深層学習　目的話者抽出　SpeakerBeam

# 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

Marc Delcroix[†1] /Katerina Zmolikova[†2]
きのした　けいすけ[†1]　あらき　しょうこ[†1]
木下　慶介　/荒木　章子
おがわ　あつのり[†1]　なかたに　ともひろ[†1]
小川　厚徳　/中谷　智広

NTTコミュニケーション科学基礎研究所[†1]
Brno University of Technology[†2]

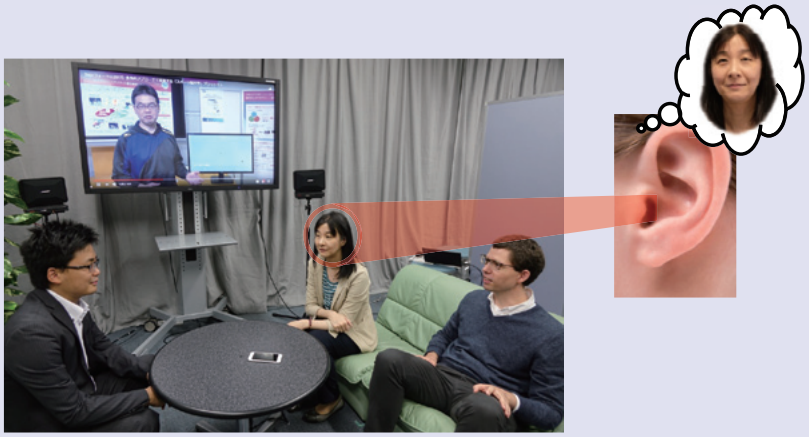
**パーティ会場などの騒がしい環境の中でも，人は，聞きたい人（目的話者）の声に注目して，その声を聞き取ること（選択的聴取）ができます．一方，従来のコンピュータでも，話者の位置が分かっていれば，その位置から来る音だけを抽出することはできました．これに対し，本稿では，目的話者の声の特徴だけが分かっているときに，深層学習技術を用いて，その特徴に合致する声を抽出する新しい技術SpeakerBeamを紹介します．**

## はじめに

近年，コンピュータによる自動音声認識技術が急速に発展し，スマートフォンやスマートスピーカなどの音声インタフェースで利用されるようになってきました．しかし，日常のさまざまな場面では，複数の人が会話をしていたり，TVの音声が背景で流れていたりするなど，目的話者以外の声が混ざって収録されることが，しばしば起きます．現在の音声認識技術では，目的話者だけに注目してその声を聞き取ることができないため，このような状況にうまく対応することができませんでした．

一方，音声認識技術と違い，人間の聴覚は，複数の人の声やほかの音が聞こえている状況でも，目的話者の声の特徴（声の高さ，声質，抑揚，強勢，音長，リズムなど）やその音の到来方向に注目して，その他の音を無視して，目的の音声だけを聞き取ることができます（**図1**）．これは，人間の聴覚の優れた能力の1つで，選択的聴取と呼ばれます．この能力のおかげで，騒がしい環境の中でも，人間は，特定の話し相手と話を続けることができます．これまで，選択的聴取の能力のうち，話者の位置に注目して声を聞き取る技術は，すでにコンピュータでも実現されていました[(1), (2)]．しかし，これらの従来技術でも，目的話者の位置が不明であったり，話者の位置が動いたりする場合には，目的話者の声の抽出は困難でした．このため，従来技術で対処できる状況は，大きく制限されていました．

これに対し，私たちは，人間の聴覚による選択的聴取の能力のうち，目的話者の声の特徴に基づき，その声だけを聞き取る能力に関して，それと同等の機能を実現できる技術を，コンピュータを用いて世界で初めて実現しました（**図2**）．この技術を，SpeakerBeamと呼びます．SpeakerBeamでは，複数の人の声の中でどれが目的話者の声であるかを定めるために，事前に収録した約10秒程度の目的話者の声（適応発話）を手掛かりとして利用します．そして，適応発話から抽出した声の特徴に基づき，その特徴に合致した音声だけを取り出すことで，目的話者の音声を抽出します．この方法を用いることで，目的話者がどの場所で話すか分か



聞きたい人の声の特徴（声の高さ，声質など）と音の到来方向に注目し，その声のみを聞くことができる

図1　人間の聴覚による声の選択的聴取

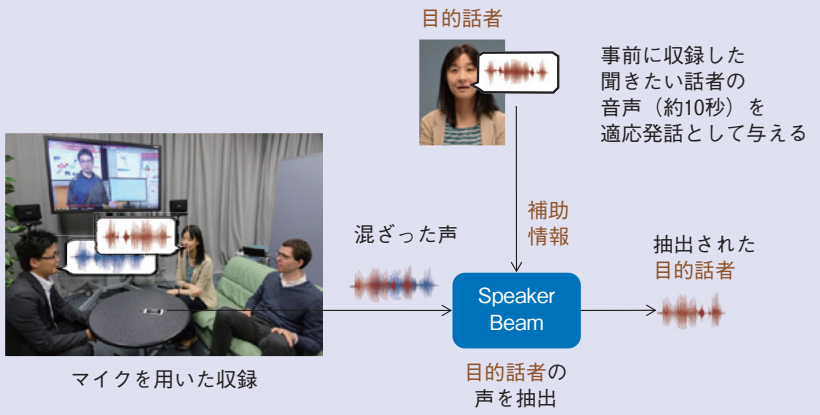



図 2　SpeakerBeamによる声の選択的聴取

らないような状況などでも，その声の特徴に注目することで，目的話者の声を抽出することができるようになります．このため，SpeakerBeamは，今後，人の会話を理解する音声認識・ロボット技術の新しい可能性を拓く技術になると期待されます．

本稿では，選択的聴取と関連の深い従来技術として音源分離について最初に簡単に紹介したのちに，音源分離との対比で，SpeakerBeamの特長を説明します．さらに，現在のSpeakerBeamの性能を実験結果を用いて示します．最後に，SpeakerBeamの将来の利用シーンを概観するとともに，今後の研究の方向性について述べます．

## 選択的聴取と関連の深い従来技術：音源分離

コンピュータを用いて騒がしい環境の中で目的話者の声を抽出する選択的聴取の実現のために，これまで，数多くの研究が行われてきました．しかし，その中で，実際に行われてきた研究のほとんどは，混ざった音声を個々の音声に分解する，音源分離に関するものでした[(1), (2)]．音源分離は，選択的聴取とよく似ているけれども異なる機能を実現する技術です．音源分離では，収録音に含まれている話者の数が既知であるとの前提の下で，何らかの音の特徴（音の到来方向など）に基づき，収録音を話者数と同じ数の音に分解します．音源分離は，すべての音を取り出せる利点がある一方で，それを実現するためには，すべての音の詳細な状態を推定する必要があるという問題がありました．具体的には，例えば，何人の話者が同時に話しているかの数の情報が必要，すべての話者の位置や雑音の統計量の推定が必要，分離音のどれが目的話者かの推定が必要などの課題がありました．このため，現時点で，音源分離の適用範囲は必ずしも大きくありません．

## SpeakerBeamの特長

すべての音を分解する音源分離のアプローチと違い，SpeakerBeam では，目的話者の音声のみを抽出します．そのため，SpeakerBeamでは，同時に話している話者の数や，その位置，また背景雑音の統計量を推定する必要がありません．収録音の中で，目的話者以外にどのような音が含まれているかによらず，目的話者の声の特徴に合致する音声に注目して，その音を抽出するというシンプルな処理を行うだけで，選択的聴取を実現することができます．しかも，10秒程度の発話を適応発話として与えるだけで目的話者の声の特徴を得ることができるため，現実の多くの収録環境で，目的話者の抽出を行うことができます．

SpeakerBeamのためのニューラルネットワークの構造を**図 3** に示します．主ネットワークと，補助ネットワー

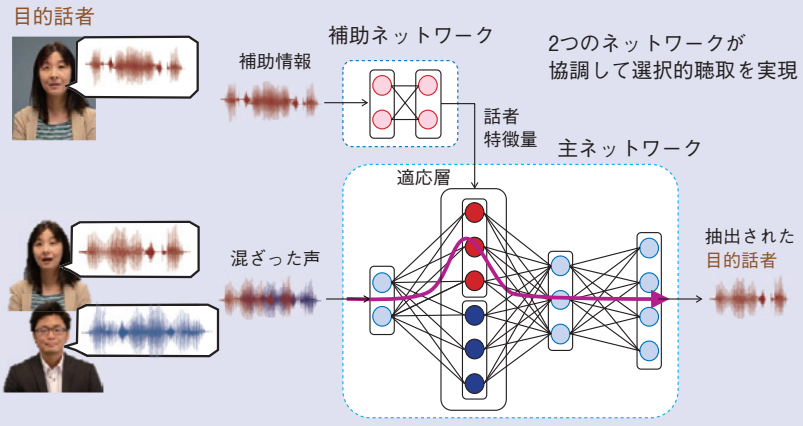



図 3　SpeakerBeamのための深層学習の新技術

クの 2 つで構成されており，それぞれ，以下の機能を担当します．

① 主ネットワーク：さまざまな音が混ざった入力音声を受け取り，その中に含まれる目的話者の音声を抽出して出力します．多層のネットワークで構成されており，その中に，適応層[3),(4)]と呼ばれる特別な層を含みます．適応層は，ネットワークの制御情報として，補助ネットワークが抽出した目的話者の声の特徴を受け取り，その特徴に合わせて，ネットワーク全体として目的話者の声の抽出ができるように，適応層の中のネットワークの重みを修正できる仕組みを持っています．

② 補助ネットワーク：入力音声とは別に収録した目的話者の声を適応発話として受け取り，多層のネットワークを用いて，その声の特徴を抽出して出力します．

これら 2 つのネットワークは，相互に結合された状態で，音声抽出の精度を最大化するように学習されます．主ネットワークと同時に補助ネットワークを学習することで，目的話者の音声を抽出するために必要な音声の特徴が何であるかを，データから自動的に学習できます．その結果，最適な特徴を，人手で見つけ出すというような面倒な作業を行う必要がありません．さらに，さまざまな話者の声や背景雑音を含んだ大量の学習データを用いてネットワークを学習することで，SpeakerBeam は，学習に含まれていない目的話者に対しても，選択的聴取が行えるようになります．ネットワークの構成と学習の手続きに関する詳細は，文献⑸でご確認いただけます．

## SpeakerBeamによる音質改善と音声認識精度の改善

SpeakerBeamによる，音声抽出性能と音声認識に与える影響を実験により確認しました．英語新聞の読み上げ音声からなる音声データベースを用いて，2 人の話者をコンピュータ上で混ぜた音声を作成し，入力音として用いました．SpeakerBeamは 1 つのマイクでも動作しますが，より多くのマイクを用いる場合のほうがより高い性能を実現できます．実験では，8 個のマイクを用いて，SpeakerBeamによる処理とマイクアレイによる処理（ビームフォーマ処理）を組み合わせて目的音声の抽出を行いました．

SpeakerBeamで処理した結果の音声波形の例と，SpeakerBeamを用いたときと用いないときの音声認識精度を**図 4** に示します．波形の例より，SpeakerBeamは，適切に目的話者を抽出できていることが確認できます．また，音声認識結果より，SpeakerBeamにより，約60％の音声認識精度改善が得られたことが確認できます．

さらに，SpeakerBeam は人が音声を聞く場合の聴感上の音声品質を改善する目的でも利用することができます．実際の環境（残響や背景音楽がある環境での実録音）で，SpeakerBeam

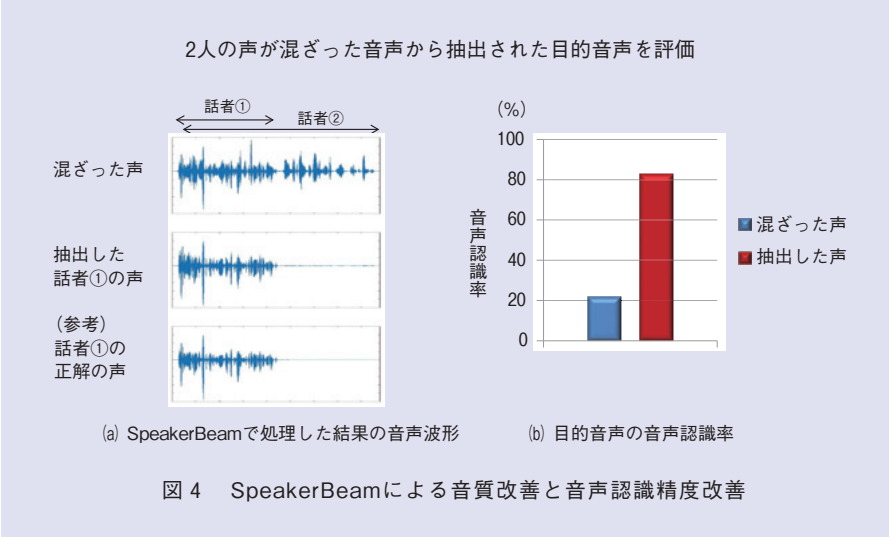



図 4　SpeakerBeamによる音質改善と音声認識精度改善

が，どのように目的話者の声の抽出を実現できるかについては，ビデオ[6]でご確認いただけます．

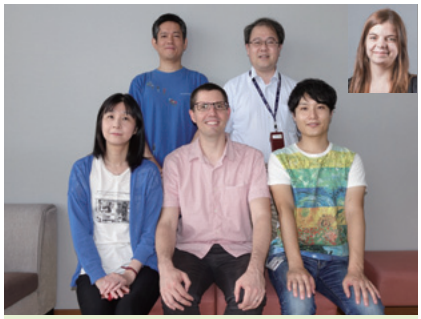
## 今後の展開

SpeakerBeamはコンピュータによる選択的聴取を実現する新しい技術です．従来から広く研究が行われてきた音源分離の技術と比べて，背景でどんな音が聞こえているか（同時に話している人の数やその位置，背景雑音の種類など）によらず，目的話者の声の抽出を行うことができます．今後，この技術は，多人数会話の音声認識，スマートスピーカのようなアシスタントデバイスの音声インタフェース，目的話者の声のみを抽出できるICレコーダや補聴器など，さまざまな新しい音声インタフェース技術を実現する要素技術として利用が期待されます．

SpeakerBeamを実際の状況で利用できるようにするためには，まだいくつかの課題が残されています．例えば，現在の技術では，声の似た人が同時に話す場合に，話者抽出性能が低下する問題があります．この問題を解決するために，より正確に話者の区別を行うことができる特徴抽出の方法，音の到来方向などの話者の位置情報を，声の特徴と組み合わせて目的話者抽出を行う方法などについて，研究を進めていく予定です．

### ■参考文献


⑴ 堀・荒木・中谷・中村：“みんなの会話を聞き取るコンピュータを目指して，” NTT技術ジャーナル，Vol.25，No.9，pp.18-21，2013.
⑵ 牧野・荒木・向井・澤田：“ブラインドな処理が可能な音源分離技術，” NTT技術ジャーナル，Vol.15，No.12，pp.8-12，2003.
⑶ Delcroix・木下・小川・苅田・樋口・中谷：“コンテキスト適応型ニューラルネットワークを用いた音声インタフェースのパーソナライズ化，” NTT技術ジャーナル，Vol.29，No.9，pp.25-28，2017.
⑷ M. Delcroix, K. Kinoshita, A. Ogawa, C. Huemmer, and T. Nakatani：“Context Adaptive Neural Network Based Acoustic Models for Rapid Adaptation，” IEEE/ACM Trans. Audio, Speech and Lang. Proc., Vol.26, No.5, pp.895-908, 2018.
⑸ K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani：“Learning speaker representation for neural network based multichannel speaker extraction，” Proc. of ASRU 2017, Okinawa, Japan, Dec. 2017.
⑹ https://youtu.be/BM0DXWgGY5A


（後列左から）小川 厚徳/ 中谷 智広/ Katerina Zmolikova（右上）
（前列左から）荒木 章子/ Marc Delcroix/ 木下 慶介


ロボットやコンピュータなどが，今までより自然に，私たちの音声を理解できるようになるよう，日々研究進めています．その要素技術として，SpeakerBeamが，さまざまな話者や雑音環境に対応できるよう検討を進めていきます．是非ビデオ[6]も見てください！

### ◆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