Computational selective hearing based on deep learning

16 聞きたい人の声に耳を傾けるコンピュータ

～深層学習に基づく音声の選択的聴取～

どんな研究

会話の中で複数の人が同時に話している時でも、人間は聞きたい人の声に集中し、聞き分けること（選択的聴取）ができます。一方、コンピュータにはその能力がなく、聞きたい人の声だけをうまく聞き取ることはできません。コンピュータによる選択的聴取の研究を進めています。

どこが凄い

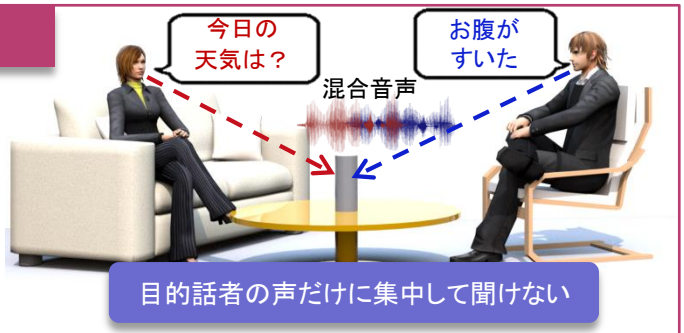
事前に収録した聞きたい人の声を補助情報として利用し、複数人が同時に話している時に、その人の声だけを聞き取ることができる技術『適応型ニューラルネットワーク』を実現しました。これにより、聞きたい人の声に耳を傾けることができるコンピュータを実現しました。

めざす未来

ロボット・ホームアシスタント・スマートスピーカなどの遠隔音声収音装置が、注目すべき話者の声だけを聞き取ることができるようになります。それにより、例えば、ロボットが特定の人の声にのみ反応するなど、人とより自然に会話できるようになります。

課題

- 複数人での会話や、テレビの音が背景で流れている時など、日常の様々な場面において、人の声同士が混ざることはよく起こる
- 従来のコンピュータ・音声入力装置は、混ざった音声が入力されてくると、その中から、聞きたい人（目的話者）の声だけに集中して聞き取るということはできなかった

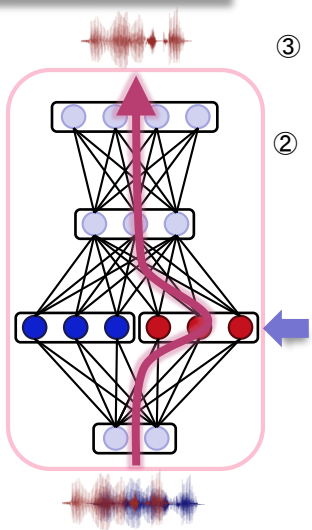


Picture designed with Sweet Home 3D. Includes 3D models created by Reallusion, Pencilart, Scopia and eTeks.

深層学習に基づく選択的聴取

聞きたい話者の声の特徴に基づき、混合音声からその話者の音声のみを取り出すようにニューラルネットワークを学習する

- ✔ 大量のデータを用いた学習により、学習データにない話者の音声であっても、その特徴量を与えることで、その人の声を取り出し可能
- ✔ 話者数や目的話者の方向など、従来技術で必要とされる情報が不要な、汎用性の高い手法



③ 聞きたい話者の声のみが出力される

② 聞きたい話者の特徴量に適応して、ニューラルネットワークがその振る舞いを自動的に変え、その話者を取り出すようになる

話者特徴量抽出

① 事前に収録した聞きたい話者の音声データ（数秒程度の事前録音）から、その話者の声を表す特徴量を抽出する

聞きたい話者

Picture designed with Sweet Home 3D. Includes 3D models created by Reallusion

関連文献

[1] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, T. Nakatani, “Speaker-aware neural network based beamformer for speaker extraction in speech mixtures,” in *Proc. Interspeech*, 2017.

[2] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, T. Nakatani, “Single channel speaker extraction and recognition with SpeakerBeam,” in *Proc. of 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’18)*, 2018.

担 当 者 デルクロア マーク（Marc Delcroix） メディア情報研究部 信号処理研究グループ