画像や音を見聞きするだけで賢くなるAI ――クロスモーダル情報処理の展開

NTTコミュニケーション科学基礎研究所では，画像，音，テキストといった種類の異なるメディア情報にまたがる情報処理の研究を進めています．これをクロスモーダル情報処理と呼ぶことにします．クロスモーダル情報処理のポイントは，複数種類のメディアデータが対応付けられている共通の場所である「共通空間」をつくることです．これにより，これまでにはなかった新しい機能を実現できる可能性が示されつつあります．音から画像や説明文をつくるといった異種メディア間の新たな変換や，メディア情報に含まれる物事についての概念獲得などです．

かしの　くにお
柏野　邦夫
NTTコミュニケーション科学基礎研究所

クロスモーダル情報処理とは

近年のAI（人工知能）の発展を支える立役者は深層学習の技術です．例えば，さまざまな物体を撮影した画像と「りんご」「みかん」といった物体の名前（クラスラベル）とを組（ペア）にしたデータを大量に用意して深層学習を行うと，画像中の物体が何であるかを高い精度で認識できるようになることが知られています．その優れた特性のためにさまざまな分野で研究や活用が進む深層学習ですが，私たちが特に着目している能力の１つは，異種のメディア情報（例えば，画像と音）の対応付けができることです．画像，音，テキストといった情報の種類のことをモダリティ（modality）と言いますので，異なるモダリティにまたがる情報の対応付けをクロスモーダル（cross-modal）情報処理と呼ぶことにします．このクロスモーダル情報処理とはどのようなもので，どんなメリットがあるのでしょうか．

新しい情報変換

⑴　音から画像をつくる

クロスモーダル情報処理のメリットの１つは，異種のメディア情報が対応付けられた共通の場所である「共通空間」を介することで，従来では考えられなかったような情報の変換が可能なことです（**図１**）．その１つとして，私たちの研究チームでは，音から画像を推定する課題に取り組んでいます．

私たち人間は，目を閉じていても周囲の音からその場の情景を思い浮かべることができます．そこで，マイクで拾った音からその場の情景を表す画像をつくってみようというわけです．例えば，室内に複数のマイクを設置し，数人の会話を数秒間録音します．４本のマイクを用いたとすると，それぞれのマイクでとらえた音の周波数成分の時間変化を表す「スペクトログラム」が４枚と，音の到来方向を表現した「角度スペクトル」の情報が１枚得られますので，これらをシステムに入力します．システムでは，これらの情報をそれぞれニューラルネットワークで処理し，低次元の空間にマッピングします．その情報を基に，ニューラルネットワークを用いて画像をつくり出します．この画像には，室内のどの場所でどのような属性の人物が発話しているかが表現されていますので，室内の大まかな様子を把握することができる，というわけです（**図２**）．このように，いったん入力を低次元空間にマッピング（エンコード）して，そこから高次元の情報を生成（デコード）する処理

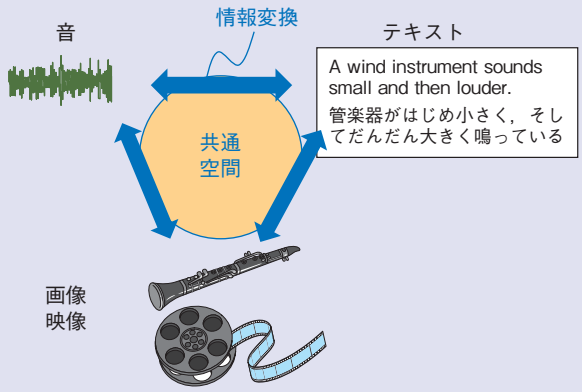


図１　クロスモーダル情報処理による情報変換の概念図

は，一般に「エンコーダ・デコーダモデル」と呼ばれ，入出力のペアを学習用データとして与えることで，深層学習によって構成することが可能です．

NTTコミュニケーション科学基礎研究所では，現在までにシミュレーション実験や実際の音を発する物体を使った実験を行って，一定の条件下で，どこに何があるかを画像として示すことが実際に可能であることを確認しています[1]．このような音から画像への変換は，これまで試みられたことがない新しい情報処理を提案したものとなりました．この技術が発展すると，カメラを置くことが望まれない場所やカメラがとらえきれない状況（物陰や暗闇など）での安全確認などにも応用できると考えています．

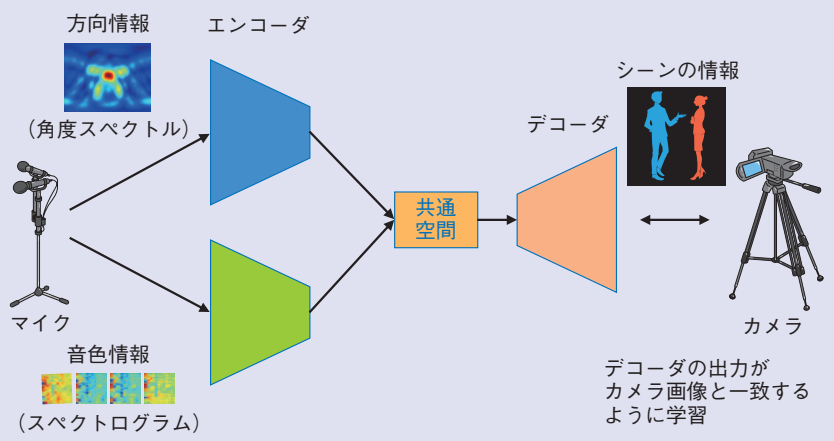


図 2　音からの画像の生成

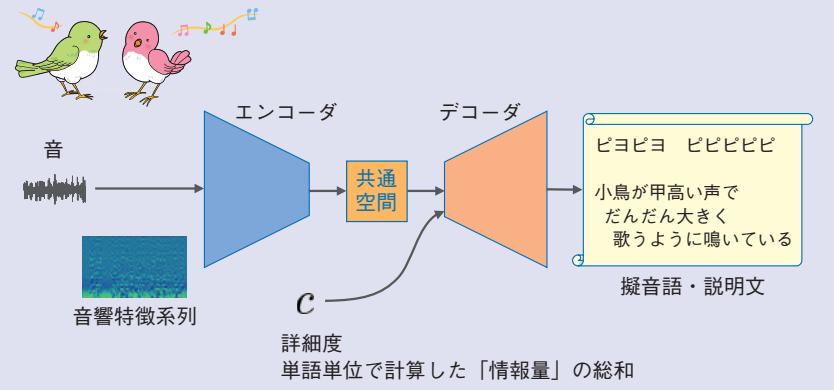


図 3　音からの説明文の生成（条件付系列変換型説明文生成法）

⑵　物音を言葉で説明する

異種情報の変換のもう１つの例は，音からテキストへの変換です．音声認識システムを用いると話し言葉をテキストに変換できますが，これまでの音声認識システムでは，話し言葉以外の物音などを適切なテキストに変換することはできませんでした．これに対し私たちは，マイクで拾った音から，その音を表現する擬音語や，その音を記述する説明文を生成する技術を開発しました[2]．

条件付系列変換型説明文生成法（CSCG: Conditional Sequence-to-sequence Caption Generation）と呼ぶこの手法も，エンコーダ・デコーダモデルに基づいています（**図 3**）．今度は系列から系列への変換（系列変換）を行います．まず，入力音響信号から抽出した特徴を時系列としてニューラルネットワークでエンコードし，低次元空間にマッピングします．次に，その情報からニューラルネットワークで音素系列（擬音語）または単語系列（説明文）をデコードします．

説明文の生成においては，どのような説明文を生成するのが適切かは場合によって異なり，唯一の正解を定めることはできません．例えば，「車が近づいている，危ない」といったように端的に短文で表現すべき場面もあれば，車種や車速などによるエンジン音の微妙なニュアンスの違いを詳細に表現したい，といった場面も考えられます．このような要請にこたえるため，

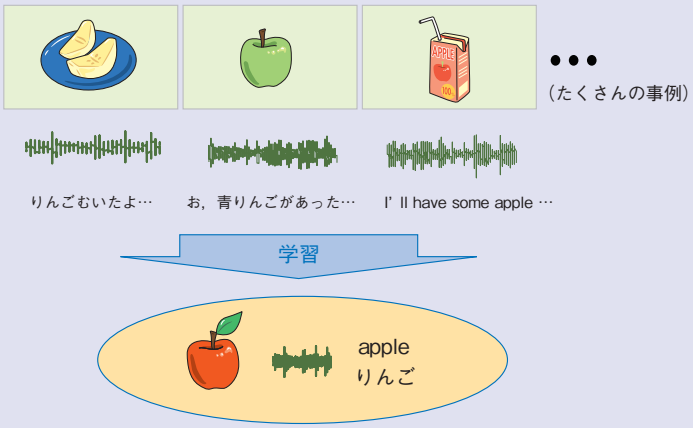


図 4　クロスモーダル情報からの概念獲得

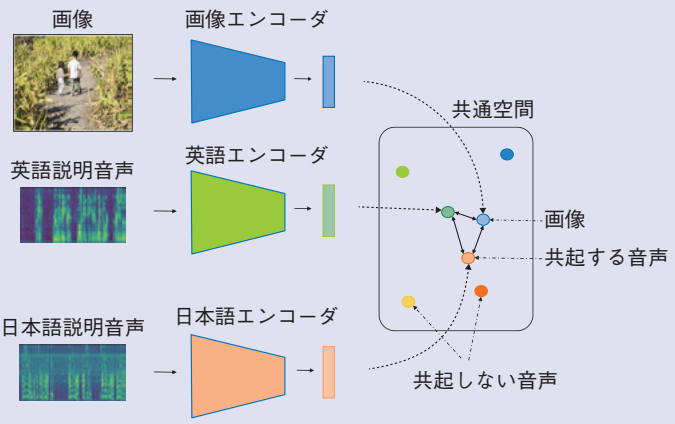
デコーダの働きを「詳細度」と呼ぶ補助入力で制御し，表現の詳しさ（説明文に含まれる単語の持つ情報量の和）を調節できるようにしました．小さな値の詳細度を指定すると端的な説明文を生成し，大きな値の詳細度を指定するほど，より具体的で，より長い説明文を生成するようになります．所定の条件における実験において，擬音語生成では人手による擬音語よりもむしろ受容度（あてはまっていると判断される割合）が高い擬音語の生成が可能であること，説明文生成や詳細度の制御も有効に機能すること，などを示しています．

本技術は，動画や実環境に対する字幕生成や，メディアの検索などに有効であると考えています．従来，音に対して「発砲音」「叫び声」「ピアノの音」などといったように既知のクラスラベルを与えることは試みられていました．しかし，画像の場合に比べても，音の場合には，音の信号と「何の音か」との対応が明らかであるとは限らず，「何かは分からない初めて聞くような音」に遭遇することは日常少なくありません．このような場合にはクラス分類だけでは有効性に限界があります．また，本技術では，音と説明文とが紐付くことにより，説明文による音の検索が可能となります．実際，共通空間においては音と擬音語や説明文との間の距離を直接測定することができ，擬音語や説明文を用いて音を検索することが可能です．このような場合，目的とする音のニュアンスを説明文で詳しく指定したい場合もあるでしょう．本技術を用いると，「車」「風」などといったクラスラベルだけではなく，音の高さや大きさ，変化の様子なども含めて，文字によって目的の音を指定することが可能になります．このような，音に対する説明文の生成も，私たちが世界で初めて提案した情報処理です．

概念獲得－未知の概念を自ら学習する

クロスモーダル情報処理のもう1つのメリットは，「共通空間」において異種情報の対応を見出すことで概念獲得が可能になることです．深層学習に必要とされる大量のデータの準備には，手間がかかったり，データの入手自体が難しかったり，クラスラベルの付け方を事前に設計することが難しかったりといった困難さを伴うことが少なくありません．そこで私たちは，メディア情報の中に含まれるひとまとまりのもの，つまり概念を自動的に獲得し，認識や検索に活用することをめざした研究に取り組んでいます．

異種のメディア情報の「共起」，つまり現実世界の中で，同じものに端を発する異種のメディア情報がランダムにではなく特定の関係性を持って現れることなどをうまく利用すると，人手でメディアデータどうしをペアリングすることなしに，共通空間を介したメディアデータのペアリングが可能になります．これを用いると，事前に「りんご」の画像とクラスラベルのペアを与えなくても，「皆がこの物体を指して“りんご”と言っているようだ．これは“りんご”というものなのだな」といった方式での学習が可能になるのです（**図 4**）．画像や音を見聞きするだけで賢くなる，というわけです．し



共起する情報を共通空間で近傍に配置するようにエンコーダを学習することで，音の各部分と画像の各部分との関連性を抽出できる

図5　画像および複数の言語情報の共通空間へのマッピング

かも周囲が“りんご”と言えばりんご，“apple”と言えば apple であると学習するといったように，周囲の人間の感じ方や振る舞い方を習得していきます．これは，私たち人間が，生まれてから成長するにつれて日常生活の中でさまざまなことを学んでいく過程に例えることができるでしょう．

私たちは，実際に，多数の写真に対して英語と日本語で何が写っているかを説明したもの（上記の，画像と音の共起を人工的に発生させたもの）を用いて，各言語における単語と，写真の中に写っている物体との対応付け（セグメンテーション）が可能であることや，画像を介した言語間の翻訳知識の自動獲得が可能であることを確認しています[3]（**図5**）．

今後の展開

本稿では，「クロスモーダル情報処理」が切り拓く新しい情報処理の最先端について，その一部を紹介しました[4]．これらの一連の研究に共通する考え方は，音，画像，テキストといったさまざまなメディア情報に対して，私たちの目や耳に触れる表層の表現形式と，その背後にある共通空間，つまり特定の表現形式には依存しない本来的な情報とを，分離して取り出してそれぞれを活用しようということです．これは多様な可能性を秘めた新しい情報処理の試みといえます．このような研究が発展すれば，私たち人間とともに暮らし，感じ方や振る舞い方を共有しながら，自ら学習していくAIも実現できそうに思われます．そのようなAIは，今よりもっと親しみを感じられるパートナーになり得るのではないでしょうか．

■参考文献

(1) G. Irie, M. Ostrek, H. Wang, H. Kameoka, A. Kimura, T. Kawanishi, and K. Kashino: “Seeing through sounds: Predicting visual semantic segmentation results from multichannel audio signals,” in Proc. ICASSP 2019, Brighton, U.K., May 2019.

(2) S. Ikawa and K. Kashino: “Generating sound words from audio signals of acoustic events with sequence-to-sequence model,” in Proc. ICASSP 2018, Calgary, Canada, April 2018.

(3) 大石・木村・川西・柏野・Harwath・Glass：“画像を説明する多言語音声データを利用したクロスモーダル探索，” 信学技報，Vol.119, No.64, PRMU 2019-11, pp.283-288, 2019.

(4) Hot News: “NTTの「クロスモーダル」幼児のように世界を理解―生成AIの急激な発展で実現可能に，” 日経エレクトロニクス，pp.14-15, 2019. 7.

柏野　邦夫

ロボットがたくさんの動画を見たり，周囲を見回しながら人の会話や環境中のさまざまな音を聞いたりするだけで，言語を覚え，世界を理解していくようになるのは，そう遠くない将来かもしれません．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
企画担当
TEL　0774-93-5020
E-mail　cs-liaison-ml@hco.ntt.co.jp