Key words:
Acoustic monitoring, Birds, Night flight calls, Deep Learning, Convolutional Neural Networks

Hanna PAMUŁA[(1)], Maciej KŁACZYŃSKI[(1)], Magdalena REMISIEWICZ[(2)], Wiesław WSZOŁEK[(1)], Dan STOWELL[(3)]

[(1)] AGH University of Science and Technology, Faculty of Mechanical Engineering and Robotics, Department of Mechanics and Vibroacoustics, Al. Adama Mickiewicza 30, 30-059 Kraków
[(2)] University of Gdańsk, Faculty of Biology,Bird Migration Research Station, ul. Wita Stwosza 59, 80-308 Gdańsk
[(3)] Queen Mary University of London, Centre for Digital Music, Machine Listening Lab, Mile End Rd, London E1 4NS

ADAPTATION OF DEEP LEARNING METHODS TO NOCTURNAL BIRD AUDIO MONITORING

Trends in the size of bird populations are good indicators of the general state of the environment. Different bird monitoring methods have been developed over the years, but surveys based on observations and bird-ringing programmes are the most common techniques. However, counting nocturnal migrant birds remains a particularly difficult task so other techniques must be applied. One possible method is automatic acoustic monitoring, which can supplement standard bird-monitoring schemes.

We investigated the automatic detection of migrant birds' night flight calls. Long-term recordings were collected during birds' autumn migration along the Baltic Sea coast. A deep learning method, convolutional neural network working on spectrograms, was adapted to detect the migrants' calls. The first results are promising (AUC=96.6%), showing the potential of acoustic methods to supplement standard bird-monitoring techniques and to suggest a directions for research.

1. INTRODUCTION

Monitoring is a crucial tool for bird conservation, and is also useful to observe changes in biodiversity and environment health because birds are good biological indicators [2,4]. The most popular current monitoring methods are direct observation and bird ringing [8]. These techniques can be used only in specific circumstances, so

many attempts have been made to develop and extend traditional methods using new technologies. Monitoring birds migrating at night is a particular case in which standard approaches have limited utility, so methods such as radar detection, thermal imaging and acoustic recordings are used [6].

Long-term bioacoustic monitoring projects using passive recording stations has many advantages, such as greater standardisation in data collection, reduced subjectivity and observer bias compared with human experts, not to mention less disturbance of the surveyed birds [9]. However, long-duration sound recording is an asset and a drawback – the large amount of audio data is produced, often measured in terabytes, which is impractical to process manually. Without reliable automated detection and classification systems the method is unlikely to be widely accepted and put into widespread use, so many efforts are being made to improve audio monitoring and detection techniques.

Three main approaches to the detection of bird sounds are: methods based on simple thresholding energy levels in short frames, template matching using spectrogram representation and the application of hidden Markov models (HMMs) [10]. Convolutional neural networks (CNN) have recently been studied to improve recognition and detection in audio data [7]. This is also true for bird detection in audio: almost all contestants in the recent *Bird Audio Detection Challenge* [10] used CNNs with promising results. However literature on audio detection techniques has focused mainly on bird songs, signals which are typically longer in duration with higher signal-to-noise ratio (SNR) than flight calls, which are typically softer and less known. Thus we wanted to verify if widely-used methods to detect bird song would perform as well with sporadic, impulsive, quiet, high-pitched flight calls recorded in the harsh environment of the Baltic Sea coast during autumn migration.

We recorded almost two months of calls over the autumn migration season with a set of five microphones. Part of the data we collected was manually annotated to check the performance of a classifier. We chose the leading CNN algorithm from the *BAD Challenge* and we applied it to our dataset. Finally we explored modifications of splitting dataset to find the best use of the method in detecting nocturnal avian migrants.

2. MATERIALS AND METHODS

2.1. DATA ACQUISITION

One of the authors (HP) recorded our data during the autumn migration from September to November 2016 on the Baltic Sea coast, then reviewed and manually annotated the bird calls. Five recording stations were deployed on the 150 metre-wide

spit between Lake Bukowo and the Baltic Sea in Dąbkowice, West Pomeranian Voivodeship, Poland (54°20'16"N 16°14'38"E). The recording area was selected near the long-standing Operation Baltic bird migration research station at Bukowo, on an important migration route. We chose our recording sites to minimise confounding anthropogenic noise and artificial light, which can induce unnatural behaviour in migrant birds [11].

We used Song Meter SM2s with weather-resistant, directional Night Flight microphones from Wildlife Acoustics Inc. The Night Flight microphones are manufactured on a 30 cm x 30 cm flat perspex baffle to attenuate sounds below the plate. The microphones were mounted on 3–5 m poles to reduce the effect of noise from vegetation like rustling leaves and insect calls from the ground. Recordings were collected every night, from sunset to sunrise, at a 44100 Hz sampling frequency with 16-bit depth in 30-minute WAV format files.

2.2. DATA STRUCTURE

We collected >3200 h of recordings, more than a human expert can annotate manually in reasonable time. Therefore we selected 22 half-hour recordings from 15 nights as a representative subset of the data we collected. Nights with different weather conditions and background noise were selected, including strong wind, rain, insect calls and sea noise (Fig. 1). The recordings came from three microphones to get a representation of soundscapes in different locations (dune, dune with vegetation, forest glade).

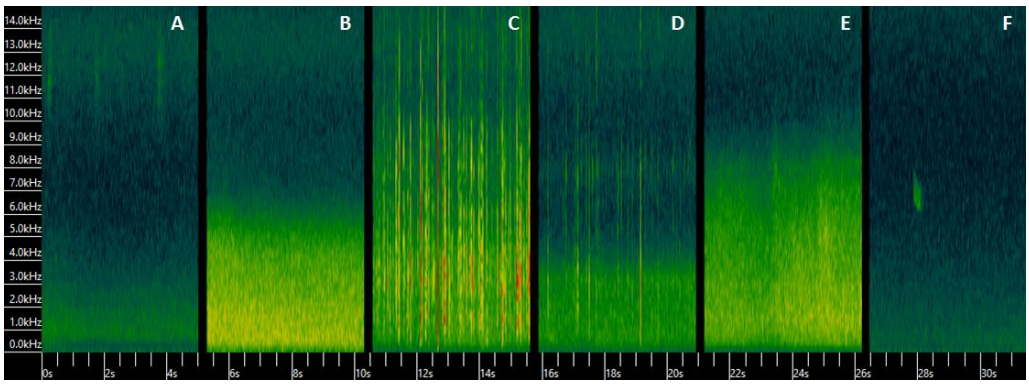

Fig. 1. 5-s random extracts from testing set recordings, showing different weather and background conditions. A) Calling insects; B) and E) different levels of sea and wind noise; C) strong and D) light rain; F) low background noise with easy-to-detect bird call (blackbird *Turdus merula*)

HP selected 11 hours of recordings, manually inspected the data and annotated precisely all the passerine calls. Audacity(R) software version 2.0.5[1] to display recordings and to annotate the presence or absence of bird calls. Spectrograms were limited to 0–14 kHz, with an FFT size of 512 in a Hamming window. The gain and range of colour bars were adjusted to improve the discrimination and detection rate in every 30-min recording. About 3–4 s of recording was displayed on the screen each time because the target calls were as short as 10 ms. We needed this magnification to ensure that we overlooked no calls because of inadequate resolution.

After manual annotation the recordings were split into two non-overlapping sets: one for training the classifier and one for testing its performance. The training dataset consisted of excerpts from 16 30-min recordings collected over 11 nights, the testing dataset had six recordings from four nights. We tested three variants of the training dataset to determine how different permutations might affect the classifier: the first version had an equal balance of positive and negative cases, with and without bird calls in the training set, the other two did not. The imbalance reflected the rate of occurrence recorded in the training data. The extracts were 10 s long in the second variant and 1 s in the third so that we could explore how this resolution affected performance. Migrant birds use night flight calls irregularly to maintain contact with their flocks [3]; we usually recorded less than 10 s of these calls in any hour. To obtain a roughly balanced training set for variant 1, 10 s excerpts were manually picked from 16 annotated baseline recordings, with an additional two recordings to provide more positive samples (other hour from 3/11). The algorithm we used to split the recordings to fixed durations added zeroes to the last extract so each set of recordings consisted of files of equal duration. Only if an onset of a manually annotated call appeared within excerpt, the algorithm labelled it as a positive. The training and testing data variants are shown in Table 1.

Tab. 1. Variants of annotated data files split into different durations. The difference in N^o of examples split to 10 s and 1 s arose because the recording devices took about 1 s to save each file

Variant	**Set**	**Description**	**N^o of examples**	**% bird presence**
Variant 1 Balanced training	Training	10 s excerpts, chosen manually from 16 recordings (+2 extra excerpts from the same date) to balance the dataset	818	49.88%
	Testing	6 recordings split into 10 s excerpts	1 080	17.32%

[1] Audacity(R) software is copyright (c) 1999-2014 Audacity Team. http://audacity.sourceforge.net/. Thename Audacity(R) is a registered trademark of Dominic Mazzoni.

Variant 2 10-s training-testing rec	Training	16 recordings split into 10 s excerpts	2 880	12.36%
	Testing	6 recordings split into 10 s excerpts	1 080	17.32%
Variant 3 1-s training-testing rec	Training	16 recordings split into 1 s excerpts	28 784	1.64%
	Testing	6 recordings split into 1 s excerpts	10 793	3.19%
Training set nights: 13/09, 17/09, 18/09, 25/09, 02/10, 4/10, 5/10, 10/10, 18/10, 24/10, 3/11				
Testing set nights: 21/09, 7/10, 15/10, 31/10				

We expected that the third variant of the dataset would be better suited to the data in terms of detected signal length (10–300 ms), but the low proportion of positively labelled recordings (1.64%) in training set might cause problems. Thus the other two dataset were split to create better balanced training datasets.

2.3. CONVOLUTIONAL NEURAL NETWORK MODEL

We used the winning solution from the *Bird Audio Challenge* [10], a competition to construct a classifier that could predict if a bird sound was present in each 10 s recording of a provided dataset. The most successful approach was proposed by Thomas Grill [5], obtaining an Area Under the Receiver Operating Characteristic Curve (AUC) of 89%. AUC is a standard measure of a classifier's performance used in machine learning that accounts for scores' rank order [1]. An AUC equal to 100% means that the classifier ranked all positives examples (bird calls) before negatives (no calls), so it is an ideal classifier; an AUC of 50% classified samples randomly.

The approach was based on convolutional neural networks (CNN) working on mel-scaled log-magnitude spectrograms. The spectrograms (FFT size=window size=2048 samples, 44100 Hz sample rate, hop size=630 samples) were calculated, then mel-scaled and filtered with 80 triangular filters ranging from 50–11000 Hz. The magnitudes were then scaled logarithmically. The features were normalised and the mean over time was subtracted from the obtained modified spectrograms.

The neural network architecture and training was used as described in [5], but we summarise it here briefly. The neural network consisted of alternating four convolution and four pooling layers, followed by three dense layers [5]. The first two pairs of layers used a 3x3 filter, the next two a 3x1 filter. As an input, the modified spectrogram was used at a size of 1x1000x80, condensed by the neural network into 16 11x8 feature maps. Then three dense layers were applied with 256, 32 and 1 unit.

All of the layers were followed by a leaky rectifier, except for the last sigmoid output layer.

Training was done by stochastic gradient descent on mini-batches with a size of 64. The size of the training data set was augmented by pitch shifting. A second stage of training was then performed using pseudo-labelled data, where some of the computed predictions were added to the training data and the model was trained once again (most confidently predicted recordings, threshold of 0.1). Final predictions were averaged over five networks trained on five cross-validation subsets from the training set.

3. RESULTS

The AUC was calculated to check the performance of each model on our night flight call dataset. The first- and second-stage prediction results are presented in Figure 2.

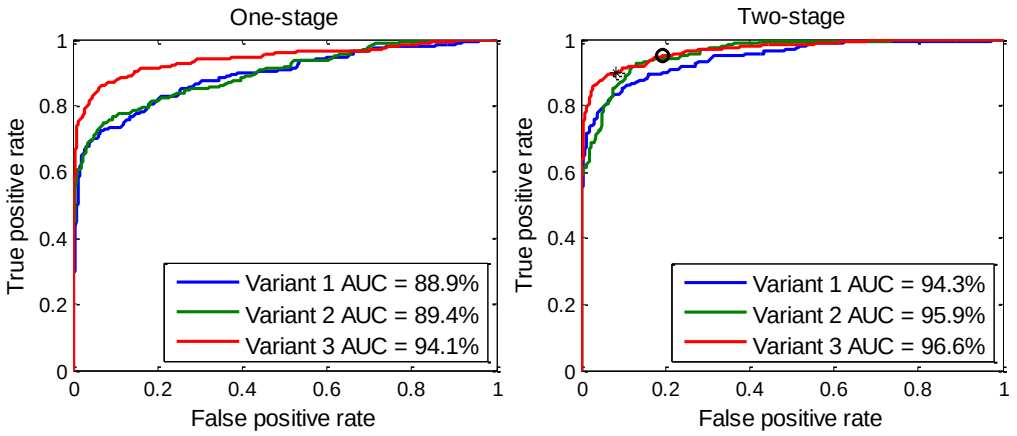

Fig. 2. Area under ROC curves for three split variants of the dataset. Results for standard one-stage training (left) and two-stage training with pseudo-labelled data (right). Detections of 90% and 95% of all calls annotated in the training set for variant 3 are marked with a black asterisk and a circle, respectively

The results show that balancing the positive and negative extracts did not increase the performance of the model, resulting in 88.9% for the balanced training set (variant 1) and 89.4% for a non-balanced set (variant 2) for one-stage training. The third variant, with the shortest recordings, improved the results significantly to 94.1%, despite the low rate of positive data in the training set (1.64%). Second-stage training using the most confidently predicted test examples improved the results, reaching a high AUC value of 96.6% in variant 3. Scores for recordings obtaining high

probability remained high, but more false positives obtained lower score and true positives higher. Small variations can be seen in Figure 3, e.g. for recordings N^{o} 24 and 38.

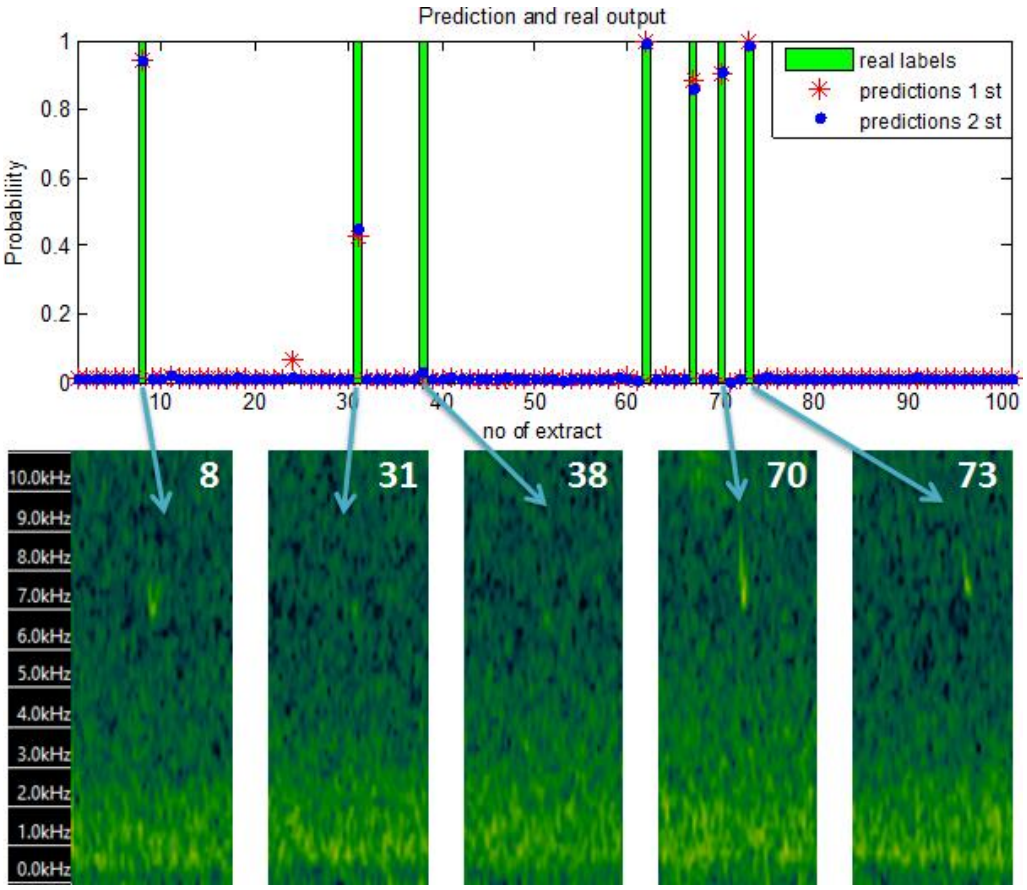

Fig. 3. Example of 1- and 2-stage prediction for variant 3 test dataset (upper part). The call that was not found in the model prediction is extremely faint (lower part of a plot, extract N^{o} 38)

5. DISCUSSION AND CONCLUSIONS

Our results demonstrated that a convolutional neural network provided acceptable performance in detecting night flight calls. Randomly checked false negatives – extracts with a low probability that should get high scores – showed that

misprediction was usually associated with faint and indistinct calls, which probably could not be classified to species by any method (Fig. 3).

The attained high AUC score seems a really great result, but we need to remember that our testing set is not balanced. If we set a threshold to detect 90% of all calls in our testing set from variant 3 (Fig. 2 right plot, asterisk), we would get 8.4% of negative samples marked as positives. That might appear to be a good result, but negative samples were overrepresented in our dataset, where more than 96% of recordings had no bird calls, so the 8.4% would produce 881 false positive samples. The 90% of positively marked calls would reflect 309 true positives. Applied method can thus reduce three hours of test recordings to less than 20 minutes of data (1190 1 s recordings), where more than a quarter of the 1 s recordings contained calls. Thus, the dataset can be reduced by 89% while preserving 90% of the calls. Manual annotation of our dataset takes approximately 1 to 3 times longer than the duration of the recordings, depending on the background noise and the expertise of the annotator. We estimated that annotating the test dataset would take about six hours. Using the proposed classifier could significantly reduce that time, so validation should not take more than double the duration of the detectors' output. In our case that would not take more than 40 minutes, probably less, because we would need only true or false discrimination. Annotating our basic training set took more time because calls had to be marked precisely to be split into different length extracts.

However, choosing a stricter requirement, such as detecting 95% of calls (Fig. 2 right plot, circle), would yield ~40 min of recordings with 17 more calls detected, but would double the number of false positives. So the threshold should be carefully chosen, depending on the aim of the study and the acceptable precision of detection.

Modifications of Grill's proposed network should also be considered, starting with a change of the 2048 FFT size. This FFT length produced the high frequency resolution we needed for mel representation with many bins. However, the temporal resolution might be inadequate. We obtained a window size of ~46 ms, but many of our calls of interest were as short as 10 ms. Choosing an FFT size of 512 would produce a temporal resolution of less than 12 ms, which seemed a more practical choice, though it would limit the representation of the mel spectrogram. A linear spectrogram rather than a mel could be used for smaller FFTs, but the current version of the classifier might not use this equally successfully.

We showed that the winning solution from the Bird Audio Detection Challenge based on convolutional neural networks gave good results for recordings from long-duration audio monitoring of nocturnal bird migration, very different datasets from those it was designed for. Recording soft, short contact calls is a particularly challenging scenario, due to noise and to the nature and rarity of the signals to be detected. The 96.6% area under the ROC curve we measured means we can be confident of good call detection with not so many false positives if the threshold is chosen wisely. The proposed method provides a good first pre-processing step for

classification, which is the next stage in developing an automated bird call recognition method. We plan to experiment further by modifying the input spectrogram representation and the FFT resolution, but the result we obtained is already exceptional and probably more effort should be put into effective classification of detected voices.

6. ACKNOWLEDGEMENTS

DS is funded by EPSRC Fellowship EP/L020505/1. HP research is supported by AGH *Dean's Grant* 2017 N°15.11.130.642 and by Erasmus+ grant.

REFERENCES

[1] BRADLEY P.A., *The use of the area under the ROC curve in the evaluation of machine learning algorithms*, Pattern Recognition, 1997, 30, 7, 1145-1159

[2] BUTCHART S.H.M. et al, *Global Biodiversity: Indicators of Recent Declines*, Science, 2010, 328(5982), 1164-1168

[3] FARNSWORTH A. ,*Flight calls and their value for future ornithological studies and conservation research*, The Auk, 2005,122,3,733-746

[4] GREGORY R.D., van STRIEN A., *Wild Bird Indicators: Using Composite Population Trends of Birds as Measures of Environmental Health*, Ornithological Science 2010, 9(1), 3-22.

[5] GRILL T., SCHLUTER J., *Two Convolutional Neural Networks for Bird Detection in Audio Signals*, EUSIPCO Conference Proceedings (accepted), August 2017
Source code: https://jobim.ofai.at/gitlab/gr/bird_audio_detection_challenge_2017
[visited 10.07.2017]

[6] HORTON K.G., SHRIVER W.G., BULER J.J. *A comparison of traffic estimates of nocturnal flying animals using radar, thermal imaging, and acoustic recording*, 2015, Ecological Applications, 25, 390–401

[7] LECUN Y., BENGIO Y., HINTON G., *Deep Learning*, Nature, 2015,521, 436–444

[8] NEWTON I., *The Migration Ecology of Birds*, Academic Press, London, 2010

[9] PAMUŁA H., KŁACZYŃSKI M., *Monitoring bioakustyczny: cele, metody i problemy związane z automatyczną identyfikacją głosów ptaków*, Studium badawcze młodych akustyków, 2016, red. A. Pilch, Wydawnictwo AGH, 71–80

[10] STOWELL D., WOOD M., STYLIANOU Y., GLOTIN H., *Bird detection in audio: a survey and a challenge*, Machine Learning for Signal Processing, 2016 IEEE 26th International Workshop on. IEEE, 2016, 1–6
Bird Audio Detection Challenge Website: http://machine-listening.eecs.qmul.ac.uk/bird-audio-detection-challenge/ [visited 10.07.2017]

[11] WATSON M.J., WILSON D.R., MENNILL D.J., *Anthropogenic light is associated with increased vocal activity by nocturnally migrating birds,* The Condor, 2016, 118, 2, 338- 344

ZASTOSOWANIE METOD UCZENIA GŁĘBOKIEGO DO AKUSTYCZNEGO MONITORINGU PTAKÓW MIGRUJĄCYCH NOCĄ

PAMUŁA Hanna [(1)], KŁACZYŃSKI Maciej [(1)], REMISIEWICZ Magdalena [(2)], WSZOŁEK Wiesław [(1)], STOWELL Dan [(3)]

[(1)] AGH Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Wydział Inżynierii Mechanicznej i Robotyki, Katedra Mechaniki i Wibroakustyki, Al. Adama Mickiewicza 30, 30-059 Kraków
[(2)] Uniwersytet Gdański, Stacja Badania Wędrówek Ptaków, Wydział Biologii, ul. Wita Stwosza 59, 80-308 Gdańsk
[(3)] Queen Mary University of London, Centre for Digital Music, Machine Listening Lab, Mile End Rd, London E1 4NS

Zmiany liczebności i terminów pojawów ptaków są dobrymi wskaźnikami stanu środowiska. Przez lata zostało wypracowanych wiele różnych metod monitoringu ptaków, lecz najpopularniejsze są wciąż bezpośrednie obserwacje i programy regularnego obrączkowania ptaków. Jednak w niektórych przypadkach metody te są nieskuteczne, m.in. w badaniach ptaków wędrujących nocą, więc inne techniki monitoringu również powinny być rozwijane. Jedną z nich jest automatyczny monitoring akustyczny, który mógłby służyć jako wspomaganie standardowych metod monitoringu ptaków.

W pracy przedstawiamy podejście do automatycznej detekcji głosów nocnych ptasich migrantów z długookresowych nagrań ciągłych jesiennej wędrówki ptaków migrujących wzdłuż wybrzeża Morza Bałtyckiego. Do wykrywania nocnych ptasich głosów zastosowano i zaadaptowano metodę uczenia głębokiego – konwolucyjne sieci neuronowe. Wyniki są obiecujące (AUC=96.6%) i pokazują potencjał metod akustycznych w rozwijaniu i uzupełnieniu tradycyjnych technik monitoringu ptaków.