Deep Learning Methods for Acoustic Monitoring of Birds Migrating at Night

Hanna Pamula, Agnieszka Pocha, Maciej Klaczynski

DEEP LEARNING METHODS FOR ACOUSTIC MONITORING OF BIRDS MIGRATING AT NIGHT

Hanna Pamula[1] Agnieszka Pocha[2] Maciej Klaczynski[1]
[1] AGH University of Science and Technology, Krakow, Poland
[2] Jagiellonian University, Krakow, Poland
pamulah@agh.edu.pl

ABSTRACT

The project aims to supplement nocturnal bird migration research with long-term audio recordings. We checked how different deep learning methods – namely convolutional neural networks (CNN) and residual neural networks (ResNets) – perform in the nocturnal flight calls detection task. Moreover, we used transfer learning from image classification models to determine if we can improve automatic detection performance. The best model obtained the AUC score over 95% (area under the receiver operating characteristic curve). Although higher bird detection accuracy is still needed because of the strong dataset imbalance, the results are very promising and suggest that deep learning models applied to audio data have great potential in supplementing bird migration studies.

1. INTRODUCTION

Billions of birds migrate between breeding and wintering sites every year. Traditionally, migration studies are concentrated on day migration. This is primarily determined by the fact that the most popular methods – such as visual observations and bird ringing – are most appropriate for daytime research. However, as most bird species migrate at night, much information about migration is lost, so novel techniques are used to fill the knowledge gaps [1]. One of these methods is acoustic monitoring.

As autonomous recording devices are getting cheaper and more accessible, the bioacoustics monitoring projects become the “big data” research area, generating terabytes of audio data [2]. However, manual analysis of long-term recordings is a very time-consuming and inefficient process requiring expert knowledge. For example, the "scanning" of the spectrograms in search for impulsive nocturnal birds’ calls can take up to 2-3 times more time than the duration of the recording. For such big datasets, manual annotation is not feasible, so effective automatic detection of birds’ calls is crucial.

While there is plenty of research focused on bird songs, the nocturnal avian flight calls projects are not yet common, regardless of the recent interest in this topic [3,4]. In this article, we present the results of research on flight calls detection, based on the well-established deep learning methods – CNNs and ResNets.

2. MATERIALS AND METHODS

2.1 Field Recordings

In the 2016 - 2019 seasons, field recordings at the site of increased migration of birds were performed. Long-term recordings of the night calls of bird migrants were conducted every season for 50-60 days, from the first half of September to the beginning of November. The recording sets were deployed on a narrow spit between Bukowo Lake and the Baltic Sea in Dabkowice near Darlowo (54°20'16 "N, 16°14'38" E). 4-6 microphones were set up in different places: on a dune, in the reeds, and in a more quiet place – a clearing near the lake. The tests were performed with SM2+ recorders connected to directional SMX-NFC microphones, placed on a flat plastic plate (Wildlife Acoustics Inc., USA), and mounted on poles 3-5 m high. The recordings were performed every night, from sunset to sunrise. In total, over 3,000 hours of recordings were collected in each season. In addition, a weather station was deployed for the two last seasons.

2.2 Training and Testing Set

At first, two automatic voice detection programs were checked; unfortunately, they did not extract the voices from the recordings accurately [5]. In search of effective bird calls detection methods, we created training and testing sets for deep learning models. We manually annotated over 50 hours of recordings from different nights and various weather conditions, for which we used the Audacity software[1]. The annotations were performed only on a time scale, not on the frequency axis. Only passerines' calls were annotated, as we assumed that other bird groups might contain calls of local or resting individuals (owls, ducks, geese, etc.).

As avian flight calls are sporadic and short (10-300 ms), a balanced training set was created. It consisted of clips derived from 94 30-min long recordings. The clips were 500 ms long, overlapping by 150 ms. The training set comprised all positive clips and the same number of randomly picked negative clips. If no call was present in the recording, a small number of negative frames was added to a subset for background noise representation. As a result, the training set consisted of more than 15.9k clips with a 45% bird call presence.

[1] Audacity®. Version 2.0.5. http://audacity.sourceforge.net.

Representation type **R**	Architecture parameters					
	Batch size *[32,64]*	Dropout **D** *[0.1, 0.35,0.5]*	No of blocks **N** *[3,4]*	Filters *[10,20]*	Dense layer size **DLS** *[128,256]*	Epochs *[20,50, 100]*
Spectrogram	32	0.5	4	20	256	20
Mel-spectrogram	32	0.5	3	10	256	50
Multitaper	64	0.35	4	10	128	100

Table 1. Chosen hyperparameters for CNN architecture

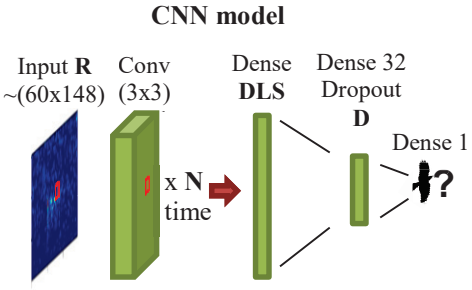

Figure 1. Architecture of CNN models

There were 20 annotated full recordings (30-min long) in the test set, prepared as in the training set. As the test set comprised full continuous recordings, it was strongly unbalanced, resulting in over 103k clips with only 1.2% bird call presence.

2.3 Deep Learning Detection Methods

Three different approaches were tested – convolutional neural networks (CNN), residual neural networks (ResNet-18 and ResNet-50 architecture [4]), and transfer learning from pre-trained ResNet models.

2.3.1 Convolutional Neural Networks

To fully use the potential of CNNs, we used image-like representations of the sound: spectrogram, mel-spectrogram, and multitaper[2] (with two tapers). We have chosen representation parameters in such a way, that each of them had a similar size, about 60x148 px. Each time 4.5-9kHz range was covered. The hyperparameters of the CNNs were optimized by grid-search over a parameter grid with 2-3 values for each parameter. ReLU activation was used for all but last layer, where we used sigmoid. Table 1 presents the chosen CNN hyperparameters, while Figure 1 illustrates the schematic network architecture.

2.3.2 Residual Neural Networks

To adapt ResNets for bird call detection, we changed the number of classes from 1000 to 2 (bird – no bird) and the input size to 1 channel (mel-spectrogram), as opposed to 3-channel in the original model (colour images).

2.3.3 Transfer Learning

ResNet-18 and ResNet-50 were used, initialised with random and with weights pre-trained on ImageNet [7]. Models were trained five times, and their predictions on the test set were averaged. For ResNet-18 the additional experiments were performed, with freezing the weights of the first 1 or 2 layers (all ResNet architectures are composed of four layers, sometimes called stages. Every layer is composed of several blocks). As an input, a mel-spectrogram was used in each channel, or – for chosen experiments – different signal representations were put in each channel (spectrogram, mel-spectrogram, multitaper).

[2] https://pypi.org/project/libtfr/

3. RESULTS AND DISCUSSION

For each experiment, the receiver operating characteristic curve (ROC) and the area under that curve (AUC) was calculated.

3.1.1 CNNs

Figure 2 shows the results for the CNN models with different input representations. Although the training process for spectrogram and mel-spectrogram architectures gave varying outcomes – some models were not trained at all, resulting in AUC oscillating around 50% – creating an ensemble of five independent runs led up to higher AUC score than for any of the individual models.

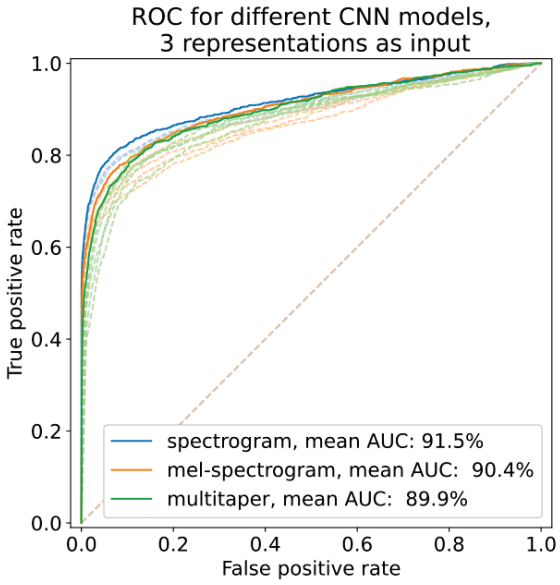

Figure 2. ROC curves for three CNN architectures. The dashed lines represent the result of individual five runs of the same algorithm. Each solid curve is a voting ensemble of five models.

3.1.2 ResNets

For ResNet models, the best AUC score was obtained by ResNet-50 with different signal representations in each channel as input, initialised with pre-trained weights, and by ResNet-18 without pre-training (Figure 3). However, the differences between the models were negligible; only the ResNet-50 without using the weights pre-trained on ImageNet got lower result. Thus, using the pre-training weights seems beneficial for ResNet-50, but not for ResNet-18.

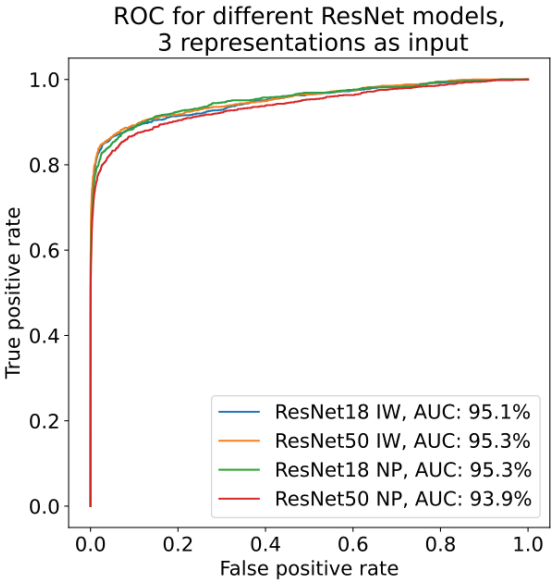

Figure 3. ROC curves for ResNet-18 and ResNet-50 models. IW- model initialised with pre-trained weights, NP – no pre-training. 3 channel input: spectrogram, mel-spectrogram, multitaper. Each curve is a voting ensemble of five models. For the sake of clarity, the individual five runs were not shown.

3.1.3 Comparing representations

The results obtained for mel-spectrogram representation in each channel were very similar to 3 representations (AUC = 95.1% for ResNet-18 IW, 95.0% for ResNet-50 IW, 95.2% for ResNet-18 NP, 93.8% for ResNet-50; compare with Figure 3). Because of these negligible differences, we do not present the separate chart. Hence, the three representations instead of one, did not significantly improve the performance, and the neural network did not profit from the additional data representations.

3.1.4 Transfer learning

In the last experiment, we checked how freezing some of the weights of the pre-trained ResNet-18 model influenced the performance. The obtained results clearly indicate that freezing the layers was degrading for the performance in all three variants, resulting in the AUC measure in the 75.9-93.3% range (Figure 4).

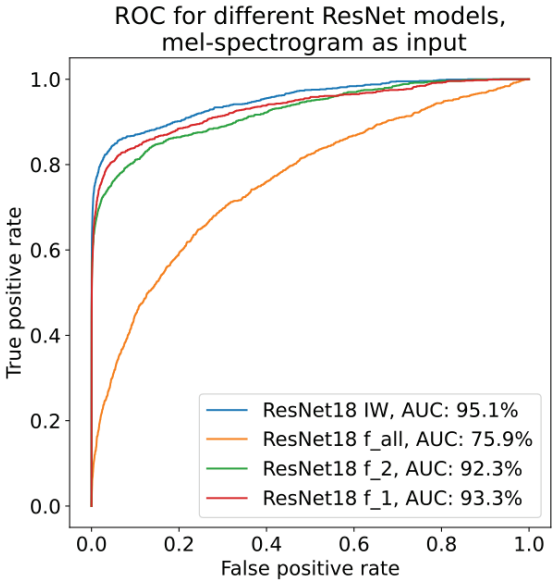

Figure 4. ROC curves for ResNet-18 models. IW- model initialised with pre-trained weights, f_all – freezing all layers apart from the last one, f_2 – freezing two layers, f_1 freezing the first layer only. Input: mel-spectrogram. Each curve is a voting ensemble of five models. For the sake of clarity, the individual five runs were not shown.

3.1.5 Comparison of the best performing models

Finally, two best models of each neural network type were compared. To put the obtained AUC measures in a wider context, the data reduction on the test set was calculated (Table 2). The data reduction was defined as reducing the time of the recordings, retaining high percent of the avian flight calls. Two recall values were chosen: retaining 80% and 90% of calls present in the test set.

Architecture	FP	TP	Precision	Data reduction
Retaining 80% of calls (*recall* 80%)				
CNN spectro	6911	984	12.5%	~1h 6min/14h
ResNet18 NP	1873	984	34.4%	~24 min/14h
Retaining 90% of calls (*recall* 90%)				
CNN spectro	33635	1107	3.3%	~4h 50min/14h
ResNet18 NP	12580	1107	8.1%	~2h 54min/14h

Table 2. The data reduction table for the best performing CNN and ResNet models. Results for retaining 80% and 90% of the calls from the test set. FP – false positive, TP – true positive.

Depending on the goal and application of the bird call detection, stricter conditions may be selected. The test set can then be reduced from 4 hours to less than 3 hours while retaining 90% of calls. However, if we choose the recall of 80%, the set may be condensed to only 24 minutes, and around one out of every three samples will contain the bird call while the initial rate was 1:82. The ResNet model with such settings could greatly reduce the effort of bird detection and annotation in long-term audio recordings.

4. SUMMARY

In this work we have shown that using deep learning methods – especially ResNet architecture – is beneficial for the presented research topic: detecting nocturnal avian flight calls. However, we could not find evidence that transfer learning from the models pre-trained on ImageNet has positive effects on the results. The outcomes are promising, and they suggest that deep learning methods applied to audio data have great potential to support nocturnal bird migration research.

5. REFERENCES

[1] KG. Horton, WG. Shriver, JJ. Buler: "A comparison of traffic estimates of nocturnal flying animals using radar, thermal imaging, and acoustic recording", *Ecological Applications,* Vol. 25, No. 2, pp. 390–401, 2015

[2] D. Stowell, M. Wood, Y. Stylianou, H. Glotin: "Bird detection in audio: a survey and a challenge", *Proc. of 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2016

[3] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, JP. Bello: "Birdvox-full-night: A dataset and benchmark for avian flight call detection", *Proc. of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 266–270, 2018

[4] BM. Winger, BC. Weeks, A. Farnsworth, AW. Jones, M. Hennen, DE Willard: "Nocturnal flight-calling behaviour predicts vulnerability to artificial light in migratory birds." *Proc. of the Royal Society B*, Vol. 286, No. 1900, pp 20190364, 2019

[5] H. Pamula, M. Klaczynski, W. Wszolek, M. Remisiewicz: "Monitoring akustyczny ptaków migrujących nocą – zagadnienia związane z automatyczną detekcją głosów", *Ornis Polonica*, Vol. 58, No. 3, pp. 187–196, 2017

[6] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016

[7] J Deng, W Dong, R Socher, LJ Li, K Li, Fei-Fei L: "Imagenet: A large-scale hierarchical image database." *Proc. of the 2009 IEEE conference on computer vision and pattern recognition*, pp. 248-255, 2009