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

Marc Delcroix 落合 翼
佐藤 宏 大石 康智
木下 慶介 中谷 智広
荒木 章子
NTTコミュニケーション科学基礎研究所

人は，騒がしい環境の中でも，聞きたい人（目的話者）の声の特徴などの手掛かり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」と呼ばれる能力を持っています．本稿では，この選択的聴取をコンピュータ上で実現することをめざした研究として，目的話者の音声の特徴を示す手掛かりに基づき，混ざった音声の中から目的話者の音声のみを抽出する技術SpeakerBeamについて紹介します．

はじめに

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）の手掛かり（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．私たちは，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．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どうしが混ざるため，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であるかが不明，といった制約があり，さまざまな応用先で利用する際の課題となっていました．

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されています．目的話者抽出は，聞きたい話者の手掛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)〜(3)]．話者の手掛かりとしては，例えば，目的話者の発話から推定された声の特徴（音響手掛かり）や唇の映像データ（映像手掛かり）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よらずに目的話者の音声のみを抽出することができ，また抽出音声と話者の対応関係も明らかであるため，BSSが持つ課題を回避することが可能です．

本稿では，以前紹介した目的話者の声の特徴に基づく目的話者抽出SpeakerBeam[(1), (2)]をレビューし，混合音声の中の話者の声が似ている際に抽出が難しくなるSpeakerBeamの問題点について実験結果を示しながら説明します．次に，この問題を回避する方法の1つとして，マルチモーダル（Multimodal:MM）SpeakerBeamを紹介します．最後に，目的話者抽出技術の他の音声処理タスクへの拡張と，さらに人間の選択的聴取の能力に近づけるための今後の取り組みについて述べます．

ニューラルネットワークによる目的話者抽出SpeakerBeam

私たちは，目的話者抽出技術として，ニューラルネットワーク（Neural Network:NN）を用いた新技術SpeakerBeam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掛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，事前録音された10秒程度の目的話者の音声（音響手掛かり）からその声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

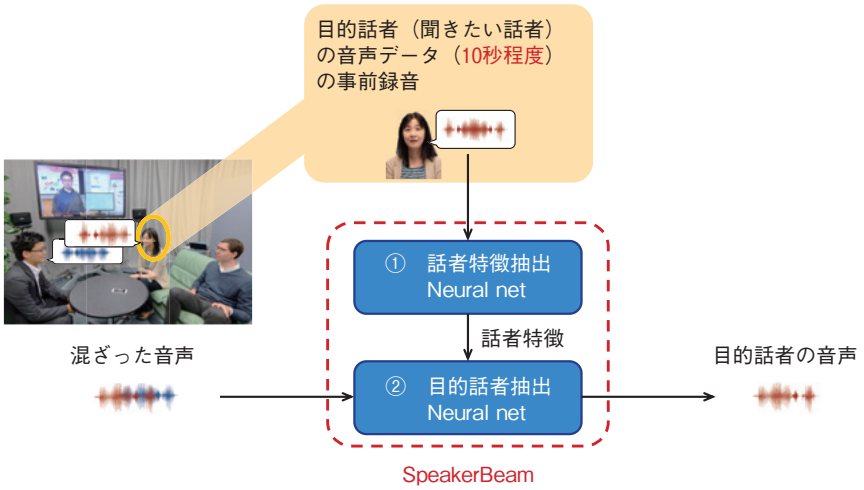

図 1　目的話者抽出技術 SpeakerBeam の仕組み

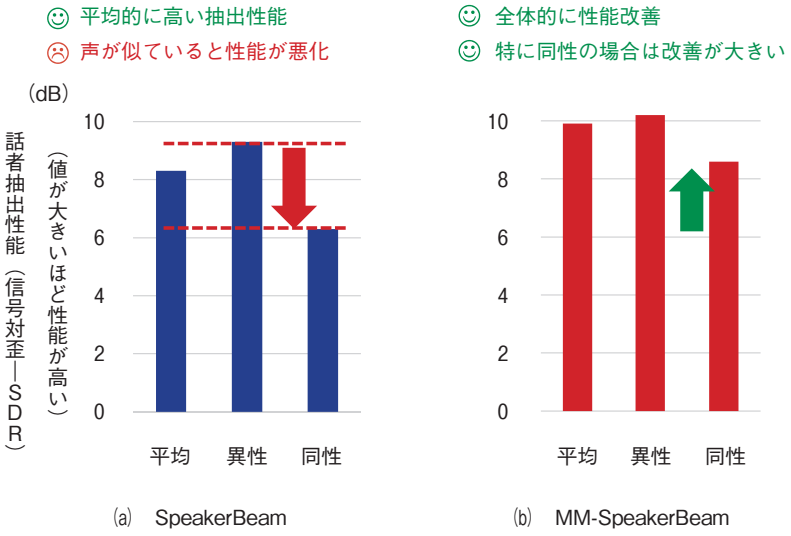

図 2　SpeakerBeam の性能（２人の話者の英語音声を混ぜた実験）

NN），の２つのNNによって構成されています．SpeakerBeamは，目的話者の手掛かり（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．

SpeakerBeamの性能を評価するために，英語発話の話者２人の混合音声を用いた，目的話者抽出実験を行いました．その実験結果を**図 2** (a)に示します．目的話者抽出性能を測るための評価指標としては，信号対歪み比（Signal-to-Distortion Ratio：SDR）を採用しました．SDRは，その数値が高いほど，抽出性能が高いことを示します．図から，SpeakerBeamは平均して 8 dB以上の高い抽出性能を達成していることが確認できます．一方，この結果を同性どうしと異性どうしの混合に分析してみると，同性どうしの性能は大きく劣化することが分かります．これは，同性の混合音声では，話者の声の特徴が互いに似通っていて，目的話者の特定や抽出を行うことが困難になる場合があるためです．この問題に対処するための，方向性の１つとしては，声の特徴に依存しない，音響手掛かりとは別の手掛かりを用いることが考えられます．

MM-SpeakerBeam

これまでにも，音響手掛かり（声の特徴）以外の，映像情報を手掛かりとした話者抽出はいくつか検討されています．例えば，参考文献(3)では，ビデオカメラを利用して，話している目的話者の顔や唇の動きを録画し，映像特徴抽出NN（例えばFaceNetという学習済みの顔認識ネットワーク）を使って，その顔画像の時系列から唇の動きを表す時系列特徴量を抽出します．目的話者抽出NNは，これを映像の手掛かりとして利用し，唇の動きに合致した音声を抽出します(3)．映像情報を手掛かりに用いた目的話者抽出は，似た声の人どうしであっても抽出が可能であると期待されます．例えば，極端な例では，同じ人が別の発話を話している混合音声でも，唇の動きが違っていれば，目的音声の抽出が可能であるとの報告もあります(3)．このように，映像情報を使う目的話者抽出は，声が似ている場合にSpeakerBeamの抽出性能が劣化する問題に対処する１つの方法です．一方で，実際の映像収録状況では，顔や唇が隠れていることはよくあり，このような状況では映像情報の手掛かりは役に立たないという問題があります．

そこで私たちは，音響手掛かりと映

像手掛かりに基づく方法のそれぞれの利点を活用するため，SpeakerBeamを複数の手掛かりを選択的に利用可能なかたちに拡張したMM-SpeakerBeamを提案しています[4]，[5]．MM-SpeakerBeamの仕組みを**図3**に示します．MM-SpeakerBeamはこの図のように，音響の手掛かりに加え，映像の手掛かりも利用する方法になります．映像の手掛かりは，参考文献⑶と同様に，ビデオカメラを用いて録画した目的話者の顔（唇の動き）の時系列になります．MM-SpeakerBeamは，複数の手掛かりから，信頼度に応じて手掛かりの選択ができる仕組みを導入しています．これを実現するために，ニューラル機械翻訳をはじめとしたさまざまな分野で，近年よく使われているAttention機構を利用しました．また，図にある目的話者抽出NNは従来のSpeakerBeamと同様のNNです．

MM-SpeakerBeamの最大の特長は，片方の手掛かりが役に立たない場合でも，他の手掛かりを使うことによって，高精度な目的話者抽出が可能になることです．例えば，声の性質が似た話者のときには，映像の手掛かりを主に利用し，逆に，唇が画面に映らない場合には，音響の手掛かりを主に利用するようになります．このように，その時々で利用できる情報源を活用することで，多様な状況でもより安定に，より高性能に動作します．

MM-SpeakerBeamの抽出結果を図2⑹に示します．マルチモーダル手掛かりを使うことによって，性能が大幅に改善していること，特に同性の場合に大きな改善が得られることを確認できます．このように，MM-SpeakerBeamは，複数モダリティ（例えば音声と映像による）の手掛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が大きく向上し，安定した抽出を可能とします．デモサイト[6]にて実際の処理音声の例を聞くことができます．

他の音声処理タスクへの拡張

SpeakerBeamの枠組みは，目的話者抽出の問題以外にも，さまざまな場面でその応用が研究されています．

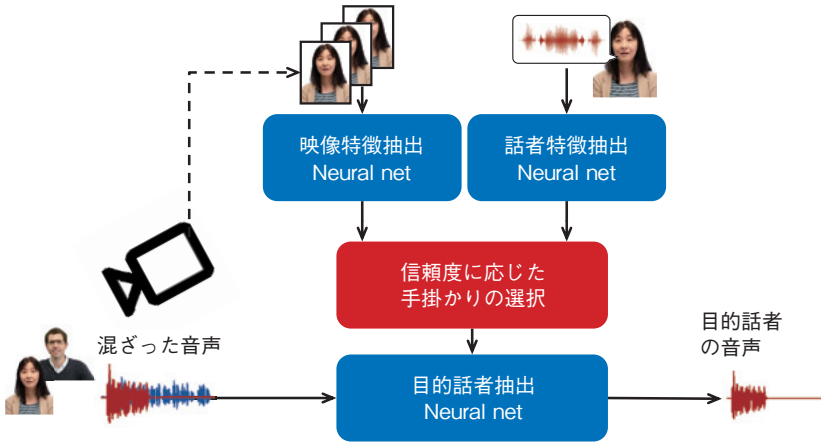

図3　MM-SpeakerBeamの仕組み

実際，SpeakerBeamの登場を受けて，①目的話者の手掛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[7]や，②信号の抽出を経ずに目的話者の手掛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[8]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，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，さまざま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[9]（**図4**）．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掛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．この技術のデモ音は，デモサイト[10]にて聞くことができます．

また，人間は音や映像といった手掛かりのほかにも，自身の聞きたい「話題」に注目し，その人の声を聞き取る能力を持っています．すなわち人間は，話している内容や概念といった，より抽象度の高い手掛かりに基づく選択的聴取もできるのです．SpeakerBeam技術を，より抽象度の高い手掛かりをも扱えるように拡張することができれば，私たちの長年の研究目標である人

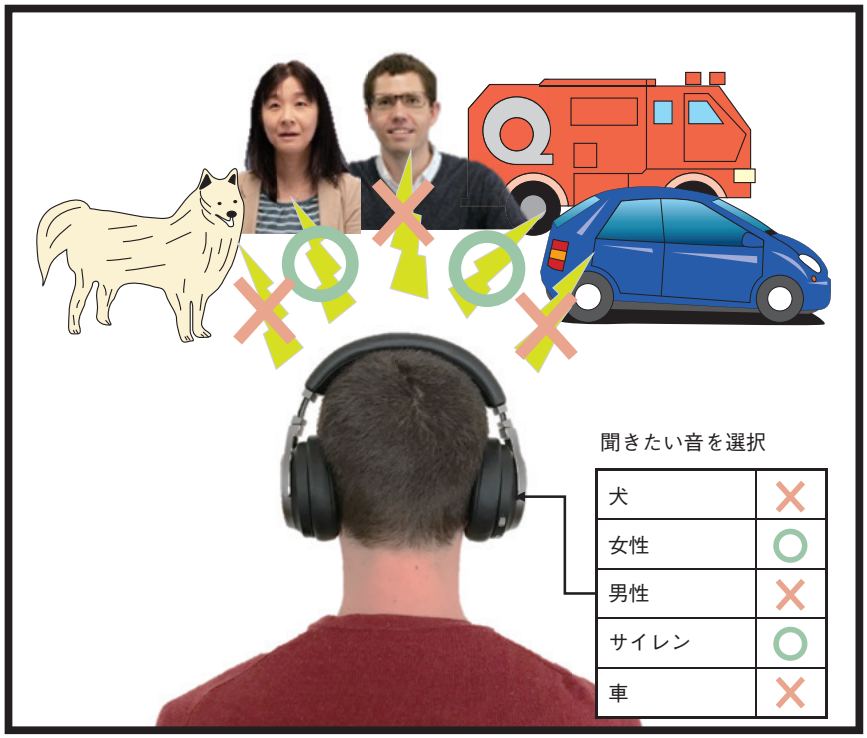

図 4　音声以外の音も扱えるユニバーサルサウンド抽出

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．

これを実現するためには，①どうやって概念を表現・獲得するか，②概念を表現できたとしても，どうやってそれを使って，聞きたい会話を抽出するか，の２つの研究課題があります．①については，すでに音声と映像から内容・概念を獲得する研究を行っています[11]．②については，今度の研究課題の１つとして取り組んでいきたいと考えています．

■参考文献

(1) Delcroix・Zmolikova・木下・荒木・小川・中谷："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 JSTSP, Vol. 13, No. 4, pp. 800–814, 2019.
(3) A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein："Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation，" ACM Trans. Graph. , Vol. 37, No. 4, Article 112, pp.1-11, August 2018.
(4) T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani："Multimodal SpeakerBeam: Single channel target speech extraction with audio-visual speaker clues，" Proc. of INTERSPEECH 2019, Graz, Austria, Sept. 2019.
(5) H. Sato, T. Ochiai, K. Kinoshita, M. Delcroix, T. Nakatani, and S. Araki："Multimodal Attention Fusion for Target Speaker Extraction，" Proc. of SLT 2021, pp. 778-784, 2021.
(6) http://www.kecl.ntt.co.jp/icl/signal/member/demo/audio_visual_speakerbeam.html
(7) 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，" Proc. of INTERSPEECH 2020, Shanghai, China, Oct. 2020.
(8) 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，" Proc. of INTERSPEECH 2019, Graz, Austria, Sept. 2019.
(9) T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki："Listen to what you want: Neural network-based universal sound selector，" Proc. of INTERSPEECH 2020, Shanghai, China, Oct. 2020.
(10) http://www.kecl.ntt.co.jp/icl/signal/member/tochiai/demos/universal_sound_selector/index.html
(11) 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，" Proc. of INTERSPEECH 2020, Shanghai, China, Oct. 2020.

（上段左から）Marc Delcroix/ 落合　翼/ 佐藤　宏/ 大石 康智
（下段左から）木下 慶介/ 中谷 智広/ 荒木 章子

私たちは，ロボットやコンピュータなどが，人間と同様に私たちの会話を理解できるように，日々研究を進めています．目的話者抽出は，それを達成するための重要な要素技術と考えています．SpeakerBeamをもっと発展させ，人間の選択的聴取の能力により近づけていきたいと考えています．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
メディア情報処理研究部
信号処理研究グループ
TEL　0774-93-5030
FAX　0774-93-5026
E-mail　cs-liaison-ml＠hco.ntt.co.jp