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

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.

Speaker 1

How are you

Model

How are you

Speaker A
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.
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:

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

\[
M'_c = \frac{M_c - \mu_c}{\sigma_c}
\]

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

\[
M'_c = \gamma_c \frac{M_c - \mu_c}{\sigma_c} + \beta_c
\]

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

\[
M'_c = \frac{1}{T} \sum_{t=1}^{T} M^t_c
\]

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

**Feature map Channel**

**AdINA**
Model - training

Problem: how to factorize the representations?

**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

Source speaker’s utterance

Speaker Encoder $E_s$

Content Encoder $E_c$

Decoder $D$

$Z_s$

$Z_c$

ADIN

GAN

IN

AVG

Converted

$\chi$

$\lambda$

$\gamma$, $\beta$

$\mu$, $\sigma$

- **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.

\[ E_c \text{ (content representation)} \rightarrow \text{Speaker Classifier} \rightarrow \text{Predict speaker} \]

<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 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.