One-shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization
Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee
National Taiwan University

1. Overview

Training: text and speaker labels are not used, only speech signal.
One-shot: only provide one utterance during inference, and speakers during inference can be unseen during training.
Idea: separately encode speaker and content representation with some customized layers (w/o GAN).

2. Model

Training

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

Testing

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

Minimize reconstruction error

3. Effect of IN

- Train another speaker classifier to see how much speaker information in content representations.
- Accuracy ↓, speaker information ↓.

<table>
<thead>
<tr>
<th>Content Classifier</th>
<th>Speaker Encoder $E_s$</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>With IN</td>
<td>$E_c$ Without IN</td>
<td>0.658</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.375</td>
</tr>
</tbody>
</table>

Content encoder + IN: less speaker information.

4. Speaker embedding visualization

- Visualize speaker embedding to see whether $E_s$ learns meaningful representations.

Unseen speakers' utterances

Speaker Encoder $E_s$

Point: utterance, Color: speaker

5. Subjective evaluation

- Experiments are conducted on converting unseen source speakers to unseen target speakers.
- Ask subjects to score the similarity between 2 utterances in 4-scales.

Conclusion: our model is able to generate voice similar to target speakers.


Demo: