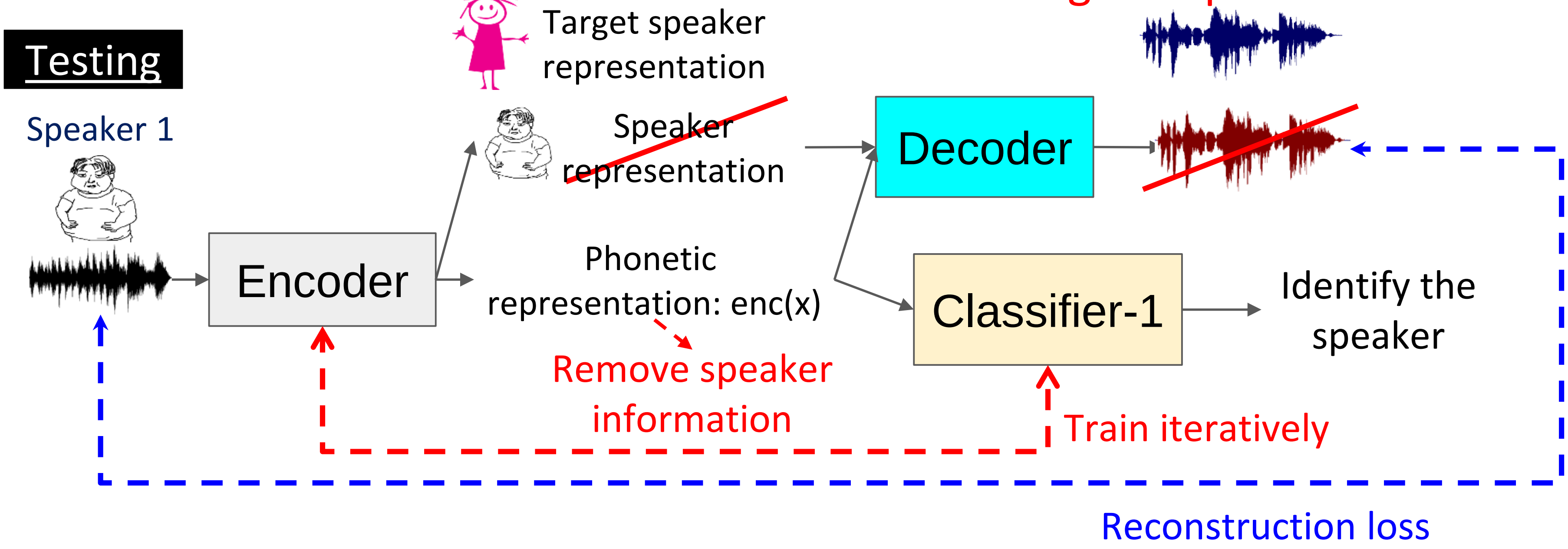
Stage 1: disentanglement between phonetic and speaker representation

- Goal of classifier-1: maximize the likelihood being the speaker.



Stage 1: disentanglement between phonetic and speaker representation

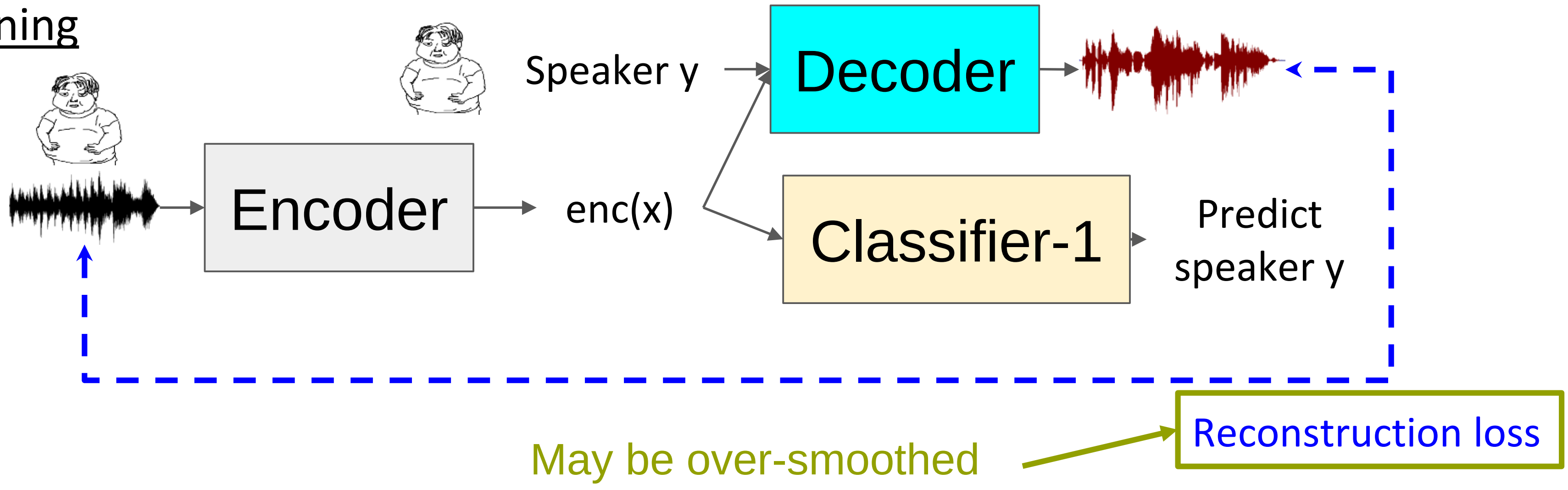
- Goal of classifier-1: maximize the likelihood being the speaker.
- Goal of encoder: minimize the likelihood being the speaker.



Problem of stage 1: over-smoothed spectra

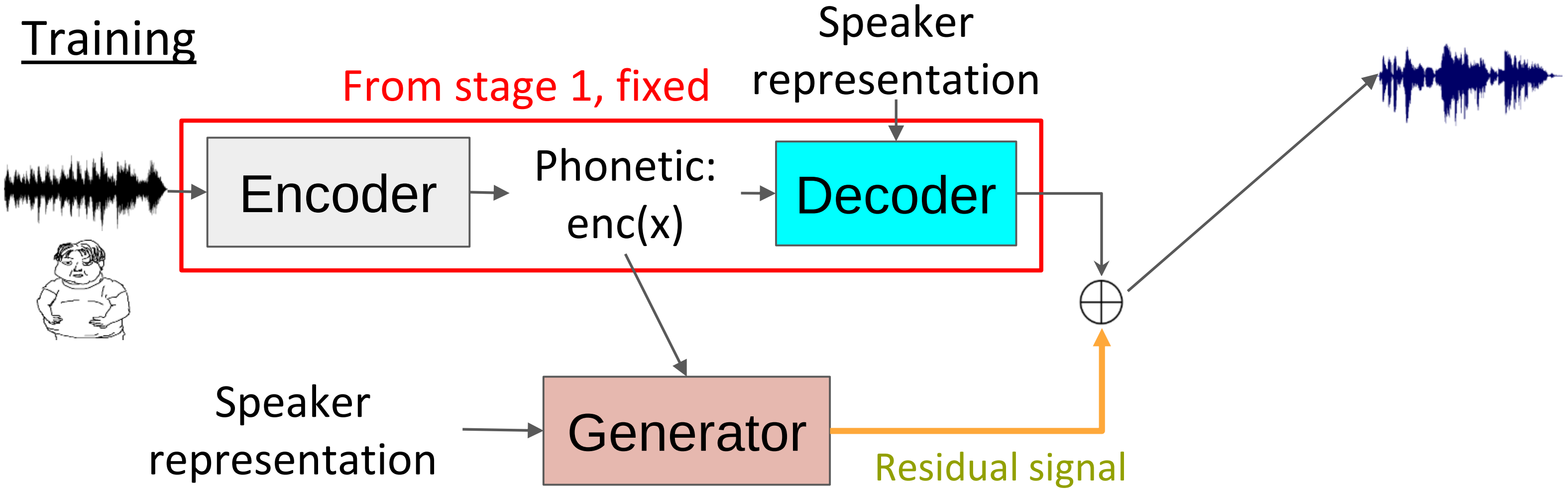
- Stage 1 alone can synthesis target voice to some extent.
- Reconstruction loss encourages the model to generate average value of the target. Leads to over-smoothed spectra, and result in buzzy synthesized speech.

Training



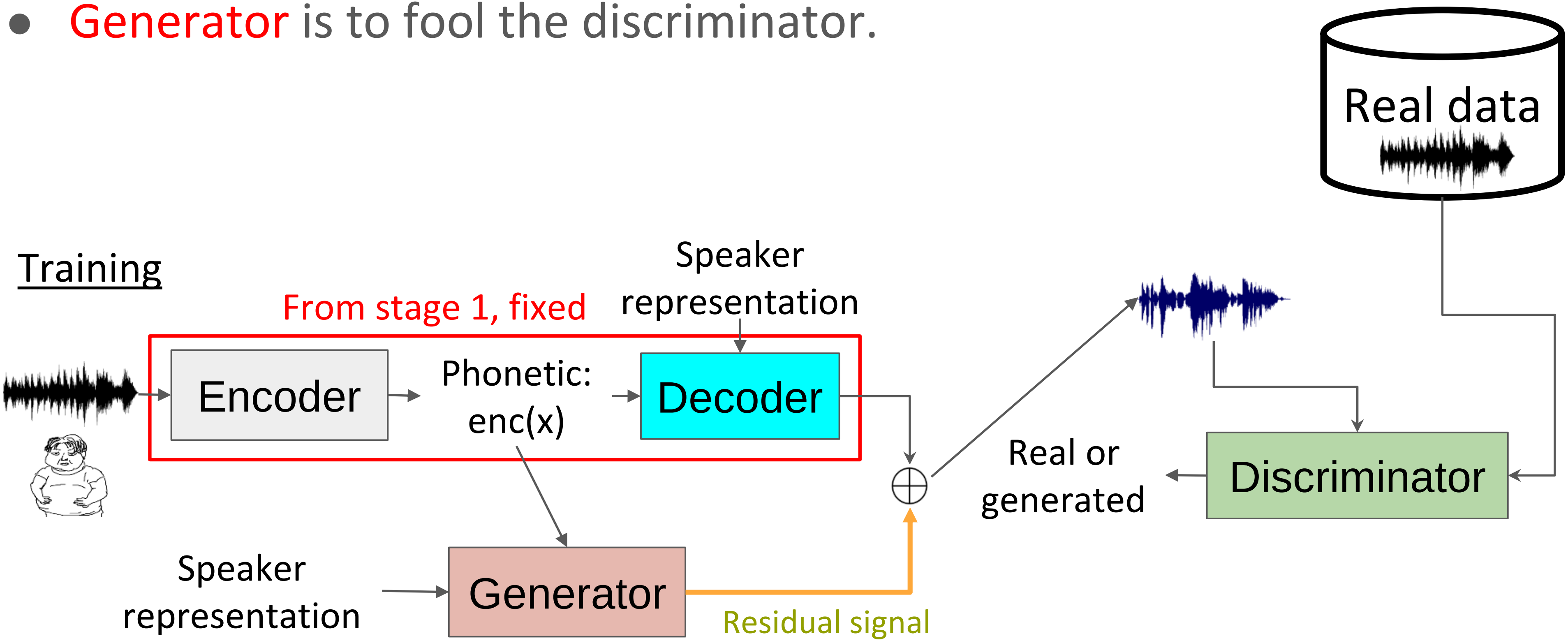
Stage 2: patch the output with a residual signal

- Train another generator to produce residual signal, making the output more natural.



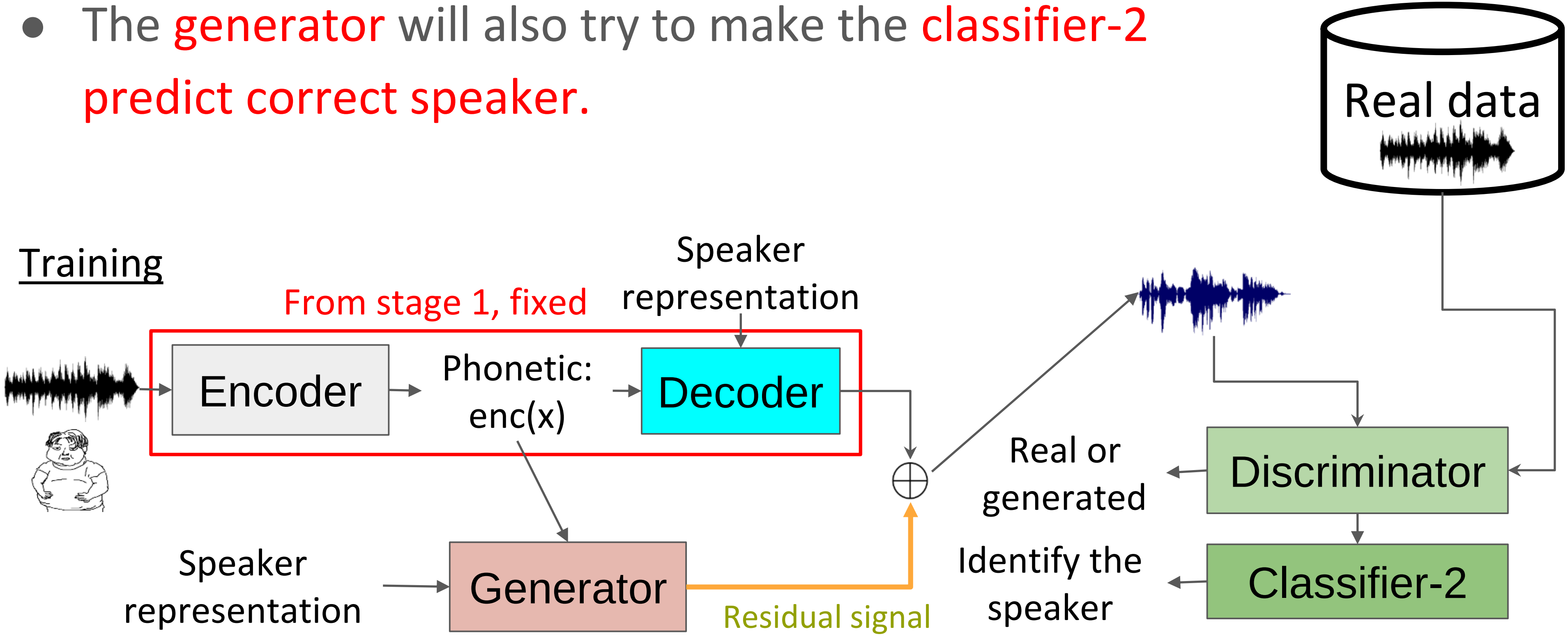
Stage 2: patch the output with a residual signal

- Discriminator is to discriminate whether synthesized or real data.
- Generator is to fool the discriminator.



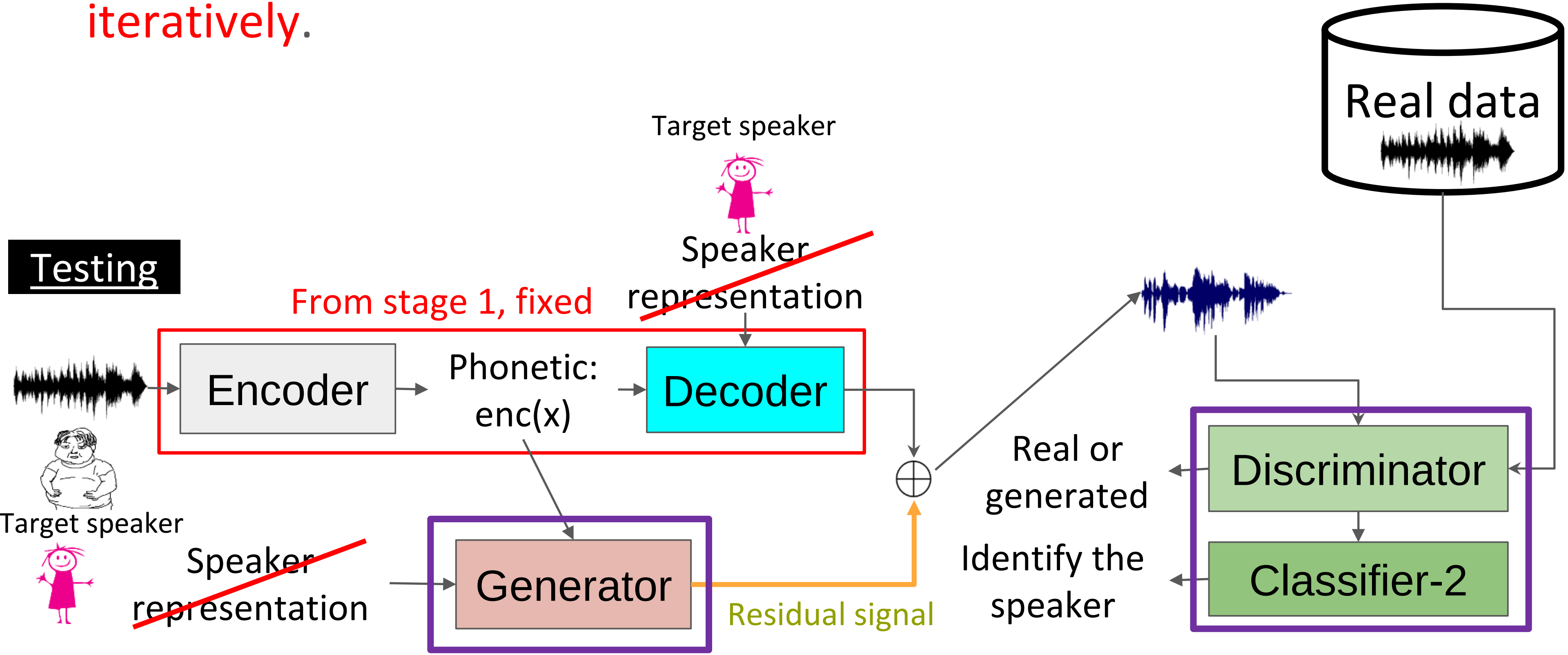
Stage 2: patch the output with a residual signal

- Classifier-2 is to identify the speaker.
- The generator will also try to make the classifier-2 predict correct speaker.



Stage 2: patch the output with a residual signal

- Generator and discriminator/classifier-2 are trained iteratively.

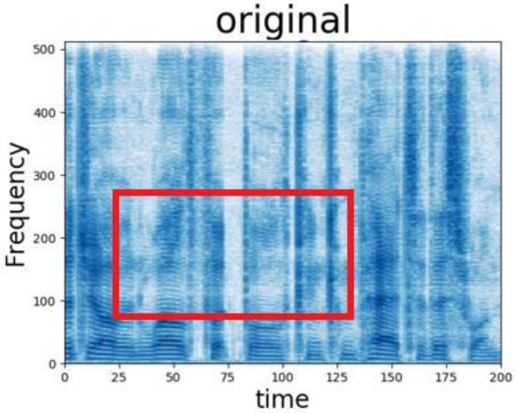


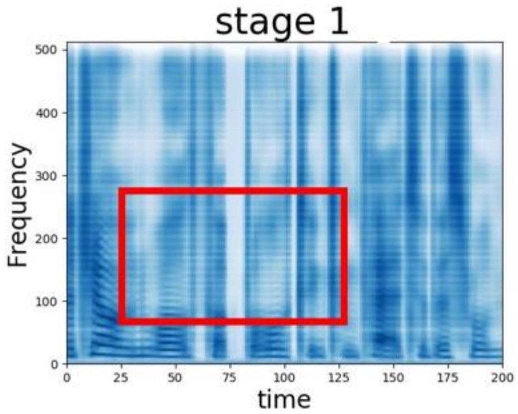
Experiments - setting

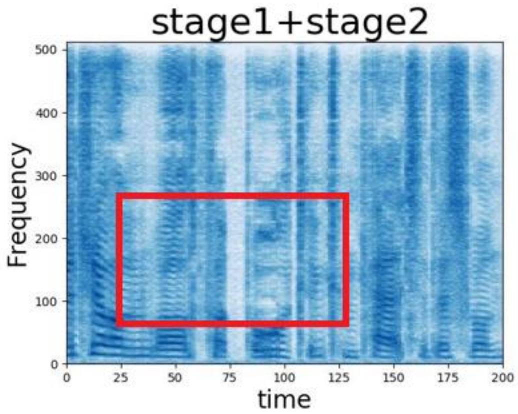
- Feature: Short-time Fourier Transform (STFT) spectrograms.
- Corpus: 20 speakers from CSTR VCTK Corpus (for TTS). 90% training, 10% testing.
- Vocoder: Griffin-Lim (non-parametric method).

Experiments – spectrogram visualization

- Is stage 2 helpful?
- Sharpness of the spectrogram is improved by stage 2.



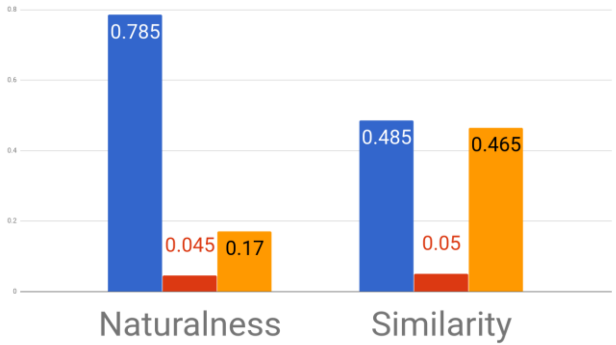




Experiments – subjective preference

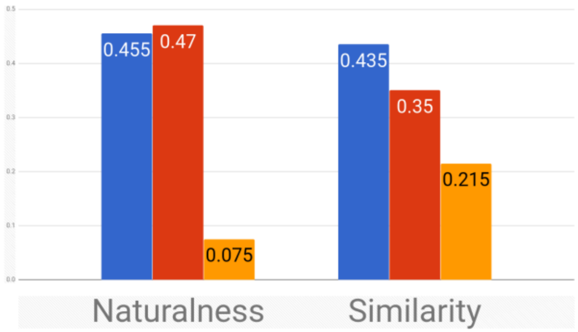
- Ask users to choose their preference in terms of naturalness and similarity.
- Stage 2 improved.
- Comparable to baseline approach.

Is stage 2 helpful?



“Stage 1 + stage 2” is better.
“Stage 1 alone” is better.
Indistinguishable.

Comparison to baseline [1].



“Stage 1 + stage 2” is better.
“CycleGAN-VC” [1] is better.
Indistinguishable.

CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks. Kaneko et.al. EUSIPCO 2018 [1]

Demo

Male to Female

Source: Target: Converted:

Female to Female

Source: Target: Converted:

Advisor(male, never seen in training data) to Female

Source: Target: Converted:

https://jjery2243542.github.io/voice_conversion_demo/

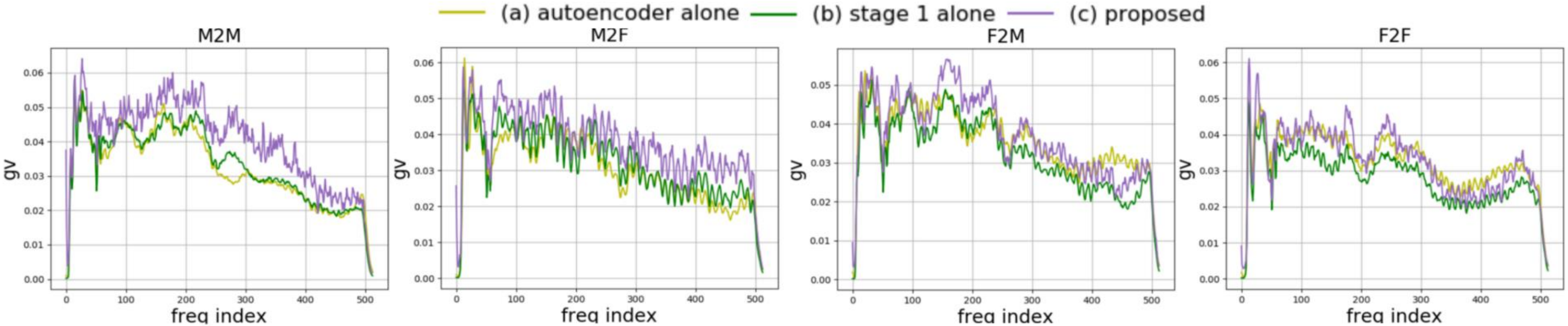
Conclusion

- A multi-target unsupervised approach for VC is proposed.
- Stage 1: disentanglement between phonetic and speaker representation.
- Stage 2: patch the output with residual signal to generate more natural speech.

Thanks for listening

Experiments – sharpness evaluation

- Speech signals have diversified distribution => high variance.
- Model with stage 2 training have highest variance.



Network architecture

- CNN + DNN + RNN
- Recurrent layer to generate varied length output.
- Dropout after each layer to provide noise for GAN-training.

Encoder	
conv-bank block	Conv1d-bank-8, LReLU, IN
conv block $\times$ 3	C-512-5, LReLU C-512-5, stride=2, LReLU, IN, Res
dense block $\times$ 4	FC-512, IN, Res
recurrent layer	bi-directional GRU-512
combine layer	recurrent output + dense output

Decoder/Generator	
conv block $\times$ 3	$\mathrm{emb}_l(y)$, C-1024-3, LReLU, PS C-512-3, LReLU, IN, Res
dense block $\times$ 4	$\mathrm{emb}_l(y)$, FC-512, IN, Res
recurrent layer	$\mathrm{emb}_l(y)$, bi-directional GRU-256
combine layer	recurrent output + dense output

Classifier-1	
conv block $\times$ 4	C-512-5, LReLU C-512-5, IN, Res
softmax layer	FC-$N_{speaker}$

Discriminator	
conv block $\times$ 5	C-K-5, stride=2, LReLU, IN
conv layer	C-32-1, LReLU, IN
output layer	scalar output, FC-$N_{speaker}$(classifier-2)

Problem - training-testing mismatch