Unsupervised Voice Conversion by Separately Embedding Speaker and Content Information with Deep Generative Model

以分別嵌入語者及語言內容資訊之深層生成模型達成無監督式語音轉換

Speaker: 周儒杰 (Ju-Chieh Chou)  
Advisor: 李琳山 (Lin-shan Lee)
Outline

1. Introduction
   • Voice Conversion
   • Branches
   • Motivation

2. Proposed Approach
   • Multi-target Model
     • Model
     • Experiments
   • One-shot Model
     • Model
     • Experiments

3. Conclusion
Outline

1. Introduction
   • Voice Conversion
   • Branches
   • Motivation

2. Proposed Approach
   • Multi-target Model
     • Model
     • Experiments
   • One-shot Model
     • 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: focus on speaker identity conversion.

Model

Speaker 1: How are you

Model

Speaker A: How are you
Conventional: supervised with parallel data

- Same sentences, different signal from 2 speakers.
- Train a model to map from speaker 1 to speaker A.
- Problem: require parallel data, which is hard to collect.

Parallel data
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.

Don’t have to speak same sentences.
Voice Conversion Branches

- Voice Conversion
  - Parallel Data
  - Non-parallel Data

Yeh et al.
- With Transcription (phonemes)
- Without Transcription (phonemes)
- This work

Data Efficiency
- Single-target
- Multi-target
- One-shot

- ch3
- ch4
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
   • Voice Conversion
   • Branches
   • Motivation

2. Proposed Approach
   • Multi-target Model
     • Model
     • Experiments
   • One-shot Model
     • Model
     • Experiments

3. Conclusion
Multi-target unsupervised with non-parallel data

3 models are needed for 3 target speakers.

$N^2$ models for N speakers.
Stage 1: disentanglement between content and speaker representation

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

Training

Speaker 1

Encoder

content representation: $\text{enc}(x)$

Classifier-1

Identify the speaker

Decoder

Speaker id

Remove speaker information

Reconstruction loss
Stage 1: disentanglement between content 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: training-testing mismatch

Training

Same speaker

Speaker y

Classifier-1

Predict speaker y

Testing

Different speaker

Speaker y’
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 (lack details). Leads to **over-smoothed spectra**, and result in buzzy synthesized speech.

Training

![Diagram of the training process involving Encoder, Decoder, Speaker id, Classifier-1, and Reconstruction loss.](image)
Stage 2: patch the output with a residual signal

- Random sample a speaker id as condition.
- Train another generator to produce residual signal (spectra details), 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 the correct speaker.
Stage 2: patch the output with a residual signal

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

![Diagram of the process]

**Testing**

1. **Encoder**
   - Content: $\text{enc}(x)$
   - Target speaker

2. **Decoder**
   - Speaker id
   - Real or generated
   - Identify the speaker

3. **Generator**
   - Residual signal
   - From stage 1, fixed

4. **Discriminator**

5. **Classifier-2**

**Real data**
Experiments – spectrogram visualization

- Is stage 2 helpful?
- Sharpness of the spectrogram is improved by stage 2.
Experiments – subjective preference

- Ask subjects 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: Kaneko et.al. EUSIPCO 2018 [1]
Demo

Male to Female

Source: [Audio] Target: [Audio] Converted: [Audio]

Female to Female

Source: [Audio] Target: [Audio] Converted: [Audio]

Prof. Hung-yi Lee (male, never seen in training data) to Female

Source: [Audio] Target: [Audio] Converted: [Audio]

Demo page: https://jjery2243542.github.io/voice_conversion_demo/
Outline

1. Introduction
   • Voice Conversion
   • Branches
   • Motivation

2. Proposed Approach
   • Multi-target Model
     • Model
     • Experiments
   • One-shot Model
     • Model
     • Experiments

3. Conclusion
One-shot unsupervised with non-parallel data

Prior work: only able to convert to speakers in training data

Speaker id one-hot encoding
(training data includes those of all target speakers)

This work: source/target speakers unseen during training.

Target speaker reference utterance (one-shot)
Idea

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

Special Designed Layers:

**IN** Instance Normalization Layer: normalizing speaker information \((\mu, \sigma)\) while preserving content information.

**AVG** Average Pooling Layer: calculating speaker information \((\gamma, \beta)\).

**AdaIN** Adaptive Instance Normalization Layer: provide speaker information \((\gamma, \beta)\).
Intuition

- Normalize global information out (ex. high frequency), retain changes across time.
Model - training

Problem: how to factorize the representations?

\[ \begin{align*}
\chi & \rightarrow \text{Content Encoder } E_c \rightarrow Z_c \\
\chi & \rightarrow \text{Speaker Encoder } E_s \rightarrow Z_s
\end{align*} \]

**AVG** calculating speaker information \((\gamma, \beta)\).

**IN** normalizing speaker information \((\mu, \sigma)\) while preserving content information.

**AdaIN** provide speaker information \((\gamma, \beta)\).
Model - testing

Target speaker’s utterance → Speaker Encoder $E_S$ → $Z_S$

Source speaker’s utterance → Content Encoder $E_C$ → $Z_C$ → Decoder $D$ → Converted

AVG calculating speaker information $(\gamma, \beta)$.

IN normalizing speaker information $(\mu, \sigma)$ while preserving content information.

AdaIN provide speaker information $(\gamma, \beta)$. 
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.

<table>
<thead>
<tr>
<th></th>
<th>$E_c$ With IN</th>
<th>$E_c$ Without IN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc.</td>
<td>0.375</td>
<td>0.658</td>
</tr>
</tbody>
</table>
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.

- Same, absolute sure
- Same, not sure
- Different, not sure
- Different, absolute sure
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 (unseen)

Male to Male

Source:  
Target:  
Converted:  

Female to Male

Source:  
Target:  
Converted:  

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

- We proposed two unsupervised VC model by the idea of “separately embedding speaker and content information”.

- Multi-target VC
  - We proposed a multi-target VC model by removing speaker information with adversarial training.
  - GAN training mitigate the problem of over-smoothing and improve the result.

- One-shot VC
  - 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.
Instance Normalization

- Instance Normalization: \( M'_c = \frac{M_c - \mu_c}{\sigma_c} \)

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

- Adaptive Instance Normalization: \( M'_c = \gamma_c \frac{M_c - \mu_c}{\sigma_c} + \beta_c \)

Provided by speaker encoder (control th global information)
Demo

Male to Female

Source: Male
Target: Female
Converted: Female

Female to Female

Source: Female
Target: Female
Converted: Female

Prof. Hung-yi Lee (male, never seen in training data) to Female

Source: Male
Target: Female
Converted: Female

Demo page: https://jjery2243542.github.io/voice_conversion_demo/
One-shot unsupervised with non-parallel data

Prior work: only able to convert to speakers in training data

This work: use one utterance as reference (one-shot)

Speaker id (fixed length)