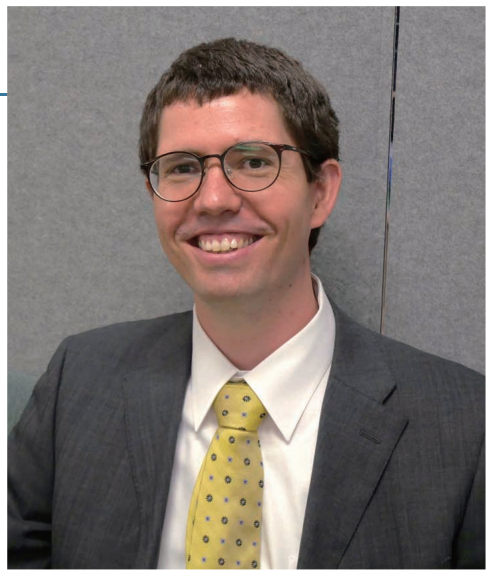
URL https://journal.ntt.co.jp/article/34173
DOI https://doi.org/10.60249/25061301

NTTコミュニケーション科学基礎研究所
特別研究員

マーク・デルクロア Marc Delcroix



ニューラルネットワークを活用し，コンピュータでも人間のように音声の選択的聴取を可能にする

私たちはたくさんの音に囲まれて生活しています．人間は複数の人が同時に話しても特定の声を聴き分ける「選択的聴取」ができますが，コンピュータでは困難でした．もしコンピュータで特定の音や声だけを抽出できれば，雑音を消し，呼び鈴や電話の音だけをヘッドホン越しに聞くことができるなど，在宅ワークの方や視覚や聴覚のサポートが必要な方など多方面への貢献につながります．今回は複数の話者の中から特定の声を抽出する技術の開発に成功したマーク・デルクロア特別研究員にお話を伺いました．

◆PROFILE：2007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年日本電信電話株式会社入社．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．

複数人の会話音声の中から特定の人の声だけを抽出する技術

■どのような研究をされているのか教えてください．

大きな研究テーマとしては，複数の音の中から特定の音だけを抽出する「音声の選択的聴取」の研究をしています．私たち人間は，さまざまな場面において「音」から情報を収集しています．また，PCやスマートフォン，ボイスレコーダなどに話し掛け，録音して音声の文字起こしをすることもあると思います．私たちの研究グループでは，ユーザ1人のみの話を理解するだけではなく，人々の自然な会話の中で，コンピュータが人のコミュニケーションを理解し，サポートしてくれるような技術を実現したいと考えています．

私たちの日常の生活ではいろいろな音が頻繁に飛び交っています．その中には聞きたい音もあれば，聞かなくてもいい音も含まれていますが，すべての音が耳に入ってきます．例えば，オンラインで会議をしているときに，電話や外のサイレンの音，犬の鳴き声や家族の生活音など，そのときは聞こえなくてもいい不要な音が聞こえてくることがありますが，通常私たちは会議音声だけを聞きたいものです．このような場合でも，私たち人間は，基本的に聞きたい人や音に注目して，必要な音を優先して聞くことができます．これは選択的聴取と呼ばれていますが，私はその能力をコンピュータでも再現したいと考え，複数人が話している中で聞きたい人の声だけを取り出す技術の開発に取り組んでいます．

この技術はいろいろな技術と組み合わせて，さまざまなユースケースに応用することができます．例えば音声認識技術と組み合わせることで，複数の話者の声の中から1人の声だけを抽出し，その人が話した言葉だけを文字にすることができます．さらに，こうしたユースケースのほかにも，多くの音が鳴っている中で人の声に限らず電話や犬の声など，聞きたい音だけを抽出する技術にも拡張しています．

私はこの研究を2017年ごろから行っており，「声の特徴に基づいて，複数人が話している中で特定の話者の声を抽出する」ことに成功しました．例えば2人で話している音声のデータから特定の話者の音声を抽出する場合，まず聞きたい話者（目的話者）の音声を事前に録音し，ニューラルネットワークを用いてその話者の「音声の特徴」を分析し学習します．そして複数人の音声が混ざったデータから，このニューラルネットワークを使って，目的話者の声の特徴に合致した音声を抽出します．

これは複数人の声や雑音が混ざった音の中から，声の特徴に基づいて目的話者の声だけを抽出できる「SpeakerBeam」という技術です（**図1**）．この技術の特長としては，何人が話しているのか分からないような場合にも適用することが可能です．すなわち，話者が2人ではなく3人以上いた場合でも，その混合音声の中の特定の人の声を抽出することが可能です．

また，この技術では，どのような音声でも平均的に良い精度で抽出できている一方で，女性どうしや男性どうしなどの似たような音声の場合に性能が落ちることも判明しています．開発当初から技術は進化しているため，このような場合でもある程度の性能改善はできているのですが，どうしても声が似ていると区別が難

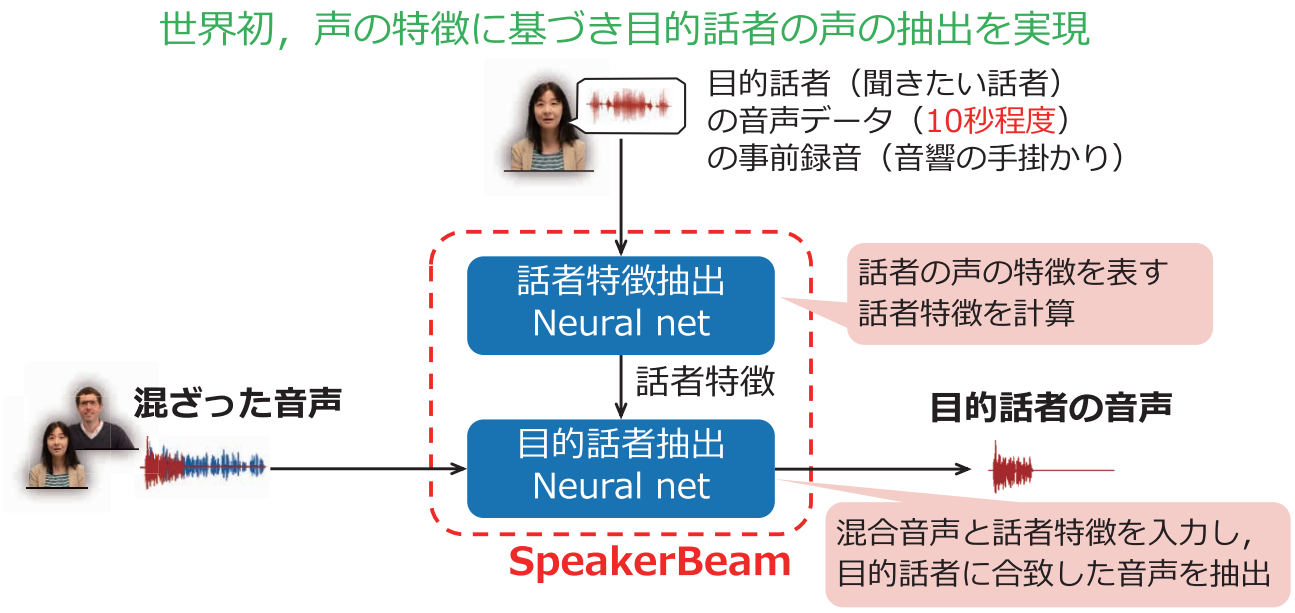


図1　声の手掛かりに基づく話者抽出SpeakerBeam

しくなります．その打開策として，録音した音声データだけではなく撮影した動画データも利用し，唇の動きも手掛かりにするマルチモーダルバージョンも提案しています．これにより似たような音声が混ざっている場合でも抽出したい話者の声を安定して抽出できるようになりました．

■ご自身の研究ならではの強みを教えてください．

私は元々，特定の話者の声に音声認識システムを適応させて，音声の文字書き起こしの性能を上げるという，音声認識の研究に取り組んでいました．私たちの研究グループは，音声認識のチームと音声強調（雑音除去や音源分離など）のチームで構成されています．私の音声認識における話者適応技術と音源分離のニューラルネットワーク技術を組み合わせれば，特定の人の声を抽出できるのではと考えて新たな研究を始め，SpeakerBeamを発明することができました．異なる研究分野の人が集まって同じグループで研究をしたからこそ世界初の技術が生まれたと考えています．

さて，私たちのSpeakerBeamは汎用性のある技術であり，SpeakerBeamの仕組みを使うことで，聞きたい音声だけではなく特定の（音声以外の）音も抽出できることも強みです．その強みを活かし「SoundBeam」という技術も提案しました．例えば在宅勤務のときに，外でサイレンが鳴っていたらその音は雑音になってしまいます．一方で，車を運転しているときにはサイレンは重要な音ですので聞こえるようにしなければなりません．このように状況に合わせて快適な選択的聴取を行うため，多くの音を学習して扱える音の種類を拡張させ，コンピュータに「このような場合にはこの音が聞きたい」と入力しさえすれば，その音は聞こえて他の音は消す，ということを実現しようとしています（**図2**）．

この技術が実現できれば多くの応用先があると考えており，例えばイヤホンやヘッドホン，補聴器などのヒアリングデバイスのノイズキャンセリング機能をカスタマイズすることができます．また，車の中で音楽を聞きながらクラクションやサイレンの音を強調させる，という設定も可能になります．ほかにも動画などの音声トラック編集で，強調したい音を抽出させて聞こえやすくするようなことも，ニーズがあるのではないかと思っています．

一般的にSoundBeamは，音の種類を特定することさえできれば，抽出できる精度は高くなるのですが，「学習した音の種類」しか抽出できないというデメリットがあります．例えば犬や猫の鳴き声を学習しておけば音声データから犬や猫の鳴き声を抽出できますが，未学習のライオンの鳴き声は抽出できません．これに対して私たちは，未学習の音を適宜追加していく技術も提案して

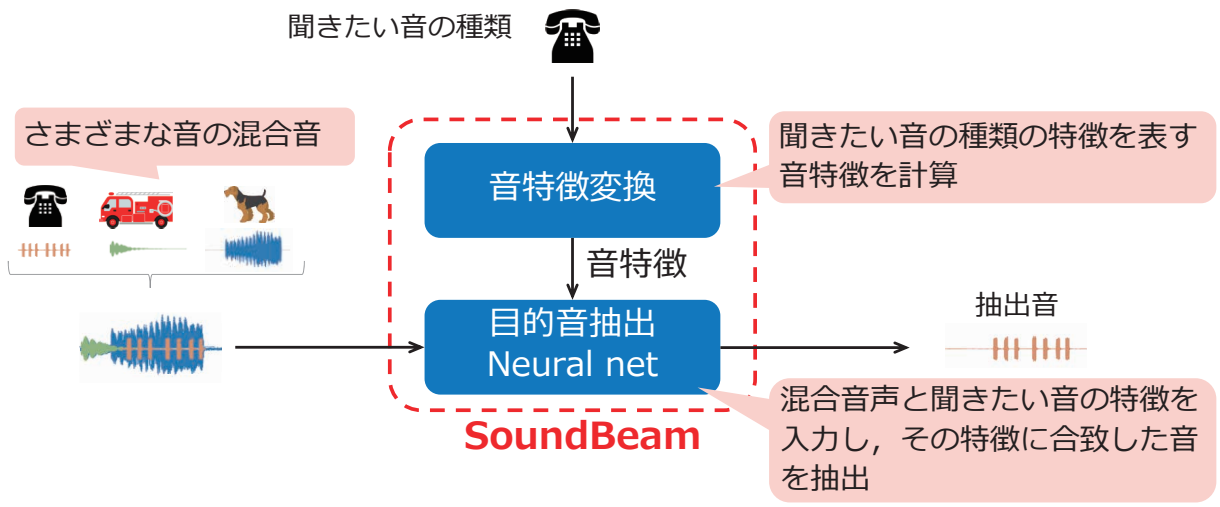


図 2　任意の音の選択的聴取 SoundBeam

います．この技術は，研究を始めてから 5 年も経っていない新しい研究領域なので，まだまだ発展の余地があると考え，研究に取り組んでいます．

SpeakerBeam や SoundBeam で雑音環境下でも快適に

■この研究の成果や，これからの展望を教えてください．

この研究が，人の生活に役に立つ研究になればと考えています．いろいろな音が飛び交っている中でも重要な音を選別し，その音で表現される情報をしっかりと認識できるようにする，ということが目標です．

例えば赤ちゃんが泣いているのは聞こえるけど，テレビの音は聞こえないようにしたり，電話やテレビ会議などで自分の声が伝わりやすくなったりすることで，日々の生活で負担に感じていたことが解消されより快適に暮らせるのではないかと思います．

生活以外の場面でも，例えば記者の方と話していると「ボイスレコーダでいろいろな会話や講演を録音した際に，聞きたい人の声だけを綺麗に抽出できたら楽なのに」という意見を聞きます．

近年では自動翻訳など多くの AI（人工知能）技術が生まれ日々進歩していますが，音が多く賑やかな環境で使うと，どうしても翻訳の精度が落ちてしまいます．もちろん各技術の性能は良くなっていますが，雑音と人の声は簡単に区別ができても「複数人の声」は背景雑音の特徴とは異なるため，システムが目的話者の音と複数人の声を区別することは必ずしも容易ではありません．しかし，私たちが研究している技術をうまく使うことができれば，将来的には，賑やかな場所で自動翻訳を使うときでも相手の声を綺麗に聞き取って正しく翻訳できたり，会議で誰が何を話したかを正確に区別して議事録を作成できるなど，さまざまな応用拡大の可能性があると考えています．

このように音を聞く多くの場面において，さまざまの人の役に立つにはどうすればよいのかを考えながら，音声抽出の精度向上に取り組んでいます．

■研究における課題やポイント，解決すべき問題を教えてください．

この技術の研究を始めた当初は，基本理論を確立させることがまずは重要でした．その後，実際に2018年のNTT コミュニケーション科学基礎研究所オープンハウスで私たちの音声抽出のデモンストレーションをすることになりました．そこでは，お客さま

がたくさん入って雑音も多いリアルな環境においてもうまく動作することが確認され，私たちの技術は理論上だけでなく実環境でも動作することが分かりました．

しかし，まだいくつかの課題が残っており，先ほど説明したように「扱える音の種類」を拡張することが，今後さらに必要になります．すでにある程度は拡張できたり，対応できたりすることは分かっているのですが，精度・性能向上のためにはもっと多くの音の種類を学習できるようにしていかなければならないと思っており，今後の課題の１つです．

ほかにも録音環境の問題もあると思います．静かな場所や，かなり賑やかな場所，特殊な音が鳴る場所など，環境がダイナミックに変わることを想定しなければなりません．実験環境に限らず，さまざまな録音環境でも正しく動作する技術にしなければ，実際に活用することはできません．そのため，録音環境によって自動的に処理を変える必要があることも考えて研究しています．例えば家ならばこういう音が想定されるし，オフィスであればまた別の音が想定される，といったことに対応するため，聞きたい音の種類に加えて，どのようなシーンなのかを条件付けられる技術も必要です．

また，声の抽出についても，私たちとしてはまだ品質に満足しているわけではなく，より高品質な音・リアルに近い音で抽出することもめざしています．

もう１つ，音声抽出の活用を想定するデバイスとして，イヤホンや補聴器が考えられますが，耳にはめるような小さいデバイスで私たちの技術を使うためには，ニューラルネットワークのパラメータ数も重要になってきます．近年ニューラルネットワークの性能は向上しているとはいえ，小さなデバイスの計算処理能力は限られているので，小さいデバイスでも高い精度での動作を保てるように技術を向上させていかなければと思っています．

■最後に，若き研究者・学生の方々へメッセージをお願いします．

研究には，理論の検討から実システムでの検証など，いろいろなフェーズがあります．私は，SpeakerBeamやSoundBeamにおいて，理論の創出とデモンストレーションシステムの開発の両方に携わりました．音声抽出の理論ができたこともうれしかったのですが，さらにそれが実環境でも理論どおりに動くことが分かったことは，格別の喜びがありました．このように研究では，世の中に対して理論から実践まで多様な貢献の仕方があり，いろいろな達成感が味わえます．

また，私はNTT研究所の同僚やインターン生，複数の大学との共同研究など，これまで多くの人と一緒に研究してきました．現在も，チェコやドイツ，米国などの大学と共同研究を行っています．海外の大学だけではなく，NTT内でもさまざまな研究所があり，音声を扱っている研究所はいくつもあるので，密接に連携しながら一緒に研究をしています．NTTの研究所は実装力も高く，前述のオープンハウスのデモンストレーションの対応も研究所のサポートで実現できました．研究も開発もできる能力の高い研究者と一緒に研究ができるのは，とても恵まれた環境だと感じています．

さらに，NTTは日本の中でもグローバルな環境で研究ができることも特徴です．私たちのグループもグローバルに研究活動をしており，海外からの優秀なインターン学生も多く来てくれますし，国際的にも有名で優秀な研究者が何人も在籍しています．この状況はすごいことだと思います．若い研究者や学生の皆さんには，困ったことがあっても１人で抱え込まないで，たくさんの人と話してほしいです．研究は１人で作業するイメージが強いように感じます．人によりますが，１人で研究して成功する人もいれば，いろいろな人と話して，いいアイデアが生まれて成功する人もいます．私としては悩んでいるときこそいろいろな人と話して，そのときに生まれた多くのアイデアをうまく使い，楽しんで研究できることが一番大切なのではと思います．

（今回はリモートにてインタビューを実施しました）