研究講演

聞きたい人の声に耳を傾けるAI

～深層学習に基づく音声の選択的聴取技術SpeakerBeam～

Developing AI that pays attention to who you want to listen to －Deep learning based selective hearing with SpeakerBeam－

メディア情報研究部 信号処理研究グループ

マーク　デルクロア

Marc Delcroix

●プロフィール

NTT コミュニケーション科学基礎研究所　メディア情報研究部　特別研究員。2008年 北海道大学　大学院情報科学　情報科学研究科　博士課程修了。博士（情報科学）。2010年にNTTに入社以来、音声強調、音声認識、目的話者抽出等の音声・音響信号処理の研究に従事。IEEE、日本音響学会の各会員。

人は、パーティ会場などの騒がしい環境の中でも、聞きたい人（目的話者）の手がかり（声の特徴、話している内容など）に注目してその人の声を聞き取ることができる、選択的聴取の能力を持っています。我々は、この選択的聴取をコンピュータ上で実現するため、長年にわたって研究を進めてきました。しかし例えば、複数人が同時に話す状況では、互いに似た特徴を持つ音声同士が混ざるため、その中から聞きたい話者の声を取り出すことは難しい課題です。これに対する従来技術として、混ざった音声を各話者の音声へ分離する音源分離技術（Blind Source Separation: BSS）があり、近年、高品質な分離が実現できるようになってきました。しかしBSSは（1）混合音声に含まれる話者数に関する事前知識もしくはその推定が必要（2）各分離音声と各話者との対応関係が不定なため、どの分離音声が目的話者の音声かが不明、といった制約があり、様々な応用先で利用する際の課題となっていました。

目的話者抽出

BSSに代わる新たな枠組みとして、混合音声から聞きたい話者の音声のみを抽出する、目的話者抽出技術が最近注目を浴びています。目的話者抽出は、聞きたい話者の手がかりを補助情報として活用し、混合音声の中からその話者の音声のみを抽出します[1、2]。話者の手がかりとしては、例えば目的話者の声の特徴や唇の映像データなどが考えられます。目的話者抽出は混合音声中の話者数に依らず目的話者音声のみを抽出できますし、抽出音声と話者の対応関係も明らかなため、BSSが持つ課題を回避できます。

ニューラルネットワーク（NN）による目的話者抽出 SpeakerBeam

我々は、目的話者抽出技術として、（Neural network:NN）を用いた新技術 SpeakerBeam[1、2]を提案しました（図1）。SpeakerBeamの特徴は、NNの挙動を制御するために、目的話者に関する何らかの手がかりを与える仕組みを導入したことにあります。SpeakerBeam は、図1にあるように、10秒程度の事前録音された目的話者音声からその特徴量を抽出するNN（①話者特徴抽出NN）と、抽出した特徴量を補助入力として混合音声から目的話者の音声を抽出するNN（②目的話者抽出NN）、の二つのNNによって構成されています。SpeakerBeamは、目的話者の手がかり（声の特徴）に基づく目的話者抽出、すなわち選択的聴取を、世界で初めて実現した手法です。ここで目的話者の手がかりとしては声の特徴以外にも様々なものが考えられるため、我々はさらにSpeakerBeamを複数の手がかりを選択的に利用可能なマルチモーダル（Multi-modal:MM）-SpeakerBeamも提案しました[3]。MM-SpeakerBeamでは、例えば音声と映像による複数モダリティの手がかりを活用することで、互いに似た声の話者の混合音声からの目的話者抽出性能を大きく向上することに成功しました。

SpeakerBeamの枠組みは、目的話者抽出の問題以外にも、様々な場面でもその応用が研究されています。実際、SpeakerBeamの登場を受けて、（1）目的話者の手がかりに基づいてその人の発話区間を推定する（目的話者発話区間推定）問題[4]や、（2）信号の抽出を経ずに目的話者の手がかりに基づいてその人の発話内容を直接的に推定する（目的話者音声認識）問題[5]などの新たな試みがなされています。

今後の展開

SpeakerBeamの応用先としては例えば、目的話者の声を聞き取りやすくする補聴器やボイスレコーダ、特定の人のみに反応するスマートデバイス、会議音声の議事録システムなど、様々なものが考えられます。さらに最近の進展として、音声だけでなく任意の音を抽出できる、ユニバーサル音抽出の研究にも取り組んでいます[6]（図2）。これは、音声や映像に基づく話者情報ではなく、聞きたい音の種類に関する手がかりを用いることで、該当する種類の音のみを抽出する枠組みです。この技術により、例えば、消防車の音と女性の声に注意し、犬の声や他の音を無視できるような未来の音声デバイスを実現できるようになると期待されます。また、人間は音や映像といった手がかりの他にも、話している内容（概念）といったより抽象度の高い手がかりに基づいて、自身の聞きたい会話に注目してその人の声を聞き取る能力を持っています。音や映像のような具体性のある手がかりを超えて、そうしたより抽象度の高い手がかりをも扱えるように拡張することができれば、我々の長年の研究目標である人間の選択的聴取の能力の実現により近づいていけるものと考えています。

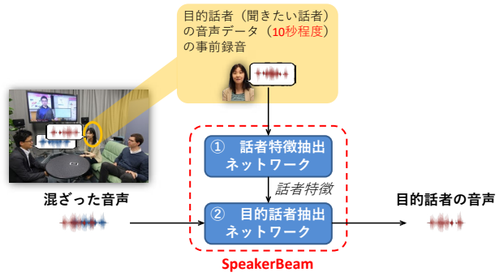

図1：目的話者抽出技術 SpeakerBeam の仕組み

図2：音声以外の音も扱えるユニバーサルサウンド抽出

●関連文献

[1] M. Delcroix, K. Zmolikova, 木下 慶介, 荒木 章子, 小川 厚徳, 中谷 智広, “SpeakerBeam: 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”, *NTT技術ジャーナル*, vol. 30, no. 9, pp. 12-15, 2018.
[2] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Cernocky, “SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.
[3] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, “Multimodal SpeakerBeam: Single channel target speech extraction with audio-visual speaker clues,” *in Proc. Interspeech*, 2019.
[4] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, and A. Romanenko, “Target-Speaker Voice Activity Detection: A Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario,” *in Proc. Interspeech*, 2020.
[5] M. Delcroix, S. Watanabe, T. Ochiai, K. Kinoshita, S. Karita, A. Ogawa, and T. Nakatani, “End-to-end speakerbeam for single channel target speech recognition,” *in Proc. Interspeech*, 2019.
[6] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, “Listen to what you want: Neural network-based universal sound selector,” *in Proc. Interspeech*, pp. 2718–2722, 2020.
[7] Y. Ohishi, A. Kimura, T. Kawanishi, K. Kashino, D. Harwath, and J. Glass, “Pair Expansion for Learning Multilingual Semantic Embeddings Using Disjoint Visually-Grounded Speech Audio Datasets,” *in Proc. Interspeech*, 2020.