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

- To 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.

![Diagram showing voice conversion with 'How are you' from Speaker 1 as input to the model, and 'How are you' as output from Speaker 2.]
Conventionally: supervised VC with parallel data

- Same sentences spoken by 2 speakers.
- Collected multiple pairs to train the model.
- Problem: requires parallel data, which is not easily accessible.
Recently: unsupervised VC with non-parallel data

- Trained on non-parallel corpus, which can be collected easily.
- Prior work: utilize deep generative models to map from one domain (speaker) to another, ex. VAE, GAN, CycleGAN.
- Problem: cannot convert to speakers not in the training data.
- Our goal: train a model can convert to speakers not in the training data.

Don’t have to speak same sentences.
Outline

1. Introduction
2. Proposed Approach
   • Model
   • Experiments
3. Conclusion
Model overview

- **Motivation**: speech signals inherently carry both **content** and **speaker** information.
- **Encoder**: learn the **content/speaker representation** separately.
- **Decoder**: synthesize the target voice by combining the **source content representation** and **target speaker representation**.
Previous approach - adversarial training

- Speaker classifier and encoders are learned iteratively.
- Drawbacks: training instability, hyper-parameter tuning.

[Ju-chieh Chou, et al., Interspeech 2018]
Idea

- Separately encode speaker and content information with some customized layers (IN, AdaIN, AVG) for this task.
- **Speaker information** - invariant within an utterance.
- **Content information** - varying within an utterance.
Customized layers – instance normalization

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

= instance normalization  (remove global information)
Customized layers – instance normalization

Each channel has zero mean and unit variance

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

Normalize through time for each channel

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

Content encoder

(ConvNet)
Customized layers – average pooling

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

- **IN** = instance normalization (remove global information)
- **AVG** = 1d average pooling over time (calculate global information)
Customized layers – adaptive InsNorm

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

- **IN** = instance normalization (remove global information)
- **AVG** = 1d average pooling over time (calculate global information)
- **AdaIN** = adaptive instance normalization (add global information)
$z'_i = \gamma \odot z_i + \beta$

Output of Speaker Encoder

Decoder

AdaIN = adaptive instance normalization (add global information)
Outline

1. Introduction
2. Proposed Approach
   • Model
   • Experiments
3. Conclusion
Experiments – effect of IN

Training from VCTK

How are you?

Content encoder

Speaker encoder

Speaker Classifier

With IN | Without IN
Acc.    | 0.375   | 0.658

20 unseen speakers

Conclusion:
with IN -> less speaker information
Experiments – speaker representation visualization

Content encoder

Speaker encoder

Training from VCTK

How are you?

Unseen Speaker Utterances

male

female
Experiments - subjective

- Ask subjects to score the similarity between 2 utterances in 4-scales.

- Same, absolutely sure
- Same, not sure
- Different, not sure
- Different, absolutely 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)

Convert an unseen source speaker to an unseen target speaker with one utterance provided respectively.

Male to Male

Source:  
Target:  
 Converted:  

Female to Male

Source:  
Target:  
 Converted:  

Demo page: https://jjery2243542.github.io/one-shot-vc-demo/
The speakers are unseen during training (one-shot VC).
Demo (unseen)

The speakers are **unseen** during training (one-shot VC).

新垣結衣
(Aragaki Yui)

![Image](509x232 to 529x305)

![Image](510x72 to 529x145)

![Image](418x225 to 477x310)

![Image](667x141 to 727x238)

![Image](210x72 to 293x157)

![Image](174x98 to 206x130)

![Image](167x175 to 314x329)

![Image](767x170 to 799x202)

![Image](418x64 to 477x150)
Conclusion

• We proposed a one-shot VC model, which is able to convert to unseen speakers with only one reference utterance from the target speaker.
• With IN and AdaIN, our model learns to factorize representation.
Thank you for your attention.