One-shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization

Ju-Chieh Chou, *Hung-yi Lee, Interspeech 2019.*

National Taiwan University

Outline

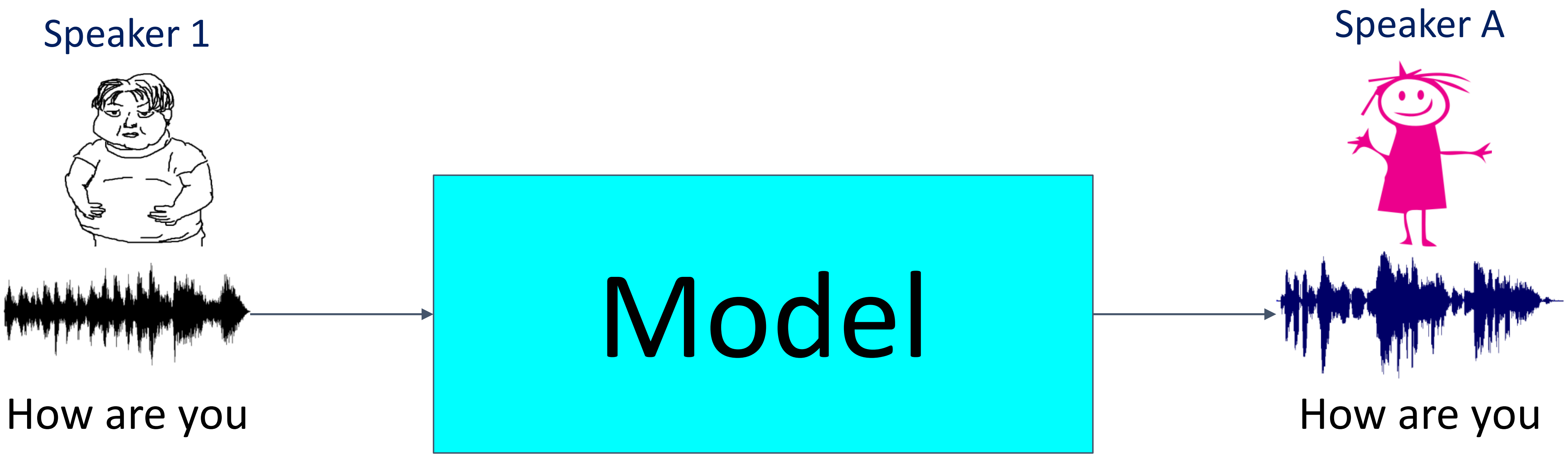
1. Introduction
2. Proposed Approach
    - Model
    - Experiments
3. Conclusion

Outline

1. Introduction
2. Proposed Approach
    - Model
    - Experiments
3. Conclusion

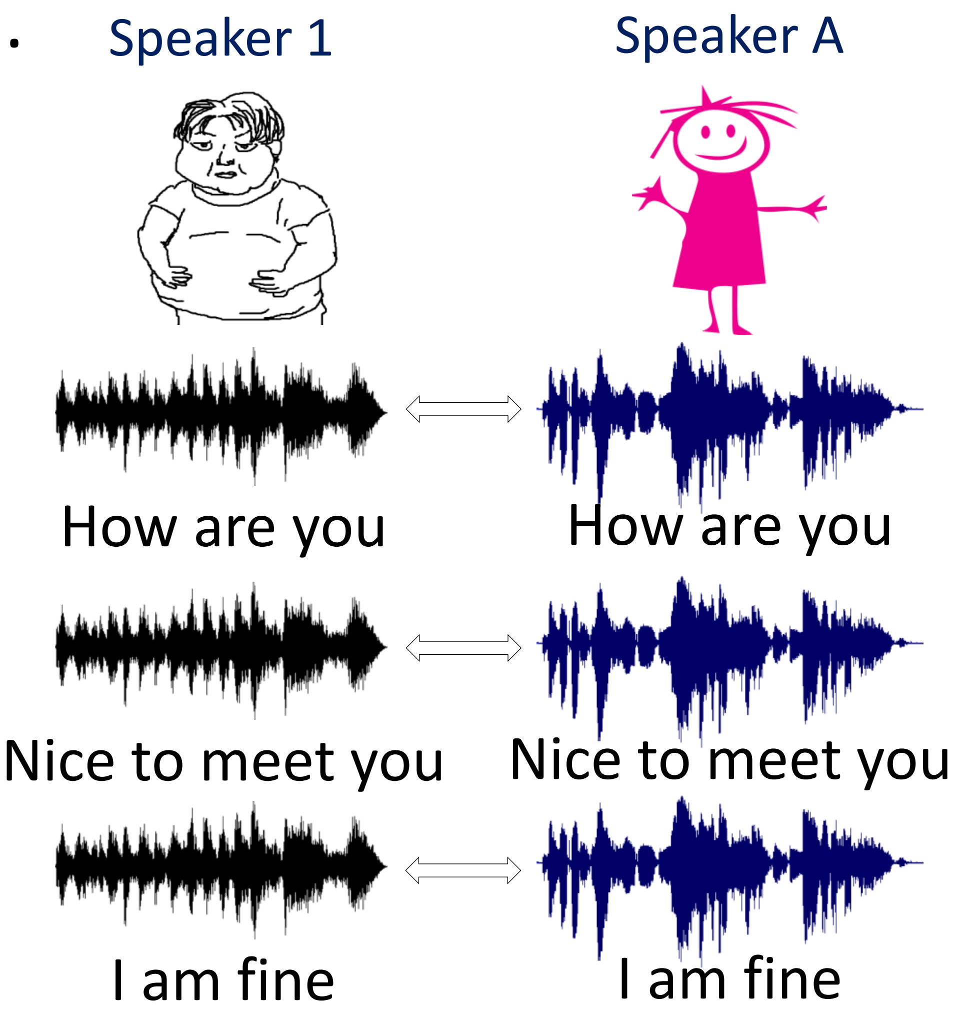
Voice conversion

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



Conventional: supervised VC with parallel data

- Same sentences, different signal from 2 speakers.
- Formulated as a supervised learning problem.
- Problem: require parallel data, which is hard to collect.

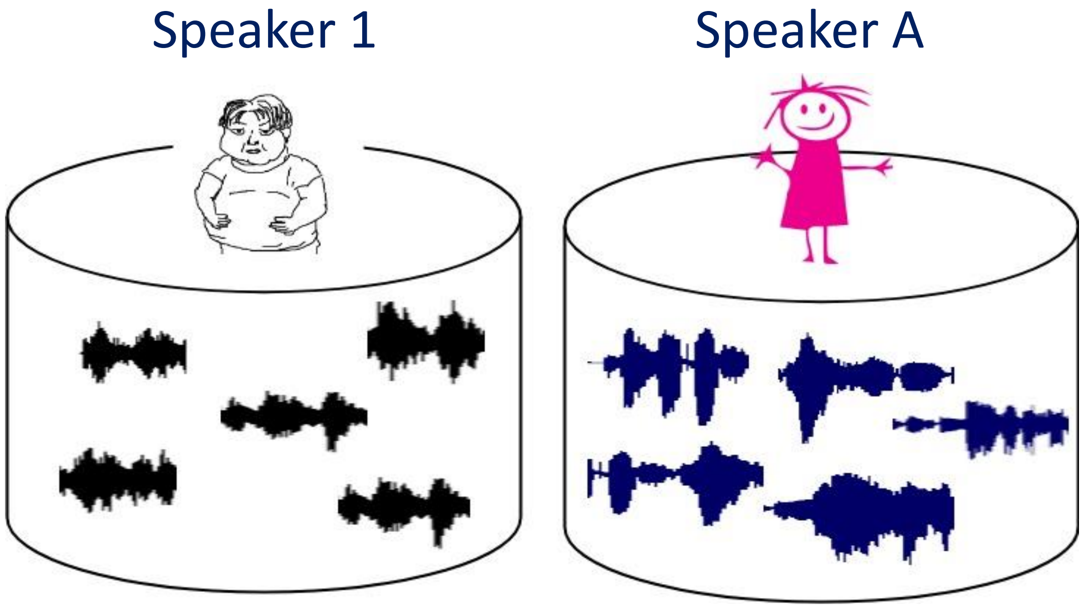


Parallel data

Recently: unsupervised VC with non-parallel data

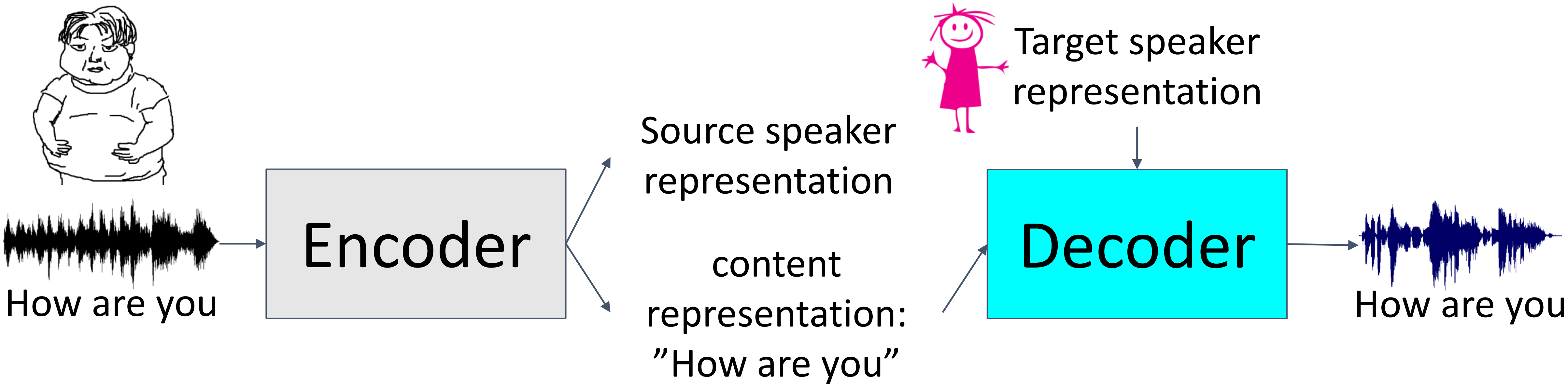
- Trained on non-parallel corpus, which is more attainable.
- Prior work: utilize deep generative model, ex. VAE, GAN, cycleGAN.
- Problem: cannot convert to speakers not in the training data.
- **Our goal: train a model which is able to convert to speakers not in the training data.**

Don’t have to speak
same sentences.



Motivation

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

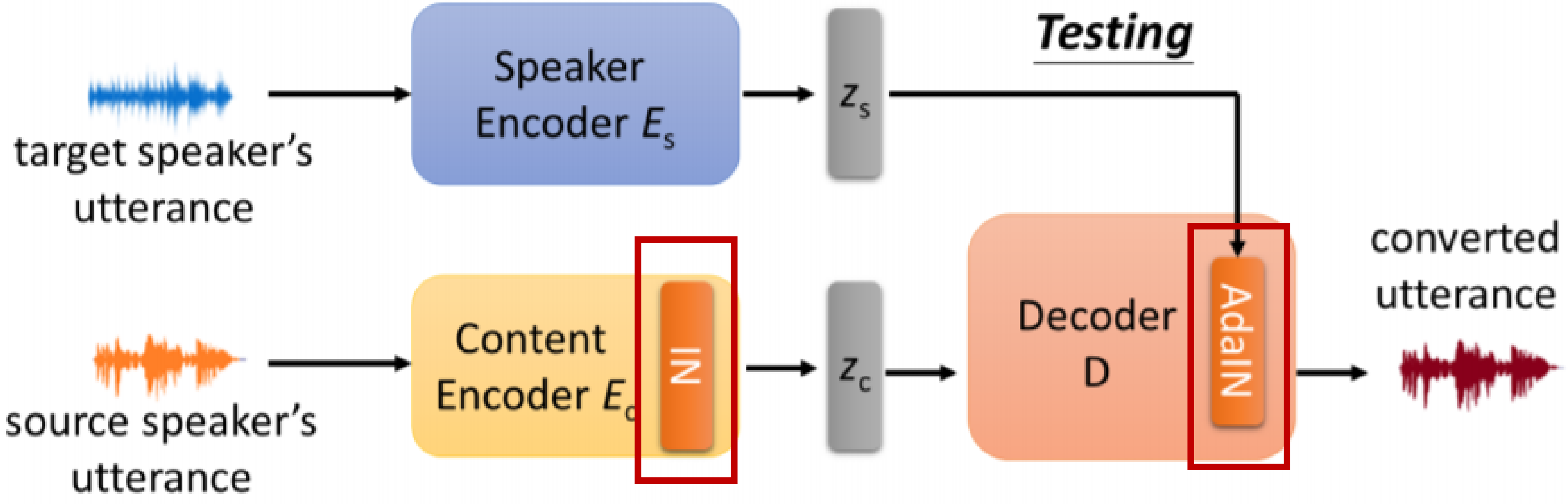


Outline

1. Introduction
2. Proposed Approach
    - Model
    - Experiments
3. Conclusion

Model overview

- One-shot VC: use a utterance from target speaker as reference, and synthesize this reference speaker's voice.
- Idea: separately encode speaker and content information with some special designed layers.



Idea

- Speaker information - invariant within an utterance.
- Content information - varying within an utterance.

Special Designed Layers:

Feature map

Channel

IN Instance Normalization Layer: normalizing speaker information (μ, σ) while preserving content information.

$$\mathrm{M}'_c = \frac{\boxed{M_c} - \mu_{\boxed{c}}}{\sigma_c}$$

Intuition: normalize global information out (ex. high frequency), retain changes over time.

AVG Average Pooling Layer: calculating speaker information (γ, β).

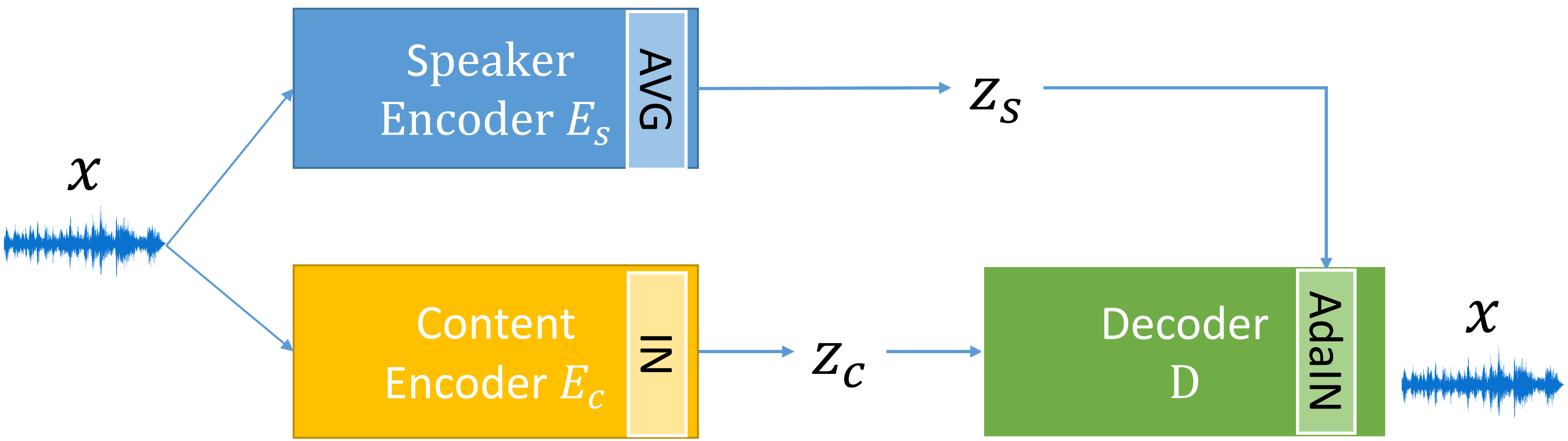
$$M'_c = \sum_{t=1}^{T} \frac{M_c^t}{T}$$

AdaIN Adaptive Instance Normalization Layer: provide speaker information (γ, β).

$$\mathrm{M}'_c = \boxed{\gamma_c} \frac{M_c - \mu_c}{\sigma_c} + \boxed{\beta_c}$$

Model - training

Problem: how to factorize the representations?

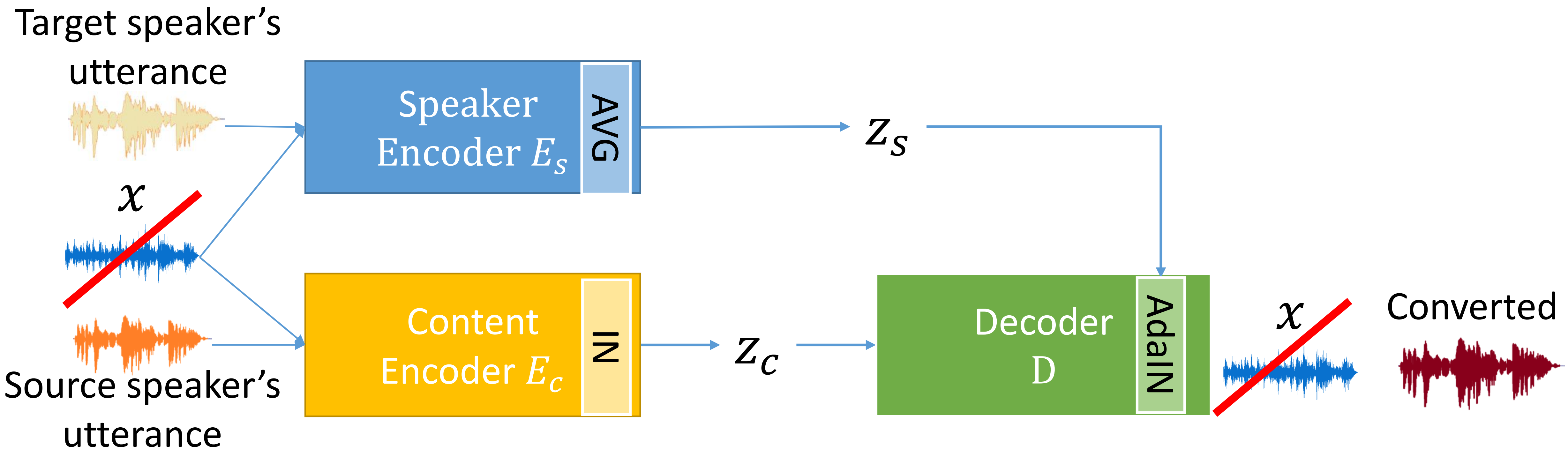


AVG calculating speaker information(γ, β).

IN normalizing speaker information (μ, σ) while preserving content information.

AdaIN provide speaker information (γ, β).

Model - testing



AVG calculating speaker information(γ, β).

IN normalizing speaker information (μ, σ) while preserving content information.

AdaIN provide speaker information (γ, β).

Experiments – effect of IN

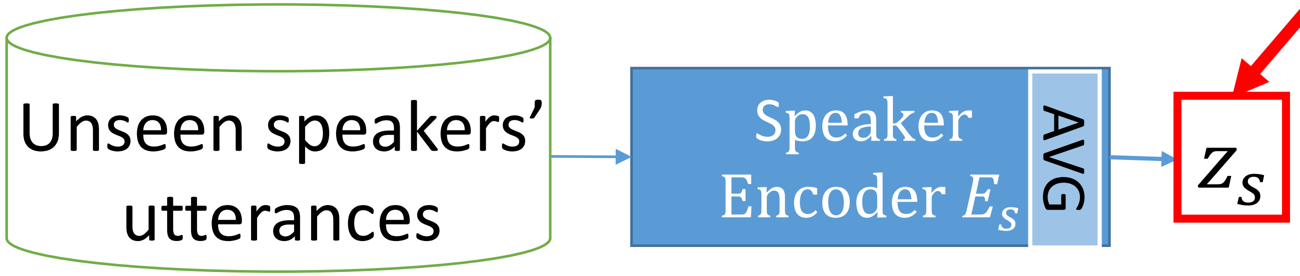
- Train another speaker classifier to see how much speaker information in content representations.
- The lower the accuracy is, the less speaker information it contains.
- Content encoder + IN: less speaker information.

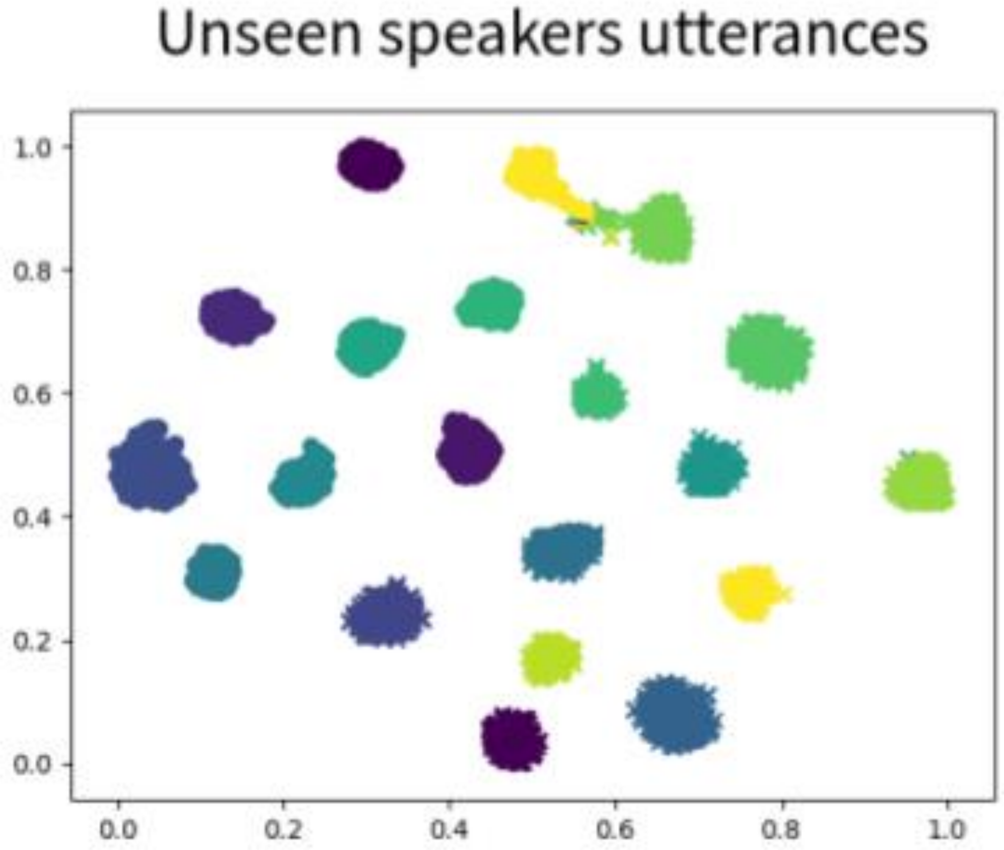
z_c (content representation) → Speaker Classifier → Predict speaker

	E_c With IN	E_c Without IN
Acc.	0.375	0.658

Experiments – speaker embedding visualization

- Does speaker encoder learns meaningful representations?
- One color represents one speaker's utterances.
- z_s from different speakers are well separated.



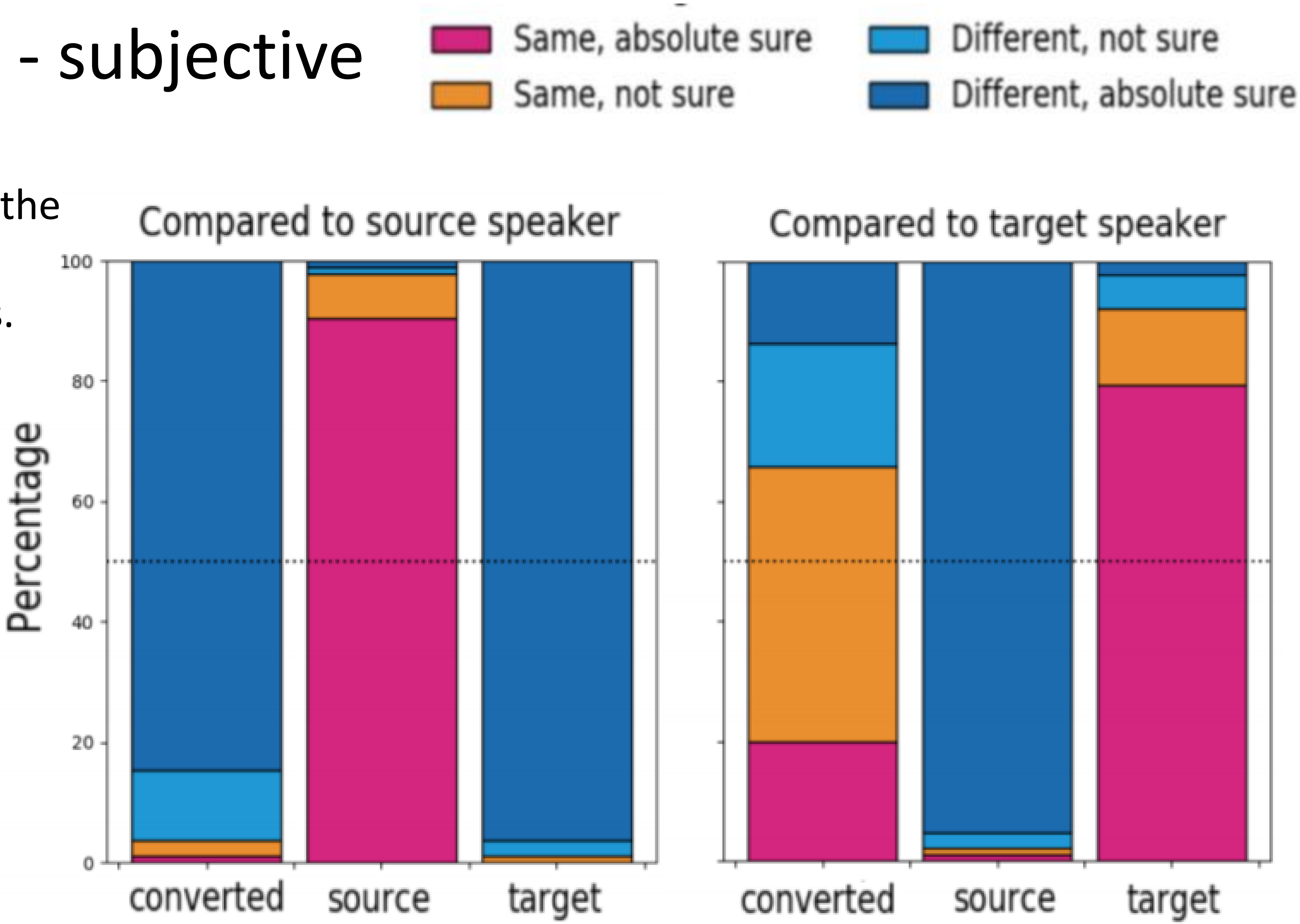


Experiments - subjective

- Ask subjects to score the similarity between 2 utterances in 4-scales.

Experiments - subjective

- Ask subjects to score the similarity between 2 utterances in 4-scales.
- Our model is able to generate the voice similar to target speaker’s.



Demo page: https://jjery2243542.github.io/one-shot-vc-demo/

Demo (unseen)

Male to Male

Source: Target: Converted:

Female to Male

Source: Target: Converted:

Conclusion

- We proposed a one-shot VC model, which is able to convert to unseen speaker with one reference utterance.
- By IN and AdaIN, our model is able to learn factorized representations.

Thank you for your attention.