17

Emotional voice conversion via facial expression recognition

Real-time speech emotion contololler using face

Abstract

There are many kinds of physical or mental barriers that prevent individuals from smooth verbal communication. One key technique to overcome some of these barriers is voice conversion (VC), a technique to convert para/non-linguistic information contained in a given utterance without changing the linguistic information. Here, we propose a crossmodal voice control system, which offers a way to control the vocal expression of emotion in speech through the facial expression in a face image. The proposed system consists of performing facial expression recognition (FER) followed by VC. For VC, we have developed a method based on sequence-to-sequence (S2S) learning, which is designed to convert the prosodic features as well as the voice characteristics in speech conditioned on the output of the FER system. We believe that this work can provide some insight on what it is like to be able to control our voice through different modalities.

Communication augmentation system

Using voice conversion (VC) technique to help overcome barriers that prevent us from smooth communication

Core techniques

Sequence-to-sequence VC

■Sequence-to-sequence (S2S) learning

- Offers a general framework for transforming one sequence into another variable length sequence
- Encoder/decoder structure and attention mechanism make it possible to learn conversion rules that reflect long-term dependencies in input/output sequences
- Usually requires large-scale parallel corpora

■ VC based on S2S learning (S2S-VC)

- Voice expressions are characterized by prosodic features (e.g., intonation and rhythm)
- S2S-VC is able to convert prosodic features as well as voice characteristics in input speech with limited amount of training data

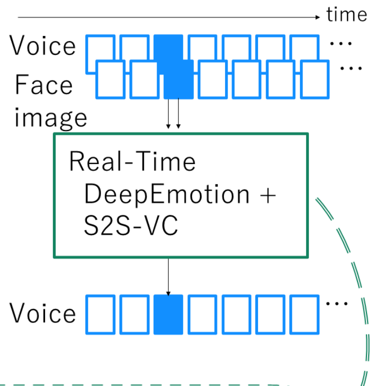
Facial expression recognition (FER)

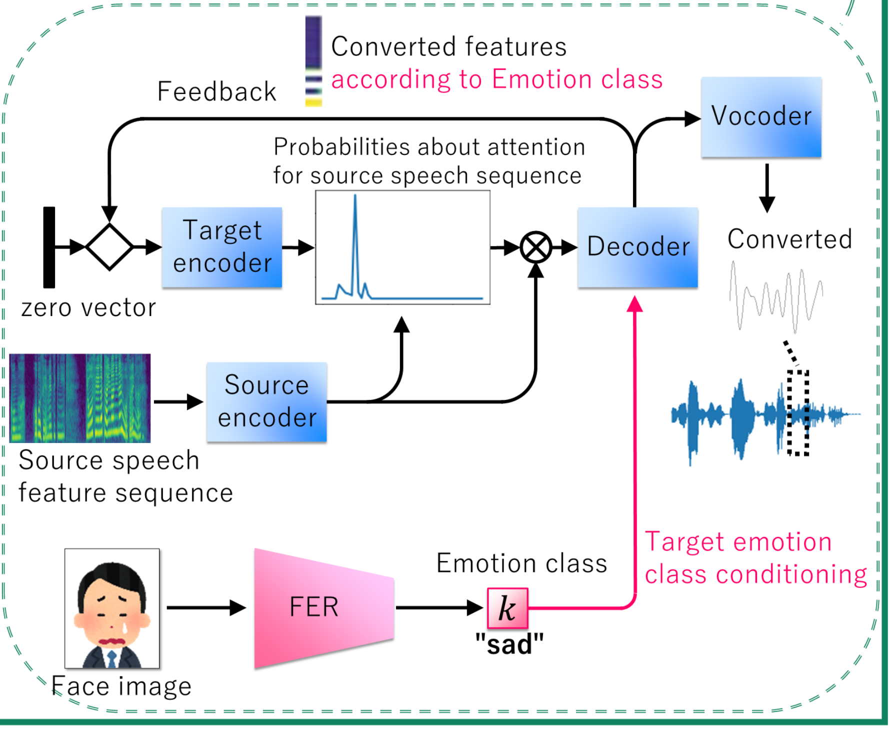
■FER using attentional convolutional network

- After prediction, output is passed to VC system

Voice expression control through face

Crossmodal voice control consisting of facial expression recognition and VC





References

[1] H. Kameoka, K. Tanaka, T. Kaneko, N. Hojo, “ConvS2S-VC: Fully convolutional sequence-to-sequence voice conversion,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1849-1863, June 2020.

[2] 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)*, pp. 6805-6809, May 2019.

[3] M. Shervin, M. Minaei, and A. Abdolrashidi, “Deep-emotion: Facial expression recognition using attentional convolutional network.” *Sensors 21.9:3046*, 2021.

Contact

Kou Tanaka / Recognition Research Group, Media Information Laboratory
Email: cs-openhouse-ml@hco.ntt.co.jp