Deep speaker attribute estimation with speaker clustering

15 この声、何歳？

話者クラスタリングを用いた深層話者属性推定

どんな研究

音声から発話者の年齢や性別といった話者に関する情報を推定する研究です。顔画像や動画といった視覚的な情報からこれら情報を高い精度で推定する手法は既にいくつか知られていますが、音声のみしか利用できない場合、最新の深層学習技術をもってしても未だ解決が困難な問題です。

どこが凄い

高精度な年齢推定を行うためには各年代の話者の膨大な学習データが必要です。しかし実際には年代毎にデータ量の違いがあり、特にデータが少ない年代の推定が困難でした。そこで、声が似た他の話者の推定結果を用いて補正することで、従来よりも高い精度で年齢推定できる技術を考案しました。

めざす未来

本技術は年齢のみならず感情など話者に関する様々な属性推定へ応用できます。今後は、各属性推定のための深層学習モデルと共に更なる性能改善を行い、話者属性を推定する汎用的な枠組を実現し、ユーザに特化した新たな音声インタフェース開発やマーケティングへの応用をめざします。

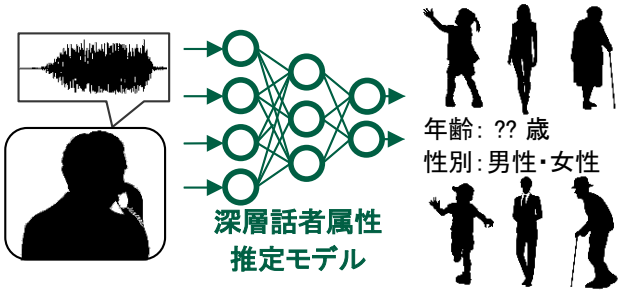
深層話者属性推定

問題設定

音声から発話者の年齢や性別などの話者属性情報を深層学習モデルにより推定

応用分野

コールセンタの応答決定やマーケティングの支援、ユーザの属性により挙動を変える音声対話システムの実現など



問題の難しさ

- 年代ごとに学習データ量の偏りが大きい（図1）
- 特定の話者・年代に対し過学習してしまう（図2）

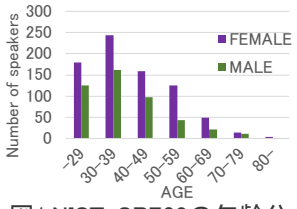


図1.NIST-SRE08の年齢分

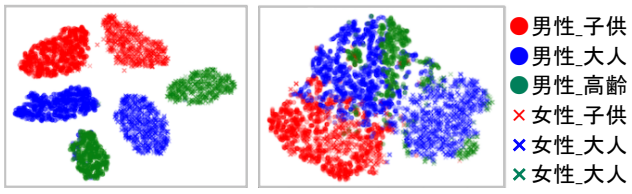
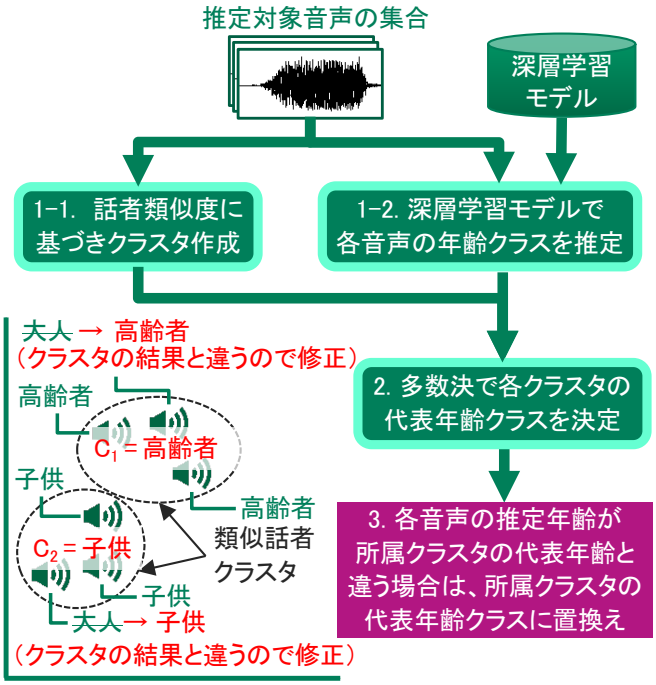


図2. 年齢推定器の出力の可視化（[1]より引用）
（左図：学習データ、右図：評価データ）

話者クラスタリングに基づく性能改善法

深層学習モデルによる年齢推定の推定結果を類似話者の推定結果の多数決により補正する



:推定対象音声とその年齢推定結果
C_*:クラスタ毎に多数決で決定した代表年齢クラス

評価尺度	従来手法	提案手法
年齢クラス分類精度（子供・大人・高齢者クラス）	59 %	72 %
年齢推定誤差	± 10.9 歳	± 8.7 歳

関連文献

[1] N. Tawara, H. Kamiyama, S. Kobashikawa, A. Ogawa, “Improving speaker-attribute estimation by voting based on speaker cluster information,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020 to appear.

[2] N. Tawara, H. Kamiyama, S. Kobashikawa, A. Ogawa, “話者クラスタリングに基づく話者年齢・性別推定精度改善法,” *日本音響学会研究発表会講演論文集*, pp. 815-816, 2019.

連 絡 先

俵 直弘（Naohiro Tawara） 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