H-015

奄美希少野鳥の鳴き声特徴抽出および自動識別について
Acoustic Feature Extraction and Automatic Identification of Rare Birds Vocalizations in Amami

上村 優介† 福元 伸也† 鹿嶋 雅之† 渡邊 睦† 榮村 奈緒子‡ 鵜川 信‡
Yusuke Uemura Shinya Fukumoto Kashima Masayuki
Mutsumi Watanabe Naoko Emura Shin Ugawa

1. はじめに

1.1 研究背景

2021 年, 奄美大島は「奄美大島, 徳之島, 沖縄島北部及び西表島」として世界自然遺産に登録された. 登録の主な理由は, 「生物多様性」とされている. これは, 現地に生息する固有種や希少生物の存在が高く評価されていることの影響が大きい. 確かに奄美大島には, ルリカケスやアマミノオオウサギをはじめとする様々な固有生物が生息している.

また, 屋外環境での長時間音声録音を用いた鳥の鳴き声解析は, 生態系の保全などにおいて幅広い応用が期待されている. 特に, 鳥の生息分布は, 気温やエサとなる生物の有無などの影響により広大であるため, 複数種の鳴き声が含まれる環境下での認識精度向上が大きな課題とされている.

したがって, 奄美大島に生息する希少種を継続的に確認し続けることは重要な意味を持つ. そこで, 警戒心が強いため近距離からの映像撮影が難しく, 存在を確認するための人的負担が大きい「野鳥の自動識別」に取り組むことにした. 野鳥は, 距離がある場合は双眼鏡による目視での観測が行われるが, 長時間飛行する鳥や夜行性の鳥に関しては鳴き声による観測が一般的である.

1.2 研究の目的

現在, 録音機器の設置および音源の回収から, 大量の音声データをもとにした野鳥の存在の確認, 種識別に至るまで, ほぼすべての工程で人力による調査が行われている. 数千時間の自然音の中から, 数秒の鳴き声を特定するという作業を, 対象種の数だけ研究者が行わなければならない. 本研究ではこのような作業負担を軽減し, 自動で野鳥を識別できるシステムの開発を行うことを目的とする. 具体的には, 共同研究者が長時間録音の音源データを入力したら, 音源に前処理, 鳴き声箇所の抽出処理, 種識別処理が順次適応され, 鳴き声箇所の時間や音源ファイル名, 種などを出力するシステムである. これにより, 18000 時間の録音データ中の野鳥の鳴き声箇所および種についての解析を行うことを目的としている. なお, 研究者がラベル付けを行う教師データには時間や数に限りがあり, 将来的には未知の音源データに対して長期的に自動識別を行うため, 可能な限り高い識別精度が要求される. したがって, 自動化システム全体の設計・運用は鳴き声箇所の検出処理および種識別処理の開発後に予定している.

1.3 奄美大島に生息する野鳥

本研究は, 奄美大島に生息する野鳥 10 種(希少種を含む)を対象とする(表 1).

表 1 解析対象種の一覧

解析対象(※希少種)	
アオバズク	**アカヒゲ**
アカショウビン	**オオトラツグミ**
ズアカアオバト	**カラスバト**
リュウキュウキジバト	**リュウキュウコノハズク**
キツツキ(ドラミング)	**ルリカケス**

表 1 にて太字で示された種(右列)は希少種であり, 黒文字で示された種(左列)は奄美大島に豊富に生息する一般種だ. 録音は, 奄美大島各地に設置したマイクから収集している. 録音の収集と, 鳥の種別特定作業は、農学部の共同研究者によって行われる.

1.4 鳴き声のスペクトログラム画像

鳴き声による解析において, 周波数の情報を用いることは当然である. 一方で, 機械学習や図形特徴を利用した識別手法への応用が利くことから, 画像へ変換して解析を行う手法がある. スペクトログラム画像はその一例である(図 1). スペクトログラム画像とは, 音声信号や周波数成分の時間変化を可視化するために使用される表現方法だ. 横軸が時間, 縦軸が周波数, 色の濃淡が周波数成分の強度を表している. 野鳥の鳴き声は破線や曲線で, ある程度の特徴を持って現れる(図 2). Kaleidoscope Pro によると近しいことも, スペクトログラム画像の特徴の一つといえる.

また, スペクトログラム画像はカラーマップが選択可能である. 本研究では, 予備実験により最も識別の精度が良好だった「jet」をカラーマップとして採用することにした.

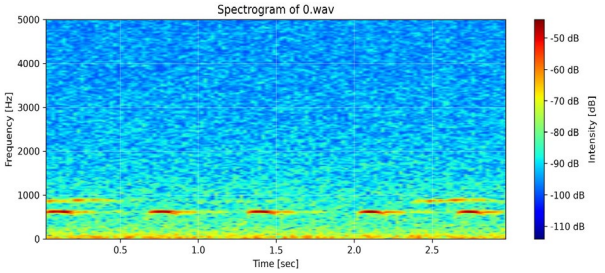

図 1 アオバズクのスペクトログラム画像

† 鹿児島大学工学部 Faculty of Engineering, Kagoshima University, Japan
‡ 鹿児島大学農学部 Faculty of Agriculture, Kagoshima University, Japan

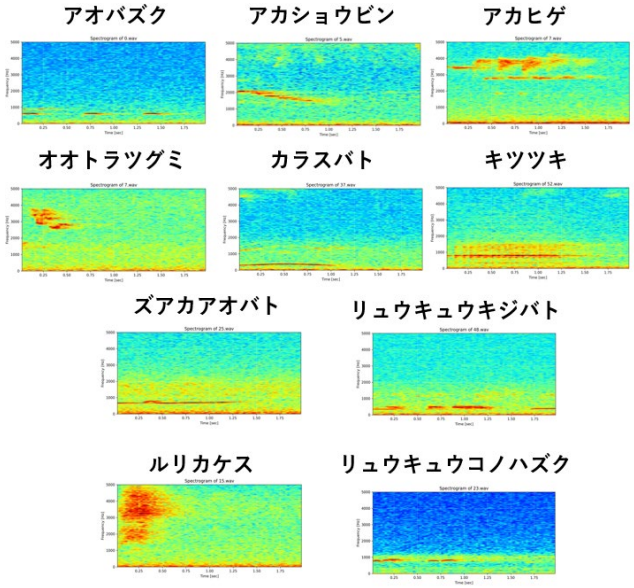

図 2　対象種のスペクトログラム画像一覧

2. 関連研究

2.1 野鳥の識別に関する関連研究

まずは, 野鳥の識別に関する研究について紹介する.

前川らは, 畳み込みニューラルネットワーク(CNN)で, 前処理を施したスペクトログラム画像の分類を行った[1].学習の結果, ランダムフォレスト, 多層パーセプトロンが 80%前後の精度であったのに対して, 90%という高い精度で識別できた. ただし, 対象は絶滅危惧種のサシバ 1 種のみである.

東谷らは, 周波数帯域パワーを用いたニューラルネットワークによる識別を行っている[2]. 野鳥 12 種につき全体の識別率が 94.29%と高い精度を出せているほか, 60%を下回っているものがないため, 周波数帯域パワーによる識別について有用性があるといえる. ただし, 対象としたのは夜行性野鳥であり, 鳴き声特徴が似ている鳥や希少種に関しては検証が必要だと述べている.

鹿児島大学における先行研究では, スペクトログラム画像を用いて畳み込みニューラルネットワーク(CNN)による識別を試みた[3]. 複数音が重複している音源環境に対して, 周波数情報をもとにしたサポートベクタマシンでの識別精度は著しく低いものであった. そこで, スペクトログラム画像に変換し, 類似した特徴を持つ鳥(カラスバト・ズアカアオバト・リュウキュウキジバト)を 2 段階に分けて識別することで, 識別精度向上を図った. その結果, 全種平均 95%の識別精度を達成した. ただし, 教師データ不足で識別できなかった種があることや, 重複環境から鳴き声音源の分離を行うことが課題として残ったとしている.

Lasseck は, Kaggle の野鳥分類課題である BirdCLEF において, 畳み込みニューラルネットワーク(CNN)と周波数帯域への注意機構(attention)を組み合わせた新しい手法を提案した[4]. 特徴抽出には EfficientNet, 分類には周波数アテンション付きの Sound Event Detection ヘッドを用い, 約 14 万件の音声を使って学習し, 5 秒単位のスペクトログラム画像を CNN に入力した. 時間方向ではなく, 周波数帯域への注意機構を導入することで, 複数種の鳴き声が重複しているときに別のクラスとしてクラスタリングすることができるようになった. その結果, 評価指標である cmAP で 76.3%を達成することができた.

Aytar らは, 学習済みのシーン認識のモデルを音声認識の教師ラベルにするという手法「SoundNet」を提案した[5]. 画像内の物体を分類する CNN の出力と画像内のシーンを分類する CNN の出力を半教師データとし, 音声を入力してその音声に含まれていると予想される物体やそのシーンを上位 3 種まで予測することを可能とした. その結果, それぞれのシーンや物体のための教師データが用意されている環境と比較して高性能な識別性能を得ることができた. この研究は, 莫大な教師データを必要とする画像識別において, 画像を教師データとして音声でマルチモーダルに学習することで良好な精度が得られるというアプローチを示した研究である.

これらより, 周波数情報の特徴を画像として表現するスペクトログラム画像の有用性に期待が持てる. また, それをもとに CNN で識別を行うことで高い精度を得ることができそうだと考えられる. なお, 本研究において, 周波数成分の強度は, スペクトログラム画像の色の濃淡として活用される.

2.2 鳴き声箇所の特定に関する関連研究

続いて, 鳴き声箇所の特定および抽出に関する研究を紹介する.

牧野らは, 長時間録音から野鳥の鳴き声箇所を抜き出すために, 雑音の平均値および相関を利用した[6]. 前処理には, 音声波形の分散や自己相関を計算しそこから雑音の平均値を定めることで対応した. それを音声全体の平均とみなし分散を再計算し平均からのズレが大きい時間帯を鳴声, 小さい時間帯を雑音とみなした. その結果, ピンクノイズ, ホワイトノイズの印加によって検出箇所特定の精度は安定することが分かった. また, 計算の簡素化によってよりリアルタイム性を高めることが計画されている.

Priyadarshani らは, 対象野鳥 6 種にそれぞれマスクを作成し, 連続ウェーブレット変換(CWT)した長時間録音を走査し, 閾値以上の相関が出た箇所を鳴き声として検出した. その結果, 平均精度 85%, シジュウカラとクロウタドリで 90%を達成した[7].

Potamitis らは, スペクトログラム画像に対して, 鳥の発声箇所のみマスクし, それ以外をゼロとするセグメンテーションを施すことで鳴き声箇所の抽出を試みた. このアイデアは, 電子顕微鏡画像での特徴がブロブ(塊)としての特徴を持つことに着想を得ている. その結果, 0.06~0.76 秒/1 画像と高速な鳴き声箇所の特定を行うことができるモデルを作成した[8].

金田らは, 楽器の演奏区間検出に機械学習を用いている[9]. 今まで一般的であった GMM-HMM (ガウス混合―隠れマルコフモデル)に代わって, DNN-HMM(Deep Neural Network－隠れマルコフモデル)を利用した. その結果, ドラム, ギターなど対象 5 種の識別で GMM-HMM の性能を 10%上回った. つまり, メルケプストラム係数を特徴量として使用する場合, 混合正規分布よりも尤度をもとにした検出のほうが高精度で識別できる傾向が確認された. ただし, 検出対象は楽器であり, 音圧が高く特徴の差異が大きい. 野鳥で有用かどうかの検証が必要である.

これらより, 鳴き声箇所の特定には, 畳み込みニューラルネットワークでの学習時に種と検出箇所を同時に特定するアプローチと, マスク処理やスペクトログラム画像のクロージングなどの前処理によって鳴き声箇所を特定し, 識別処理と分けて行うアプローチが存在することが分かる.

本研究では, 奄美大島に生息する野鳥 10 種について, 鳴き声箇所を包含する 2 秒で抽出された音源およびスペクトログラム画像を用いて識別を行う. 識別には, 4 種の機械学習モデル(GMM-HMM, DNN-HMM, RNN, CNN)を用いて, 識別精度を比較する. なお, スペクトログラム画像を用いるのは, 時間的変化や周波数特性を画像特徴としてみることができるという性質を利用するためである.

また, 最終的に自動化システム全体を構築し, 学習に未使用のラベル無し音源での精度検証が必要である. したがって, システム全体の設計・構築は, 検出・識別の各処理の精度検証が完了次第行うものとする.

3. 鳴き声特徴抽出および自動識別について

本研究では, 野鳥の自動識別に向けて, 音響に施す前処理および機械学習モデルの識別精度に与える影響について調査する(図 4). 3.2 および 3.3 で述べるスペクトルサブトラクションおよび音響マスクの作成については, 論文執筆段階で十分な検証ができていないので概要にとどめる.

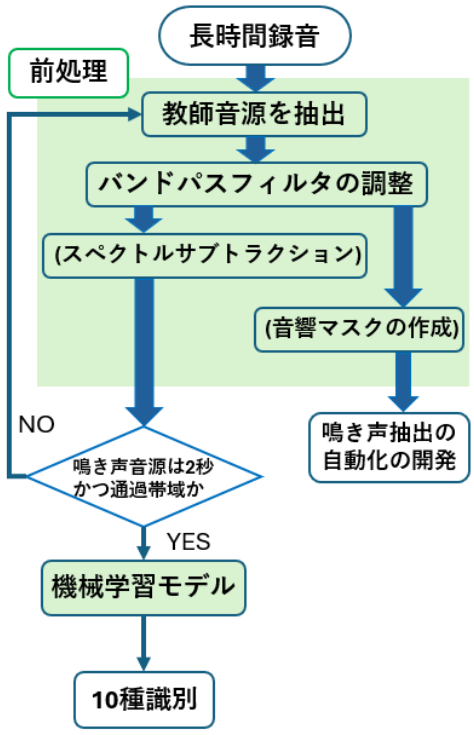

図 3　提案システム

3.1 バンドパスフィルタの作成

システム全体の自動化を検討する際, 対象種の鳴き声と機械学習に適した前処理が重要となる. 本研究では, メルケプストラム係数(MFCC)とスペクトログラム画像の元となる.wav ファイルに対して適切なフィルタを適用し, 対象種の特徴を鮮明にすることを試みる. そこで, 対象種の特徴を整理した. 周波数帯域は, リュウキュウキジバトの 200Hz～700Hz が最も低く, ルリカケスやオオトラツグミの 1500Hz～4500Hz が最も高かった. また, 鳴き声時間に関しては, 繰り返しなく種もいることから, 2~3 秒で鳴き声を包含できると考えた. しかし, 予備実験の結果, 3 秒の音源には, 非常に高い確率でほかの生物の音や対象音以外の音が含まれてしまうことが確認できたため, 使用する音源はすべて 2 秒で統一した. このことから, 長時間録音を 2 秒で分割し, 200Hz から 4500Hz のバンドパスフィルタを作成することにした.

しかし, バンドパスフィルタにも種類がある. Butterworth フィルタと Cutoff フィルタである(図 5). Butterworth フィルタは, 通過帯域内で波形を可能な限り維持しながら減衰させるフィルタである. 一方, Cutoff フィルタは, 指定された境界周波数で波形情報を切り落とすフィルタである. 200Hz 以下には風切り音や録音装置に雨水が当たる音などが常にちらついており, 識別において 1kHz 程度のハト類（カラスバト・ズアカアオバト・リュウキュウキジバト）の誤識別に大きく影響している. 本研究では, 自然音源と 2 種類のフィルタ処理された音源(Butterworth フィルタと Cutoff フィルタ), あるいは, それらから変換されたスペクトログラム画像を使用して機械学習モデルを作成し, その精度を比較する.

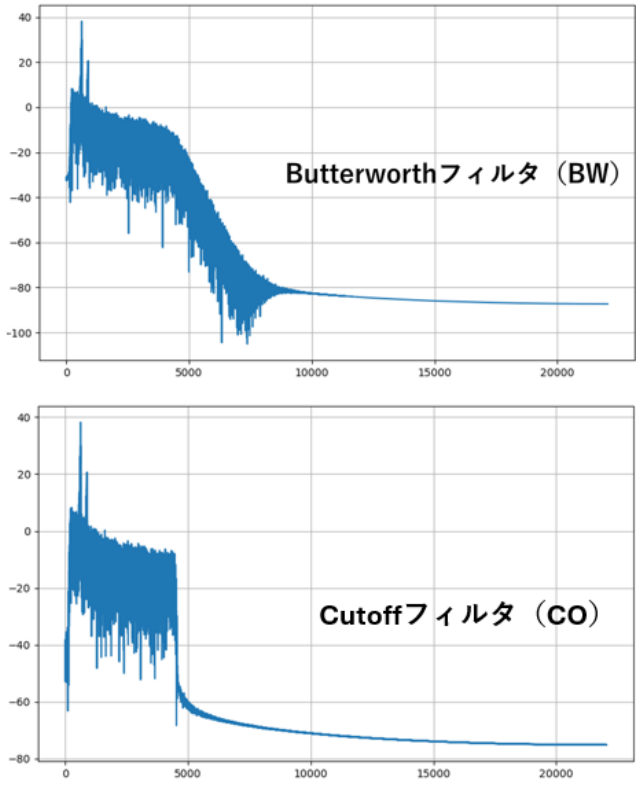

図 4　バンドパスフィルタの減衰特性の違い

3.2 スペクトルサブトラクション

現在, 鳴き声抽出に関する前処理および抽出手法について複数の案を検討している. ここでは, 今後検証を進めていく予定である処理について紹介する.

まず, 雑音の除去に関して, スペクトルサブトラクション法という手法がある. スペクトルサブトラクション法は, その理論的単純さと実装の容易さから, 音声強調分野において広く採用されている基本的手法である. 適切なパラメー

タ設定により, 実用的な雑音除去効果を得ることができる. しかし, ミュージカルノイズや音声歪みといった固有の問題があるため, 応用に際してはこれらの制限事項を十分に考慮する必要がある.

3.3 音響マスクの生成

鳴き声箇所の抽出について, 音響マスクを作成する手法だ. 音響マスクとは, スペクトログラム画像上で鳥の鳴き声が存在する時間-周波数領域を白色で, それ以外の領域を黒色で表現したバイナリ画像である. 鳥の鳴き声は種固有の周波数特性を持つため, スペクトログラム上で「スペクトラルブロブ(塊)」として現れる. これらのブロブの形状, 位置, 強度が種の同定や行動解析の重要な手がかりとなると考えている.

3.4 機械学習モデル

機械学習モデルの訓練は, 対象種ごとに 100 データの教師音源またはスペクトログラム画像を使用して行われる. 訓練に使用される音源とスペクトログラム画像は, 80 データを訓練用, 10 データを検証用, 10 データをテスト用に分割し, クロスバリデーション用に 10 個のモデルを作成する. 訓練には, GMM-HMM モデル, DNN-HMM モデル, RNN モデル, CNN モデルの 4 種類のモデルを使用する. 実験で用いられたモデルの説明は以下の通りだ(図 6).

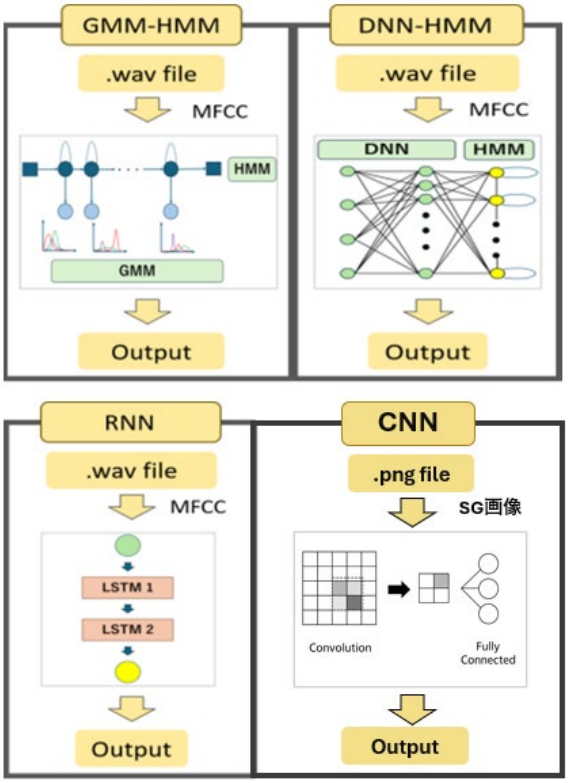

図 5　4 種の機械学習モデル

3 つの音響処理モデル（GMM-HMM, DNN-HMM, RNN）の入力は.wav ファイルであり, 使用される特徴量は MFCC だ. GMM-HMM(ガウス混合―隠れマルコフモデル)は, 混合ガウス分布の尤度に基づいて, 入力音源がどの鳥の音源に類似しているかを分類する. 音響認識や音声認識の分野で一般的に使用されてきた処理手法で, このプロセスを利用しているソフトウェアも多く存在する. DNN-HMM(Deep Neural Network―隠れマルコフモデル)は, 音響特徴量が DNN に入力され, DNN の出力層のノードが HMM の各状態にマッピングされる. つまり, GMM で求められた尤度に代わって, 事後確率が HMM に伝達される. RNN(Recurrent Neural Network)は, 時系列データやシーケンスデータを処理するために設計されたニューラルネットワークだ. 入力データと前の隠れ層の状態を基に、新しい隠れ層の状態が計算されることで再帰性が実現されている. 本研究では, MFCC に基づく 2 層の LSTM で構成され, 128 の隠れ層を有している.

CNN(Convolutional Neural Network)モデルは, 画像処理で高い識別精度を出すことで知られている. 入力は.png ファイルであり, 使用される特徴量はスペクトログラム画像である. 畳み込みでは Keras の Conv2D にて 3×3 のフィルタを 64 枚使用する. その後, MaxPooling(ここでは 2×2 の最大値抽出)を行い, 平滑化が実行され, 全結合層を通過した後, SoftMax 関数を用いて各出力の期待確率が計算される.

4. 実験

4.1 データの準備

共同研究者が付与したラベルに基づいて, Python を使用して鳥の鳴き声を含む特定音源を抽出する. 種ごとの訓練データの量の違いによる影響を回避するため, ラベル付きデータが最も少なかった「アカショウビン」のデータの量（100 の自然音源）を, 本研究における基準として使用した. これらの 100 の音源を基に, 訓練データ, 検証データ, テストデータに分割し, 組み合わせを変更しながら, 合計 1000 ファイルでモデルを構築する. 用意した合計 1000 ファイル（対象種 10 種それぞれに 100 音源）について, 1000 ファイルは自然音源データ、1000 ファイルは Butterworth フィルタを適用した音源、1000 ファイルは Cutoff フィルタを適用した音源となるようにコピーを用意する. 併せて, それらの音源をもとに, 対応するスペクトログラム画像を作成する.

4.2 バンドパスフィルタの調整

ここでは, 自然音源, Butterworth フィルタを適用した音源, Cutoff フィルタを適用した音源から作成されたスペクトログラム画像を比較する(図 7). 自然音源では, 赤で強調表示された鳥の鳴き声特徴に加え, 黄緑色と淡い青色の周波数成分が存在することが分かる.

Butterworth フィルタを適用した後の画像では, 通過帯域(200Hz-4500Hz)外の領域で黄色がかった緑色と淡い青色の周波数成分が減少しており, 通過帯域内でも鳴き声の部分がより明確になったことが確認できる.

Cutoff フィルタを適用した後の画像では. 通過帯域外の領域の成分が消失していることが確認できる.

これらの前処理は, 存在した周波数情報を削いでいると考えることもできる一方で, 画像特徴は鮮明になっている.

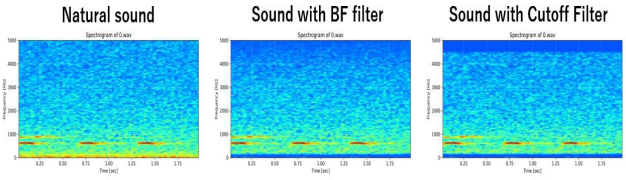

図 6　バンドパスフィルタ適応前後の変化

4.3 スペクトルサブトラクション

3.3 で紹介した検討中の手法に関して, 現在判明している内容を示す.

スペクトルサブトラクションを適応した際の変化の例を以下に示す(図 8). スペクトルサブトラクションは, 音声活動が検出されない区間（通常は録音開始部分）において雑音スペクトラムを推定するため, 2 秒という短いクリップでの有効性および適切なパラメータを模索する必要がある.

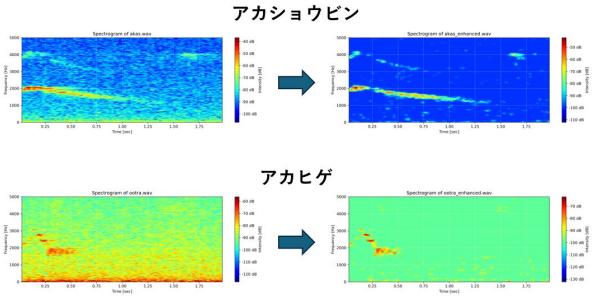

図 7　スペクトルサブトラクションによる変化

4.4 機械学習のパラメータ

ここでは, 各モデルの処理の流れと学習に使用した各パラメータを記載する.

4.4.1 GMM-HMM

GMM-HMM では, 音声の前処理および特徴抽出, GMM-HMM の学習, 正答率の計算という大きく 3 段階に分けられる. 使用したパラメータを以下に示す(表 2).

表 2　GMM-HMM のパラメータ

パラメータ名	値	説明
sampling_rate	16000	音声情報を変換
n_fft	2048	Librosa デフォルト
hop_length	512	Librosa デフォルト
n_mfcc	13	次元数を指定
n_components	5	隠れ状態を指定
n_mix	3	音のばらつきに対応
class	10	対象種は 10 種
n_iter	100	繰り返し回数

4.4.2 DNN-HMM

DNN-HMM では, 音声の前処理および特徴抽出, DNN によるフレームごとの状態確率の出力, HMM による時系列モデリング・Viterbi デコーディング, 10 クラス分類という処理フローが行われる. 使用したパラメータを以下に示す(表 3).

表 3　DNN-HMM のパラメータ

パラメータ名	値	説明
sampling_rate	16000	音声情報を変換
n_fft	512	Librosa デフォルト
n_mfcc	13	次元数を指定
n_components	5	隠れ状態を指定
dropout	0.3	過学習を抑制
learning_rate	0.001	Adam による最適化
epoch	100	学習の繰り返し回数

4.4.3 RNN

RNN では, 音声の前処理および特徴抽出, 系列化, RNN モデルによる時系列モデリング, 最後の隠れ状態の確認, 分類器(Dense)による 10 クラス分類という工程で処理が進む. 使用したパラメータを以下に示す. (表 4).

表 4　RNN のパラメータ

パラメータ名	値	説明
sampling_rate	16000	音声情報を変換
n_fft	13	次元数を指定
RNN_type	LSTM	RNN の 1 種
n_layers	2	LSTM2 層で学習
Dense_units	64	特徴の圧縮と分類
Output_units	10	10 種を分類
learning_rate	0.001	Adam による最適化
epoch	200	学習の繰り返し回数

4.4.4 CNN

CNN では, 画像の入力が行われ, 畳み込みと MaxPooling をそれぞれ 2 回繰り返し, 平滑化を行い, 全結合処理を経て, 分類が行われる. 使用したパラメータを以下に示す(表 5).

表 5　CNN のパラメータ

パラメータ名	値	説明
target size	(64, 64)	画像の解像度を設定
画像のチャネル	3	1 は Gray, 3 は RGB
フィルタ数	64 (×2 層)	フィルタを作成
kernel サイズ	(3, 3)	畳み込みを実行
Pooling サイズ	(2, 2)	プーリングを実行
全結合層	256, (10)	256 ユニットから 10 へ
出力クラス数	10	10 種の識別
learning_rate	0.001	Adam による最適化
epoch	50	学習の繰り返し回数

4.5 結果

ここでは識別モデルに関する結果を示す. 以下に, 各モデルおよび各入力に対する識別精度の結果を示す(表 6).

表 6　野鳥識別の結果

Model	Source	Accuracy
GMM-HMM	Natural	0.827
	BW Filter	0.853
	CO Filter	0.838
DNN-HMM	自然音源	0.868
	BW フィルタ	0.875
	CO フィルタ	0.896
RNN	自然音源	0.724
	BW フィルタ	0.774
	CO フィルタ	0.770
CNN	自然音源(SG)	**0.925**
	BW フィルタ(SG)	**0.930**
	CO フィルタ(SG)	**0.935**

GMM-HMM モデルは, 自然音源に対して 0.827, Butterworth フィルタを適応した音源に対して 0.853, Cutoff フィルタを適応した音源に対して 0.838 の識別率を示した.

DNN-HMM モデルは, 自然音源に対して 0.868, Butterworth フィルタを適応した音源に対して 0.875, Cutoff フィルタを適応した音源に対して 0.896 の識別率を示した.

RNN モデルは, 自然音源に対して 0.724, Butterworth フィルタを適応した音源に対して 0.774, Cutoff フィルタを適応した音源に対して 0.770 の識別率を示した.

CNN モデルは, 自然音源に対して 0.925, Butterworth フィルタを適応した音源に対して 0.930, Cutoff フィルタを適応した音源に対して 0.935 の識別率を示した.

GMM-HMM, DNN-HMM, RNN は, 各モデルあたり 3~10 分の訓練時間で, 比較的高速だった. 一方, スペクトログラム画像を特徴量として使用する CNN は, 訓練に 50 分以上かかるが, 最も高い識別精度を達成した.

5. 考察

結果から見て, バンドパスフィルタを適用することで、すべてのモデルにおいて精度が向上したことが示されたため, フィルタの適応が精度向上に良い影響を与える傾向があることが示唆された. ただし、Butterworth フィルタと Cutoff フィルタ間の性能差を評価するためには, より詳細な調査が必要である.

学習モデルに関しては, 識別精度 93.5%は期待している性能を満たしてしていると考えている. ただし、自動抽出された未知の音源に対しての識別性能の検証とは別の問題として把握すべきだ.

対象種別の識別精度も分析した. 例えば, 「ルリカケス」や「アカショウビン」のような, スペクトログラム画像の特徴が他の鳥と異なる種の場合, 風切り音を含むスペクトログラム画像に基づいた CNN でも高い識別精度を示した. 一方, スペクトログラム画像の特徴が類似している場合や, 鳴き声が重複する場合（例えば「カラスバト」と「キツツキ」）は, 識別精度が低下した.

本研究で用いられた音源には, 鳴き声以外の特徴, 例えば雨の音や小動物の足音などが含まれる. 誤識別されたスペクトログラム画像を確認したところ, ほとんどの誤識別は鳴き声の重なりが原因であることが判明した. 本研究においては, まずより適切な前処理について検討を重ねていく. 具体的には, 鳴き声重複環境における音源分離による対象種音の鮮明化を目指す. その際, 大雨や強風環境での音源は省き, 平常環境下での識別精度向上に努める.

さらに, エポック数を増やしても精度が向上しなくなったことから, ノイズの抑制には限界があると考えられる.一部の研究では, 一定距離離れて配置された 2 つのマイクを用いた空間情報分析の有効性が示唆されているが[10], 本研究では空間情報は使用することができない.

また, 本研究においては 18000 時間の自然音を扱うため, 誤識別によるデータが蓄積され続けることを無視できない. そのため, 識別性能を優先し, 一部を研究者が判断する半自動システムとすることも視野に研究を進める.

6. まとめと今後の課題

6.1 まとめ

本研究では, 自然音源から野鳥の鳴き声箇所を特定・抽出し, その種を識別するシステムを構築することを目的としている. そのため, ラベル付きデータをもとに野鳥の鳴き声箇所を識別する処理について, 複数の機械学習モデルを利用して識別精度を比較した.

識別処理については, 音響処理モデルとして GMM-HMM, DNN-HMM, RNN を用意し, 画像処理モデルとして CNN を利用した. その結果, CNN で得られた 93.5%の識別精度が最も高く, 全モデルで最低だったのが RNN の 73.4%だ.音源には, 小動物の足音や雨音, 風の音, 複数種の鳴き声などが含まれるため, より最適な前処理や学習パラメータを模索することで精度の向上が期待できる.

6.2 今後の課題

今後の課題としては, 自然音源から鳥の鳴き声箇所を自動的に抽出するプロセスの実装だ. これは, システム全体の実行時間に 最も大きな影響を与えると予想される.

次に, RNN や CNN の層数を含む機械学習パラメータの模索だ. 標精度は全種平均 90%である.

続いて, 自動的に抽出された音源を使用した分類の精度検証だ. このプロセスでは, 対象種以外の鳴き声や対象種と無関係な音の場合, その音源を分類から除外できるかどうかを確認することが重要である.

最後に, サーバーの設置とシステムの自動化のためのプログラムの改善に取り組む.

謝辞

本研究は，鹿児島大学における文部科学省ミッション実現戦略分事業「奄美群島を中心とした『生物と文化の多様性保全』と『地方創生』の革新的融合モデル」の一環として，支援を受けて遂行された.

参考文献

[1] 前川 侑子, 田口 華麗, 牛込 祐司, 佐藤 匠, 小林 啓悟, 芳賀 智宏, 町村 尚, 東海 明宏, 松井 孝典, “AI 技術による鳥類の鳴き声モニタリング手法の検討~サシバを事例として~”, Bird Research, Vol.18 (2022).

[2] 東谷 幸治, 三田 長久, 牧野洋平, “音声情報によるニューラルネットワークを用いた夜行性野鳥の識別”, FIT2006(第 5 回情報科学技術フォーラム) (2006).

[3] 眞島京音, 福元伸也, 鹿嶋雅之, 渡邊睦, 鵜川信, 榮村奈緒子, “奄美大島に生息する希少種の鳴き声自動認識に関する研究”, 火の国情報シンポジウム, A8-1 (2023).

[4] Mario Lasseck, “Bird Species Recognition using Convolutional Neural Networks with on Frequency Bands”, Conference and Labs of the Evaluation Forum (2023).

[5] Yusuf Aytar, Carl Vondrick, Antonio Torralba, (MIT), “SoundNet: Learning Sound Representation from Unlabeled Video”,30th Conference on Neural information Processing Systems (NIPS) (2016).

[6] 牧野 洋平, 三田 長久, 岩崎 祐介, 高橋 幸司, カムケオシーパーチャン, “雑音の平均値を用いた長時間録音からの野鳥の鳴き声の抜出”, FIT2007(第 6 回情報科学技術フォーラム), pp.331-332 (2007).

[7] Nirosha Priyadarshani, Stephen Marsland, Julius Juodakis, Isabel Castro, Virginia Listanti, “Wavelet filters for automated recognition of birdsong in long-time field recordings”, Methods in Ecology and Evolution, vol.11 (2020).

[8] Ilyas Potamitis, “Deep learning for detection of bird vocalizations”, arXiv:1609.08408[cs.SD] (2016).

[9] 金田 響, 岩野 公司, “Deep Neural Network を用いた楽器の演奏区間検出”, 情報処理学会第 78 回全国大会, pp501-502 (2016).

[10] 古山 諒, 鈴木 麗璽, 炭谷 晋司, 有田 隆也, “鳴禽類の雌のさえずりの役割の理解に向けた音源定位手法の活用に関する一検討”, 人工知能学会第二種研究会資料 AI チャレンジ研究会, 2021 巻 Challenge-058 号 (2021).