Real-time selective listening of everyday sounds

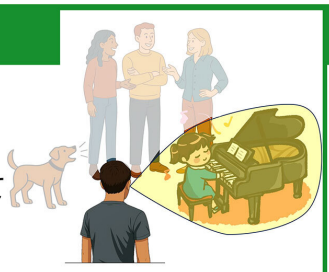
08 深層学習に基づく任意の音の低遅延選択的聴取

聴きたい音だけをリアルタイムで抜き出す！

どんな研究	人間は、様々な音が混在する中でも、聴きたい種類の音に注目して聴くことができる、選択的聴取と呼ばれる能力を持っています。本研究では、人間が持つ選択的聴取の機能を計算機上で実現し、混ざった音から聴きたい音を取り出す技術(目的音抽出) をめざします。
どこが凄い	高い目的音抽出精度を保ったまま、一般的なPCでもリアルタイムで動作する目的音抽出手法を考案しました。また、音の汎用表現モデルを用いた高精度化・高品質化や、音の到来方向を検知可能なバイノーラル処理も実現し、人間の選択的聴取能力に近づけることができました。
めざす未来	子どものピアノの音や生活の音は、在宅勤務のWeb会議では不要な音ですが、実家の祖父母との通話では大事な音かも知れません。本技術によって、周囲で鳴っている音の中から、状況に応じて聴きたい音・聴きたくない音の制御を可能とし、より快適なコミュニケーションを実現します。

選択的聴取

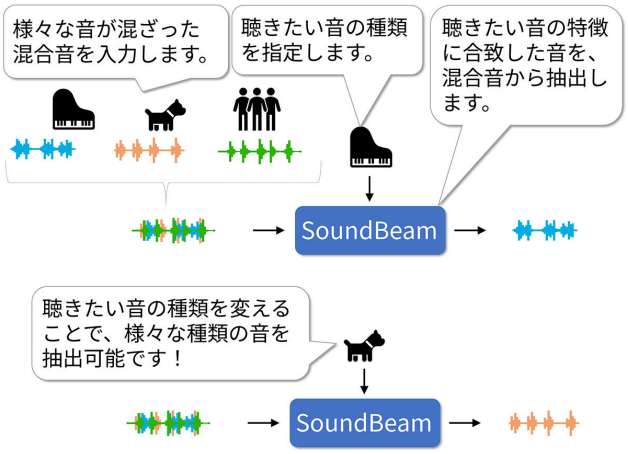
- 日常生活では様々な音が、同時に聞こえています。
- 人間は、状況に応じて聴きたい音だけに集中して聴くことができます。（＝選択的聴取）



コンピュータを用いて選択的聴取を実現する仕組みを研究しています。

SoundBeamの仕組み [1]

深層学習・ニューラルネットワーク(NN）に基づき、任意の音の選択的聴取を実現します。



→本展示では、SoundBeamによる音の選択的聴取について、さまざまな*応用に向けて*3つの新しい機能・性能拡張を紹介します!

図の一部は生成AIで作成したものを使っています。

①リアルタイム処理[2]

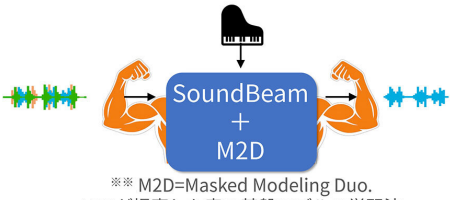
録音と同時にリアルタイム処理を行い、効果的に選択的聴取機能を実現するNNの設計手法を考案しました。(CD研・人間研と連携)

※SDR＝信号対歪比
抽出精度の尺度

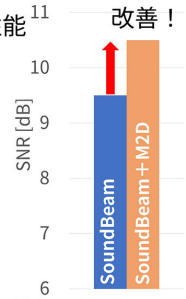
応用例：Web会議などにおいて、相手に聴かせたい音だけを選択的に送信することができるようになります。

②高性能かつ高品質な音の抽出[3]

音の詳細な特徴を捉えられるNTTの音の基盤モデルM2D※※と組み合わせることで、聴きたい音の選択性能の向上と抽出した音の品質改善を実現しました！



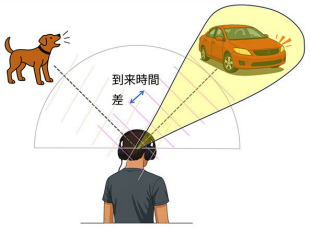
※※ M2D=Masked Modeling Duo.
NTTが提案した音の基盤モデルの学習法

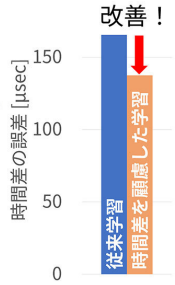


応用例：映画・音楽・ホームビデオなどの収録音に対し、高品質なポストプロダクションが可能になります。

③方向情報を正確に再現[4]

左右の耳に到達する音の時間差を正確に抽出することで、音源方向の情報も正確に保存できる選択的聴取を実現しました。





応用例：外出時にイヤホン等を使用中でも、危険を知らせる重要な音をその方向情報も含めて聴けるようになります。

関連文献

[1] M. Delcroix, J. B. Vázquez, T. Ochiai, K. Kinoshita, Y. Ohishi, S. Araki, “SoundBeam: Target sound extraction conditioned on sound-class labels and enrollment clues for increased performance and continuous learning,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 31, pp.121-136, 2022..

[2] K. Wakayama, T. Ochiai, M. Delcroix, M. Yasuda, S. Saito, S. Araki, A. Nakayama, “Online target sound extraction with knowledge distillation from partially non-causal teacher,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 561-565, 2024.

[3] C. Hernandez-Olivan, M. Delcroix, T. Ochiai, D. Niizumi, N. Tawara, T. Nakatani, S. Araki, “SoundBeam meets M2D: Target sound extraction with audio foundation model,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[4] C. Hernandez-Olivan, M. Delcroix, T. Ochiai, N. Tawara, T. Nakatani, S. Araki, “Interaural time difference loss for binaural target sound extraction,” in *Proc. 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 210-214), 2024. IEEE.

連絡先

デルクロア マーク（Marc Delcroix）メディア情報研究部 信号処理研究グループ

共同研究先・外部資金

本研究のうち、M2DとSoundBeamの組み合わせおよび方向情報を保存できる選択的聴取に関する部分はJST SICORP (JPMJSC2306)の支援を受けて実施しました。