Self-Supervised Pre-Trained Voice Conversion

Chung-Ming Chien

* Work done at National Taiwan University
* Collaborated with Yist Y. Lin, Jheng-Hao Lin, Hung-yi Lee and Lin-shan Lee
* Published at IEEE ICASSP 2021 & InterSpeech 2021
Background
Voice Conversion

Source Speaker: “How are you”

Voice Conversion

Output: “How are you”

Target Speaker: “The whether is cold”

Why?

Text-to-Speech

Too little data...

And more applications…
Self-Supervised Learning (SSL) Representations

- Original Speech Signals
- Masked Speech Signals
- Self-Supervised Model
- Speech Representations
- Automatic Speech Recognition
- Speech Resynthesis
- Generation
- Speaker Verification
- Else

SOTA performance with 1 linear layer ➔ text-like
Proposed: Encoding & Generation in One Model

Voice Conversion Model

Speech Representations

Encoding → Text-Like Features → Conversion of speaker identity → Generation

Any-to-any conversion
Prior Arts
Prior Art 1: Exemplar-Based Voice Conversion

Heavily handcrafted $\rightarrow$ end-to-end + self-supervised representations
Prior Art 2: Any-to-Any Voice Conversion

Source Speaker: “How are you”

Content Encoder

Source Features
(speaker information removed)

Decoder

Output: “How are you”

Target Speaker

Speaker Encoder

Fixed-dimensional speaker embedding

Insufficient to encode speaker information?
Proposed Methods
Model Architecture

SSL Model

Source Waveform

SSL Model

Bottleneck Source Encoder

Source Features

Target Waveforms

SSL Model

Target Encoder

Attention Module

Decoder

Output Waveform

Mel-Spectrogram
(time-frequency speech features)

Pre-Trained Vocoder

Prior Art: Any-to-Any VC

Target Features

Content Encoder

Decoder

Output

Target Waveforms

The use of SSL models

Prevent speaker leakage

Similar to exemplar-based methods

Target features in a sequence

The use of SSL models

Content

Speaker

Encoder

Encoder

Decodes

Output

Mel-Spectrogram
(time-frequency speech features)
Exemplar-based Voice Conversion

Attention Module

Source: “Have some fun!”

Search

Extract

Attention Map

Phonetically similar fragments

“Sometimes.” “Have you” “Funny!”

Fuse

Detailed speaker information

Output: “Have some fun!”

Exemplar-based Voice Conversion

Search

Extract

Join
Training

Reconstruction Loss

Same utterance

Speaker information

Free of parallel-data

Training

Testing
Experiments
Experimental Setup

- **Training**
  - VCTK corpus (109 speakers)

- **Testing**
  - seen speaker (VCTK)
  - unseen speakers (CMU)
    - one-shot conversion

- **Compared SSL Features**
  - CPC (contrastive predictive coding)
  - APC (autoregressive predictive coding)
  - Wav2Vec 2.0

- **Non SSL Features**
  - Mel spectrograms
  - PPG (phoneme posteriorgram trained with text annotations)
Automatic Speaker Similarity Evaluation

- Off-the-shelf speaker verification system

  - the percentage of outputs passing the system (the higher the better)

Target features affect speaker similarity more
Subjective Evaluation

- 5-scale Mean Opinion Score (MOS) of synthetic utterances
  - Speaker similarity
  - Naturalness
Compared with Previous Works

- Compared with previous works that are also
  - One-shot
  - Any-to-any voice conversion
  - Parallel-data-free

[1] Chou et al., One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization
[2] Qian et al., AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss
Attention Analysis

- Same sentence, different speakers
- Attention map alignment from the Transformer block

Phonetically similar units are aligned

Source Speaker
“Please call Stella.”

Target Speaker
“Please call Stella.”

Converted
“Please call Stella.”
Conclusion
Conclusion

- A SOTA approach to any-to-any voice conversion
  - One-shot and parallel-data-free
  - Show the advantage of sequence speaker features over fixed-dimensional embeddings
- Combine SSL encoding & generation in a voice conversion task without any annotation
  - Compare different SSL features
  - SSL features are better than traditional features
Future Work

- The bottleneck has to be carefully monitored to balance the content correctness and speaker information leakage
  - Better disentanglement of speaker and content information
  - Will discrete SSL features be more text-like?
Questions?