BIRDVOX-FULL-NIGHT: A DATASET AND BENCHMARK FOR AVIAN FLIGHT CALL DETECTION

Vincent Lostanlen[⋆†] *Justin Salamon*[†] *Andrew Farnsworth*[⋆] *Steve Kelling*[⋆] *Juan Pablo Bello*[†]

[⋆] Cornell Lab of Ornithology, Cornell University, Ithaca, NY, USA
[†] Music and Audio Research Laboratory, New York University, New York, NY, USA
vincent.lostanlen@nyu.edu

ABSTRACT

This article addresses the automatic detection of vocal, nocturnally migrating birds from a network of acoustic sensors. Thus far, owing to the lack of annotated continuous recordings, existing methods had been benchmarked in a binary classification setting (presence vs. absence). Instead, with the aim of comparing them in event detection, we release BirdVox-full-night, a dataset of 62 hours of audio comprising 35402 flight calls of nocturnally migrating birds, as recorded from 6 sensors. We find a large performance gap between energy-based detection functions and data-driven machine listening. The best model is a deep convolutional neural network trained with data augmentation. We correlate recall with the density of flight calls over time and frequency and identify the main causes of false alarm.

***Index Terms*—** Acoustic signal detection, audio databases, ecosystems, multi-layer neural network, supervised learning.

1. INTRODUCTION

Migratory birds face an increasing number of threats, for example from rapidly changing climate, habitat loss, and human alteration of the environment [1, 2]. In this context, assessing the spatial and temporal distributions of bird populations represents a critical need for creating appropriate conservation plans. Yet, most birds migrate at night [3, 4], which severely limits the efficacy of most existing monitoring methods for assessing their movements, e.g. diurnal citizen scientist observations of eBird [5]. A potential solution for monitoring nocturnally migrating birds is to deploy a network of low-cost acoustic sensors in desired study areas to record vocalizations of birds in sustained nocturnal migratory flights, known as flight calls [6, 7]. Yet, the processing and analysis of audio data to extract flight calls is a time-consuming and inefficient process, requiring costly efforts by a small number of experts with experience in identifying flight calls [8]. If bioacoustic analysis could be made scalable, automating the detection of flight calls in audio recordings, achieving the potential for automated monitoring would be possible [9]. Such methodology would represent a sea change in the monitoring of nocturnal movements of birds [10].

The lack of available datasets hinders the development of full-fledged systems for species-agnostic avian flight call detection. On one hand, energy-based detection functions and template matching algorithms have mostly been evaluated on near-field recordings in the presence of a single species [11, 12]. On the other, deep learning systems have recently achieved state-of-the-art results in species classification [13] and activity detection [10, 14], but their performance in event detection remains unknown. In addition, existing datasets are inadequate for context-adaptive algorithms, as they consist of intermittent recordings.

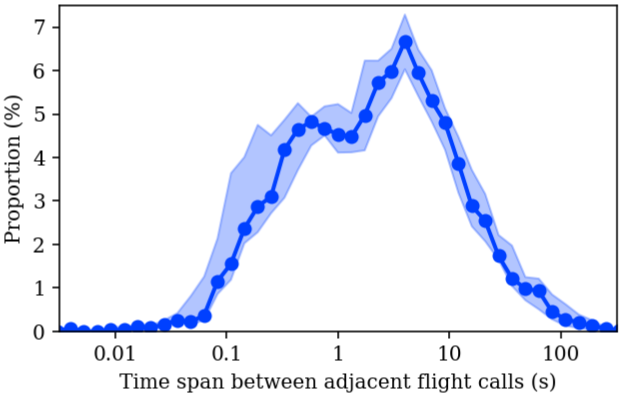

Fig. 1. Histogram of time spans between adjacent flight calls in the BirdVox-full-nightdataset. The shaded area corresponds to the interquartile range across 6 full night recordings.

In this article, we present BirdVox-full-night, a collection of 6 full night recordings comprising 35k nocturnal flight calls annotated in time and frequency. We compare 4 methods from the existing literature, including a re implementation of the "Old Bird" software, widely used among practitioners. We conduct a post hoc analysis of detection recall according to time (dusk, night, and dawn) and acoustic frequency.

We invite the reader to visit the companion website[1] of this paper, where we release the BirdVox-full-nightdataset under Creative Commons Attribution International (CC BY 4.0) license; the Python source code to reproduce experiments and figures under MIT license; and a pre-trained deep learning model under MIT license.

2. DATASET

In this section, we provide an overview of existing datasets for the acoustic detection of bird vocalizations, explain the need for a new dataset, and describe the specificities of BirdVox-full-night.

First, the CLO-WTSP and CLO-SWTH [10] datasets were collected by running a low-precision detector of flight calls whose false

This work is partially supported by NSF awards 1633259 and 1633206, Leon Levy Foundation, and a Google faculty award.

[1] Companion website: https://wp.nyu.edu/birdvox/birdvox-full-night

alarms were manually labeled as negative examples, yielding a task of binary classification of clips. Although this approach requires less human effort than inspecting continuous recordings, the low-precision detector influences the sampling of the clips, which biases the training of binary classifiers towards a restrained set of confounding factors. Furthermore, since the low-precision detector has an unknown recall, the accuracies of binary classifiers do not reflect their usability in continuous monitoring. Secondly, the Bird-DB datasets [15] consist of single-species continuous recordings of bird vocalizations, not necessarily flight calls, annotated by experts at the phrase level. These datasets do not have an official split between training set, validation set, and test set, which hinders the reproducibility of machine learning research. Thirdly, the freefield1010 dataset [16] consists of 10-second soundscapes, labeled at the clip level with species-agnostic presence of bird vocalizations, and gathered into 10 folds for cross-validation.

With BirdVox-full-night, we provide a dataset of 6 far-field, full night recordings, containing 35k flight calls from 25 species of passerines, individually annotated in time and frequency by an expert, along with an official evaluation methodology.

During the fall migration season of 2015, we deployed 10 ROBIN autonomous recording units around Ithaca, NY, USA [10]. This resulted in 966 recordings (6600 hours), among which 548 are at least 8 hours long. For 6 full night recordings, corresponding to different recording units active at the same date (September 23^{rd} and 24^{th}, under mild weather conditions), one of us (AF) pinpointed the center of every flight call in the time-frequency domain. Bird chatter and non-passerine utterances (e.g. speech, geese, dogs) were ignored. In total, the annotator pinpointed 35402 flight calls. This annotation campaign took 102 hours in total.

Figure 1 shows the distribution of time spans between adjacent flight calls in BirdVox-full-night. We find that 80% of these time spans are between $100\,\text{ms}$ and $10\,\text{s}$. With the aim of deriving abundance estimates from the output of the benchmarked systems, we formulate the task as sound event detection instead of coarse-scale binary classification of bird presence.

3. METHODS

In this section, we present four methods for avian flight call detection: a domain-specific detector ("Old Bird"); spectral flux; a shallow learning pipeline; and a deep convolutional neural network.

3.1. Energy-based detectors: "Old Bird" and spectral flux

We evaluate the "Old Bird" system of [17], which combines a detector of warblers and sparrows, *Tseep*; with a detector of thrushes, *Thrush*. *Tseep* (resp. *Thrush*) applies a real-valued filter of passband $6-10\,\text{kHz}$ (resp. $2.8-5\,\text{kHz}$) to the waveform, followed by squaring, low-pass filtering at $11\,\text{Hz}$ (resp. $5.5\,\text{Hz}$), logarithmic transformation, and differentiation at the scale of $20\,\text{ms}$. The Python reimplementation of Old Bird, named Vesper [18], runs 20 times faster than real time.

This results in a detection function on which clips are selected as flight calls if they meet the following criteria: the detection function is above a fixed threshold τ at the onset; the detection function is below $1/\tau$ at the offset; the time lag between onset and offset is between $100\,\text{ms}$ and $400\,\text{ms}$. While the author recommends an ad hoc value for τ, we apply 100 different values to match the evaluation setting of machine learning algorithms, and select τ maximizing F_1-scoreon a hold-out validation set. As a post-processing step, we follow the original implementation by applying a "clip suppressor" heuristic, which discards any sequence of 15 or more (resp. 10 or more) consecutive clips selected by *Tseep* (resp. *Thrush*) within a time span of 20 seconds or less.

We also evaluate spectral flux [19], a common method in music onset detection which comprises the same processing steps as *Tseep* and *Thrush*, yet with a sum of responses from 40 mel-frequency bands instead of domain-specific passbands.

3.2. Spherical k-means and support vector machines

We evaluate the "shallow learning" pipeline of [13], originally designed for species classification [20]. This pipeline consists of a time-frequency representation, here log-mel-spectrogram; an unsupervised feature learning stage, here principal component analysis (PCA) and spherical k-means (SKM); and a supervised classification stage, here a support vector machine (SVM).

The log-mel-spectrogram consists of 40 bands between $2\,\text{kHz}$ and $11.025\,\text{kHz}$, and is computed with the librosa library [21] with a Hann window of duration $12\,\text{ms}$ (256 samples at $22.050\,\text{kHz}$) and hop length of $1.5\,\text{ms}$ (32 samples). We extract non-overlapping patches of width $46\,\text{ms}$ (32 frames) in the time-frequency domain, leading to $32 \times 40 = 1280$ features.

Principal component analysis (PCA) projects patches into a lower-dimensional space keeping 99% of the variance in the training set, and standardizes each dimension to null mean and unit variance.

In the space of standardized principal components, the spherical k-means (SKM) algorithm learns $k = 256$ clusters maximizing intra-cluster cosine similarity on the training set [22]. We scale the centroids of these clusters to unit L^2 norm and gather them into a family of k vectors onto which PCA-whitened features are projected at prediction time.

We train a support vector machine (SVM) with radial basis function (RBF) kernel to discriminate positive from negative clips in the space of 256 standardized SKM features. Out of the 5M clips of duration $150\,\text{ms}$ in BirdVox-full-night, only 35k (0.8%) are positive. In order to compensate for this class imbalance, we restrict the number of negative clips to a subset of 35k clips. To select this subset, we start by training a shallow learning model (PCA-SKM-SVM) on an external dataset of clips collected in 2012 and 2013 in various locations of North America, not including Ithaca. We retained as negative clips the false alarms of this model predicted with greatest confidence. As a result, the 35k flight calls for all 6 recording units are supplemented with 35k false alarms, summing up to a balanced dataset of 70k clips, named BirdVox-70k. A grid search on BirdVox-70k maximizing validation accuracy selects the parameters C and γ of the model. The optimal values of these parameters vary across folds and trials; typical values are $C = 1$ and $\gamma = 5 \cdot 10^{-3}$.

Platt scaling transforms the output of the SVM into probabilistic estimates of flight call activity over clips of duration $150\,\text{ms}$ with a hop size of $50\,\text{ms}$. We interpret this sequence of probabilities as a detection function. To retrieve the temporal locations of flight calls, we select peaks of the detection function above a threshold τ under the constraint that they must be at least $\Delta t = 150\,\text{ms}$ apart from each other. Both the threshold τ and the time lag Δt are optimized by grid search on the validation set. Prediction is 8 times faster than real time.

3.3. Deep convolutional network

We evaluate the deep convolutional neural network (CNN) of [13], originally designed for species classification. The network consists

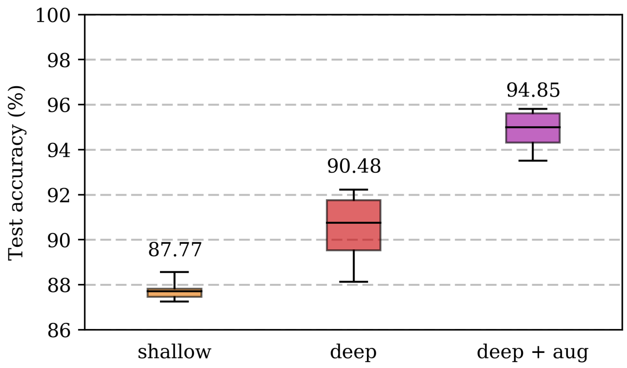

Fig. 2. Test accuracies on binary classification, averaged across folds. Each box contains 5 independent trials for each model.

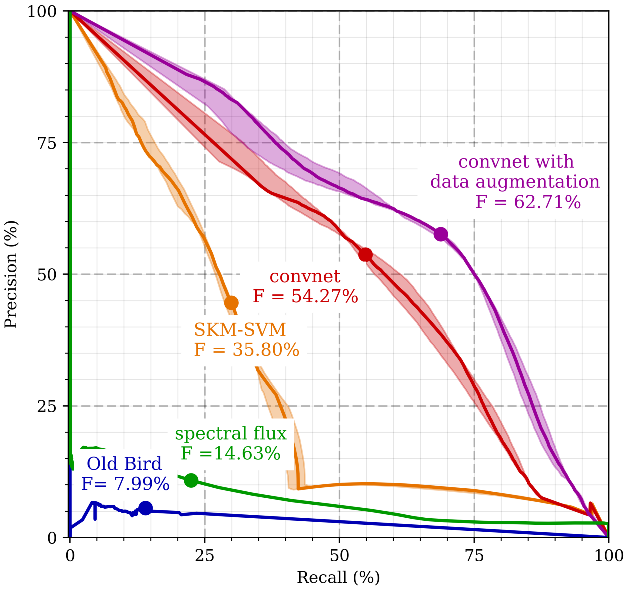

Fig. 3. Precision-recall curves on detection, averaged across folds. The shaded area corresponds to the interquartile precision range across 5 independent trials.

of 3 convolutional layers and 2 dense layers and has 677 k parameters in total. The input features are a log-mel-spectrogram with identical characteristics as the one used in the shallow learning algorithm, except that it has 128 bands instead of 40. In a preliminary experiment, we tried replacing the log-mel-spectrogram by a representation encompassing spectrotemporal modulations, namely the time-frequency scattering transform [23, 24], but did not succeed in consistently outperforming the baseline.

The training procedure is slightly modified with respect to [13]. First, because the last layer has a single output and a sigmoid nonlinearity, the loss function is binary cross-entropy instead of categorical cross-entropy. Secondly, we use the initialization method of He [25] instead of Glorot [26] for layers followed by a rectified linear unit. Thirdly, we use the Adam optimizer [27] instead of a stochastic gradient descent. Fourthly, we do not apply dropout, as we found that it consistently prevented the model to train. We use Keras [28] to train the convolutional neural network and the Pescador library [29] to stream data. Training took five hours per fold and trial on a single graphics processing unit (GPU). We interpret the output of the last layer as a detection function. Again, we select peaks above a fixed threshold τ under the constraint that they must be at least 150 ms apart from each other. Prediction time is on par with real time.

4. RESULTS

In this section, we report results for two tasks: binary classification on a balanced dataset of 70k clips and event detection on 6 full night recordings. In both cases, we split the data into 6 folds, each corresponding to a different recording unit, and run 6-fold cross-validation, with 3 folds for training, 2 for validation, and 1 for testing. To account for statistical fluctuations, we train 5 independent trials for each fold, with randomized initialization and shuffling of training data.

In binary classification, we measure global accuracy by summing the number of correctly classified clips across folds and dividing the sum by the total number of clips, that is, 70804. We find that the CNN slightly outperforms the shallow learning system, with $90.48\% \pm 1.5$ and $87.77\% \pm 0.4$ respectively, but the difference is not statistically significant ($p > 5 \cdot 10^{-3}$ after independent t-test with $n = 5$). This is in accordance with the CLO-43SD dataset for classification of flight calls into 43 species, in which the two systems perform comparably [13].

Data augmentation has been successfully employed in species classification [13] and activity detection [30] to reduce the overfitting of deep neural networks. We use the muda software [31] to deform the audio clips from the training set in 20 different ways: 12 additions of background noise (4 from each unit in the training set), 4 pitch transpositions, and 4 time stretchings. The architecture of the network remains the same, but the number of epochs is doubled to account for a slower convergence rate. Data augmentation brings the accuracy of the CNN to $94.85\% \pm 0.8$, an improvement which is statistically significant across trials ($p < 10^{-3}$ after independent t-test with $n = 5$). We do not report the accuracy of the shallow learning model on the augmented dataset because training an SVM on 700k samples is intractable in batch mode. Figure 2 summarizes binary classification results.

Once trained and validated, all systems are compared on a task of event detection over full night recordings. To match detected events with annotated events within a tolerance range of 500 ms, we use the fast implementation of maximum bipartite graph matching from the mir_eval library [32]. Varying the threshold τ allows to adjust the number of detected events, and derive true positives, false positives, and false negatives in each fold. We sum these numbers across folds before computing global metrics: precision, recall, and F_1-score. Figure 3 compares the precision of all systems as a function of their recall.

After validating the threshold τ, the Old Bird detection function and spectral flux have respective F_1-scores of 8.0% and 14.6% on average. Even with a high value of τ, yielding few or no true positives, many false alarms remain. Upon inspection, these false alarms correspond to audio artifacts ("pops") distant by exactly 30 seconds, due by the recording hardware itself. The increase in recall caused by disabling the clip suppressor in Old Bird is compensated by a decrease in precision, resulting in an F_1-score almost unchanged. Surprisingly, spectral flux outperforms the Old Bird system, which suggests that inducing hard constraints on clip durations might be detrimental.

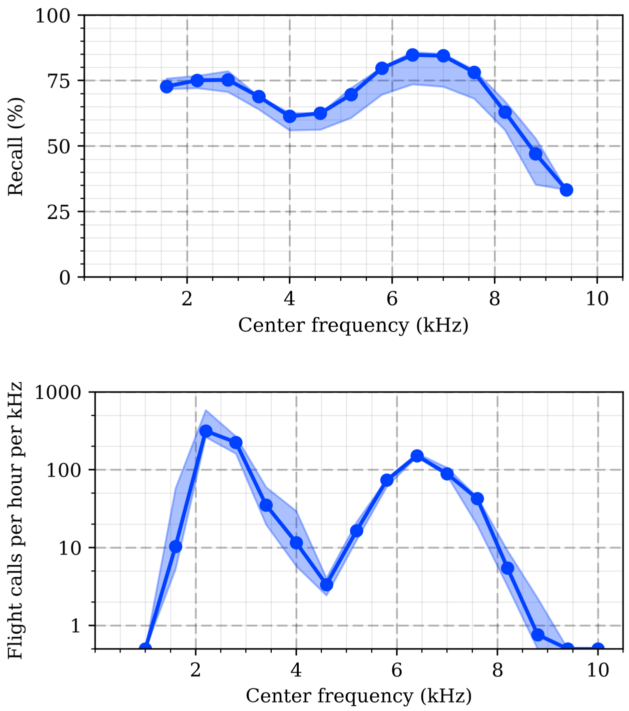

Fig. 4. Top: test recall of the CNN model, trained with data augmentation, as a function of center frequency of the flight call. The shaded area corresponds to the interquartile range across 5 independent trials. Bottom: density of flight calls in the training set as a function of their center frequency. The shaded area corresponds to the extremal range across 3 recordings.

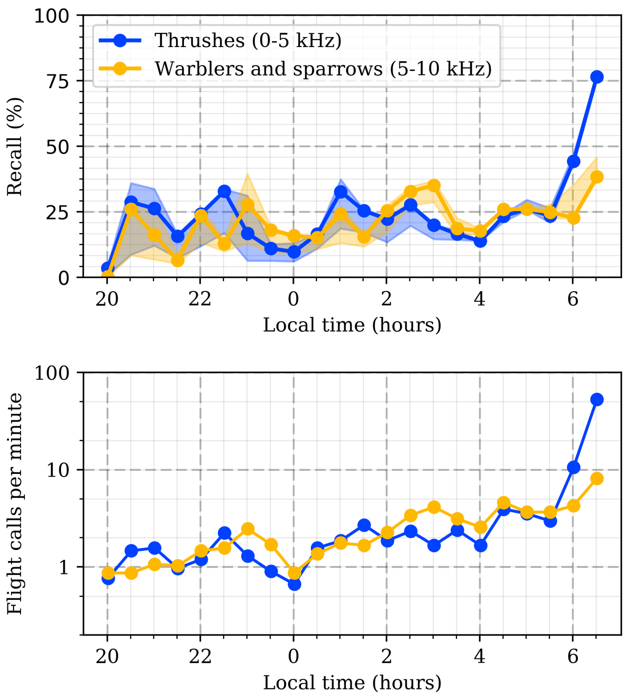

Fig. 5. Top: test recall of the CNN model, trained with data augmentation, as a function of time. The shaded areas corresponds to the interquartile range across 5 independent trials. Bottom: density of flight calls in the training set as a function of time.

Without data augmentation, the shallow and deep learning models have respective F_1-scores of 35.8% and 54.3%. Data augmentation significantly improves the F_1-score of the CNN up to 62.3% ($p < 5 \cdot 10^{-6}$ after independent t-test with $n = 5$).

Because BirdVox-full-night is annotated in frequency, we can perform a post hoc analysis of true positives and false negatives according to the center frequency of the flight call to be retrieved. In Figure 4, we cluster flight calls over bands of width 600 Hz and measure the recall of the detector in each cluster. We observe a strong positive correlation between the median recall across trials and the logarithm of the density of flight calls in the training set (Pearson's $R = 0.89, p < 2 \cdot 10^{-5}$).

Furthermore, because BirdVox-full-night consists of full night recordings, we can also cluster flight calls by local time (Figure 5). Again, at a scale of 30 minutes, we observe a strong positive correlation between the median test F_1-score across trials and species and the logarithm of the density of flight calls in the training set (Pearson's $R = 0.89, p < 5 \cdot 10^{-16}$).

5. CONCLUSION

The flight calls of migratory passerines contain valuable information for conservation science. BirdVox-full-night is the first dataset of full night recordings in which flight calls are annotated in time and frequency. As such, it provides a challenging benchmark for binary classification and event detection. Whereas energy-based detection functions, such as the "Old Bird" domain-specific system, are not robust to sonic artifacts, supervised learning systems generalize relatively well across recording locations after being trained as binary classifiers on a balanced subset of clips. In our companion website, we publish the weights of a deep convolutional network trained with data augmentation on the whole dataset. The state of the art on BirdVox-full-nightis 95% binary accuracy and 63% detection F_1-score, under a leave-one-sensor-out evaluation procedure.

Post hoc analysis highlights the biases in the state of the art. Because the traditional training procedure for supervised learning assigns the same weight to each sample, rare flight calls (e.g. at dusk or at non-typical frequencies) are less likely to be retrieved. Conversely, BirdVox-full-night offers a test bed to mitigate this problem. Indeed, since it consists of continuous recordings and is folded by sensor locations, it aims at fostering research in context-adaptive machine listening, with topics including noise reduction, recurrent models, abundance priors, weighted sampling, and transfer learning.

6. ACKNOWLEDGMENTS

We thank Jessie Barry, Ian Davies, Tom Fredericks, Jeff Gerbracht, Sara Keen, Holger Klinck, Anne Klingensmith, Ray Mack, Peter Marchetto, Ed Moore, Matt Robbins, Ken Rosenberg, and Chris Tessaglia-Hymes for designing autonomous recording units and collecting data; and Harold Mills for his help on evaluating the Old Bird detectors in the Vesper software. We acknowledge that the land on which the data was collected is the unceded territory of the Cayuga nation, which is part of the Haudenosaunee (Iroquois) confederacy.

7. REFERENCES

[1] Scott R Loss, Tom Will, and Peter P Marra, "Direct mortality of birds from anthropogenic causes," *Annual Review of Ecology, Evolution, and Systematics*, vol. 46, pp. 99–120, 2015.

[2] Franz Bairlein, "Migratory birds under threat," *Science*, vol. 354, no. 6312, pp. 547–548, 2016.

[3] Frank M Chapman, "Observations on the nocturnal migration of birds," *The Auk*, vol. 5, no. 1, pp. 37–39, 1888.

[4] Orin Grant Libby, "The nocturnal flight of migrating birds," *The Auk*, vol. 16, no. 2, pp. 140–146, 1899.

[5] Brian L. Sullivan, Jocelyn L. Aycrigg, Jessie H. Barry, Rick E. Bonney, Nicholas Bruns, Caren B. Cooper, Theo Damoulas, André A. Dhondt, Tom Dietterich, Andrew Farnsworth, Daniel Fink, John W. Fitzpatrick, Thomas Fredericks, Jeff Gerbracht, Carla Gomes, Wesley M. Hochachka, Marshall J. Iliff, Carl Lagoze, Frank A. La Sorte, Matthew Merrifield, Will Morris, Tina B. Phillips, Mark Reynolds, Amanda D. Rodewald, Kenneth V. Rosenberg, Nancy M. Trautmann, Andrea Wiggins, David W. Winkler, Weng-Keen Wong, Christopher L. Wood, Jun Yu, and Steve Kelling, "The eBird enterprise: an integrated approach to development and application of citizen science," *Biological Conservation*, vol. 169, pp. 31–40, 2014.

[6] Andrew Farnsworth, "Flight calls and their value for future ornithological studies and conservation research," *The Auk*, vol. 122, no. 3, pp. 733–746, 2005.

[7] Murray G Efford, Deanna K Dawson, and David L Borchers, "Population density estimated from locations of individuals on a passive detector array," *Ecology*, vol. 90, no. 10, pp. 2676–2682, 2009.

[8] Julia Shonfield and Erin Bayne, "Autonomous recording units in avian ecological research: current use and future applications," *Avian Conservation and Ecology*, vol. 12, no. 1, 2017.

[9] Mathieu Marcarini, Geoffrey A. Williamson, and Luis de Sisternes Garcia, "Comparison of methods for automated recognition of avian nocturnal flight calls," in *Proc. IEEE ICASSP*, 2008.

[10] Justin Salamon, Juan Pablo Bello, Andrew Farnsworth, Matt Robbins, Sara Keen, Holger Klinck, and Steve Kelling, "Towards the automatic classification of avian flight calls for bioacoustic monitoring," *PLoS One*, vol. 11, no. 11, 2016.

[11] Kantapon Kaewtip, Abeer Alwan, Colm O'Reilly, and Charles E Taylor, "A robust automatic birdsong phrase classification: a template-based approach," *The Journal of the Acoustical Society of America (JASA)*, vol. 140, no. 5, pp. 3691–3701, 2016.

[12] Anshul Thakur, Vinayak Abrol, Pulkit Sharma, and Padmanabhan Rajan, "Rényi entropy based mutual information for semi-supervised bird vocalization segmentation," in *Proc. MLSP*, 2017.

[13] Justin Salamon, Juan Pablo Bello, Andrew Farnsworth, and Steve Kelling, "Fusing shallow and deep learning for bioacoustic bird species classification," in *Proc. IEEE ICASSP*, 2017.

[14] Dan Stowell, Mike Wood, Yannis Stylianou, and Hervé Glotin, "Bird detection in audio: a survey and a challenge," in *Proc. MLSP*, 2016.

[15] Julio G. Arriaga, Martin L. Cody, Edgar E. Vallejo, and Charles E. Taylor, "Bird-DB: A database for annotated bird song sequences," *Ecological Informatics*, vol. 27, Supplement C, pp. 21 – 25, 2015.

[16] Dan Stowell and Mark D. Plumbley, "An open dataset for research on audio field recording archives: freefield1010," in *Proc. AES International Conference on Semantic Audio*, 2013.

[17] William R Evans and David K Mellinger, "Monitoring grassland birds in nocturnal migration," *Studies in Avian Biology*, vol. 19, pp. 219–229, 1999.

[18] Harold Mills, "Vesper v0.3.12," https://github.com/HaroldMills/Vesper, 2017.

[19] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler, "A tutorial on onset detection in music signals," *IEEE Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, 2005.

[20] Dan Stowell and Mark D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, 2014.

[21] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in Python," in *Proc. SciPy*, 2015.

[22] Inderjit S Dhillon and Dharmendra S Modha, "Concept decompositions for large sparse text data using clustering," *Machine learning*, vol. 42, no. 1, pp. 143–175, 2001.

[23] Joakim Andén, Vincent Lostanlen, and Stéphane Mallat, "Joint time-frequency scattering for audio classification," in *Proc. MLSP*, 2015.

[24] Vincent Lostanlen, *Convolutional operators in the time-frequency domain*, Ph.D. thesis, École normale supérieure, 2017.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. ICCV*, 2015.

[26] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, 2010.

[27] Diederik Kingma and Jimmy Ba, "Adam: a method for stochastic optimization," *Proc. ICLR*, 2015.

[28] François Chollet, "Keras v2.0.0," https://github.com/fchollet/keras, 2018.

[29] Brian McFee, Christopher Jacoby, and Eric Humphrey, "pescador v1.1," https://github.com/pescadores/pescador, 2017.

[30] Thomas Grill and Jan Schlüter, "Two convolutional neural networks for bird detection in audio signals," in *Proc. EUSIPCO special session on bird audio signal processing*, 2017.

[31] Brian McFee, Eric J. Humphrey, and Juan Pablo Bello, "A software framework for musical data augmentation," in *Proc. ISMIR*, 2015.

[32] Colin Raffel, Brian McFee, Eric J. Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, and Daniel P. W. Ellis, "mir_eval: a transparent implementation of common MIR metrics," in *Proc. ISMIR*, 2014.