動物音声認識のための教師なし機械学習

研究ノート

森田　　尭*

Unsupervised Machine Learning for Animal Vocal Recognition

Key Words：speech recognition, unsupervised learning, deep learning, clustering, animal vocalization

はじめに

深層学習技術の発展を中心に、機械学習は2010年代以降大きな躍進を続けている。深層学習の特徴としてしばしば取り沙汰されるのはその汎用性であり、自然言語処理用に開発された手法[1]が画像認識において従来手法を上回る性能を示すような事例も生じている[2]。その一方で、新規性の高いデータに対して深層学習技術を応用することは必ずしも容易ではない。本稿では動物音声に対する機械学習技術の応用に着目し、その課題と関連研究を紹介する。

音声認識学習と教師データの制約

ヒトが発する言語音声を自動的に文字起こしすることを目指した機械学習課題を音声認識と呼ぶ。音声は身体的特徴に起因する要素を始めとした個体差をふんだんに含み、また同一人物が同じ文章を読み上げた場合でも観測される音声波形はその都度異なる。そのような「無視すべきばらつき」に左右されることなく、テキストデータのような一貫した言語表現に音声を変換することが音声認識の目的となる。

現在主流の音声認識技術は、音声データとそれに対応する文字起こしデータの組を用いた機械学習によって実現されている。この文字起こしデータは予め人間の手で用意され、学習器は与えられた音声と文字起こしの対応関係を学習し、最終的に初聴の音声でも正しく文字起こしすることを目指す。このように学習器の出力目標となるデータを教師データと呼び、教師データを用いる機械学習を教師あり学習と呼ぶ。

教師あり学習の形式を取る音声認識技術を直接的に応用するためには、教師データとなる文字起こしが手に入ることが前提となる。しかしながら、動物音声を認識したい場合、動物自身が音声の文字起こしをしてくれることはない。言い換えると、観測される様々な音声のうち、異なる認識結果を与えるべきものはどれで、逆にばらつきを無視して同一視すべきものはどれなのか、動物は教示してくれない。ヒトである研究者が教師データを用意し、教師あり音声認識技術を動物音声に応用した事例[3][4]も存在するが、科学的分析としては客観性と再現可能性の問題が伴う。加えて、一般的な教師あり学習と同様に教師データ作成は作業負担が大きく、特に少人数の専門家による作成となるため要求される時間的コストが問題となる。

教師なし音声認識への取り組み

妥当な教師データが入手困難である事情から、動物音声認識には教師データを用いず、音声データのみから学習する手法が望ましい。そのような機械学習手法を教師なし学習と呼ぶ。

教師なし学習は動物音声だけでなく、ヒトの音声認識においても重要な役割を持つ[5]。第1に、文字を持たない言語に対しては教師あり学習の適応が難しく、音声データから自動的に適切なテキスト表現を発見し、文字起こしする手法の開発が求められる。特に消滅危機に瀕する少数言語の記録においては、言語学者によるフィールドワーク研究の人手不足問題を解決する手段として期待される。第2に、教師

* Takashi MORITA
1990年4月生まれ
Massachusetts Institute of Technology, Dept. of Linguistics & Philosophy, Ph.D program in Linguistics（2018年）
現在、大阪大学産業科学研究所　助教、同大学院　情報科学研究科　情報数理学専攻　助教　Ph.D in Linguistics
専門／機械学習・データサイエンス
TEL：06-6879-7608
FAX：06-6879-7612
E-mail：tmorita.sanken@osaka-u.ac.jp

図1　教師なし音声認識学習の構成要素。伝統的にはそれぞれが独立に行われてきたが、近年の深層学習を利用した研究では特徴量抽出と分類、ないしは分節から分類までを統一の基準下で最適化されるend-to-end学習が採用される。

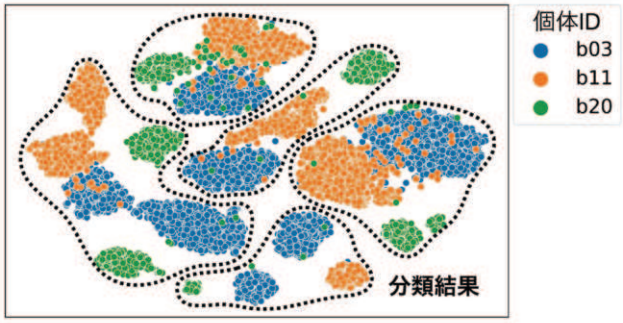

図2　ジュウシマツの音節の教師なし分類[13]。散布図は変分オートエンコーダ[9]によって抽出した各音節の16次元特徴量を個体毎に色分けし（全18個体中3個体を抜粋）、t-SNEで2次元空間上に可視化した。点線枠は特徴量抽出と音節分類のend-to-end深層学習で得られた分類結果を示す。

なし音声認識学習は認知科学的なヒトの音声言語学習モデル開発の側面を持つ。ヒトの音声言語学習は読み書きの学習に先行し、また先述のとおり文字を持たない言語も存在するため、音声とテキストを照らし合わせる教師あり学習はヒトの音声言語学習モデルとして受け入れられない。個人差を始めとする多様な音声変動の中から、ヒトの乳幼児が学習対象言語において適切な音声カテゴリを発見する様は、まさに教師なし学習の形態を取る。

典型的な教師なし音声認識学習は、(1) 音声の分節、(2) 各音節の特徴量抽出、(3) 特徴量に基づく音節の教師なし分類（クラスタリング）の3要素を含む（図1）。伝統的にはこれら3つの分析がそれぞれ独立に行われてきた。鳴禽類のさえずり分析を例に取ると、まず録音データから音声区間検出を行い、検出された音声区間がそのまま分節結果として採用される（すなわち、無音・非音声区間によって音声の句切れが定義される）[3)6)]。次に、得られた音節毎に特徴量を抽出する。特徴量には基本周波数や音節長、スペクトルエントロピーのような手動指定のものが採用される他[3)]、スペクトログラムデータをそのまま低次元ベクトルに圧縮する手法も試みられており[6)7)8)]、後者では変分オートエンコーダ[9)]を中心に深層学習の活用も行われている[7)8)]。最後に得られた特徴量ベクトルに対してクラスタリングを行うことで、各音節を離散的カテゴリに分類する[6)]。

上記のように段階毎に分割された分析とは対照的に、深層学習においては全ての段階を一貫して行うend-to-endな分析が良しとされることが多い。教師なし学習の場合、end-to-end分析の最も大きな利点の1つは最適化方針の一貫性である。例えば、特徴量抽出と分類を別々のアルゴリズムで実装した場合、それぞれが異なる方針の下で最適化されるため、得られる特徴量は必ずしも分類器の挙動に沿った分布を示さない。実際、近年生物学分野で広く用いられる特徴量抽出・次元削減法であるUMAP[10)]は、可視化（入力空間の距離的性質の保存）目的とクラスタリング目的とで異なるパラメータ設定を推奨しており[11)]、特徴量抽出とその後に続く分類の接続が一筋縄ではいかないことを顕著に表している。また複数の最適化方針が混在することで、全体としてどのような基準で分類が得られたのかを解釈することが困難となる。これに対し、end-to-end学習では特徴量抽出と分類が単一の目的関数に沿って最適化されるため、接続の問題を回避でき、分類基準の解釈も容易となる。

最適化方針の一貫性の他、特徴量抽出と分類のend-to-end深層学習によって分類に反映すべきでない情報を除外することが可能となる[12)]。例えば、鳴禽類の一種であるジュウシマツの音声は特徴量空間で明瞭なクラスタ分布を示すが、一方で個体差が顕著なためクラスタリング適用時に個体毎に独立したカテゴリが推定されてしまう（図2）。筆者らの研究では、end-to-end深層学習によってこの個体差を除去した音声分類を実現した[13)]。

ジュウシマツのさえずりの場合、音声区間検出によって得られる音節が明瞭なクラスタ分布を示すため、さらに詳細な音声分節の必要性が議論されることは少ない。一方、ヒトを含むその他多くの動物音声では、連続的に発せられた音声の中に複数の構成要素が含まれる[6)]。そして各要素の区切りは録音記録上では不明瞭である場合が多く、特徴量抽出・分類に適切な分節を独立に行うことは難しい。このため、ヒトの教師なし音声認識では分節から分類まで

を全て一貫して行うend-to-end学習が主流となっている。こうして得られた教師なし音声認識器は、一部のベンチマークにおいて教師あり学習を上回る成績を達成するまでに至っているが、一般的に言語学で仮定される音素数（～60種類程度）を大幅に上回る音声カテゴリ（512種類程度）が仮定されており、これらが細かなスペクトルパターンに対応付けられているため、細切れの分節が推定されてしまっている[14]。すなわち、適切な記号列表現が得られているというよりも、音声を一定長のフレーム毎に9ビットで圧縮表現しているような状態であり、目標とする音声認識結果とは依然大きな乖離があると言える。

おわりに

本稿では、動物用音声認識の実現に向けた機械学習応用に関する取り組みを紹介した。音声の文字起こしを予め教師データとして用意できない動物音声では、教師なし学習技術の開発・応用が重要となる。その際、深層学習を活用したend-to-end分析は、一貫した方針の下での最適化や個体差の除去等を可能にするため、今後も重要な役割を担うと考えられる。一方で、音声分節については課題が顕著であり、今後の開発において一層の注力が求められる。

参考文献

1) Vaswani et al., *NeulIPS*, pp. 5998－6008, (2017)
2) Dosovitskiy et al., *ICLR*, (2021)
3) Tachibana et al., *PLOS ONE*, Vol. 9, No. 3, pp. 1－8, (2014)
4) Oikarinen et al., *JASA*, Vol. 145, No. 2, pp. 654－662, (2019)
5) Zero Resource Speech Challenge, https://www.zerospeech.com
6) Sainburg et al., *PLOS Comp. Biol.*, Vol. 16, No. 10, p. e1008228, (2020)
7) Coffey et al., *NPP*, Vol. 44, No. 5, pp. 859－868, (2019)
8) Goffinet et al., *eLife*, Vol. 10, p. e67855, (2021)
9) Kingma & Welling., *ICLR*, (2014)
10) McInnes et al., *arXiv*, 1802.03426, (2020)
11) UMAP公式実装, https://umap-learn.readthedocs.io/en/latest/clustering.html#umap-enhanced-clustering
12) Chorowski et al., *IEEE/ACM Trans. Audio, Speech, Language Process.*, Vol. 27, No. 12, 2041－2053, (2019)
13) Morita et al., *PLOS Comp. Biol.*, Vol. 17, No. 12, p. e1009707, (2021)
14) van Niekerk et al., *INTERSPEECH*, pp. 4836－4840, (2020)