NIPS4BPLUS: A RICHLY ANNOTATED BIRDSONG AUDIO DATASET

Veronica Morfi[⋆]   *Yves Bas*[†,‡]   *Hanna Pamuła* [††]   *Hervé Glotin* [‡‡]   *Dan Stowell* [⋆]

[⋆]Machine Listening Lab, Centre for Digital Music (C4DM), Queen Mary Univ. of London, UK
[†]Centre d'Ecologie et des Sciences de la Conservation (CESCO),
Muséum National d'Histoire Naturelle, CNRS, Sorbonne Univ., Paris, France
[‡]Centre d'Ecologie Fonctionnelle et Evolutive (CEFE), CNRS, Univ. de Montpellier,
Univ. Paul-Valéry Montpellier, France
[††]AGH Univ. of Science and Technology, Department of Mechanics and Vibroacoustics, Kraków, Poland
[‡‡]Univ. Toulon, Aix Marseille Univ., CNRS, LIS, DYNI team, SABIOD, Marseille, France

ABSTRACT

Recent advances in birdsong detection and classification have approached a limit due to the lack of fully annotated recordings. In this paper, we present NIPS4Bplus, the first richly annotated birdsong audio dataset, that is comprised of recordings containing bird vocalisations along with their active species tags plus the temporal annotations acquired for them. Statistical information about the recordings, their species specific tags and their temporal annotations are presented along with example uses. NIPS4Bplus could be used in various ecoacoustic tasks, such as training models for bird population monitoring, species classification, birdsong vocalisation detection and classification.

***Index Terms*—** audio dataset, ecosystems, bird vocalisations, rich annotations, ecoacoustics

1. INTRODUCTION

The potential applications of automatic species detection and classification of birds from their sounds are many (e.g. ecological research, biodiversity monitoring, archival) [1, 2, 3, 4, 5]. In recent decades, there has been an increasing amount of ecological audio datasets that have tags assigned to them to indicate the presence or not of a specific bird species. Utilising these datasets and the provided tags, many authors have proposed methods for bird audio detection [6, 7] and bird species classification, e.g. in the context of LifeCLEF classification challenges [8, 9] and more [10, 11]. However, these methods do not predict any information about the temporal location of each event or the number of its occurrences in a recording.

Some research has been made into using audio tags in order to predict temporal annotations, labels that contain temporal information about the audio events. In [12, 13], the authors try to exploit audio tags in birdsong detection and bird species classification, in [14], the authors use deep networks to tag the temporal location of active bird vocalisations, while in [15], the authors propose a bioacoustic segmentation based on the Hierarchical Dirichlet Process (HDP-HMM) to infer song units in birdsong recordings. Furthermore, some methods for temporal predictions by using tags have been proposed for other types of general audio [16, 17, 18]. However, in all the above cases some kind of temporal annotations were used in order to evaluate the performance of the methods.

Annotating ecological data with temporal annotations to train sound event detectors and classifiers is a time consuming task involving a lot of manual labour and expert annotators. There is a high diversity of animal vocalisations, both in the types of the basic syllables and in the way they are combined [19, 20]. Also, there is noise present in most habitats, and many bird communities contain multiple bird species that can potentially have overlapping vocalizations [21, 22, 23]. These factors make detailed annotations laborious to gather, while on the other hand acquiring audio tags takes much less time and effort, since the annotator has to only mark the active sound event classes in a recording and not their exact boundaries. This means that many ecological datasets lack temporal annotations of bird vocalisations even though they are vital to the training of automated methods that predict the temporal annotations which could potentially solve the issue of needing a human annotator.

Recently, BirdVox-full-night [24], a dataset containing some temporal and frequency information about flight calls of nocturnally migrating birds, was released. However, BirdVox-full-night only focuses on avian flight calls, a specific type of bird calls, that usually have a very short duration in time. The temporal annotations provided for them don't include any onset, offset or information about the duration of

the calls, they simply contain a single time marker at which the flight call is active. Additionally, there is no distinction between the different bird species, hence no specific species annotations are provided, but only the presence of flight calls through the duration of a recording is denoted. Hence, the dataset can provide data to train models for flight call detection but is not efficient for models performing both event detection and classification for a variety of bird vocalisations.

In this paper, we introduce NIPS4Bplus, the first ecological audio dataset that contains bird species tags and temporal annotations [25], and can be used for training supervised automated methods that perform bird vocalisation detection and classification and can also be used for evaluating methods that use only audio tags or no annotations for training. The rest of the paper is structured as follows: Section 2 describes the process of collecting and selecting the recordings comprising the dataset, Section 3 presents our approach of acquiring the tags and temporal annotations and provides statistical information about the labels and recordings comprising the dataset followed by example uses of NIPS4Bplus, with the conclusion in Section 4.

2. AUDIO DATA COLLECTION

The recordings that comprise the Neural Information Processing Scaled for Bioacoustics (NIPS4B) 2013 training and testing dataset were collected by recorders placed in 39 different locations, which can be summarised by 7 regions in France and Spain, as depicted in Fig. 1. 20% of the recordings were collected from the Haute-Loire region in Central France, 65% of them were collected from the Pyrénées-Orientales, Aude and Hérault regions in south-central France along the Mediterranean cost and the remaining 15% of the recordings originated from the Granada, Jaén and Almeria regions in eastern Andalusia, Spain. The Haute-Loire area is a more hilly and cold region, while the rest of the regions are mostly along the Mediterranean coast and have a more Mediterranean climate.

The recorders used to acquire the recordings were the SM2BAT using SMX-US microphones. They were originally put in the field for bat echolocation call sampling, but they were also set to record for 3 hours single channel at 44.1 kHz sampling rate starting 30 minutes after sunrise, right after bat sampling. The recorders were set to a 6 dB Signal-to-Noise Ratio (SNR) trigger with a window of 2 seconds, and acquired recordings only when the trigger was activated.

Approximately 30 hours of field recordings were collected. Any recording longer than 5 seconds was split into multiple 5 second files. SonoChiro, a chirp detection tool used for bat vocalisation detection, was used on each file to identify recordings with bird vocalisations.[1] A stratified random sampling was then applied to all acquired recordings,

---

[1] http://www.leclub-biotope.com/fr/72-sonochiro

based on locations and clustering of features, to maximise the diversity in the labelled dataset, resulting in nearly 5000 files being chosen. Following the first stage of selection, manual annotations were produced for the classes active in these 5000 files and any recordings that contained unidentified species' vocalisations were discarded. Furthermore, the training set and testing set recordings were allocated so that the same species were active in both. Finally, for training purposes, only species that could be covered by at least 7 recordings in the training set were included in the final dataset, the rest were considered rare species' occurrences that would make it hard to train any classifier, hence were discarded. The final training and testing set consist of 687 files of total duration of less than an hour, and 1000 files of total duration of nearly two hours, respectively.
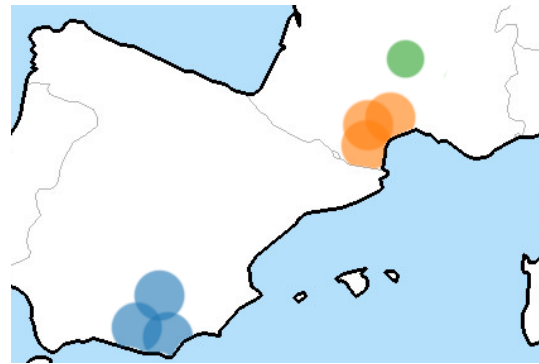

Fig. 1. Regions where the dataset recordings were collected from. Green indicates Central France region Haute-Loire. Orange indicates Southern France regions Pyrénées-Orientales, Aude and Hérault. Blue indicates Southern Spain regions Granada, Jaén and Almeria.

3. ANNOTATIONS

3.1. Tags

The labels for the species active in each recording of the training set were initially created for the NIPS4B 2013 bird song classification challenge [26]. There is a total of 61 different bird species active in the dataset. For some species we discriminate the song from the call and from the drum. We also include some species living with these birds: 7 insects and an amphibian. This tagging process resulted in 87 classes. A detailed list of the class names and their corresponding species English and scientific names can be found in [25]. These tags only provide information about the species active in a recording and do not include any temporal information. In addition to the recordings containing bird vocalisations, some training files only contain background noise acquired from the same regions and have no bird song in them, these files can be used to tune a model during training. Fig. 2 depicts the number of occurrences per class for recordings collected in each of the 3
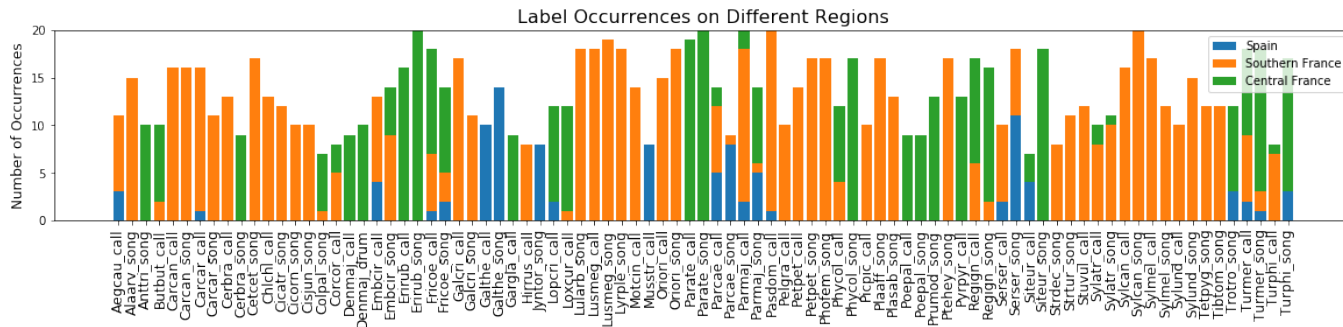
Fig. 2. Number of occurrences of each sound type in recordings collected from Spain, Southern France and Central France.

different general regions of Spain, South France and Central France. Each tag is represented by at least 7 up to a maximum of 20 recordings.

Each recording that contains bird vocalisations includes 1 to 6 individual labels. These files may contain different vocalisations from the same species and also may contain a variety of other species that vocalise along with this species. Fig. 3 depicts the distribution of the number of active classes in the dataset.
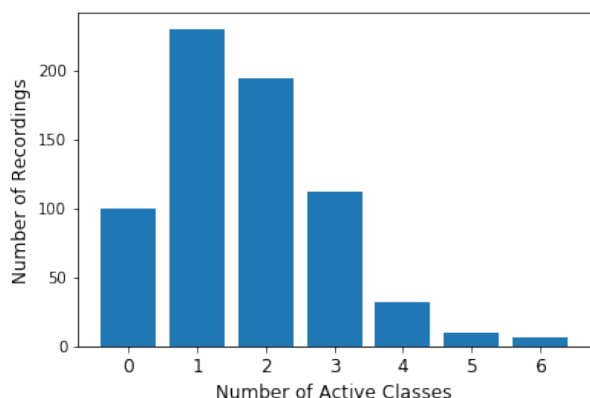

Fig. 3. Distribution of number of active classes in dataset recordings.

3.2. Temporal Annotations

Temporal annotations for each recording in the training set of the NIPS4B dataset were produced manually using Sonic Visualiser.[2] The temporal annotations were made by a single annotator, Hanna Pamuła, and can be found in [25]. Table 1 presents the temporal annotation format as is provided in NIPS4Bplus and Fig. 4 depicts the visual representation of the temporal annotations.

In concern to the temporal annotations for the dataset, we should mention the following:

- The original tags were used for guidance, however some files were judged to have a different set of species than the ones given in the original metadata.

- In a few rare occurrences, despite the tags suggesting a bird species active in a recording, the annotator was not able to detect any bird vocalisation.

- An extra 'Unknown' tag was added to the dataset for vocalisations that could not be classified to a class.

- An extra 'Human' tag was added to a few recordings that have very obvious human sounds, such as speech, present in them.

- Out of the 687 recordings of the training set 100 recordings contain only background noise, hence no temporal annotations were needed for them.

- Of the remaining 587 recordings that contain vocalisations, 6 could not be unambiguously labelled due to hard to identify vocalisations, thus no temporal annotation files were produced for them.

- An annotation file for any recording containing multiple insects does not differentiate between the insect species and the 'Unknown' label was given to all insect species present.

- In the rare case where no birds were active along with the insects no annotation file was provided. Hence, 7 recordings containing only insects were left unlabelled.

- In total, 13 recordings have no temporal annotation files. These can be used when training a model that does not use temporal annotations.

- On some occasions, the different syllables of a song were separated in time into different events while in other occasions they were summarised into a larger event, according to the judgement of the expert annotator. This variety could help train an unbiased model regarding separating events or grouping them together as one continuous time event.

As mentioned above, each recording may contain multiple species vocalising at the same time. This can often occur in wildlife recordings and is important to be taken into account
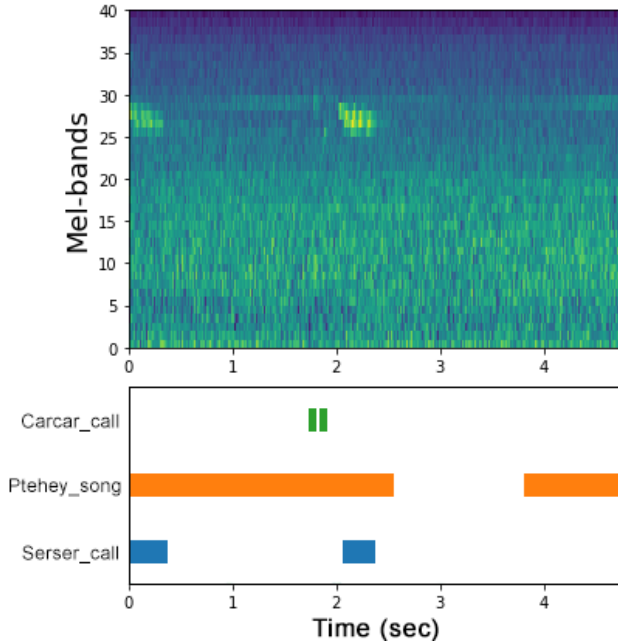
---

[2]https://www.sonicvisualiser.org/

Fig. 4. Mel-band spectrogram of a recording in NIPS4Bplus and the visual representation of the corresponding temporal annotations as noted in Table 1.

Table 1. NIPS4Bplus temporal annotations of the recording depicted in Fig. 4

Starting Time (sec)	Duration (sec)	Tag
0.00	0.37	Serser_call
0.00	2.62	Ptehey_song
1.77	0.06	Carcar_call
1.86	0.07	Carcar_call
2.02	0.41	Serser_call
3.87	1.09	Ptehey_song

when training a model. Fig. 5 presents the fraction of the total duration containing overlapping vocalisations as well as the number of simultaneously occurring classes.

A few examples of the NIPS4Bplus dataset and temporal annotations being used can be found in [27] and [28]. First, in [27], we use NIPS4Bplus to carry out the training and evaluation of a newly proposed multi-instance learning (MIL) loss function for audio event detection. And in [28], we combine the proposed method of [27] and a network trained on the NIPS4Bplus tags that performs audio tagging in a multi-task learning (MTL) setting. Additional applications using NIPS4Bplus could include training models for bird species audio event detection and classification, evaluating how generalisable of method trained on a different set of data is, and many more.
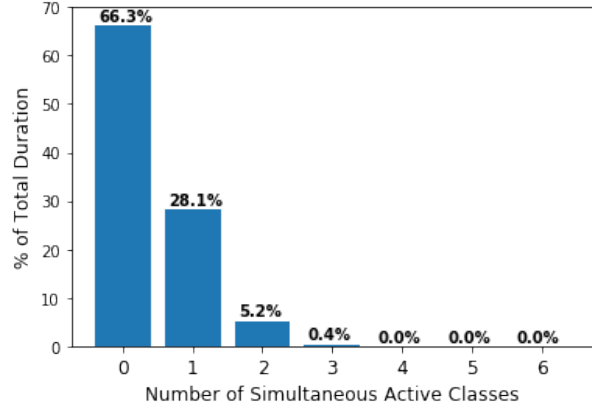

Fig. 5. Distribution of simultaneous number of active classes on the total duration of the recordings.

4. CONCLUSION

In this paper, we present NIPS4Bplus, the first richly annotated birdsong audio dataset. NIPS4Bplus is comprised of the NIPS4B dataset and tags used for the 2013 bird song classification challenge plus the newly acquired temporal annotations. We provide statistical information about the recordings, their species specific tags and their temporal annotations.

5. REFERENCES

[1] D. K. Dawson and M. G. Efford, "Bird population density estimated from acoustic signals," *Journal of Applied Ecology*, vol. 46, no. 6, pp. 1201–1209, 2009.

[2] K. T. A. Lambert and P. G. McDonald, "A low-cost, yet simple and highly repeatable system for acoustically surveying cryptic species," *Austral Ecology*, vol. 39, no. 7, pp. 779–785, 2014.

[3] K. L. Drake, M. Frey, D. Hogan, and R. Hedley, "Using digital recordings and sonogram analysis to obtain counts of yellow rails," *Wildlife Society Bulletin*, vol. 40, no. 2, pp. 346–354, 2016.

[4] S. G. Sovern, E. D. Forsman, G. S. Olson, B. L. Biswell, M. Taylor, and R. G. Anthony, "Barred owls and landscape attributes influence territory occupancy of northern spotted owls," *The Journal of Wildlife Management*, vol. 78, no. 8, pp. 1436–1443, 2014.

[5] Tiago A. Marques, Len Thomas, Stephen W. Martin, David K. Mellinger, Jessica A. Ward, David J. Moretti, Danielle Harris, and Peter L. Tyack, "Estimating animal population density using passive acoustics," *Biological Reviews*, vol. 88, no. 2, pp. 287–309, 2013.

[6] S. Adavanne, K. Drossos, E. Çakir, and T. Virtanen, "Stacked convolutional and recurrent neural networks for bird audio detection," in *2017 25th European Signal Processing Conference (EUSIPCO)*, Aug 2017, pp. 1729–1733.

[7] T. Pellegrini, "Densely connected cnns for bird audio detection," in *2017 25th European Signal Processing Conference (EUSIPCO)*, Aug 2017, pp. 1734–1738.

[8] H. Goëau, H. Glotin, W.-P. Vellinga, R. Planqué, and A. Joly, "LifeCLEF Bird Identification Task 2016: The arrival of Deep Learning," 2016.

[9] H. Goëau, H. Glotin, W.-P. Vellinga, R. Planqué, and A. Joly, "LifeCLEF Bird Identification Task 2017," 2017.

[10] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, pp. 279–283, 2017.

[11] E. Knight, K. C. Hannah, G. J. Foley, C. D. Scott, R. Brigham, and E. Bayne, "Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs," *Avian Conservation and Ecology*, vol. 12, 11 2017.

[12] F. Briggs, B. Lakshminarayanan, L. Neal, X. Fern, R. Raich, S. J. K. Hadley, A. S. Hadley, and M. G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *Journal of the Acoustic Society of America*, vol. 131, pp. 4640–4650, 2014.

[13] J. F. Ruiz-Muñoz, Mauricio Orozco-Alzate, and G. Castellanos-Dominguez, "Multiple instance learning-based birdsong classification using unsupervised recording segmentation," in *Proceedings of the 24th International Conference on Artificial Intelligence*. 2015, IJCAI'15, pp. 2632–2638, AAAI Press.

[14] L. Fanioudakis and I. Potamitis, "Deep networks tag the location of bird vocalisations on audio spectrograms," 2017, *arXiv:1711.04347*.

[15] V. Roger, M. Bartcus, F. Chamroukhi, and H. Glotin, "Unsupervised bioacoustic segmentation by hierarchical dirichlet process hidden markov model," in *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, pp. 113–130. Springer, 2018.

[16] J. Schlüter, "Learning to Pinpoint Singing Voice from Weakly Labeled Examples," in *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, New York, USA, 2016.

[17] S. Adavanne and T. Virtanen, "Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 12–16.

[18] A. Kumar and B. Raj, "Audio event detection using weakly labeled data," in *Proceedings of the 2016 ACM on Multimedia Conference*, New York, NY, USA, 2016, MM '16, pp. 1038–1047, ACM.

[19] T. S. Brandes, "Automated sound recording and analysis techniques for bird surveys and conservation," *Bird Conservation International - BIRD CONSERV INT*, vol. 18, pp. S163–S173, 09 2008.

[20] D.E. Kroodsma, *The Singing Life of Birds: Audio CD*, The Singing Life of Birds: The Art and Science of Listening to Birdsong. Houghton Mifflin, 2005, *isbn:9780618405688*.

[21] D. Luther, "Signaller: Receiver coordination and the timing of communication in amazonian birds," *Biology Letters*, vol. 4, pp. 651–654, 2008.

[22] D. Luther and R. Wiley, "Production and perception of communicatory signals in a noisy environment," *Biology Letters*, vol. 5, pp. 183–187, 2009.

[23] K. Pacifici, T. R. Simons, and K. H. Pollock, "Effects of vegetation and background noise on the detection process in auditory avian point-count surveys," *The Auk*, vol. 125, no. 4, pp. 998–998, 2008.

[24] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Birdvox-full-night: a dataset and benchmark for avian flight call detection," in *Proc. IEEE ICASSP*, April 2018.

[25] V. Morfi, D. Stowell, and H. Pamuła, "Transcriptions of NIPS4B 2013 Bird Challenge training dataset," 2018, doi: `https://doi.org/10.6084/m9.figshare.6798548.v4`, Accessed: 29-Oct-2018.

[26] Artières T. Mallat S. Tchernichovski O. Halkias X. Glotin H., LeCun Y., "Proc. neural information processing scaled for bioacoustics, from neurons to big data," USA, 2013, NIPS Int. Conf.

[27] V. Morfi and D. Stowell, "Data-efficient weakly supervised learning for low-resource audio event detection using deep learning," 2018, *arXiv preprint arXiv:1807.06972*.

[28] V. Morfi and D. Stowell, "Deep learning for audio event detection and tagging on low-resource datasets," *Applied Sciences*, vol. 8, no. 8, pp. 1397, Aug 2018.