Neural selective hearing with audio-visual speaker clues

17 聞きたい人の声に耳を傾けるコンピュータ（Ⅱ）

音声と映像を手がかりとしたマルチモーダル選択的聴取

どんな研究

人間は、複数の人が同時に話している状況においても、聞きたい人の声に集中し、聞きたい人の声を聞き取る能力（＝ 選択的聴取）を持っています。本研究は、そうした人間が持つ選択的聴取の機能をコンピュータ上で実現することをめざしたものです。

どこが凄い

音声情報に加え、映像情報を手がかりとして利用する、マルチモーダル選択的聴取の技術を実現しました。人間のように複数の情報源を適切に活用することで、声の性質が似た話者の会話といった、音声情報だけでは困難であった状況でも安定して動作可能な技術へと発展しました。

めざす未来

複数の人の声が混ざった音声から「聞きたい人の声のみを抽出する技術」は、人の音声を入力とする様々なデバイスの基盤となる技術です。人を認識して対応を変えるロボットやスマートスピーカーの実現といった、人とより自然に対話するコンピュータの実現に寄与します。

音声情報に基づく選択的聴取

□ 選択的聴取とは

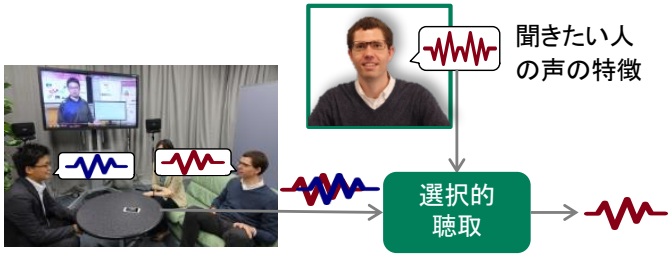
- 複数の人の声が混ざった状況においても、聞きたい人の声に注意し聞くことが出来る能力
- 日常的な会話シーンでは、複数の人の声が混ざった状況は自然に起こる

⇨ 選択的聴取は人間には容易であるが、従来のコンピュータには困難な問題であった

⇨ 音声情報に基づく選択的聴取の初提案（オープンハウス2018）

□ 課題

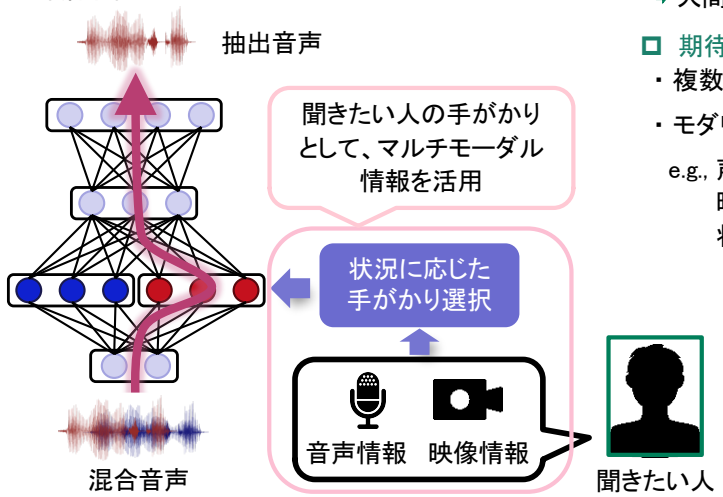
音声情報のみを手がかりとした選択的聴取は、声の性質が似た話者同士の場合には困難



音声と映像双方に由来する話者手がかりの利用

□ SpeakerBeam（= 深層学習に基づく選択的聴取モデル）

聞きたい人に関する手がかりを与えることで、混合音声からその話者の音声のみを取り出す深層学習モデル



□ 解決の方針：Multimodal SpeakerBeam の提案

声の特徴（音声情報）に加え、唇の動き（映像情報）を手がかりとして利用

⇨ 人間のようにマルチモーダル情報を活用

□ 期待される効果

- 複数モダリティの活用による性能向上
- モダリティの劣化や欠損に対する頑健性向上

e.g., 声の特徴が役に立たない（声の性質が似た話者）
映像データが欠損した（唇が映らない）
状況でも抽出できる

＊ 聞きたい人の手がかりについて

- 事前に録音された聞きたい人の音声データ
- 混合音声と同時に録画された聞きたい人の映像データ（唇周り）

関連文献

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

連 絡 先

落合 翼（Tsubasa Ochiai） メディア情報研究部 信号処理研究グループ

Email: cs-openhouse-ml@hco.ntt.co.jp

オープンハウス 2020