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

Kaizhi Qian*¹, Yang Zhang*²³, Shiyu Chang²³, Xuesong Yang¹, Mark Hasegawa-Johnson¹

¹University of Illinois at Urbana-Champaign
²MIT-IBM Watson AI Lab
³IBM 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

Content Encoder

Decoder

Speaker Encoder

Training
AutoVC

$X_1 \xrightarrow{\text{Encoder}} C_1 \xrightarrow{\text{Decoder}}$

Training

Content Encoder

Speaker Encoder
AutoVC

$X_1 \xrightarrow{\text{Content Encoder}} C_1 \xrightarrow{\text{Decoder}}$

$X_1 \xrightarrow{\text{Speaker Encoder}} S_1$

Training
AutoVC
AutoVC

- Train only on self-reconstruction Loss:
  \[ \mathbb{E} \left[ \| \hat{X}_{1 \rightarrow 1} - X_1 \|_2^2 + \lambda \| \hat{C}_{1 \rightarrow 1} - C_1 \|_1 \right] \]
AutoVC

• Train only on self-reconstruction Loss:

\[ \mathbb{E} \left[ \| \hat{X}_{1 \rightarrow 1} - X_1 \|_2^2 + \lambda \| \hat{C}_{1 \rightarrow 1} - C_1 \|_1 \right] \]
• Train only on self-reconstruction Loss:

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

- Train only on **self-reconstruction** Loss:
  \[
  \mathbb{E} \left[ ||\hat{X}_{1\rightarrow 1} - X_1||^2_2 + \lambda ||\hat{C}_{1\rightarrow 1} - C_1||_1 \right]
  \]
AutoVC

- Train only on **self-reconstruction Loss:**
  \[
  \mathbb{E} \left[ \| \hat{X}_{1\rightarrow1} - X_1 \|^2_2 + \lambda \| \hat{C}_{1\rightarrow1} - C_1 \|_1 \right]
  \]
AutoVC

• Train only on **self-reconstruction** Loss:
  \[
  \mathbb{E} \left[ \|\hat{X}_{1\rightarrow 1} - X_1\|_2^2 + \lambda \|\hat{C}_{1\rightarrow 1} - C_1\|_1 \right]
  \]

• 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
- M2M
- F2F
- M2F
- F2M

### Target
- MOS
- Similarity

### Converted
- AutoVC
- AutoVC-one-hot
- StarGAN\(^1\)
- Chou et. al.\(^2\)

\(^1\)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**
- M2M
- F2F
- M2F
- F2M

**Target**
- Seen to seen
- Seen to unseen
- Unseen to seen
- Unseen to unseen

**Converted**
- MOS
- Similarity

![Bar charts showing MOS and Similarity for different conversion types (M2M, F2F, M2F, F2M) and seen/unseen conditions.](chart_image)
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