17 Pay attention to the speaker you want to listen to (II)

Neural selective hearing with audio-visual speaker clues

Abstract

Human beings have the ability to concentrate on listening to a desired speaker (= selective hearing) even when multiple people are speaking at the same time. The purpose of this research is to realize the selective listening mechanism of human beings on a computer. In this research, we propose multimodal selective hearing technology that uses video information as the target speaker's clues in addition to audio information. By utilizing multiple information sources like humans, the technology become advanced that can operate stably even in situations, where audio clues are useless, such as conversations between speakers with similar voice characteristics. This technology will become fundamentals of various devices that take human voice as input. For example, it will contributes to the realization of robots and smart speakers that recognize people and change their response.

Selective Hearing with Audio Speaker Clue

□ Selective Hearing

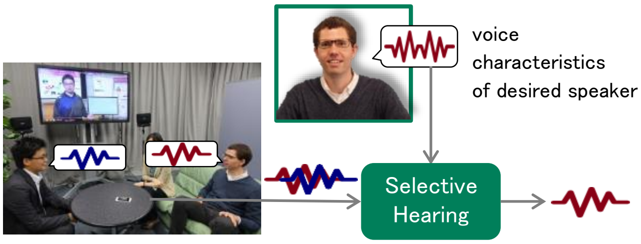
- Ability to focus on listening to desired speaker from mixture signals
- In daily conversations, multiple speakers often speak at same time

⇨ Humans easily perform such selective hearing, but it is difficult for conventional computers

⇨ First proposal of neural selective hearing with audio speaker clue（OPEN HOUSE 2018）

□ Problem

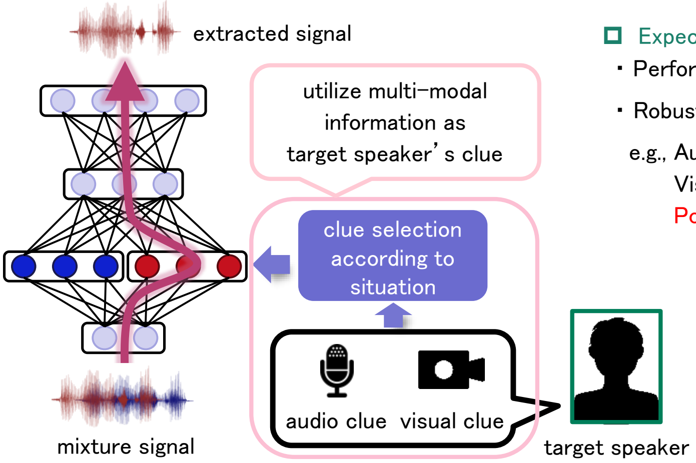
With audio clues, extraction performance degrades for mixture signals with similar voice characteristics



Utilization of Audio and Visual Speaker Clues

□ SpeakerBeam（= Selective Hearing based on Deep Learning）

Deep learning-based model, which extracts desired speaker's voice from mixture signal given by target speaker's clue


extracted signal
utilize multi-modal information as target speaker's clue
clue selection according to situation
audio clue
visual clue
target speaker
mixture signal

□ Solution: Proposal of Multimodal SpeakerBeam

In addition to voice characteristics (audio info.), use mouth motion (visual info.) as speaker clues

⇨ utilize multi-modal information like humans

□ Expected effect

- Performance improvement by utilizing multiple modality
- Robustness improvement against lack of speaker clues

e.g., Audio clue is useless (similar voice characteristics)
Visual clue is missing（face not detected)
Possible to extract even in above situations

＊ About target speaker's clue

Pre-recorded audio data of target speaker

Video data (around mouth) recorded same time as mixture signal

References

[1] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, T. Nakatani, "Multimodal SpeakerBeam: Single channel target speech extraction with audio-visual speaker clues," *Proc. Interspeech*, 2019.

[2] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, J. Cernocky, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing,* 2019.

Contact

Tsubasa Ochiai Email: cs-openhouse-ml@hco.ntt.co.jp
Signal Processing Research Group, Media Information Laboratory

オープンハウス 2020