5Q-01

深層学習による野外録音音声からの動物の鳴き声検出とその環境影響評価への応用-オオタカの鳴き声を事例として-

太田 望 * 今井 俊輔 * 延原 肇 *
Nozomu Ohta Shunsuke Imai Hajime Nobuhara
徳江 義宏 † 今村 史子 † 太田 敬一 † 夏川 遼生 ‡
Yoshihiro Tokue Fumiko Imamura Keiichi Ota Haruki Natsukawa

1. はじめに

近年、希少動物の保護、および環境影響評価の一環として、高速道路やダムなどの大規模建設の際、動物の鳴き声の野外録音による生態系の事前調査が行われている [1]。しかし、動物の鳴き声からその種類を特定するためには長時間の録音音声とそのスペクトログラムを専門家が膨大なコストを費やして確認しなければならない。これを解消するために、動物の鳴き声を音声データから自動的に検出する研究がなされている [2] が、雑音が入る環境において実用的な精度で多様な生物の鳴き声を検出できるシステムの実現には至っていない。

本研究ではこのような雑音を含む音声データから特定の動物の鳴き声を検出する手法を、対象動物の鳴き声に合わせて設計した畳み込みニューラルネットワークを用いて実現する。ここでは一例としてオオタカを対象とした畳み込みニューラルネットワークによる検出を試みる。

合計 9 時間の野外録音データから、オオタカの鳴き声を検出する実験を行い、従来の検出手法 [2] に比べて、20%の精度が向上することを示す。

2. 提案手法

提案法では野外録音の当該データからメル周波数ケプストラム係数によって抽出された特徴ベクトルを畳込みニューラルネットワークに入力し、4 クラスのどれに分類されるかを示す 4 次元の確率ベクトルを出力する (図 1)。従来研究 [2] では中間層 1 の多層パーセプトロンによる分類が行われているが、入力されるデータは雑音がなく、図 2(左) のように 1 秒間特定の野鳥の鳴き声だけが聞こえる理想状態におけるものである。実際には人の手でラベルをつけるため、図 2(右) のようにラベルのついた区間全てに対象の動物の鳴き声が含まれているとは限らない。本研究の独創的な点はこの鳴き声区間の位置ずれの問題を畳み込みニューラルネットワークによって解消することである。

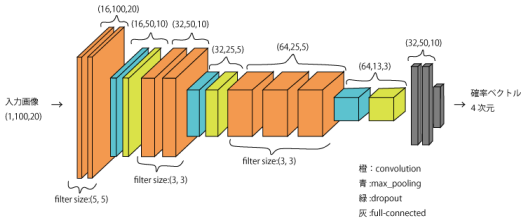

図 1: 提案手法のモデルの概形

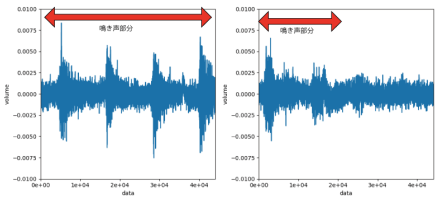

図 2: 鳴き声のサンプル、理想状態 (左) と実録音の状態 (右)

3. 実験

提案手法と従来法である多層パーセプトロンによる各ラベル分類の精度、再現度、F 値を評価の指標として比較する。

分類データとして、実際の現場で計測に従事している企業から提供された、44.1kHz のサンプリング周波数で録音された約 9 時間の野外録音音声を用いる。データ範囲は [-1,1] で値は浮動小数点型 32bit である。このデータに関しては、1 秒ごとに専門家によるアノテーションが表 1 に示すように事前に行われている。背景音の部分が

表 1: データセットの詳細

クラス番号	ラベル名	秒数	割合
0	背景音のみ	8406	0.396
1	人工音	3103	0.146
2	動物の鳴き声	8406	0.396
3	オオタカの鳴き声	1315	0.062
	合計	21230	1

* 筑波大学, University of Tsukuba
† 日本工営株式会社, Nippon Koei Co.,Ltd.
‡ 横浜国立大学, Yokohama National University

多いため、動物の鳴き声と等しくなるようにアンダーサンプリングした。このクラスのうち、背景音としてアノテーションされているものは、常に水の流れる音が録音されているものである。人工音は車や電車、人の声などを含む。また、他の動物の鳴き声は主にカラス、分類できない鳥の鳴き声である。これらのデータを学習用データとテスト用データに 8:2 の割合で分割する。

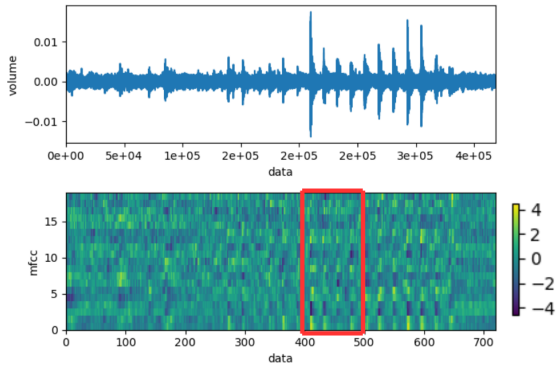

図 3: 原音声 (上) と mfcc 変換後の係数分布 (下)

本研究では、音声データを 441 サンプルごとにメル周波数ケプトラム係数により 20 次元の特徴ベクトルに変換する。特徴ベクトルをバッチごとに標準化する。元の音声データと変換された特徴ベクトルの様子を図 3 に示す。1 秒間分の特徴ベクトルを画像として畳み込みニューラルネットワークに入力するので、図 3 に示す赤枠で囲まれた 100x20 の画像が入力となる。提案手法のハイパーパラメータはバッチサイズ 32、イテレーション間隔 424、エポック数 100 と設定する。畳み込み層、全結合層の活性化関数はすべて reLu を用いる。出力層の活性化関数はソフトマックス関数、損失関数は交差エントロピーである。最適化手法は Adam を用いる。比較対象の従来手法では、1 秒間の特徴ベクトル 100 個の各要素について正規化した後に平均と標準偏差を計算し、1 つのベクトルに結合し 40 次元のベクトルとして入力する。中間層のノード数は 80、40、4 とする。中間層数以外のハイパーパラメータは提案手法と同じに設定する。

表 2 に多層パーセプトロン、表 3 に畳込みニューラルネットワークのテストデータに対する精度、再現度、F 値を示す。表のとおり、従来手法に比べ提案手法のオオタカの鳴き声の検出精度は 21%向上し、全体の検出精度は 20%向上した。提案手法のオオタカの鳴き声と人工音のラベルの再現度が他のラベルと比べて低いが、これはデータセットのラベルの偏りが影響していると考えられる。また、提案手法の学習曲線を図 4 に示す。

表 2: 多層パーセプトロンの精度、再現度、F 値

label	precision	recall	f1-score	support
0	0.52	0.68	0.59	1687
1	0	0	0	646
2	0.49	0.56	0.52	1632
3	0.56	0.41	0.47	281
avg/total	0.43	0.51	0.47	4246

表 3: 畳込みニューラルネットワークの精度、再現度、F 値

label	precision	recall	f1-score	support
0	0.65	0.68	0.67	1732
1	0.51	0.55	0.53	579
2	0.64	0.62	0.63	1670
3	0.77	0.56	0.65	265
avg/total	0.63	0.63	0.63	4246

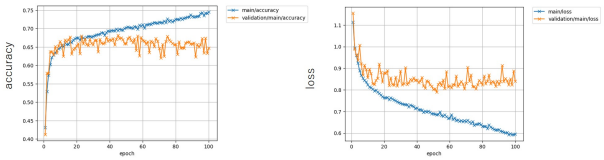

図 4: 精度 (左) と誤差（右）の学習曲線

4. おわりに

本研究では野外環境においてオオタカの鳴き声を検知するという課題に対して畳込みニューラルネットワークによる多クラス分類手法を提案した。実験では比較対象より精度は 21%、再現度は 15%、F 値は 18%の向上という結果より、提案法の有効性を確認した。今後の課題としては、より野鳥の鳴き声の検出に向いた特徴抽出手法の検討が挙げられる。また、現場では多少精度を落としてももれなく対象動物の鳴き声を検出することが求められるので、データセットのラベルの偏りをなくす、誤差関数に重みをつけるなどの再現度を上げる工夫が必要である。

参考文献

[1] Seppo Fagerlund, Automatic Recognition of Bird Species by Their Sounds, Helsinki University of Technology, 2004-11.

[2] 東谷幸治, 三田長久, 他, 音声情報を用いたニューラルネットワークによる野鳥の種識別, 電子情報通信学会総合大会講演論文集 2007 年情報・システム (1), 146, 2007-03-07.