All-neural source separation, counting and diarization model

17 いつ、誰が、何を話した？全部で何人いた？

～何人の会話でも聞き分けられる深層学習モデル ～

どんな研究

会話収録音から、「いつ、誰が、何を話したか」という情報を推定します。従来の類似技術は、収録状況に関して様々な条件（話者は移動不可、話者数は既知、等）を仮定し、その条件が満たされた時のみうまく動作するものでした。しかし、実データではこれらの条件が満たされないことも多々あります。

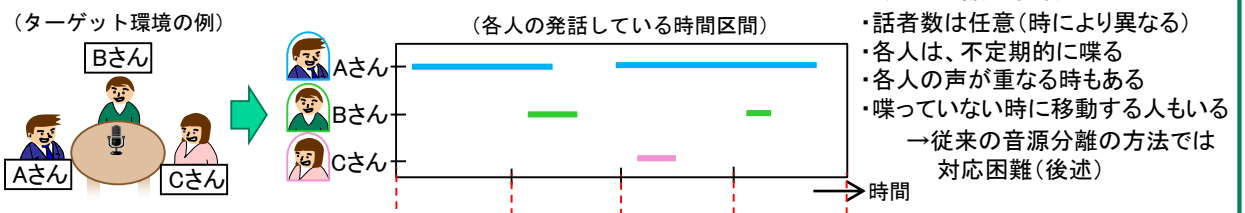
どこが凄い

提案法は、複数人の声が重なっていても、話者数を数え上げながら、「いつ、誰が話したか」という情報を、話者の声の特徴に基づき精度良く推定します。深層学習に基づく、任意の会話状況を表現できるモデルであり、適切な学習データがあれば、あらゆる実会話データに対応できることが期待されます。

めざす未来

人と人との会話から「いつ、誰が、何を話したか」という情報を自動獲得する技術は、会話を自動分析する技術の基盤となり、議事録自動作成や、私たちのコミュニケーションを助けるロボットの実現に寄与します。今後は、より実際的なデータを用い、提案原理の有効性を検証していきます。

会話分析（「いつ、誰が、何を話したか」の推定）の難しさ

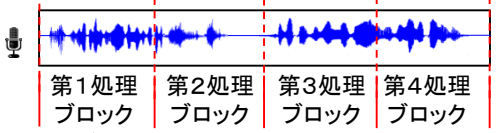


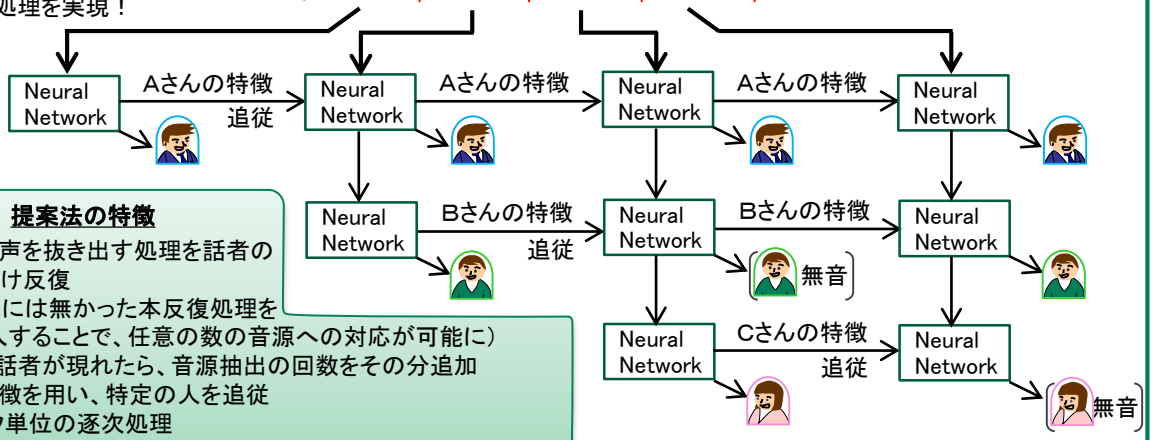
自然な会話は多様でダイナミック
- 話者数は任意（時により異なる）
- 各人は、不定期的に喋る
- 各人の声が重なる時もある
- 喋っていない時に移動する人もいる

→従来の音源分離の方法では対応困難（後述）

提案法

深層学習を用いて、収録音から「いつ、誰が、何を話したか」を推定！学習データに基づき、最適な処理を実現！





提案法の特徴
- 各人の声を抜き出す処理を話者の数分だけ反復
  →従来には無かった本反復処理を導入することで、任意の数の音源への対応が可能に）
- 新たな話者が現れたら、音源抽出の回数をその分追加
- 声の特徴を用い、特定の人を追従
- ブロック単位の逐次処理

提案法の優位性（従来の音源分離手法との比較）

- 提案法は、音源分離と話者数推定を同時に実現。一方、従来技術は、話者数を既知としていた。
- 提案法は、話者の声の特徴に注目し、目的話者の声を聴き続けられる。喋っていない時に、話者が移動しても対応可能。一方、従来技術は、方向情報を基にある話者に注目し続けるものが多く、喋っていない話者が移動すると対応が困難。

関連文献

[1] K. Kinoshita, L. Drude, M. Delcroix, T. Nakatani, “Listening to each speaker one by one with recurrent selective hearing networks,” in *Proc. IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP)*, pp. 5064–5068, 2018.

[2] T. von Neuman, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, R. Haeb-Umbach, “All-neural online source separation, counting, and diarization for meeting analysis,” in *Proc. IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP)*, 2019. (to appear)

連 絡 先

木下 慶介 (Keisuke Kinoshita) メディア情報研究部 信号処理研究グループ

Email: cs-liaison-ml at hco.ntt.co.jp

Innovative R&D by NTT

オープンハウス 2019