BIRD DETECTION IN AUDIO: A SURVEY AND A CHALLENGE

*Dan Stowell**   *Mike Wood†*   *Yannis Stylianou*   *Hervé Glotin‡*

Machine Listening Lab, Centre for Digital Music, Queen Mary University of London.

Ecosystems and Environment Research Centre, University of Salford.

Computer Science Department, University of Crete.

LSIS UMR CNRS, University of Toulon, Inst. Univ. de France

ABSTRACT

Many biological monitoring projects rely on acoustic detection of birds. Despite increasingly large datasets, this detection is often manual or semi-automatic, requiring manual tuning/postprocessing. We review the state of the art in automatic bird sound detection, and identify a widespread need for tuning-free and species-agnostic approaches. We introduce new datasets and an IEEE research challenge to address this need, to make possible the development of fully automatic algorithms for bird sound detection.

1. INTRODUCTION

Monitoring birds by their sound is important for many environmental and scientific purposes. A variety of crowdsourcing and remote-monitoring projects now record these sounds, and some analyse the sound automatically—yet there are still many issues to solve, as indicated by the number of projects that are yet to be fully automated [1, 2, 3]. In this paper we review research on the specific topic of bird *detection* in audio. The audio modality is well-suited to bird monitoring because many birds are much more clearly detectable by sound than by vision or other indicators. We overview the paradigms and techniques used for bird audio detection, and specific issues to be addressed. We then describe a data challenge which we are introducing, with new public datasets, as an initiative to advance the state of the art. First, though, we must outline the applications for which bird detection in audio is useful.

Bioacoustics has in recent years become one of the "big data" research areas, in particular with remote acoustic monitoring projects generating terabytes of audio, far more than can feasibly be inspected manually. The goal of such projects is usually to monitor population densities and migration patterns of animal species, or to monitor overall ecosystem health. For example [4] found that automatically-detected calling activity was a reliable indicator of relative abundance for monitoring a seabird colony. [1] further reviewed the use of passive acoustic monitoring to estimate animal population density and there is increasing interest in using acoustic indices for biodiversity assessments [5].

Other large-scale monitoring programmes have used an *occupancy* framework, meaning that instead of working with abundance data (i.e. estimated numbers of individuals), the simpler presence/absence of a species in a spatio-temporal window is the basic observation [6, 7]. The efficiency of collecting occupancy data motivates its use in large-scale studies [6]. Rowe [7], using an occupancy framework, found that automated recognition software improves detectability for a range of bird species' vocalizations, though also found that with current technology the manual effort required—to set parameters and to check and post-process the results—means that the efficiency in terms of person-hours was actually not reduced relative to a manual survey. This demonstrates that automatic detection is useful in practice but the automation of this requires further development. Marques et al. [1] likewise concluded that improvements in automatic detection (and classification) would be desirable, especially with respect to calibration and full automation.

Unattended monitoring is not the only application to require bird detection. Another common use case is as a pre-filtering step before other automatic analyses such as bird species classification [8, 9]. It is particularly needed in uncontrolled data collection scenarios such as crowdsourcing.

These diverse use cases have broadly similar requirements, but can differ in the exact precision of detail required. Hence, before reviewing technical approaches used to address these tasks, we must be a little more specific about the task specifications.

978-1-5090-0746-2/16/$31.00 ©2016 IEEE

Output paradigms for event detection of birds
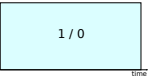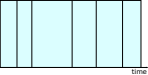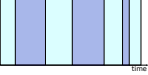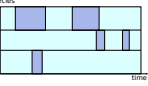
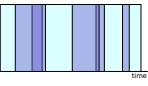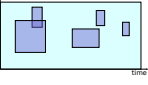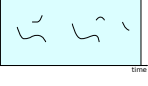| Output format		Common algorithms	Used by applications	Advantages	Disadvantages	Cmplxty
Presence/absence	1 / 0	Classifiers	Occupancy-models in statistical ecology; retrieval / data mining systems generally	Evaluation is straightforward; manual annotation can be efficient	Low temporal precision; multiple events merged	1
Onsets		Onset detectors e.g. energy slope, per-frame classifier		Overlapping events are OK	No offset/duration information	1
Monophonic segmentation / VAD		Energy thresholding; HMM decoding	VAD		Overlapping events merged	2
Polyphonic segmentation (multi-monophonic)		NMF		Joint estimation can reduce confusion between similar sound types; overlaps *between* species are OK	Overlaps in *same* species merged	3
Polyphonic segmentation (overlappable)				Overlapping events are OK		3
Time-frequency boxes		Spectrogram cross-correlation	Common where spectrogram cross-correlation used e.g. in Raven			4
Time-frequency blobs		Image-processing object-detection methods e.g. spectrogram thresholding, connected component			Harmonic stacks may separate	5
Time-frequency sinusoids		Pitch trackers	Common in whale studies		Inappropriate for non-tonal and other complex sounds; harmonic stacks may separate	5

Fig. 1. Task paradigms for bird detection. The final column gives a rough ordering in ascending complexity/difficulty.

2. TASK PARADIGMS

The idea of detecting birds in audio can be made into a concrete task in various ways, each connecting with different application tasks and implying very different output data. Do we need to know the exact start/end times? Do we need to know about each vocalisation separately? Do we need to know how many vocalisations, or how many individuals, or just an overall presence/absence? Figure 1 illustrates different task paradigms that have been studied, along with some of their characteristics. Their differing characteristics have strong implications, both for which computational approaches are appropriate and for the practical feasibility of annotation and evaluation.

The most basic is the simple estimation of presence/absence in a given sound clip: a detector outputs a zero if none of the target species are detected, and a one otherwise. This provides relatively little information—low temporal detail, no differentiation between one and many detections—yet it has practical relevance. The *occupancy modelling* framework in statistical ecology [10] uses exactly this type of data, and is recommended because it is often easier and less expensive to collect than abundance data, especially in the context of very large surveys [6, 7]. The output format is a simple binary decision, which gives scope for various classification methods to be adapted directly, and also allows for very efficient manual annotation. For applications such as filtering a large dataset, or assisting with manual data browsing, annotated fine detail is often unnecessary since the purpose is simply to help a person or a machine skip over the (typically large) number of negative instances to focus on the audio region containing the positive instances.

A common variant is to add temporal detail to the pres-

ence/absence decision: in other words, to partition the time axis into positive and negative regions (Figure 1 c). This is often the format produced by methods such as thresholding based on short-term energy levels, and is appealing for streaming applications such as real-time detectors. It is analogous to the approach commonly adopted in voice activity detection (VAD) for speech [11]. Note that in VAD applications there is typically one dominant source of interest, while in natural sound monitoring the signals of interest are often intrinsically polyphonic. Thus a positive region may contain one or many vocalisations together. This task paradigm maintains the advantage of relative simplicity while adding a little more temporal resolution.

Related methods deal with polyphony more explicitly. Template-matching methods (see next section) yield either event detections (Figure 1 b) or time-frequency boxes (Figure 1 f), allowing overlapping vocalisations to be represented as separate data points. Occasionally other methods output temporal regions which can overlap [14].

In bird sound, a paradigm that has become common and built in to standard software is to describe events via time-frequency boxes (Figure 1 f). These can be annotated relatively intuitively by drawing boxes on a spectrogram, and detected using template-matching methods. This paradigm works well when sounds are compact in time and frequency: the approach is not seen in speech and music analysis, because in those cases the signals of interest are often broadband, consisting of harmonic stacks and noises. A sizeable portion of bird sounds is relatively bandlimited; however for sounds with significant energy across a range of harmonics, there may be a tendency to exclude higher harmonics, or to create large regions containing many subbands with no energy from the signal of interest. These are not show-stopping issues, but may inhibit accuracy for some species.

Going to even more detail, some authors consider detections as arbitrarily-shaped regions on a spectrogram (Figure 1 g) [15, 16]. This approach fits with object-detection methods in image processing. Often each detected event is required to be a single fully-connected region ("blob detection"), which is problematic for harmonic sounds since harmonics may then be detected as separate events. Manually labelling data at this resolution is labour-intensive, and it is rare that the final downstream applications require such detail.

Tonal sounds can be considered as time-varying sinusoids, in which case annotation may come in the form of frequency tracks (Figure 1 h). This has been explored in marine mammal detection, and sometimes for bird sounds [17]. Again, harmonics might be detected as separate events; however some models are able to conjoin harmonics into unified tracks.

The task paradigms just discussed each have different affordances. They provide varying levels of detail for downstream tasks, but they also enable different sets of technical solutions, and imply different amounts of manual labour to annotate and evaluate. We will return to the relative impor-

tance of these task paradigms after reviewing the literature on technical methods that have been used to address them.

3. TECHNICAL METHODS

3.1. Established/baseline methods

The most common methods for detection are based on either energy, spectrogram cross-correlation, or hidden Markov models (HMMs). These well-known baselines are available in widely-used bioacoustics software (*Raven*, *XBAT*, *Song Scope*), and have been used for various surveys.

Perhaps the simplest method is energy thresholding, which yields a VAD-like segmentation output: positive if the energy in a short time-window is higher than a threshold, otherwise negative. For bioacoustic surveys it is usually preceded by some kind of noise reduction, and often applied to bandlimited frequency regions of interest [18, 19]. [20] augment it with an iterative process which estimates the background noise level as it converges.

Also common is spectrogram cross-correlation, an alternative which uses one or more example sounds as templates. These templates are species-specific, and used to scan a spectrogram for regions with strongly-matching profiles by cross-correlation. For example [4] used spectrogram cross-correlation (in XBAT) to detect calls in a seabird colony. Across millions of calls the detection accuracy was 53.6%, which the authors compare against 22%, 17% and 24% for studies detecting terrestrial birds; the authors also reported that soundscape characteristics had an impact on detection.

Hidden Markov models (HMMs) have widely been used for sound sequence analysis, especially in speech, and have a particular appeal of temporal flexibility that goes beyond template matching. HMMs have been used for various purposes in bioacoustics including bird detection, for example in *Song Scope* software [18]. Using *Song Scope*, [2] reported sensitivity ranging from 56% to 69% for detecting three seabird species. [16] report in passing that they found MFCCs and HMMs to perform poorly for detection, hence their use of other methods.

Alternatively, some investigators use template matching, but instead of cross-correlation they employ dynamic time warping (DTW) which allows for the template and the target sound to be slightly warped in time relative to each other [21]. DTW is in principle a very suitable method for natural sounds such as bird sounds which often have much organic variability. However, note that DTW remains much less widely used than cross-correlation, perhaps because in practice the flexibility does not give a strong enough boost over the simpler (and thus computationally more efficient) cross-correlation.

3.2. Recent work

Towsey et al. [16] provide a software toolbox with multiple detection methods, including: spectrogram template match-

ing; "oscillation detection" by detecting amplitude modulations in narrow frequency bands; energy-based segmentation; spectral peak tracking; and spectrogram blob detection. Note that the focus of Towsey et al. is explicitly on single-species targeted studies, and the choice of method must be chosen based on knowledge of the target species vocalisations. This is eminently possible for targeted single-species studies, but difficult to generalise to species-agnostic detection. It is an interesting open question whether the various different approaches can simply be aggregated under a meta-algorithm to produce species-agnostic output; to our knowledge this has not been attempted.

[22, 17] detect sinusoidal (pure-tone-like) signals in noise, and use these as a basis for detecting bird species. Their method is able to detect very fast-modulated pitch tracks, which sets it apart from other methods and makes it suitable for many frequency-modulated bird sounds. However, a method based on sinusoidal tracks may of course be inappropriate for the case of non-tonal bird sounds.

[14] propose a model based on detecting onsets and offsets separately, then using typical syllable durations to unify the onsets and offsets into probabilistically smoothed event detections, which may include overlaps. This method is perhaps most suitable for monosyllabic vocalisations.

[23] use a random forest (RF) classifier to make detection decisions for the presence/absence of flight calls. Sound events are first detected with simple bandlimited energy detection, and then the RF method refines these initial decisions by discarding many of the false positives. [23] argue that the random forest algorithm is appropriate for the detection task, because of various properties including its ability to handle polymorphic categories (i.e. the positive events do not all have to be of the same type).

[15] work within the relatively uncommon paradigm of detecting exactly which pixels in a spectrogram should be labelled as belonging to bird sound. They train a RF classifier to make the pixel-level decisions. This approach has the clear advantages and drawbacks of the paradigm: the system is able to output detailed estimates (spectrotemporal shapes), at the cost that the user must train the system by providing a set of pixel-wise binary mask information. However, they demonstrate that in this paradigm, the RF achieves much better results than energy-based detection. [16] also include a pixel-wise method in their toolbox, based on energy and size of 'blob' rather than on training a classifier, which should thus be more general though potentially less accurate. [24] also use a smoothing technique which sits with these image-based approaches, in their case to preprocess the spectrogram before applying energy-based segmentation.

There are of course detection systems developed and deployed for related tasks outside of bird sound, in speech, music and environmental sound. In speech, VAD is well-studied and should be a source of inspiration, although some methods may be speech-specific [11]. Note also that VAD is typically applied in a monophonic close-mic scenario, whereas we often wish to detect polyphonic and distant bird sounds. [8] use "a simple voice activity detection system, with acoustic models trained with bird vocalization data" as a preprocessing step before bird species classification. However the VAD method is not specified.

For general soundscapes, [25] review the state of the art in detecting everyday sound events in urban sound scenes, and evaluate many methods via a public data challenge using audio recorded in office environments. The challenge uses two detection paradigms: monophonic and "multi-monophonic" (in our current terminology), in both cases aiming to retrieve the start time, end time and label for each event. The generally best-performing detector in their study was a two-layer HMM approach; also strong was a combined HMM and RF method. MFCCs were not found to be useful features for event detection, generally outperformed by spectrogram or filterbank features. Regarding evaluation, the authors conclude that more work is needed to ensure that the evaluation measures used for structured data tasks match up with the task desiderata.

4. PRACTICAL CONSIDERATIONS

Noise reduction; weather noise: Background noise must be a consideration, and even simple noise reduction can help with downstream processing. The assumption of temporally constant background noise levels (see e.g. [20]) is in general unrealistic for outdoor sound recordings, but is a first step for simple noise reduction. Some approaches allow for smoothly-varying background noise. However, the bigger issue is robustness to strongly-varying noise, especially wind and rain, but also from other fauna (e.g. [3, 2]). In practical applications heavily affected sound clips may have to be removed (e.g. 2.9% of recording time in [2]). For example the toolbox of [16] performs automatic wind and rain detection as a classification task.

Manual intervention: calibration, post-processing: An important issue for large-scale studies and for general application is how much user intervention is actually needed in practice, even for nominally automatic methods. Widely-used tools such as Raven and SongScope require manual calibration of thresholds and/or templates for each species of interest before they can be used, and this can have a strong effect on precision and sensitivity [18, 7, 2].

Single-species vs. generic: An important question is whether the detector for a certain situation should be detecting individuals from a single species, or more generally such as from a whole taxon. Single-species studies can be useful for example in studying so-called "keystone species", or in cases where highly custom detectors might be used to detect idiomatic sounds (e.g. woodpecker drumming). Conversely, there are many cases in which the desire is to detect all vocalisations irrespective of species: e.g. for overall ecosystem

monitoring, or as a filtering front-end before further analysis such as classification. This is particularly the case in situations where not all species are known or well-characterised.

Some techniques are inherently more suited to one or the other: template detection is inherently specific, while energy-based detection can be very generic. For surveys that must cover a wide range of species yet with high specificity (and perhaps with a high rejection of distractor events), it may be useful to apply a range of focussed detectors and then to aggregate their outputs. This can be done straightforwardly. There is unexplored scope however for meta-algorithms to aggregate the outputs of multiple detectors intelligently, helping to mitigate "double firing" from independent detectors.

5. A RESEARCH DATA CHALLENGE

To stimulate the next research advances on species-agnostic bird detection, we present an IEEE-sponsored data challenge. For this challenge we introduce two new public datasets of annotated audio data. For the challenge tasks we have opted for the presence/absence paradigm, applied to ten-second audio excerpts. As discussed, this approach fits well with statistical applications such as the occupancy framework, is efficient for manual annotation, and has clear evaluation. (cf. [16] using the same paradigm.) It can be addressed by a wide variety of approaches.

5.1. Datasets

Our first dataset comes from a UK bird-sound crowdsourcing research spinout called Warblr.[1] From this initiative we have over 16,000 ten-second smartphone audio recordings from around the UK. The audio totals around 44 hours duration. The audio will be published by Warblr under a Creative Commons licence. The audio covers a wide distribution of UK locations and environments, and includes weather noise, traffic noise, human speech and even human bird imitations. It is directly representative of the data that is collected from a mobile crowdsourcing initiative. Annotations of the Warblr dataset are performed by a network of volunteers.

Our second dataset comes from the TREE (Transfer-Exposure-Effects) research project (www.ceh.ac.uk/TREE), which is funded by the Natural Environment Research Council (NERC), Environment Agency and Radioactive Waste Management Ltd. Dr Mike Wood's team are using unattended acoustic recorders in the Chernobyl Exclusion Zone (CEZ) to capture the Chernobyl soundscape and investigate the long-term effects of the Chernobyl accident on the local ecology. To date, the study has captured approximately 10,000 hours of audio from the CEZ. Dr Wood's team are annotating a data subset for bird species presence/absence, and approx 48—72 hours of annotated audio will be made available for the BAD Challenge. The audio covers a range of birds and includes

weather, large mammal and insect noise sampled across various CEZ environments, including abandoned village, grassland and forest areas.

5.2. Organisation and baseline results

As is typical for data challenges, we will partition the data and annotations into training, validation and testing partitions, with the testing annotations kept private for evaluation. Further, we will incentivise the development of generalisable and "tuning-free" methods by ensuring that at least one set of testing data is recorded under different conditions than the publically-available data. This helps ensure that the challenge addresses the need for methods that work with minimal manual intervention, as identified in the review we present here.

Participants will be challenged to create a system that can label the presence/absence across a diverse species range; they will not be required to identify the species. The data will be released in Summer 2016, with a deadline of late 2016 for challenge submissions. Results will be presented at a conference special session in 2017. For more detail on the timeline, please visit the challenge website.[2]

To give an indication of the general level of difficulty of this dataset, we ran a two-fold crossvalidation test using a subset of the Warblr data and a simple baseline binary classifier. We used the baseline previously created for the DCASE challenge, a generic MFCC+GMM pipeline as used for various audio tasks in the past [25]. In a previous study, this baseline system achieved 82% AUC in an auto-tagging study to detect the "birdsong" tag in audio soundscapes [26]. In the present case, the baseline attained 79% AUC, which is above the 50% chance level but with substantial headroom for the challenge. The AUC comparison hints that our dataset provides a relatively difficult test.

6. CONCLUSIONS

This survey has described current approaches to automatic bird detection in audio, including the current level of generality. Open topics include weather robustness and tuning-free methods. We have introduced a challenge giving researchers an opportunity to create a step change in these directions. A wide variety of methodological options remains open to further study, such as recent innovations in deep learning, or meta-algorithms that can automatically select detectors or combine their outputs.

7. REFERENCES

[1] T. A. Marques et al., "Estimating animal population density using passive acoustics," *Biol Reviews*, 2012.

---

[2] R. T. Buxton and I. L. Jones, "Measuring nocturnal seabird activity and status using acoustic recording devices: applications for island restoration," *J Field Ornithology*, vol. 83, no. 1, pp. 47–60, 2012.

[3] A. Digby et al., "A practical comparison of manual and autonomous methods for acoustic monitoring," *Meth Ecol Evol*, vol. 4, no. 7, pp. 675–683, 2013.

[4] A. L. Borker et al., "Vocal activity as a low cost and scalable index of seabird colony size," *Conservation Biology*, 2014.

[5] J. Sueur, A. Farina, A. Gasc, N. Pieretti, and S. Pavoine, "Acoustic indices for biodiversity assessment and landscape investigation," *Acta Acustica united with Acustica*, vol. 100, no. 4, pp. 772–781, 2014.

[6] B. J. Furnas and R. L Callas, "Using automated recorders and occupancy models to monitor common forest birds across a large geographic region," *J Wildlife Management*, vol. 79, no. 2, pp. 325–337, 2015.

[7] K. M. C. Rowe, "Automated recognition software improves detectability for a range of bird species' vocalizations," in *Int Bioacoustics Congress (IBAC)*, 2015.

[8] M. Graciarena et al., "Bird species recognition combining acoustic and sequence modeling," in *Proc ICASSP*, 2011, p. 341.

[9] D. Stowell and M. D. Plumbley, "Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning," *PeerJ*, vol. 2, pp. e488, 2014.

[10] D. I. MacKenzie et al., *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*, Elsevier/Academic Press, Burlington, MA, 2006.

[11] J. Ramírez, JM Górriz, and JC Segura, "Voice activity detection. fundamentals and speech recognition system robustness," in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds., chapter 1. 2007.

[12] G. Ferroni, R. Bonfigli, E. Principi, S. Squartini, and F. Piazza, "A deep neural network approach for voice activity detection in multi-room domestic scenarios," in *2015 Int Joint Conf on Neural Networks (IJCNN)*. IEEE, 2015, pp. 1–8.

[13] X.-L. Zhang and D. Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Trans Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 252–264, 2016.

[14] D. Stowell and D. Clayton, "Acoustic event detection for multiple overlapping similar sources," in *Proc WASPAA 2015*, 2015.

[15] L. Neal et al., "Time-frequency segmentation of bird song in noisy acoustic environments," in *Proc ICASSP*, 2011, pp. 2012–2015.

[16] M. Towsey et al., "A toolbox for animal call recognition," *Bioacoustics*, vol. 21, no. 2, pp. 107–125, 2012.

[17] P. Jančovič and M. Köküer, "Automatic detection and recognition of tonal bird sounds in noisy environments," *EURASIP J Advances in Sig Proc*, vol. 2011, no. 1, pp. 982936, 2011.

[18] S. Duan et al., "Timed probabilistic automaton: a bridge between raven and song scope for automatic species recognition," in *Proc. 25th Innovative Applications of Artificial Intelligence Conf*, 2013, pp. 1519–1524.

[19] A. L. McIlraith and H. C. Card, "Birdsong recognition using backpropagation and multivariate statistics," *IEEE Trans Sig Proc*, vol. 45, no. 11, pp. 2740–2748, 1997.

[20] S. Fagerlund, "Bird species recognition using support vector machines," *EURASIP J Applied Sig Proc*, p. 38637, 2007.

[21] S. E. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings," *J Acoustical Soc America*, vol. 100, no. 2, Part 1, pp. 1209–1219, 1996.

[22] P. Jančovič and M. Köküer, "Detection of sinusoidal signals in noise by probabilistic modelling of the spectral magnitude shape and phase continuity," in *Proc ICASSP*, 2011, pp. 517–520.

[23] J. C. Ross and P. E. Allen, "Random forest for improved analysis efficiency in passive acoustic monitoring," *Ecological Informatics*, 2013.

[24] A. G. de Oliveira et al., "Bird acoustic activity detection based on morphological filtering of the spectrogram," *Applied Acoustics*, vol. 98, pp. 34–42, 2015.

[25] D. Stowell et al., "Detection and classification of acoustic scenes and events," *IEEE Trans Multimedia*, vol. 17, no. 10, pp. 1733–1746, October 2015.

[26] D. Stowell and M. D. Plumbley, "An open dataset for research on audio field recording archives: freefield1010," in *Proc AES53*. 2014, Audio Engineering Society.