深層学習を用いた鳥類の鳴き声の短時間スペクトルの構造の分類*

☆ 中谷優太, 山田宏樹, 矢田部浩平 (農工大)

1 はじめに

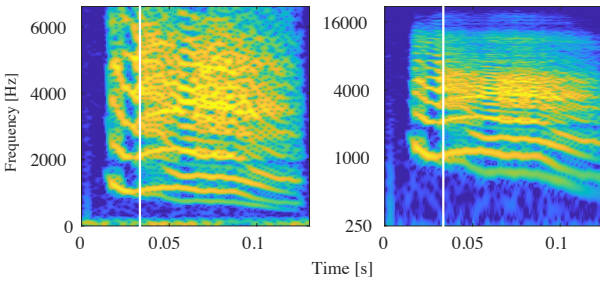
多くの鳥類の鳴き声は，気管支に位置する鳴管という器官で作られる．気嚢から空気が送られると，鳴管の左右に位置する labia と呼ばれるひだが振動し音が生じる [1, 2]．一部の鳥類は左右の labia を独立に制御し，異なる音を同時に鳴らすことができる．さらに，左右の labia の非線形な相互作用により，2 つの調波音の単純な重ね合わせでは表せない複雑な鳴き声を発する [2]．例として，キンカチョウの鳴き声のスペクトログラムを図–1 に示す．図より，1 つの調波で表せる構造を持つ区間と 1 つの調波では表せない構造を持つ区間があることが分かる．

キンカチョウは，決まったパターンの鳴き声を発してコミュニケーションをとることが知られているが，labia の振動制御と鳴き声の関係は解明されていない．labia の振動を解析するためには，鳴き声の各時刻のスペクトル構造を複雑さによって分類できることが望ましい．そこで本稿では，鳴き声のスペクトルの構造を CNN (Convolutinal Neural Network) を用いて分類する手法を提案する．また，少ない学習データで学習を行う際の過学習を防ぎ，汎化性能を高めるために，正則化やドロップアウトを適用し，テストデータに対する分類性能の変化を調査した．

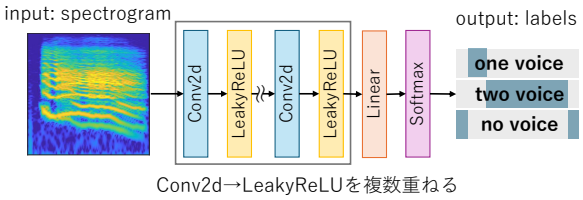
2 スペクトル構造を分類する CNN

キンカチョウの鳴き声のスペクトルには，1 つの調波で表せる構造 (one voice) と 1 つの調波では表せない構造 (two voice) が含まれる．前回，物理モデルを用いたシミュレーションにより，左右の labia の振動が異なると two voice が現れることがあると分かった [3]．このことから，one voice と two voice では，発する際の鳴管の状態が異なることが想定される．したがって，今後解析を行っていく上でこれらの分類を自動で行うことは有用である．そこで本稿では，スペクトログラムを入力し，各時刻のスペクトルの構造を one voice, two voice, no voice (無音区間) の 3 つのラベルに分類した結果を出力する CNN を作成した．

作成したネットワークの構造を図–2 に示す．ピッチに影響されずにスペクトルの構造を分類するため，図–1 の右図のように，入力するスペクトログラムの周波数軸は対数にした．時間方向のフィルタサイズは，予備実験で高い分類性能を示した 1 とした．



図–1 キンカチョウの鳴き声のスペクトログラム．左図，右図はそれぞれ周波数軸を線形，対数で示している．白い線で示した時刻を境に，1 つの調波で表せる構造から 1 つの調波では表せない構造へ変化している．



図–2 作成したネットワーク構造．Conv2d 層と LeakyReLU 層を 3, 5, 7 層にして実験を行った．ドロップアウトを適用する際には，全ての LeakyReLU 層の後にドロップアウト層を追加した．

3 実験

作成した CNN を用いて各時間フレームを分類し，その性能を評価した．キンカチョウの鳴き声 29 サンプルのうち，学習データは 23 サンプル，テストデータは 6 サンプルとした．学習用の 23 サンプルの時間フレーム数はそれぞれ異なり，合計 2606 個の時間フレームが含まれる．ラベリングは手動で行った．今回は学習サンプルが非常に少ないため，過学習を防ぎ汎化性能を高めることに焦点を当てて実験を行った．3.1 節では層数とフィルタ数について検討し，3.2 節では過学習の対策としてよく用いられる正則化やドロップアウトを適用した際の効果について検討した．

全ての実験に共通するパラメータについて記す．エポック数は 300 とし，n エポック目の学習率は $0.001 \times 0.99^{(n-1)}$ とした．損失関数には交差エントロピーを用いた．評価指標には正解率と F 値を用い，各条件において 3 回の試行の平均で評価した．

3.1 層数とパラメータ数の検討

畳み込み層と LeakyReLU 層の数が 3, 5, 7 層の 3 通り，それぞれについてフィルタ数が 3 通りの，合計 9 通りの条件で分類性能の違いを検証した．各条件に

*Short-time spectrum classification of bird song using deep learning. By Yuta NAKAYA, Koki YAMADA and Kohei YATABE (Tokyo University of Agriculture and Technology).

表–1　上から，畳み込み層が 3, 5, 7 層のときのパラメータ．フィルタサイズは (周波数方向, 時間方向) である．フィルタ数が少ない方から順に条件 A, B, C とした．

層	フィルタサイズ	フィルタ数/総パラメータ数			Dilation	Stride
		A	B	C		
1	(100, 1)	16	32	64	2	1
2	(50, 1)	32	64	128	1	2
3	(20, 1)	64	128	256	1	2
総パラメータ数		73728	280576	1093632	–	–
1	(50, 1)	16	32	64	2	1
2	(50, 1)	16	32	64	1	2
3	(30, 1)	32	64	128	1	2
4	(20, 1)	32	64	128	1	2
5	(10, 1)	64	128	256	1	2
総パラメータ数		70496	279232	1111424	–	–
1	(50, 1)	8	16	32	2	1
2	(50, 1)	8	16	32	1	2
3	(40, 1)	16	32	64	1	2
4	(40, 1)	16	32	64	1	1
5	(30, 1)	32	64	128	1	1
6	(20, 1)	32	64	128	1	1
7	(10, 1)	64	128	256	1	1
総パラメータ数		67216	264992	1052224	–	–

表–2　畳み込み層の数とフィルタ数を変えたときの結果．値は正解率/F 値であり，3 回の試行の平均である．正解率と F 値それぞれの最大値を太字にしている．

畳み込み層の数	条件 A	条件 B	条件 C
3 層	**0.851/0.711**	0.851/0.699	0.828/0.660
5 層	0.846/0.708	0.836/0.697	0.830/0.637
7 層	0.839/0.689	0.812/0.671	0.784/0.603

おけるパラメータの値を表–1 に示す．フィルタ数が少ない方から順に条件 A, B, C とした．各条件において，層数を変えたときに畳み込み層と全結合層のパラメータ数 (フィルタサイズ×チャネル数×フィルタ数) の合計が近くなるように調整した．

実験結果を表–2 に示す．表から分かる傾向として，畳み込み層の数に着目すると，層数が少ない方が性能が高かった．また，フィルタ数に着目すると，フィルタ数が少ない方が性能が高かった．これらから，単純なネットワークの方が過学習しづらく，テストデータに対して高い性能を発揮することが示唆される．

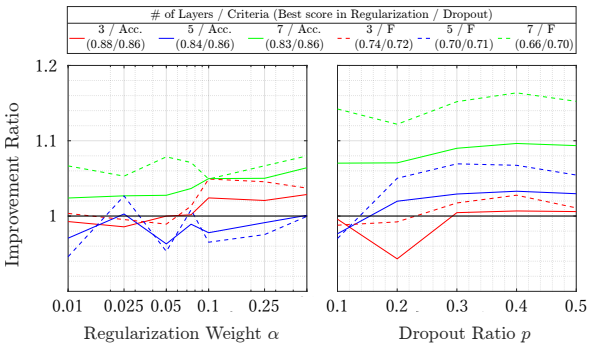
3.2　L1 正則化・ドロップアウトの適用

動物の鳴き声のクリーンなデータは十分にないことが多く，本実験でもキンカチョウの鳴き声のデータ数が限られている．そこで，過学習を防ぐ手法としてよく用いられる L1 正則化とドロップアウトを適用して，分類性能がどのように変化するかを検証した．

L1 正則化では，損失関数を

$$L_{\text{regularized}} = L + \alpha||w||_1 \qquad (1)$$

に変更する．ただし，L はもとの損失関数を表し，$||w||_1$ は畳み込み層の重みの L1 ノルムである．非ゼロ成分が少なくするように学習することで，実質的なパラメータ数を減らし，過学習を抑制する効果がある．また，ドロップアウトは，各 LeakyReLU 層の出力を確率 p でランダムに 0 にする処理である．



図–3　L1 正則化 (左) とドロップアウト (右) の適用による改善率．色は層数を表す．実線が正解率で破線が F 値である．1 より大きい値は分類性能の改善を表す．凡例の括弧内の値は各条件の正解率と F 値の最大値である．

本実験では，表–1 に示す 3 層，5 層，7 層のネットワークを用い，フィルタ数はそれぞれ条件 A, B, C とした．また，L1 正則化のパラメータ α の値を 0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5 とし，ドロップアウト率 p を 0.1 から 0.5 まで 0.1 刻みで変化させた．

L1 正則化やドロップアウトを適用しない場合との比較をするために，図–3 に各パラメータにおける改善率を示す．どちらも過学習を防ぐ手法であるため，過学習しやすいパラメータ数の多い条件から順番 (緑, 青, 赤) に改善率が高くなると予想される．左図より，L1 正則化では緑，赤，青の順に，右図より，ドロップアウトでは緑，青，赤の順に改善率が高い傾向にあることが分かる．パラメータ数の多い条件において改善率が高いことから，過学習が原因で性能が低下していたことが示唆される．なお，今回の実験において，フィルタ数を条件 A にした 3 層の CNN に L1 正則化を適用したケースで，最も高い正解率/F 値を示した．

4　むすび

鳥類の鳴き声のスペクトログラムの各時間フレームを，1 つの調波で表せる構造，1 つの調波では表せない構造，無音区間の 3 つに分類する CNN を作成した．また，L1 正則化とドロップアウトの適用により，性能が向上するが，パラメータ数が多い条件でより顕著になることを確かめた．今後はデータ拡張を行う．

参考文献

[1] 橘亮輔, “小鳥の音声が伝えるもの —さえずりと地鳴きの仕組みと機能—,” 日本音響学会誌, **79**(1), 28–33 (2022).

[2] R. Laje, D. Sciamarella, J. Zanella and G. B. Mindlin, “Bilateral source acoustic interaction in a syrinx model of an oscine bird,” *Phys. Rev. E*, **77**(1), 011912 (2008).

[3] 中谷優太, 松本和樹, 山田宏樹, 矢田部浩平, “鳴管の左右の音源を独立制御可能な鳥類の鳴き声の分析,” 音講論集, pp. 617–618 (2024.3).