あなたの声を「すぐそば」品質で聴くAI

～遠くからでも近接マイク品質で混ざった音を聞き分ける革新的音響処理技術～

AI hears your voice as if it were "right next to you" – Audio processing framework for separating distant sounds with close microphone quality –

メディア情報研究部 信号処理研究グループ
中谷　智広
Tomohiro Nakatani

●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　上席特別研究員。１９９１年　京都大学　情報学研究科　応用システム科学専攻　修士課程修了。博士（情報学）。１９９１年にＮＴＴに入社以来、音源分離、残響抑圧、雑音抑圧、音声特徴量抽出、ロバスト音声認識など、統計的信号処理を用いた音声処理の研究に従事。電子情報通信学会、日本音響学会の各会員。ＩＥＥＥフェロー。

近年、スマートフォンによる音声認識やヘッドセットを用いたリモート会議など、話者の口元のマイク（近接マイク）を用いた音声アプリケーションが広く利用されています。一方、今後、AIが、より深く私たちの生活に溶け込み、身近（＝「すぐそば」）な存在になるためには、日常生活の中で、必ずしもマイクの「すぐそば」で話されていない音声をも扱うことが求められるようになります。しかし、話者から離れたマイクでは、壁や天井からの反射である残響、複数の話者の音声、背景雑音などが混ざってしまうため、音声アプリケーションの性能は大きく劣化します。これを解決するために、マイクから離れていても近接マイク品質で複数の音声を聞き分ける音声強調技術の研究を進めています。本講演では、単一マイクよりも高品質な処理が可能な複数マイクを用いる音声強調の最新技術を紹介します。

近接マイク品質の実現のための課題

収録音から近接マイク品質の音声を抽出するには、残響抑圧、音源分離、雑音抑圧の3つの処理が必要です。まず、残響抑圧により、遠くで響いている印象のぼやけた音声を、すぐ近くにいる印象のはっきりした音声に変えます。さらに、複数の音声や背景雑音が混在している場合は、音源分離や雑音抑圧により個々の音に分解します。

複数マイク音声強調では、この残響抑圧、音源分離、雑音抑圧の各課題に対し、音が音源からマイクに伝播し混合する各過程（収録音の生成過程）を推定し、その逆変換を適用することで、近接マイク品質の音声を求めます（図1（a））。具体的には、残響が壁や天井に反射してマイクに到達する過程、複数の音声が各方向から到来し混合する過程、雑音が全方向から到来し重畳する過程、をそれぞれ推定し、逆変換を行います。

しかし、従来は、この３つの処理を同時に行うことは困難でした。このため、各処理を順番に適用するしかなく、全体として最適な処理を行えませんでした。例えば、最初に、混ざったままの音に対し残響抑圧を高精度に行うのは、困難な課題でした。全体最適な処理の実現は、音響処理における重要な未解決課題でした。

残響抑圧、雑音抑圧、音源分離の統一モデル

これに対し、私たちは、3つの処理を全体最適な形で実現できる統一モデルを考案（世界初）し、さらに、実時間処理が可能な高速アルゴリズムを構築しました［1,2］。統一モデルでは、まず、近接マイク品質の音声や雑音が満たすべき一般的な性質（音声はスパースな信号、背景雑音は定常信号など）を数理的にモデル化します。そして、3つの処理を組み合わせた結果の音がこの性質を最もよく満たすようにするという「統一基準」に基づき各処理を最適化することで、全体最適な処理を実現します（図1（b））。その結果、例えば、離れたマイクで収録した音声の認識性能を大幅に改善できるようになりました（図2（a）-(c)）。

さらに、この統一モデルを応用して、比較的少数のマイクでも高精度な推定が可能なスイッチ機構を考案しました［3,4］。従来の複数マイク音声強調では、収録音に含まれる音源数に対し十分な数のマイクがないと推定精度は大きく低下します。一方、収録音を各短時間区間でみると、多くの場合、同時に音を出している音源数は少なくなります。この性質を利用して、統一モデルに基づくスイッチ機構では、収録音を、比較的少数の同じ音源のみからなる時間区間に分類し、時間区間ごとに個別に音声強調を適用します。そして、上記の統一基準に従い、時間区間の分類を含めた全体の最適化を行うことで、少数マイクでも高精度な推定が可能になりました。

統一モデルは、これまで試行錯誤的に組み合わされてきた各処理に対し、理論的にも実用的にも優れた統合指針を与えます。今後、音響処理のデファクト技術として広く利用されていくことが期待されます。

今後の発展：深層学習との最適な統合

音声強調のもう一つの重要なアプローチに深層学習があります。深層学習では、SpeakerBeam［5］が実現した声の特徴に基づく選択的聴取や単一マイクによる音声強調など、複数マイク音声強調では困難な処理が実現できる一方で、残響があると処理音質が悪くなる、音声認識性能の改善は限定的であるなどの課題があります。これらの特徴は、複数マイク音声強調と相補的であり、例えば、深層学習で注目した音声を、複数マイク音声強調で近接マイク品質にすることが可能です（図2（d））。今後、両者の最適な統合方法を構築することで、より高機能で高品質な音声処理が実現されると考えています。

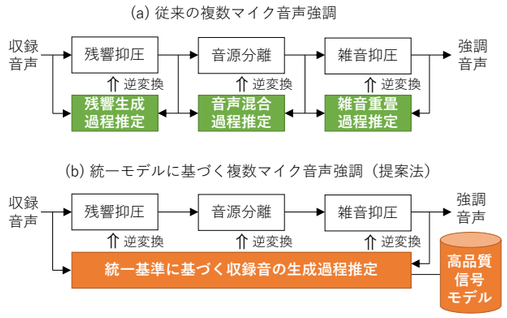

図１：複数マイク音声強調の従来法と提案法

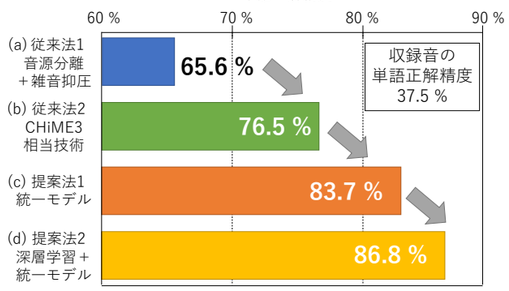

図２：複数マイク音声強調による音声認識率の改善

●参考文献

［1］T. Nakatani, C. Boeddeker, K. Kinoshita, R. Ikeshita, M. Delcroix, R. Haeb-Umbach, “Jointly optimal denoising, dereverberation, and source separation,” IEEE/ACM Trans. Audio, Speech, and Language Processing, vol. 28, pp. 2267-2282, 2020.
［2］R. Ikeshita and T. Nakatani, “Independent vector extraction for fast joint blind source separation and dereverberation,” IEEE Signal Processing Letters, vol. 28, pp. 972-976, 2021.
［3］R. Ikeshita, N. Kamo, T. Nakatani, “Blind signal dereverberation based on mixture of weighted prediction error models,” IEEE Signal Processing Letters, vol. 28, pp. 399-403, 2021.
［4］T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, N. Kamo, S. Araki, “Switching independent vector analysis and its extension to blind and spatially guided convolutional beamforming algorithms,” IEEE/ACM Trans. Audio, Speech, and Language Processing, 2022.
［5］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.