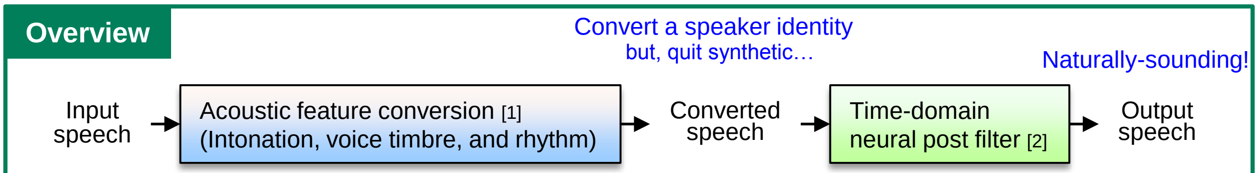
18 Changing your voice and speaking style

- Voice and prosody conversion with sequence-to-sequence model -

Abstract

We propose an voice and prosody conversion method for impersonating a desired speaker's identity and hiding a speaker's identity. The conversion method consists of acoustic feature conversion and time-domain neural postfilter. The acoustic feature conversion is based on a sequence-to-sequence learning with attention mechanism, which makes it possible to capture the long-range temporal dependencies between source and target sequences. The later post filter employs a cyclic model based on adversarial networks, which requires no assumption for the speech waveform modeling. In contrast to current voice conversion techniques, the proposed method makes it possible to convert not only voice timbre but also prosody and rhythm while achieving high-quality speech waveform generation due to the proposed time-domain neural post filter. The remaining challenge is the real-time voice conversion which is our ongoing work.

Overview



Convert a speaker identity
but, quit synthetic…

Naturally-sounding!

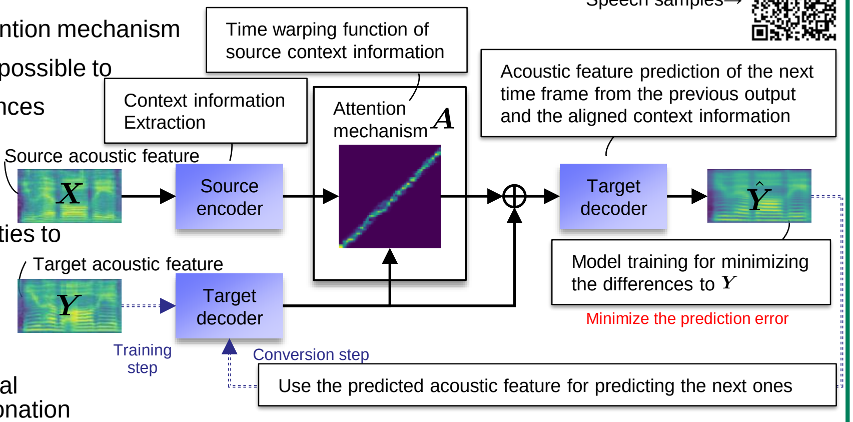
Input speech → Acoustic feature conversion [1] (Intonation, voice timbre, and rhythm) → Converted speech → Time-domain neural post filter [2] → Output speech

Acoustic Feature Conversion [1]

(e.g., Impersonating a speaker's identity and modifying pronunciation)

Speech samples→

- Train encoders, decoder, and attention mechanism
- Encoder-decoder model makes it possible to
  1. Handle input and output sequences of different lengths
  2. Convert not only voice timbre but also rhythm
- Attention mechanism has the abilities to
  1. Select critical information from the encoded representation in accordance with the output sequence representation
  2. Consider the long-range temporal dependencies for converting intonation



Time warping function of source context information

Context information Extraction

Acoustic feature prediction of the next time frame from the previous output and the aligned context information

Source acoustic feature

Target acoustic feature

Model training for minimizing the differences to $\boldsymbol{Y}$

Minimize the prediction error

Training step

Conversion step

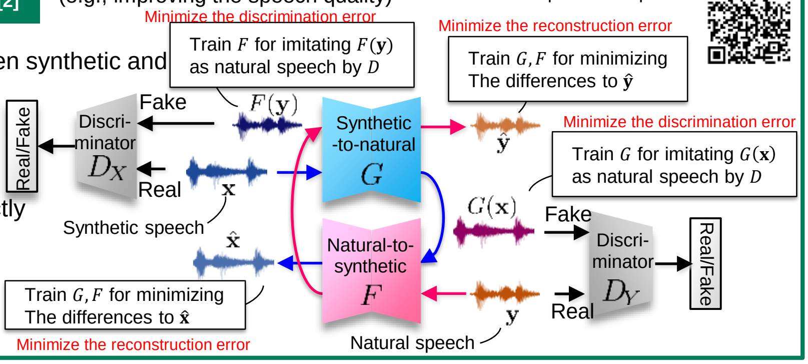
Use the predicted acoustic feature for predicting the next ones

Time-domain Neural Post Filter [2]

(e.g., improving the speech quality)

Speech samples →

- Train conversion functions G, F between synthetic and
- Cyclic model makes it possible to
  1. Train the models with non-parallel data of synthetic and natural speech
  2. Handle the phase information correctly due to need for the reconstruction of speech waveform
- Generative adversarial learning helps to generate clear speech



Minimize the discrimination error

Train F for imitating $F(\mathbf{y})$ as natural speech by D

Minimize the reconstruction error

Train G, F for minimizing The differences to $\hat{\mathbf{y}}$

Minimize the discrimination error

Train G for imitating $G(\mathbf{x})$ as natural speech by D

Train G, F for minimizing The differences to $\hat{\mathbf{x}}$

Minimize the reconstruction error

References

[1] K. Tanaka, H. Kameoka, T. Kaneko, N. Hojo, " AttS2S-VC: Sequence-to-Sequence Voice Conversion with Attention and Context Preservation Mechanisms," in *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2019),* May 2019.

[2] K. Tanaka, H. Kameoka, T. Kaneko, N. Hojo,"WaveCycleGAN2: Time-domain Neural Post-filter for Speech Waveform Generation," *arXiv:1904.02892,* Apr. 2019, (submitted to *INTERSPEECH2019.*)

Contact

Kou Tanaka Email: cs-liaison-ml at hco.ntt.co.jp
Learning and Intelligent Systems Research Group, Innovative Communication Laboratory

Innovative R&D by NTT
Open House 2019