AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss

Kaizhi Qian*[1], Yang Zhang*[23], Shiyu Chang[23],
Xuesong Yang[1], Mark Hasegawa-Johnson[1]

[1]University of Illinois at Urbana-Champaign
[2]MIT-IBM Watson AI Lab
[3]IBM Research Cambridge

Motivation

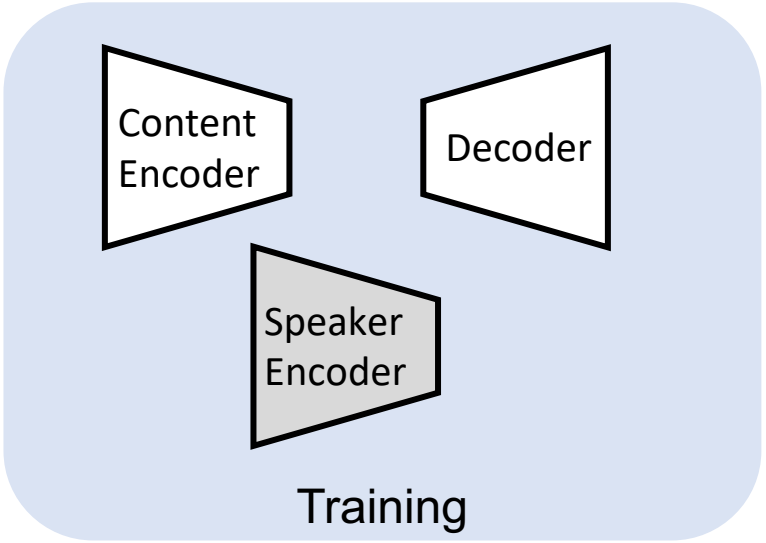
- Voice conversion aims to modify the source speech to make it sound like being uttered by another speaker.


Motivation

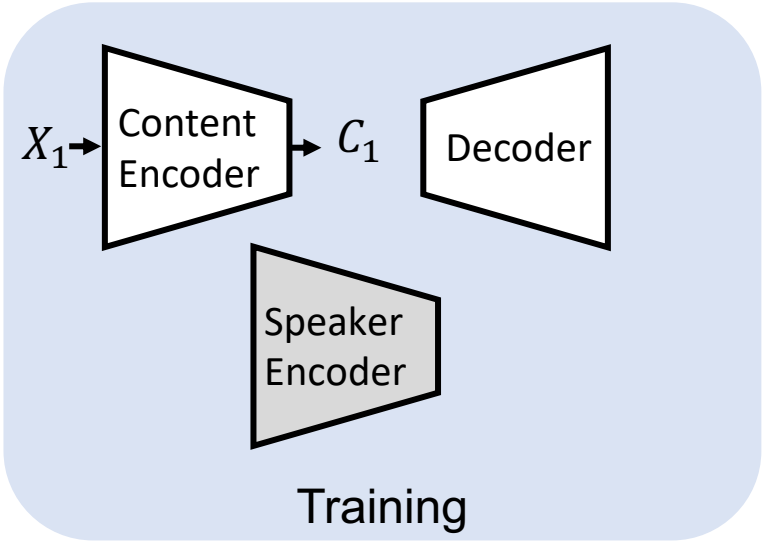
- Voice conversion aims to modify the source speech to make it sound like being uttered by another speaker.

- Existing voice style transfer techniques:
  - Use complex architectures and training schemes but do not work well for speech
  - Only convert between seen speakers

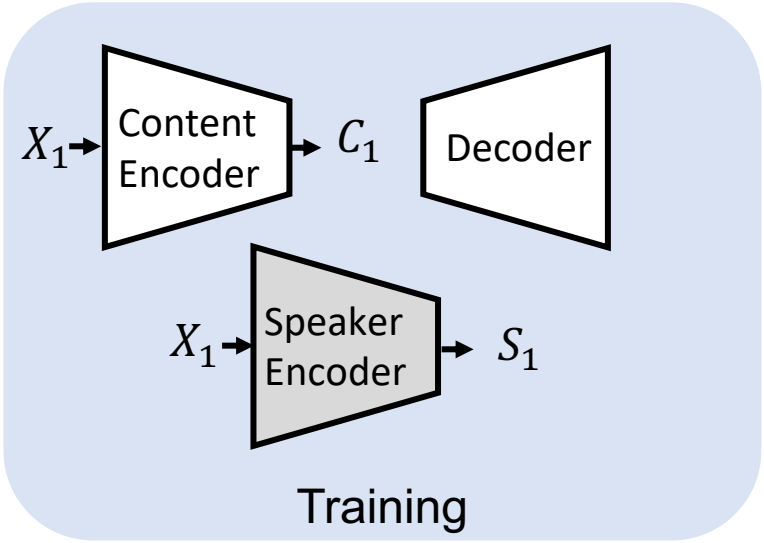

AutoVC



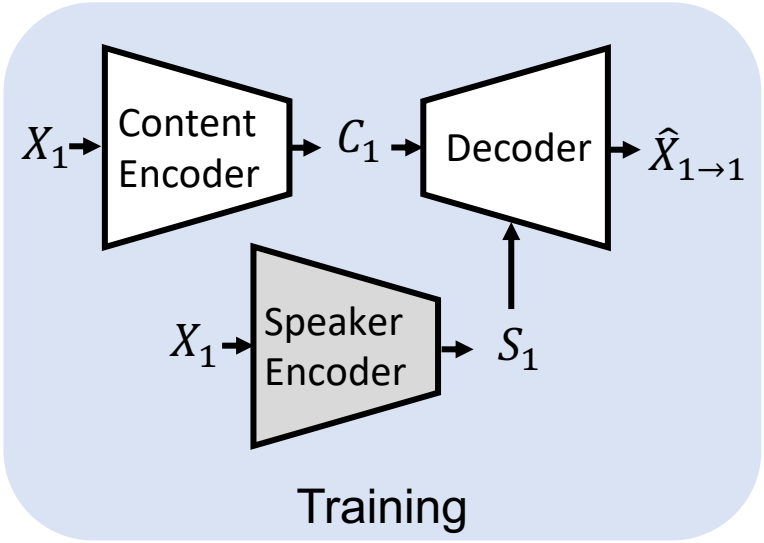
AutoVC



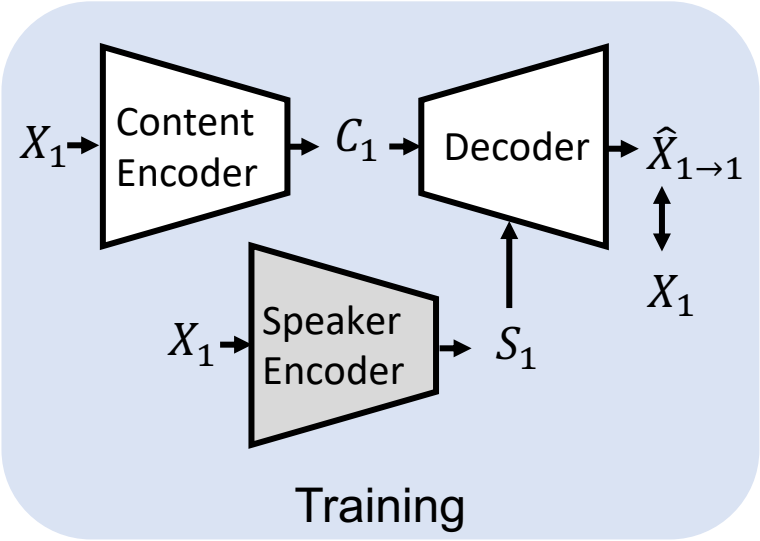
AutoVC


X_1
Content Encoder
C_1
Decoder
X_1
Speaker Encoder
S_1
Training

AutoVC


X_1
Content Encoder
C_1
Decoder
$\hat{X}_{1\to1}$
X_1
Speaker Encoder
S_1
Training

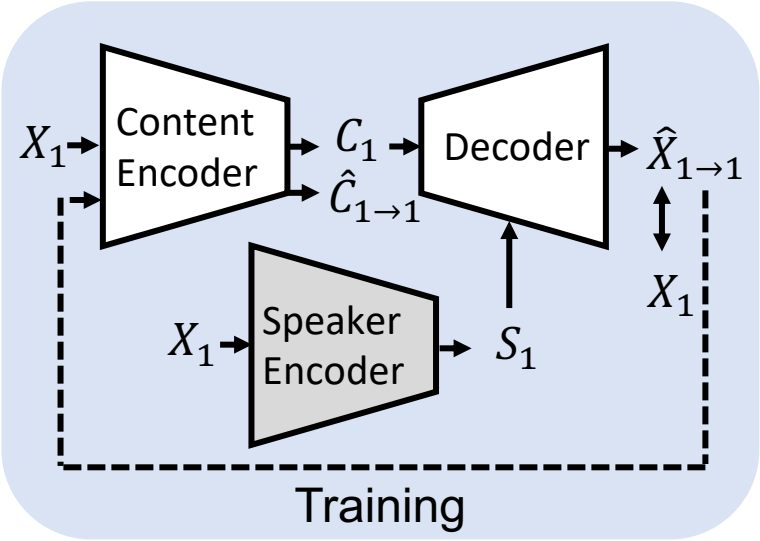
AutoVC



- Train only on self-reconstruction Loss:

$$\mathbb{E}\left[\left\|\hat{X}_{1\to1} - X_1\right\|_2^2 + \lambda\left\|\hat{C}_{1\to1} - C_1\right\|_1\right]$$

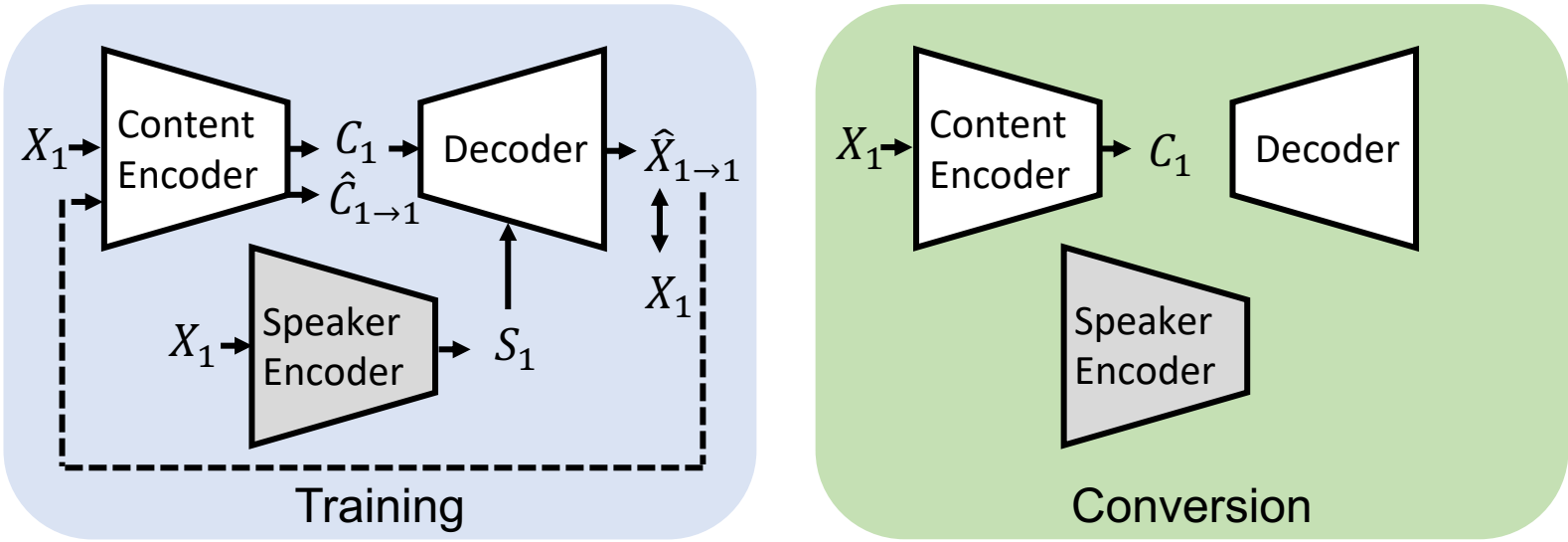
AutoVC



- Train only on self-reconstruction Loss:

$$\mathbb{E}\left[\left\|\hat{X}_{1\to1} - X_1\right\|_2^2 + \lambda\left\|\hat{C}_{1\to1} - C_1\right\|_1\right]$$

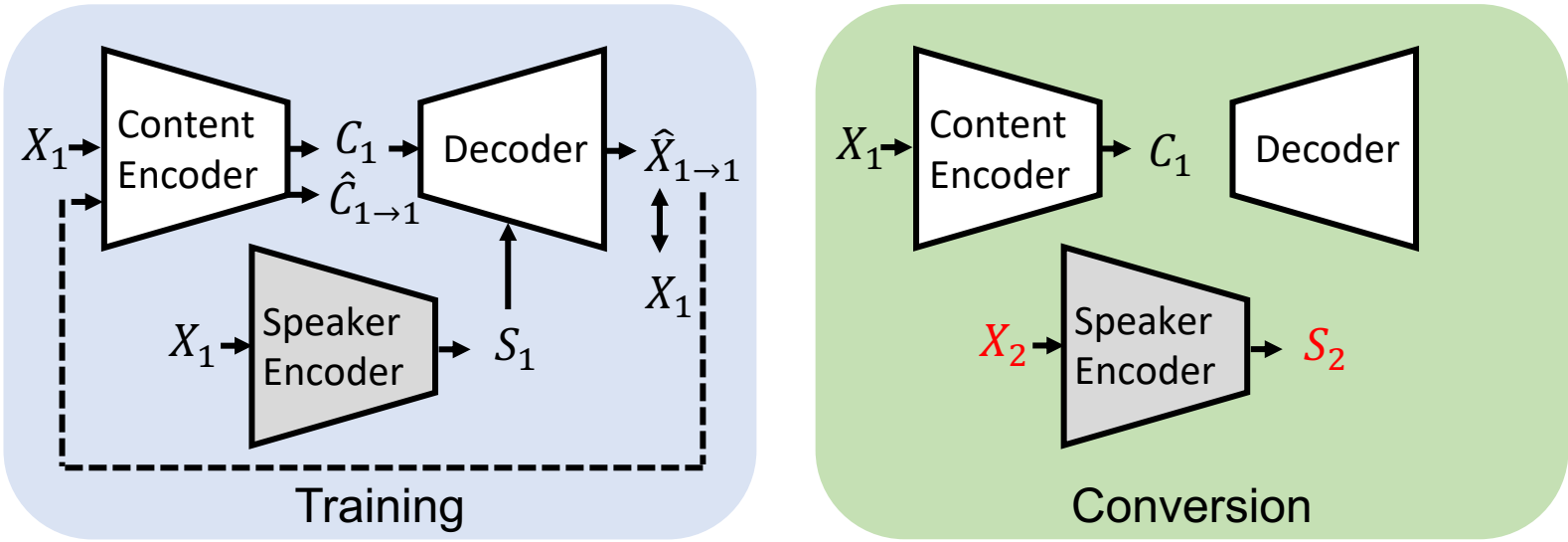
AutoVC



- Train only on self-reconstruction Loss:

$$\mathbb{E}\left[\left\|\hat{X}_{1\to1} - X_1\right\|_2^2 + \lambda\left\|\hat{C}_{1\to1} - C_1\right\|_1\right]$$

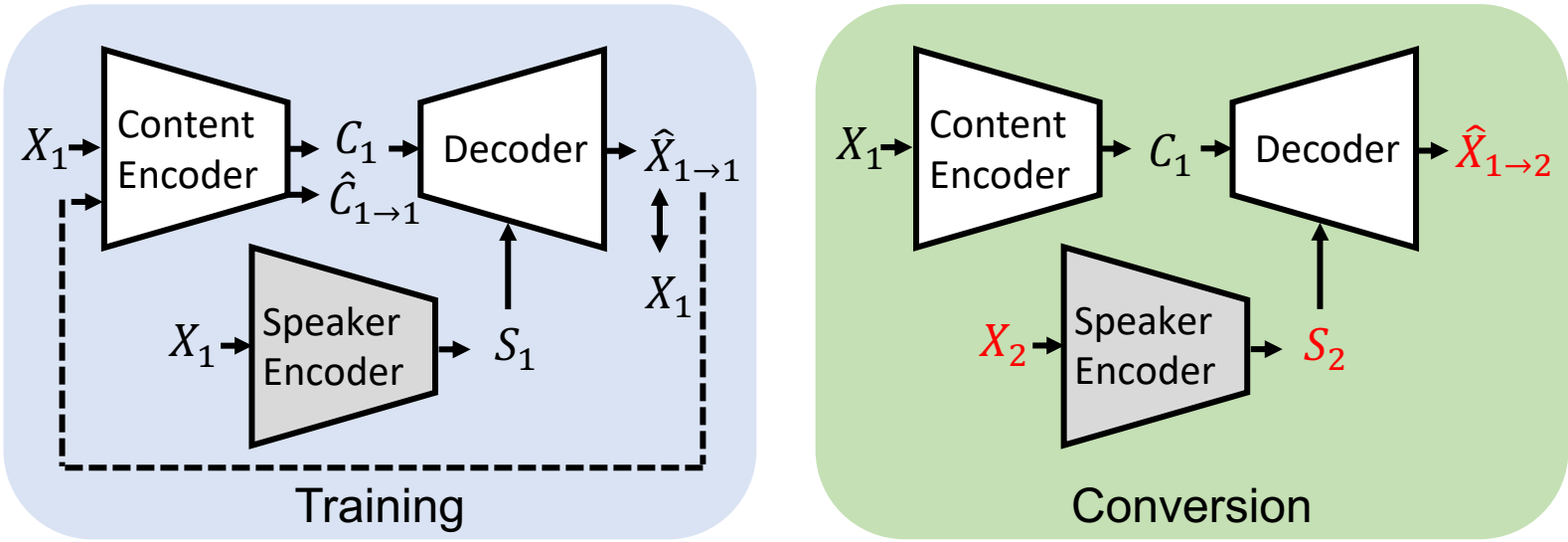
AutoVC



- Train only on self-reconstruction Loss:

$$\mathbb{E}\left[\left\|\hat{X}_{1\to1} - X_1\right\|_2^2 + \lambda\left\|\hat{C}_{1\to1} - C_1\right\|_1\right]$$

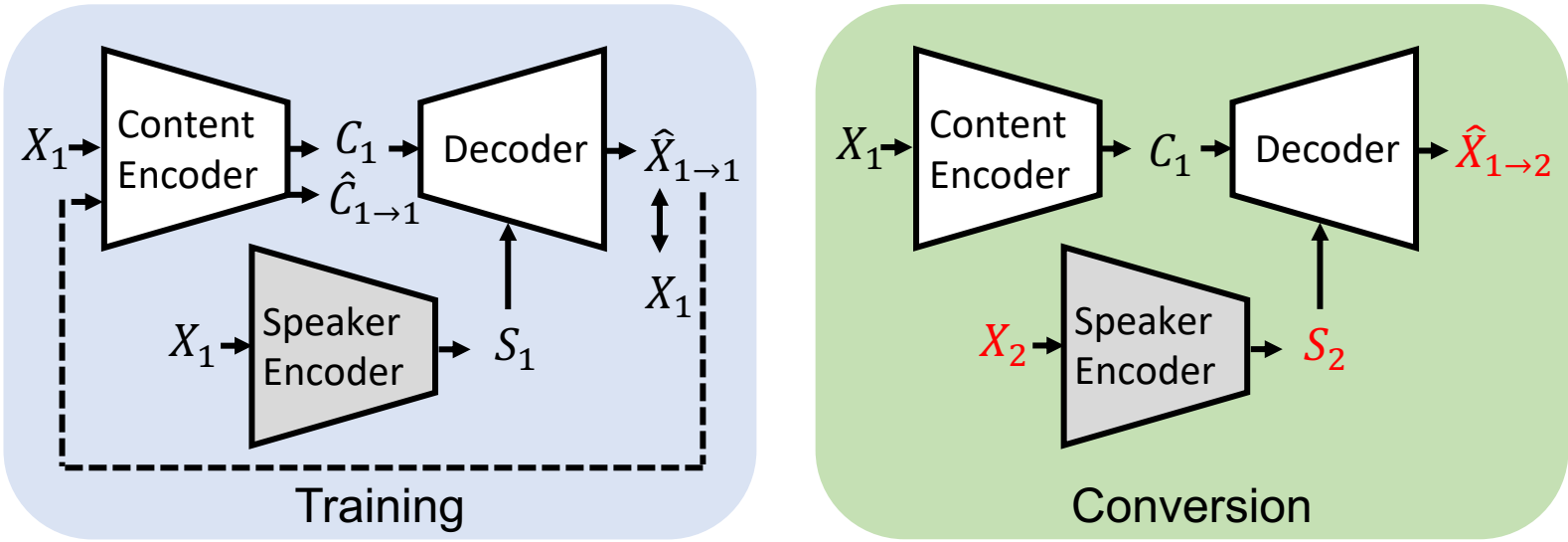
AutoVC



- Train only on self-reconstruction Loss:

$$\mathbb{E}\left[\left\|\hat{X}_{1\to1} - X_1\right\|_2^2 + \lambda\left\|\hat{C}_{1\to1} - C_1\right\|_1\right]$$

AutoVC



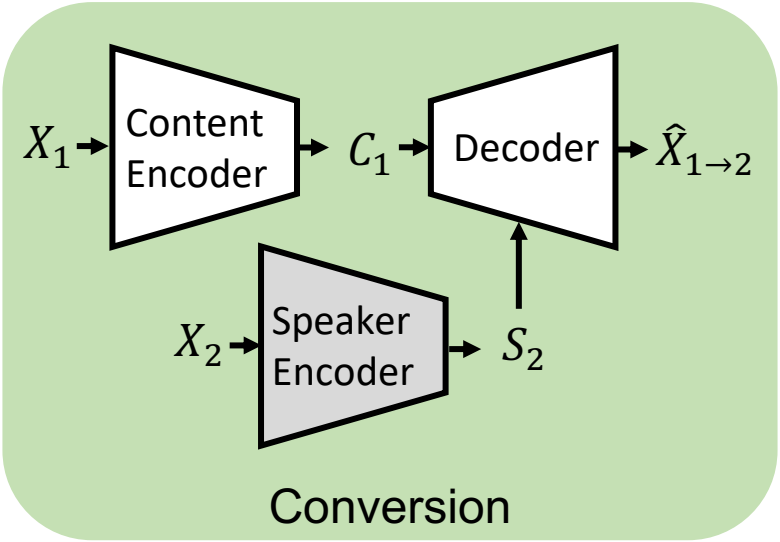
- Train only on self-reconstruction Loss:

$$\mathbb{E}\left[\left\|\hat{X}_{1\to1} - X_1\right\|_2^2 + \lambda\left\|\hat{C}_{1\to1} - C_1\right\|_1\right]$$

- With bottleneck tuning, AutoVC can match the distribution!

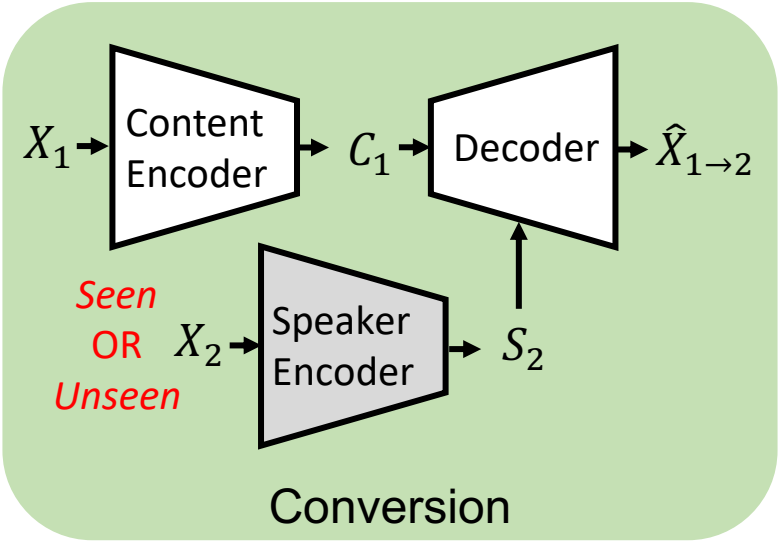


AutoVC



- Speaker encoder is pretrained

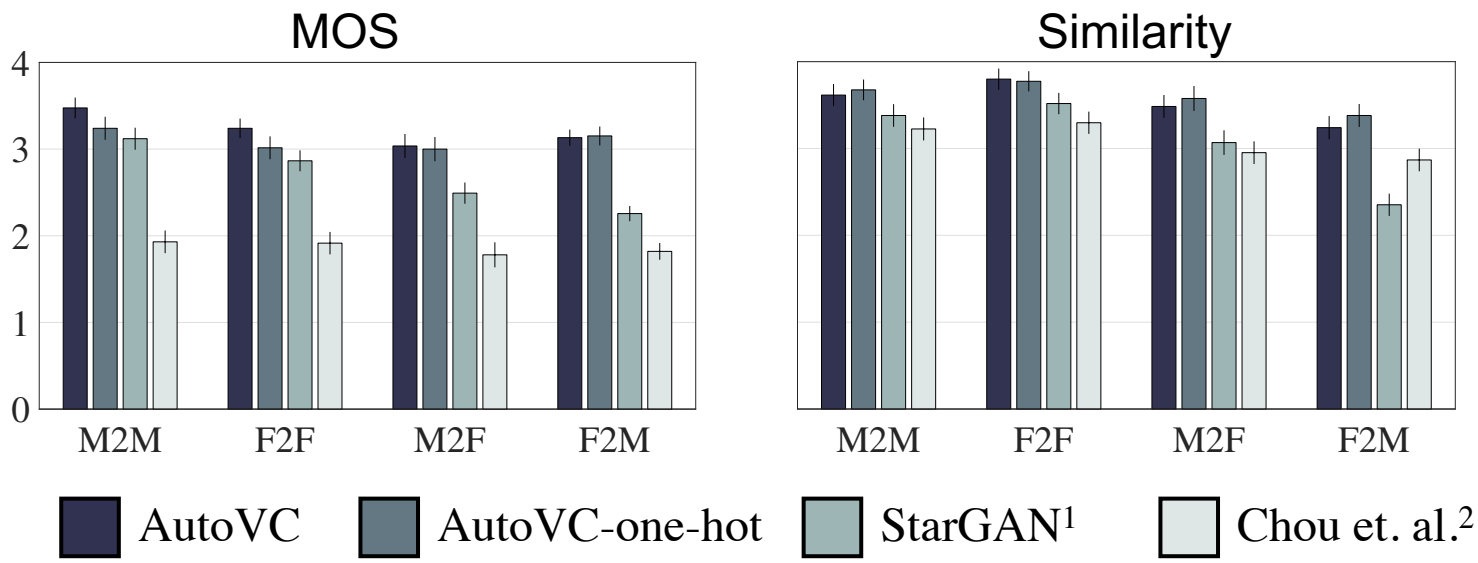
AutoVC



- Speaker encoder is pretrained
- Can generalize to unseen speakers – zero-shot conversion

Conversion Between Seen Speakers

Source Target Converted

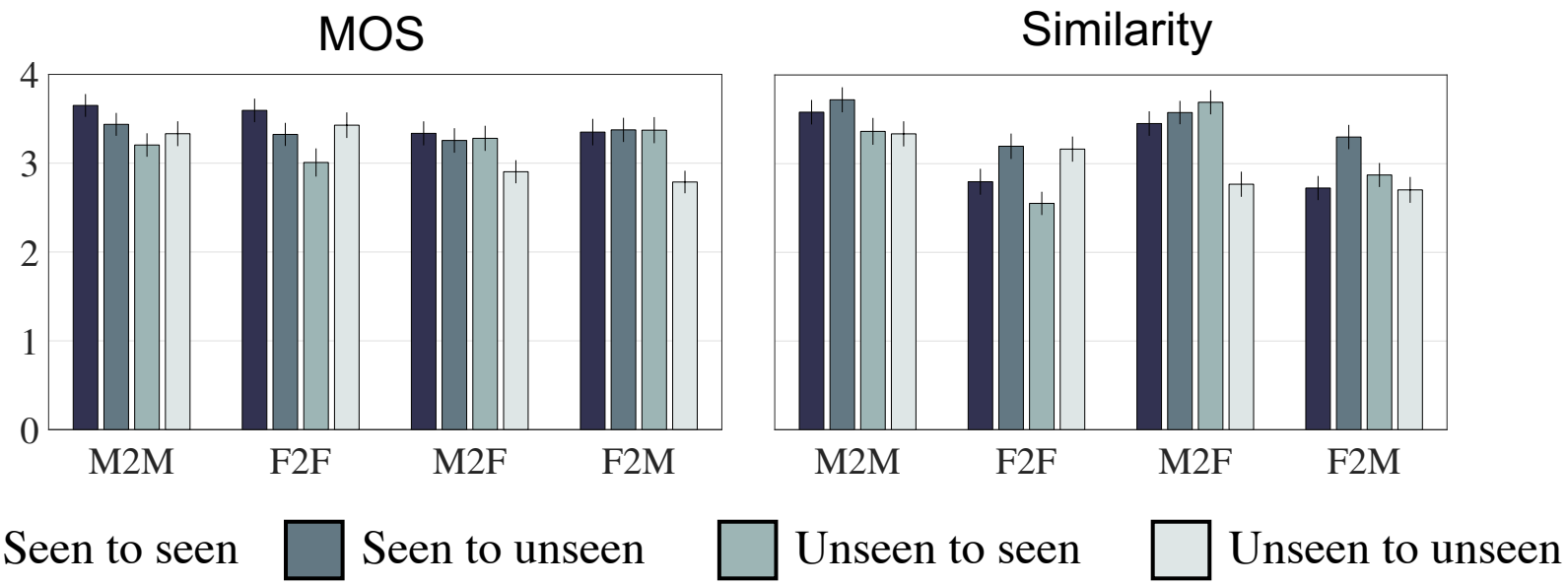


[1]StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks
[2]Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations

Conversion Between Unseen Speakers

- The first zero-shot voice conversion framework

Source Target Converted



Take Away

- Autoencoder is all you need to achieve theoretically ideal voice conversion


Take Away

- Autoencoder is all you need to achieve theoretically ideal voice conversion

- AutoVC generalizes well to unseen speakers


Thank you!

Poster #225