International journal of application or innovation in engineering and management

ijaiem.com/Jan 2023/ Volume 12/Issue 1/Article No-1/01-07
ISSN: 2319-4847

Identifying Species using Audio with Convolutional Neural Networks

Saiphaneendra Rajasthan, Saiphaneendra0823@gmail.com

Article Info

Received: 17-10-2022 *Revised: 21 -11-2022* *Accepted: 06-12-2022* *Published:15/01/2023*

Abstract*:* In this study, we use the deep learning methodology of CONVOLUTIONAL Neural Network (CNN) to provide two straightforward methods for identifying a bird's species based on its sound. We provide a 1D CNN method that uses direct signals as the neural network's inputs. The second method involves extracting characteristics from the sound sources and feeding them into a 2D CNN. This method was tested on two sets of bird call data for 35 European and 39 Indian species**.**

Keywords: CNN, 1D-CNN, 2D-CNN, Bird species identification, MFCC

INTRODUCTION

A. *Overview*

Vocalization is a vital mode of communication for birds. Bird calls and songs have distinct spectral features[1][2][3]. Spectral Analysis of bird speech is gaining attention in recent times as a means of identifying the bird species from its call or song[4][5][6][7][8][9][10]. Properties of vocalization vary based on various physiological as well as external factors. While we ignore the latter in this study, differences in the former can be exploited for identifying a bird species from its call. When done with the help of machines, this can potentially become a valuable tool for ecological surveys, ornithological studies as well as hobbyist birdwatching.

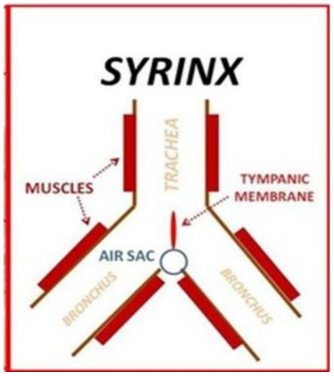

Fig. 1. Simplified model of a bird Syrinx[11]

B. *Biology*

Anatomy of bird speech differs from that of humans in some key areas. Birds don’t have vocal cords. Sound is instead produced via an organ called the Syrinx (shown in *Fig 1*) which is located near the junction of the trachea and the bronchi[12]. Its sound source is the tympani form membrane which faces to the bronchus on one side and the air sac on the other. When some of the syringeal muscles contract, the lumen of the bronchus is throttled and produces vibration in the membrane. When stretched along one dimension, the membrane produces harmonic sounds, but when stretched along two dimensions, the sound contains non-harmonic components. Syrinx can feature multiple simultaneous

METHODOLOGY

A. Overview

As shown in *Fig 2*, we start with preprocessing where we segment the sound files of each dataset (D_i) into a corresponding dataset of numerous smaller files containing a single syllable each (S_i). After that, 4 models were trained using S_i's. A 1D CNN and a 2D CNN each for Indian and European birds

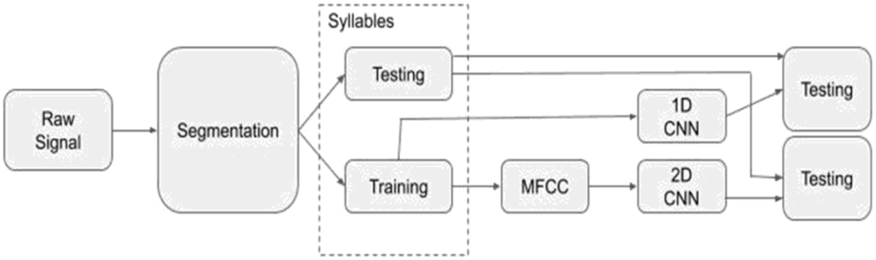

Fig. 2. Block diagram of the approach
(D1,D2,D3: Orignal sound file datasets
S1,S2,S3: Extracted corresponding
syllables $M1_{1D}$,$M1_{2D}$,$M2_{1D}$,$M2_{2D}$:
CNN Models
$MFCC_1$,$MFCC1_2$: Respective Mel Cepstrum Coefficients)

For 1D CNN, we take the raw signal directly as an input where as for 2D CNN, we extract cepstral features that are then used as input.

Data Acquisition

1) Dataset 1: On field recordings were collected for 41 different bird species totaling to 163 files. These were bought from a local ornithologist [14]. These are all Indian birds. Hereafter referred to as **D1**
2) Dataset 2: Training dataset of The ICML 2013 Bird Challenge, hosted on Kaggle[15], containing recordings for 35 species with 2801 individual data points, was procured. Based on the literature from the Kaggle website, we find that these are European birds [15]. Hereafter referred to as **D2**

For each dataset, 20% was removed at the start for testing. From the remaining 80%, 64% was used for training and 16% for validation

B. Segmentation

For auto segmentation, [17] was largely followed. Give below is a step by step procedure for the same

1) Audio files from each dataset were resampled to 22050 Hz. A spectrogram was obtained with Hanning window, column width of 512 and 75% overlap
2) This spectrogram normalized to the [0,1] interval and bottom 5 and top 30 frequency bins are removed as they predominantly contain noise
3) The resulting image was converted to a binary mask using Median Clipping. Here, each pixel is set to 1 if its value is greater than 3 times the median of its row and also its column. Else its set to 0
4) The resulting image is cleaned using standard morphological image processing techniques of closing, dilation and median filtering
5) Smaller regions (i.e. groups of connected pixels not exceeding a threshold) in the resulting are removed to get the final mask. 2 Approaches were tried here:

a) A fixed threshold was set for all files. While this worked well for files with high SNR, those with low SNR gave both false positives and false negatives
b) To avoid these cases, a dynamic threshold was selected to only choose a given number of regions with the highest

area

6) Manual inspection was done to remove any remaining false positives and false negatives. While (5.2) prevented all noise packets from being selected in all scenarios, some false positives were encountered in cased where the syllables we close enough to become a continuous area, making room for a noise pocket to be selected instead. (5.1) had instance of both

The outcome of each of these steps can be seen in *Fig 3*.

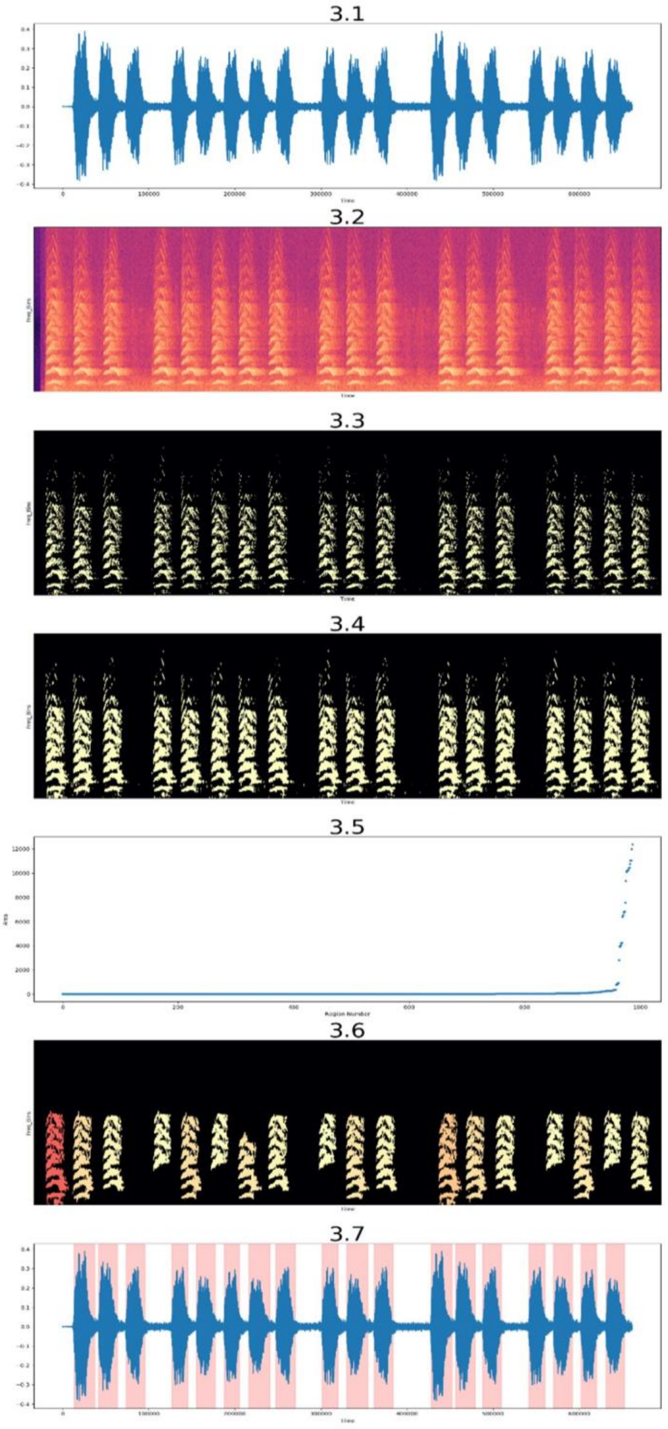

Fig 3. Segmentation Outcomes

3.1. Raw Signal (amplitude vs time)

3.2. Spectrogram (freq. bins vs time)
3.3. Median Clipped Mask (freq. bins vs time)
3.4. Post applying closing dilation & median filtering (freq. bins vs time)
3.5. Areas of regions (area vs region number)
3.6. Mask with syllables selected (freq. bins vs time)
3.7. Segmentation Output (amplitude vs time)

The above auto segmentation algorithm was able to correctly identify close to 80% of the syllables correctly as can be seen in Table 1 below.

TABLE I. AUTOSEGMENTATION ACCURACY

Dataset	No. of files	Syllables (Auto-segmentation)	Syllables (manual inspection)	Accuracy
D1	163	1317	1700	77.47%
D2	35	2338	2801	83.47%

C. *Data Augmentation*

Post segmentation, it was found that there was a disparity in the number of syllable samples per species in D2. To fix this, a minimum threshold of 50 was selected and data augmentation was done for all classes with less than the threshold number of samples. The following methods were employed for this:

1) *Adding Noise:* A random noise, 3-10% the RMS value of the signal was added to a randomly chosen sample. This helps in reducing generalization errors
2) *Pitch Shift:* Small amounts Pitch Shifts can help reduce classification errors [7] [18] [19]. 5% pitch shift was also applied to the above chosen sample

Samples for chosen for data augmentation till the total number of samples reached the threshold number

D. *Feature Extraction*

Cepstrum belongs to the class of homomorphic representations[18][20], that has been found to be useful for various recognition tasks. The Mel cepstrum[21], particularly, is popular for its robustness and simplicity. It also has the advantage of not requiring any parameter tuning to compute [22]. While Mel cepstrum is useful for analysis of human speech, it has been found to be effective for avian speech analysis as well [22][23][24]. Mel Frequency Cepstral Coefficients (MFCC's) are calculated from the logarithmic Mel cepstrum as follows:

E. *Modelling*

While RNNs/LSTMs might typically be used for classifying time series data, there is increasing application of CNN techniques in signal processing[25][26][27]. Here, we have tried two network architectures; a 1-D CNN which works directly with the raw audio signal and a 2-D CNN which takes the MFCC sequence of the audio signal as the input

The motivation behind using the raw audio signal is the simplicity of the pipeline - once trained, this method requires no pre- processing of the audio and can be used directly. The intuition behind using a 1-D, single channel CNN is that the classifier will try to learn the spatial patterns in the audio signal which are specific to a given bird species. The MFCC sequence captures the unique patterns in each audio signal, which are thus

F. *Network Architecture*

Both 1-D and 2-D CNN's consist of a sequence of three 'blocks' followed by a dense layer, as shown in *Fig 4* and *Fig 5*. Each block comprises of two convolution layers, followed by a max-pooling layer. Each layer uses ReLU non-linear activation, with the second convolution layer also using batch normalization for efficient training. For regularization, a dropout layer has been added to each block. A dropout probability of 0.2 was chosen for 1D CNN and 0.35 for 2D CNN. The final output layer uses a softmax activation with loss calculated using categorical cross-entropy

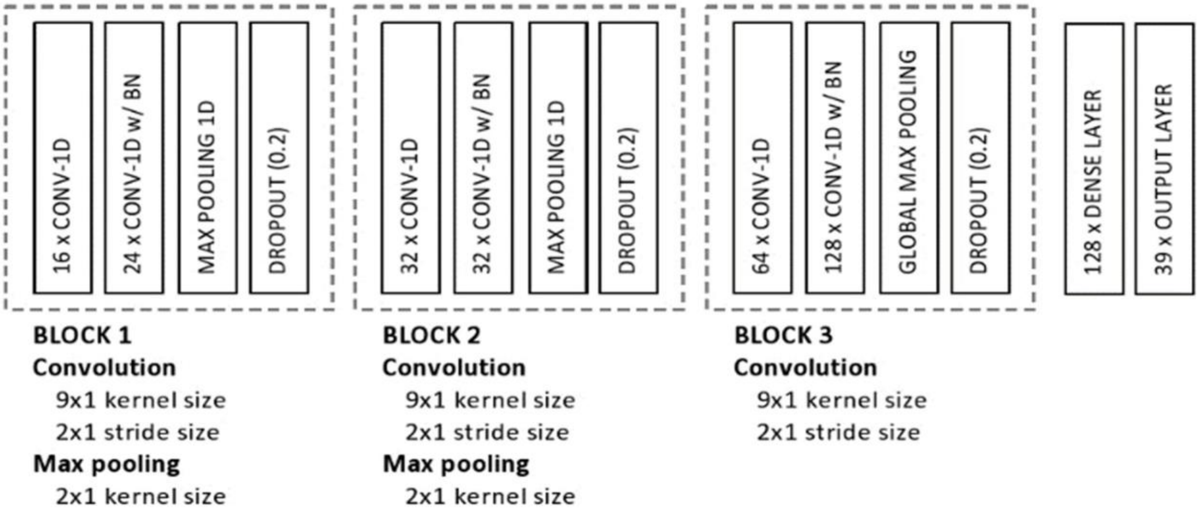

Fig 4.Block diagram for 1D CNN Network Architecture

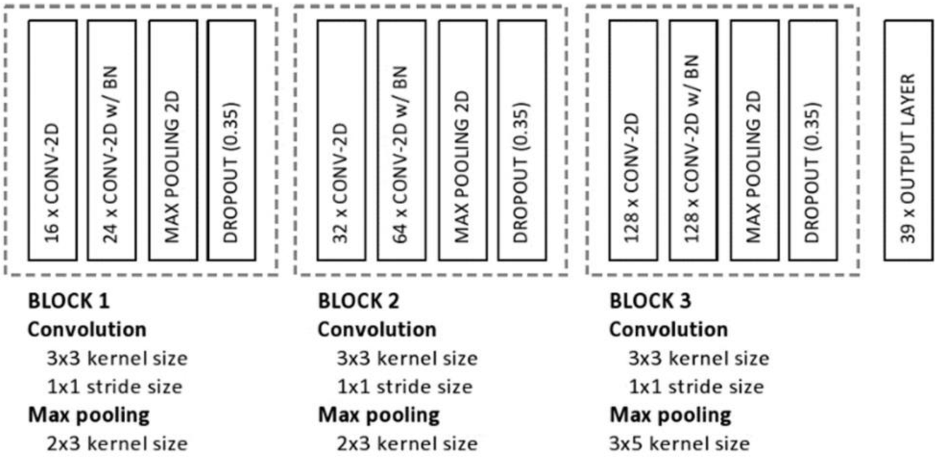

Fig 5. Block diagram for 2D CNN Network Architecture

G. *Training*

Both 1-D and 2-D classifiers were trained using a constant learning rate of 0.0001. Both models used a batch size of 32 and Adam optimizer for training. Weights in both networks were initialized with Xavier initialization, and the inputs were normalized for maximal efficiency in training. Both models were trained for 200 epochs

RESULTS

The accuracies for both models are listed in *Table II*.As seen from the results, the 1D CNN classifier is simple and would require little to no pre-processing of the raw audio. However, the 2D CNN based approach results in a better performance

TABLE II. Accuracy

Dataset	Efficiency	
	1D-CNN	*2D-CNN*
D1 (Validation Set)	73.78%	85.93%
D1 (Test Set)	67.35%	80.47%
D2 (Validation Set)	60.30%	82.54%
D2 (Test Set)	56.33%	74.78%

A. *Conclusion*

Using CNN results in a reasonably better accuracy as compared to linear classifiers or similar machine learning techniques. It is also evident that using feature extraction techniques like MFCC results in a more accurate classification when compared to using raw signal directly. The seeming disparity in the efficiency of the models generated by D1 and D2 can be explained by the fact that D1 was collected using professional equipment and has a very high SNR. Also D1 is a smaller dataset than D2

B. *Outlook*

This study was done with the broad idea that species identification can be automated. The biggest potential hurdle to building a fully automatic solution is the auto-segmentation step and ways to improve or bypass it in the future would be very valuable towards this goal. We would also like to reiterate that song birds as well as specific calls – like seasonal or mating calls we ignored in this study. A real world solution should probably handle those. Also, this study assumes only a single bird species in the foreground. However, it is quite reasonable to assume that there might be multiple birds in the background as well.

REFERENCES

[1] Harmonic structure classification in bird vocalization by A. Härmä and P. Somervuo, "IEEE International Conference on Acoustics, Speech, and Signal Processing"
IEEE Explore (https://ieee.org/abstract/document/1327207, 2004)
Classification of the harmonic structure in bird vocalization by P. Somervuo, A. Harma, and S. Fagerlund. IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, Issue 6, November 2006. Examine this link: https://ieeexplore.ieee.org/abstract/document/1709912
[3]"Automatic identification of bird calls using Spectral Ensemble Average Voice Prints," 14th European Signal Processing Conference, 2006, Hemant Tyagi, Rajesh M. Hegde, Hema A. Murthy, and Anil Prabhakar, https://ieeexplore.ieee.org/document/7071098
[4] "Audio-only bird classification using unsupervised feature learning," by Dan Stowell and Mark D.

Plumbley, CLEF 2014, http://ceur-ws.org/Vol-1180/CLEF2014wn-Life-StowellEt2014.pdf
[5] "Detection of bird acoustic activity using morphological filtering of the spectrogram" by Allan Gonçalves de Oliveira et al. (2015) in Applied Acoustics
[6] Juang, C., and Chen, T., "Protection-based recurrent neural fuzzy networks for bird song recognition," Neuro computing, vol. 71, pp. 121–120, 2007.
[7]Thomas Hofmann, Yannic Kilcher, Elias Sprengel, and Martin Jaggi, "Audio Based Bird Species Identification using Deep Learning Techniques"
[8]Briggs et al. (2012), "Classification of multiple bird species," Journal of Acoustic Society of America, vol. 131, pp. 4640-4650, Lakshminarayanan, B., Neal, L., Fern, X. Z., Raich, R., Hadley, S., et al.
[9]Fagerlund, S.: Support vector machines for bird species identification. In: Journal of Signal Processing Advancements (2007), 7, 64–71
[10]Terrence J. Sejnowski and Kenji Doya: A New Reinforcement Model for Learning the Vocalization of Birdsongs. In: Neural Information Processing Systems Advances 7 (NIPS 1994) [11]Birds.html https://echolocation-physiology-ansc3301.weebly.com
[12] D. S. Vicario (1991). Songbird vocal production neural processes. Neurobiology's Current Opinion, 1:595–600
[13]Aki Härmä. Bird species are automatically identified using sinusoidal syllable modeling. Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. V-545. IEEE
[14]In [15], https://www.birdcalls.info/The icml 2013 Bird Challenge: https://www.kaggle.com/c/the-icml;
[16]F. Deroussen (2001). Garden Oiseaux de France. French company Nashvert Production, Charenton; Deroussen, F., and Jiguet, F. (2006). The Museum's sonotheque: French Passengers and Islands. http://naturophonia.fr; Nashvert manufacturing, Charenton, France.
[17] Mario Lasseck (2013): Classifying Bird Songs in Field Recordings: The NIPS4B 2013 Competition Winning Solution
[18]In Proc. Symp. Time Series Analysis, Istanbul, Turkey, June 5–9, pp. 209–243, B. P. Bogert, M. J. R. Healy, and J. W. Tukey (1963): The frequency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking
[19]Schldeluter, J., and Grill, T.: Investigating data augmentation for enhanced neural network singing voice identification
[20]Schafer, R. and Oppenheim, A. (1975): Digital Signal Processing. Springer, Prentice-Hall, New York
[21]Parametric representations for monosyllabic word identification in continuously uttered phrases are compared by S. Davis and P. Mermelstein (1980). Speech, Signal Process, IEEE Trans. Acoust., vol. ASSP-28, no. 4, pp. 357–366
[22]Seppo Fagerlund, Aki Härmä, and Panu Somervuo (2006): Automatic Species Recognition Using Parametric Representations of Bird Sounds. IEEE Transactions on Speech, Language, and Audio Processing, Volume 14, Issue 6,
[23]Lubos Juranek, Michal Munk, and Jiri Stastny (2018): Automatic identification of bird species using vocalizations 2018:19 [24] EURASIP Journal on Audio, Speech, and Music ProcessingHervé Glotin and Julien Ricard (2016): A bag of MFCC-based terms for identifying birds. LifeCLEF [25]An Analysis of 1-D and 2-D Deep Convolutional Neural Networks in ECG Classification by Yunan Wu, Feng Yang, Ying Liu,