URL https://journal.ntt.co.jp/article/35362
DOI https://doi.org/10.60249/25085104

音の聴き方を自ら学ぶAI――自己教師あり学習によるさまざまな音の汎用表現学習技術から，大規模言語モデルを活用した音の理解の最前線へ

音や画像などメディア情報から意味の抽出に有用な特徴量を学習する表現学習は，AI（人工知能）による優れたデータの理解を可能にしました．本稿では，私たちの身の周りのさまざまな音をAIに理解させるための音の表現学習技術を紹介します．学習した表現は，動物の鳴き声の分類や音楽ジャンルの識別など，音の理解にかかわる幅広い問題に応用できます．私たちは，表現学習の中でも，音データだけを用いて学習を行う「自己教師あり学習」を活用して技術を深化させており，さらに大規模言語モデルを活用した文章の意味と対応付ける音の理解へと発展させています．

キーワード：#表現学習，#自己教師あり学習，#音の理解

仁泉（にいずみ）　大輔（だいすけ）
NTTコミュニケーション科学基礎研究所

はじめに

音や画像などメディア情報から，それらの理解に有用な特徴量の自動抽出処理を学習により獲得する表現学習*1は，AI（人工知能）による優れたデータの理解を可能にしました．本稿では，私たちの身の周りのさまざまな音をAIに理解させるための音の表現学習技術を紹介します．学習した深層学習*2モデルは人の声や動物の鳴き声といった音の種類の認識や音楽ジャンルの識別など，幅広い音の理解にかかわる問題に応用できます．表現学習の中でも，音データの内容を示すラベルを推論する従来の「教師あり学習」*3手法に代わって，データそのものから教師ラベルをつくり出す「自己教師あり学習」*4手法が注目されています．この手法を用いることで，ラベル付けされていない大量のデータを活用し，より有効な表現を学習することが可能になります．私たちはこの自己教師あり学習を活用し，音の表現学習技術を発展させてきました．その技術は，言語を介してメディア情報を取り扱うニーズの高まりを背景に，音と言語を対応付ける表現*5の学習へと発展し，さらには大規模言語モデル（LLM）の持つ文章の意味に対する深い理解を活用する学習手法へと発展しています．

データを理解するための基礎技術「表現学習」

表現学習は，音や画像などのメディア情報をコンピュータで理解するために，有用な特徴量（＝表現）の自動抽出処理を学習する技術です．従来は，信号処理により得られる音の大きさや高さ，周波数成分など基本的な情報を基に人の手で用途に特化した特徴量を設計していましたが，例えば会話音声や乗り物の音，動物の鳴き声など，多様で複雑な音の理解には限界があります．この解決として登場した深層学習は，メディア情報とそのラベルを用いる教師あり学習により，多様な音のデータを互いに区別するうえで重要となる特徴量を自動的に抽出する表現学習を実現し，より優れた理解を可能にしました．さらに近年では，大量のデータに内在する豊かなパターンから教師ラベルなしで表現を学習できる自己教師あり学習が着目されています．この学習手法は，教師ラベルをデータ自体から生成することで，ラベル付けのコストを必要とせず，効率的な表現学習を実現しています（**表**）．

音の細かな違いを吸収する表現「BYOL for Audio」

多様な音を理解するうえで有用な特徴量をどのように学習すればよいか，その答えの1つとして画像分野の手法BYOL（Bootstrap Your Own Latent）を基に自己教師あり学習手法BYOL for Audio（BYOL-A）[(1)]を提案しました．BYOL-Aは音のパターンをとらえながらも，細かな違いを抑えた数値の表現を学習します．例え

表　表現の実現方法の変遷

フェーズ	実現方法
深層学習登場以前	人の手で用途に合わせて特徴量を組み合わせて設計
教師あり学習を用いた深層学習ベース	大量のデータに分類ラベル付けを行い，分類問題を用いる教師あり学習を通じて，データを互いに区別するうえで重要となる特徴量の自動抽出モデルを獲得
自己教師あり学習を用いた深層学習ベース	大量のデータから，データそのものを手がかりに生成した教師ラベルを用いる学習を通じて，特徴量の自動抽出モデルを獲得

＊1 表現学習：表現の自動抽出処理を，深層学習などを用いる学習により獲得すること．
＊2 深層学習：多数の層で構成されるニューラルネットワークを学習することで目的とする処理を行えるようにさせること．
＊3 教師あり学習：メディア情報などのデータと対になる教師ラベルを利用して，機械学習モデルが入力データのラベルを推論できるように学習させること．
＊4 自己教師あり学習：メディア情報などのデータそのものから教師ラベルをつくり出して学習を実現すること．
＊5 表現：メディア情報の特徴を数値で表したもの．

ば犬の鳴き声は音の高さや大きさ，または犬の種類によって異なりますが，「ワン」と擬音語で表されるように，大まかには「犬の鳴き声」と理解されます．BYOL-Aは，このように音に違いがあっても同じ音のパターンであれば同じ数値表現を抽出することを学習目標とすることで，有用な表現の学習を可能にしました（**図1**）．その実現のために，BYOL-Aでは音の違いに対して頑健（ロバスト）な，すなわち多少の変化に影響されにくい表現を学習します．具体的には，背景雑音，音の長さ，音の高さ（ピッチ），音量といった特性にランダムな変化を加えた2つの音を用意し，それらから得られる数値表現ができるだけ一致するように，繰り返し学習を行います．つまり，元の音に異なる変化を加えた音に対しても，同じ数値表現を抽出できるように訓練することで，音の本質的な特徴をとらえ，その意味につながる表現を自動抽出させます．その結果，従来の手法と比べて発話キーワード認識で20ポイント以上精度を改善するなど，表現学習がタスクの飛躍的な性能向上をもたらしました．その背景には，学習した表現がさまざまな問題において音の種類ごとにクラスタ（かたまり）を構成し，クラスタごとに分類することを容易にしていることが挙げられます（**図2** (a)）．

この技術の鍵は，その学習の仕組みにあります．直接的には，音の特性に変化を加えた2つの音を，同じような数値で表現できるように学習させることが目的です．しかしその結果として，音の種類ごとに自然とクラスタが形成されるような，音の意味を反映した表現が得られるようになります．いわば「風が吹けば桶屋が儲かる」ように，一見遠回りに見える仕組みよって，AIは音をどのように「聴く」か，自ら学習することにつながるといえるのです．ここで「音の意味」とは，例えばそれが犬の鳴き声なのか，あるいは楽器の音なのかといったように，タスクや問題ごとに異なる音の概念を指しています．この学習によって得られた表現は，音の本質的な特徴を反映しており，その結果，問題ごとに異なる音の概念に応じたクラスタを形成できるようになります．こうした仕組み（アルゴリズム）の工夫こそが，自己教師あり学習の研究の面白さです．

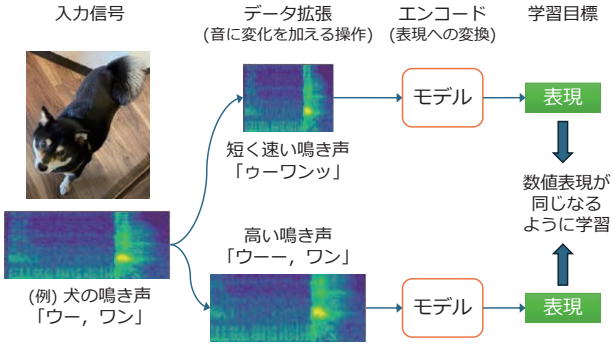
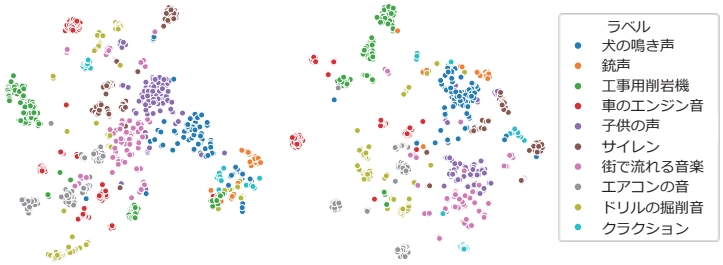


図1　BYOL for Audioの学習



(a) BYOL-A　　(b) M2D

図2　10種類の環境音データ（UrbanSound8K）に含まれる音の表現の分布

音の穴埋め問題で優れた表現を学習する「Masked Modeling Duo」

BYOL-Aでは音に変化をつけて学習を行うため，変化をつける音の特性について十分にその詳細を表現できないという問題がありました．BYOL-Aではとらえきれない細かな違いも含め，音を構成する本質に迫る特徴を抽出するため提案した新たな自己教師あり学習手法Masked Modeling Duo（M2D）[2]は，「音の穴埋め問題」，すなわち音の隠された部分の予測を通じて表現を学習します（**図3**上）．マスクによって一部を隠したデータや，未来のデータを予測する学習手法は，LLMをはじめとする自然言語処理の分野で大きな成功を収め，優れた表現を学習できることが知られています．M2Dでは，見えている部分の入力データを表現に変換し，その表現を用いて隠された部分の表現を予測します．例えば，メロディの一部が欠けた音楽データから欠損部分を予測する場合，曲の構造やリズムに関する特徴が表現に反映される必要があります．同様に，犬の鳴き声，演奏される音楽，さざ波の音など，多様な音を学習することで，それぞれの音の背後にある構造的なパターンを示唆する特徴が効果的に抽出されるよう学習されます．さらに，この学習手法はBYOL-Aとは異なり，細かな違いも区別できるように表現学習が進みます．例えば，隠されたピアノの音の一部を正確に予測するためには，音程や音の長さ，強さといった情報をできるだけ忠実に表現を保持する必要があります．このように，予測精度を高めるための最適化が促されるため，音の詳細な特徴までをとらえる表現が学習されます．

その結果，M2Dは音の種類ごと，より密に集まったクラスタを形成し（図2 (b)），結果として多くのタスクへの応用において，性能を大きく向上させました．例えば，心音の聴診において病気の兆候を示唆する心雑音の検出に適用した場合，BYOL-Aでは従来法の性能に達しなかったのに対し，M2Dはそれを上回る結果を示しました[3]．これは，心音に含まれる雑音成分，いわば微細な音の違いに関する情報を，M2Dの表現がより保持できていることが貢献していると考えられます．

音とその説明文を対応付ける「M2D-CLAP」

近年の生成AIにおいては，言語を介したAIとのやり取りが不可欠です．そのため，メディア情報を言語と対応付ける技術は，

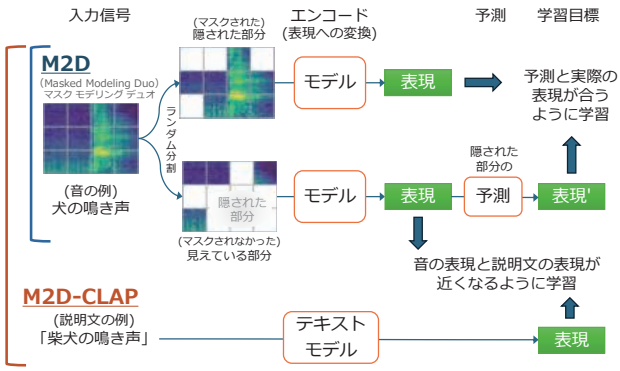


図 3　音の穴埋め問題で学習する M2D および，テキストモデルを統合し拡張した M2D-CLAP

AIによる情報の理解や生成を支える重要な要素となっています．そこで提案した M2D-CLAP[4]は，音の数値表現に加えて，音の内容をその説明文に対応付けられる表現をも兼ね備えた手法です．そのため，音の穴埋め学習を行うM2Dに，音と言語の対応関係を学習するCLAP（Contrastive Language-Audio Pre-training）を統合しました．このアプローチにより，音響信号とその言語による説明文の意味を結びつけ，内容の近さを数値で表せる表現を獲得できます（図 3 下）．例えば，「雷雨の音」という説明文とその実際の音が近い数値表現を持ち，異なる音とは異なる表現になるように学習されます．これにより，「犬の鳴き声」といったテキストに対応する音の検索や，事前に学習していない「電気自動車」「バイク」などの新しく定義されたラベルへの分類問題を解くゼロショット学習が可能になります．このように言語を用いて音の問題を解くことができるようになる一方で，教師あり学習であるCLAPを組み合わせるために，CLAPの学習において必要となる説明文を作成するコストが発生します．しかし近年，音響イベント検出など音の内容を記述する技術の発展により可能になっている音の説明文の自動生成を，音の表現学習に利用するサイクルが生まれつつあります．M2D-CLAPはこのトレンドを踏まえ，自動生成された説明文を活用した学習を実現しています．

M2D-CLAPは，CLAPにより学習される音と言語を対応付ける表現に加えて，従来の音の表現学習においても最先端の性能を実現している点が，他の手法にはない特徴です．このように，M2D-CLAPは音と言語の対応付けを含む多様な応用場面において，汎用的に幅広く活用できる表現を提供できます．

LLMの知識を学ぶ「M2D2」

LLMの知識や文章の意味に対する深い理解を音の表現学習に活用したのがM2D2[5]です．例えば「さまざまな金属を叩いた際の音」を尋ねるプロンプト文章でLLMに質問すると，素材の特性などを根拠に，各金属がどのような音を出すかを詳しく解説します．これは，LLMが音に関する豊富な知識を持っていることを示しています．さらに，LLMは文章の言い回しの違いやニュアンスの機微を理解し，適切な応答を生成できるため，文の意味に対する深い理解も備えています．そこで，M2D-CLAPによる学習においてテキストモデルにLLMを用いることで，LLMを用いた文章の表現，いわばLLMの知識が反映された表現を通じて音の表現を学習させることが可能になります．このようにLLMの知識を活用することで，M2D2は音についての理解が高度化し，音の検索や説明文生成の性能を向上させました．またさらに，LLMから学んだ音の表現は，LLM自体と組み合わせて利用することで，例えば単に説明文の生成にとどまらず，音を通じた状況の理解に基づく対話の実現など，さらに高度な応用への発展が期待されます．

ま と め

これらの研究により，多様な音を理解するための表現は，音の特徴を効果的にとらえるだけでなく，言語の持つ意味との結びつきを獲得し，LLMの知識を活用できる汎用表現へと進化してきました．しかし，これらの技術はまだ研究段階にあり，音は画像や言語と比べて非常にデータが少ないなど，実用化に向けてはさまざまな課題が残されています．将来的には，こうした課題を克服し，日常のさまざまなシーンにおける音を学習し，私たちの身の周りの音を理解するAIの実現をめざしています．それにより，例えば音を通じて日々の健康状態をモニタリングするサービスなど，音を活用した幅広い問題解決に役立つことが期待されます．

■参考文献

(1) D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino: “BYOL for Audio: Exploring Pre-Trained General-Purpose Audio Representations,” IEEE/ACM Trans. Audio, Speech, Language Process., Vol. 31, pp. 137-151, 2023.
(2) D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino: “Masked Modeling Duo: Towards a Universal Audio Pre-Training Framework,” IEEE/ACM Trans. Audio, Speech, Language Process., Vol. 32, pp. 2391-2406, 2024.
(3) D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino: “Exploring Pre-trained General-purpose Audio Representations for Heart Murmur Detection,” Proc. of EMBC 2024, pp. 1 -4, 2024.
(4) D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, M. Yasuda, S. Tsubaki, and K. Imoto: “M2D-CLAP: Masked Modeling Duo Meets CLAP for Learning General-purpose Audio-Language Representation,” Proc. of Interspeech 2024, pp. 57-61, 2024.
(5) D. Niizumi, D. Takeuchi, M. Yasuda, B. T. Nguyen, Y. Ohishi, and N. Harada: “M2D2: Exploring General-purpose Audio-Language Representations Beyond CLAP,” arXiv:2503.22104, 2025.

仁泉 大輔

急速な発展を遂げるAIが音などメディア情報を理解するための仕組みとして学習方法があり，その妙が結果として得られる効果を変えていくその面白さを知っていただければ幸いです．

◆問い合わせ先

NTTコミュニケーション科学基礎研究所
メディア情報研究部