BIOLOGICAL REVIEWS

Cambridge Philosophical Society

Automatic detection for bioacoustic research: a practical guide to the state of the art and future directions

Journal:	*Biological Reviews*
Manuscript ID	Draft
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Kershenbaum, Arik; University of Cambridge, Girton College Akçay, Çağlar; Anglia Ruskin University, Behavioural Ecology Research Group Babu-Saheer, Lakshmi ; Anglia Ruskin University, Computing Informatics and Applications Research Group Barnhill, Alex; Friedrich-Alexander-Universität Erlangen-Nürnberg, Pattern Recognition Lab, Department of Computer Science Best, Paul; Université de Toulon, Aix Marseille Univ Cauzinille, Jules; Université de Toulon, Aix Marseille Univ Clink, Dena; Cornell Lab of Ornithology, K. Lisa Yang Center for Conservation Bioacoustics Dassow, Angela; Carthage College, Biology Dufourq, Emmanuel; African Institute for Mathematical Sciences; Stellenbosch University, Mathematical Sciences Growcott, Jonathan; University of Exeter - Cornwall Campus, Centre of Ecology and Conservation, College of Life and Environmental Sciences; Recanati-Kaplan Centre, Wildlife Conservation Research Unit Markham, Andrew; University of Oxford, Computer Science Marti-Domken, Barbara; University of Oviedo Marxer, Ricard; Université de Toulon, Aix Marseille Univ Muir, Jen; Anglia Ruskin University, Behavioural Ecology Research Group Reynolds, Sam; Anglia Ruskin University, Behavioural Ecology Research Group Root-Gutteridge, Holly; University of Lincoln, Dept of Life Sciences, College of Health and Science Sadhukhan, Sougata ; Bharati Vidyapeeth (Deemed to be University) Schindler, Loretta; Charles University, Zoology Smith, Bethany; Zoological Society of London Institute of Zoology Stowell, Dan; Tilburg University; Naturalis Biodiversity Center Wascher, Claudia; Anglia Ruskin University, Behavioural Ecology Research Group Dunn, Jacob; Anglia Ruskin University, Behavioural Ecology Research Group; University of Cambridge, Archaeology; University of Vienna, Behavioral and Cognitive Biology
Keywords:	artificial intelligence, bioacoustics, machine learning, passive acoustic monitoring, acoustic monitoring, animal communication, automatic detection, classification, deep learning, neural networks

Automatic detection for bioacoustic research: a practical guide to the state of the art and future directions

Arik Kershenbaum[1], Çağlar Akçay[2], Lakshmi Babu-Saheer[3], Alex Barnhill[4], Paul Best[5], Jules Cauzinille[5], Dena Clink[6], Angela Dassow[7], Emmanuel Dufourq[8,9,10], Jonathan Growcott[11,12], Andrew Markham[13], Barbara Marti-Domken[14], Ricard Marxer[5], Jen Muir[2], Sam Reynolds[2], Holly Root-Gutteridge[15], Sougata Sadhukhan[16], Loretta Schindler[17], Bethany R. Smith[18], Dan Stowell[19,20], Claudia A.F. Wascher[2], Jacob C. Dunn[2,21,22*]

[1]Girton College and Department of Zoology, University of Cambridge, Cambridge, CB3 0JG, United Kingdom

[2]Behavioural Ecology Research Group, Anglia Ruskin University, East Road, Cambridge, CB1 1PT, United Kingdom

[3]Computing Informatics and Applications Research Group, Anglia Ruskin University, East Road, Cambridge, CB1 1PT, United Kingdom

[4]Pattern Recognition Lab, Department of Computer Science, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058, Erlangen, Germany

[5]Université de Toulon, Aix Marseille Univ, CNRS, LIS, ILCB, Toulon, France

[6]K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University, Ithaca, NY, USA

[7]Biology Department, Carthage College, Kenosha, Wisconsin, United States

[8]African Institute for Mathematical Sciences, South Africa

[9]Department of Mathematical Sciences, Stellenbosch University, South Africa

[10]National Institute for Theoretical and Computational Sciences, South Africa

[11]Centre of Ecology and Conservation, College of Life and Environmental Sciences, University of Exeter, Cornwall Campus, TR10 9FE, United Kingdom

[12]Wildlife Conservation Research Unit, Recanati-Kaplan Centre, Tubney House, Abingdon Road Tubney, Abingdon OX13 5QL, United Kingdom

[13]Department of Computer Science, University of Oxford, Oxford, OX1 3QD, United Kingdom

[14]University of Oviedo, Principality of Asturias, Spain

[15]Dept of Life Sciences, College of Health and Science, University of Lincoln, Brayford Pool, Lincoln, LN5 7TS

[16]Institute of Environment Education and Research, Bharati Vidyapeeth Deemed University, Pune, India

[17]Department of Zoology, Faculty of Science, Charles University, Prague 128 44, Czech Republic

[18]Institute of Zoology, Zoological Society of London, London NW1 4RY, UK

[19]Tilburg University, Tilburg, The Netherlands

[20]Naturalis Biodiversity Center, Leiden, The Netherlands

[21]Department of Archaeology, University of Cambridge, Cambridge, CB2 3DZ, United Kingdom

[22]Department of Behavioral and Cognitive Biology, University of Vienna, Vienna, Austria

*Correspondence: Dr Jacob C. Dunn, Behavioural Ecology Research Group, School of Life Sciences, Anglia Ruskin University, Cambridge, CB1 1PT, United Kingdom; Telephone: +44 1223 698220; Email: jacob.dunn@aru.ac.uk

ABSTRACT

Recent years have seen a dramatic rise in the use of passive acoustic monitoring (PAM) for biological and ecological applications, and a corresponding increase in the volume of data generated. However, datasets are often becoming so sizable that analysing them manually is burdensome and unrealistic. Fortunately, we have also seen a corresponding rise in computing power and the capability of machine learning algorithms, which offer the possibility of performing some of the analysis required for PAM automatically. Nonetheless, the field of automatic detection of acoustic events is still in its infancy in biology and ecology. In this review, we examine the trends in bioacoustic PAM applications, and their implications for the burgeoning amount of data that needs to be analysed. We explore the different methods of machine learning and other tools for scanning, analysing, and extracting acoustic events automatically from large volumes of recordings. We then provide a step-by-step practical guide for using automatic detection in bioacoustics.

One of the biggest challenges to greater use of automatic detection in bioacoustics is that there is often a gulf in expertise between the biological sciences and the field of machine learning and computer science. Therefore, this review first presents an overview of the requirements for automatic detection in bioacoustics, intended to familiarise those from a computer science background with the needs of the bioacoustics community, followed by an introduction to the key elements of machine learning and artificial intelligence that a biologist needs to understand to incorporate automatic detection into their research. We then provide a practical guide to building an automatic detection pipeline for bioacoustic data, and conclude with a discussion of possible future directions in the field.

KEYWORDS

Animal communication, Automatic detection, Artificial intelligence, Bioacoustics, Classification, Deep learning, Machine learning, Neural networks, Passive acoustic monitoring

CONTENTS

1. INTRODUCTION

1.1. Acoustic monitoring

The acoustic monitoring of captive and wild animals provides valuable data for many purposes, including scientific research, conservation efforts, management decisions, and the welfare of individual animals. Acoustic data can be collected using handheld microphones, on-animal devices, or autonomous recording units (ARUs) placed in the field. Such data can be collected over periods of time ranging from short, opportunistic, recordings, to long-term deployments lasting months or years. The use of handheld microphones and ARUs are non-invasive methods that do not require the capture of individual animals, and so reduce disturbance and welfare impacts (Browning *et al.*, 2017; Soulsbury *et al.*, 2020; Ross *et al.*, 2023). Acoustic data can help with monitoring of elusive, cryptic, or nocturnal species that are difficult to observe directly (Zwerts *et al.*, 2021), e.g., bats (Frick, 2013), wolves (Harrington & Mech, 1982; Kershenbaum, Owens & Waller, 2019), or cetaceans (Zimmer, 2011). Additionally, where animals use long-distance vocalisations, ARUs are beneficial in recording species over large spatial scales, e.g., crested argus pheasants (Vu *et al.*, 2023), gibbons (Vu & Tran, 2019; Dufourq *et al.*, 2021), howler monkeys (Pérez-Granados & Schuchmann, 2021), and wolves (Kershenbaum *et al.*, 2019). Such methods can offer detection ranges in the order of several kilometres for some species, compared with tens of metres for camera traps. However, as a passive technique, the obvious disadvantage of acoustic monitoring is that the animal needs to be producing sound to be detected.

Whilst the collection of acoustic data offers many benefits and opportunities, it brings with it certain challenges. First, the deployment and servicing of ARUs (e.g., replacing batteries and memory storage cards) can be costly in terms of time and labour (Metcalf *et al.*, 2023b). Second, although the tools for acoustic monitoring are now more widely available, cheaper in cost, and include larger storage capacities and longer battery life (Hill *et al.*, 2019), this has led to a very large increase in the quantity of data being stored, transferred, and analysed. Third, a major challenge is distinguishing the sound(s) of interest from background sounds which takes an enormous amount of researcher time, effort, and expertise, to recognise the calls of species accurately and annotate the recordings reliably. All of this creates long delays between data collection and the final results of a study, yet the need for real time results can be pressing, especially in the field of conservation biology. Automatic detection can solve many of these issues, as a tool to extract sounds of interest automatically, reducing or even eliminating the need for manual analysis of the data.

1.1.1. What is automatic detection?

Automatic detection is the process of extracting acoustic signals from sound recordings automatically, without human effort. Once detected, numerous properties of the acoustic signal can be determined (with or without additional human effort). For example, the acoustic signal could be classified as being produced by a particular species, its location determined, and the identity of the animal inferred. The temporal and spectral properties (e.g., fundamental frequency, harmonics, modulation, etc) of the acoustic signal can be calculated and used for additional processing or for inferring additional information about the sound production. Some approaches implicitly combine the processes of automatic detection with other tasks e.g., classification of the vocalising species, but fundamentally, the first step within an automated bioacoustic processing pipeline is detection.

Throughout this paper, we will use the term "acoustic signal" to describe any sound or acoustic event produced by an animal without regard to the purpose or intentionality of the signal.

1.2. Scope of review

In this paper we set out to highlight and describe the emerging field of automatic detection of acoustic signals as a highly interdisciplinary effort that requires expertise from both biological and computer science to move forward. We present a review and tutorial that addresses both the needs of the community of biologists using acoustic monitoring to answer ecological, evolutionary, and conservation research questions, and the needs of computer scientists developing new algorithms and implementations. As the overlap between these two needs and the overlap between domain knowledge of these two groups is often small, this review attempts to bridge that gap by addressing both groups simultaneously, enhancing the missing knowledge of both. A reader from either field will find this review to be a useful integration of both domains, providing new information to both without being inaccessible to either. The review arose from an investigative workshop held in July 2023 at Girton College, University of Cambridge, attended by 22 scientists from both the biological and computer sciences.

By way of introduction, the review first presents the perspectives on automatic detection for bioacoustics from the point of view of a biological researcher, aiming to instruct the computer scientist in the needs of the end-user. Then, we present the perspective of the computer scientist, aiming to instruct the biologist in the technologies available and their limitations. There then follows a step-by-step guide to the practical implementation of automatic detection, and finally a discussion of the potential future directions of the field.

205 2. BACKGROUND OF AUTOMATIC DETECTION IN BIOACOUSTICS

206 2.1. What is automatic detection and why do we need it?

207 To address the challenge of converting terabytes of acoustic recordings into useful
208 information, scientists have sought to develop techniques to automate the detection of
209 acoustic signals of interest. The traditional method of identifying the signals of interest from
210 longer acoustic recordings was to create a spectrogram and manually draw bounding boxes
211 around the signals of interest, incurring a significant cost in terms of time and expertise.
212 Fundamentally, the challenge is to replace the human annotator with computational methods
213 without a consequent loss in accuracy (Miller *et al.*, 2023). At its simplest, the aim of
214 automatic detection is to indicate segments or windows of audio which are likely to contain a
215 target sound of interest, substantially reducing the burden, even if the automated annotations
216 need then be checked by a human. The annotation label can simply be a binary label of
217 presence/absence of a sound, but this can also be further refined to classify by taxon, call-
218 type, number of individuals, etc., in increasing levels of precision and consequent difficulty
219 for both annotator and algorithm. For many species, it can be possible to identify an
220 individual through its unique vocal characteristics (Petso, Jamisola & Mpoeleng, 2021). In
221 addition to the class label, some systems also allow the position or bearing of the sound to be
222 estimated (Kershenbaum *et al.*, 2019; Smith *et al.*, 2021). Such information can then be used
223 in numerous downstream tasks such as occupancy monitoring, spatial habitat use, and
224 behavioural analysis, and automatic detection offers researchers the opportunity to scale to
225 larger spatiotemporal datasets.

226 2.2. The current state of the art in automatic detection

227 The use of automatic detection to accelerate acoustic monitoring has a long history (Acevedo
228 *et al.*, 2009; Aide *et al.*, 2013; Dufourq *et al.*, 2021; Oswald *et al.*, 2022). As an early

approach towards automation, simple techniques based on the energy within a particular frequency range, characteristic to the target sound, have been used to detect signals of interest (Morrissey *et al.*, 2006). However, this approach only works if the signal-to-noise ratio of the target sound is sufficiently high, and if other sounds are not present in the same frequency range which act to mask it. Subsequent techniques have used statistical modelling or classical machine learning models such as hidden Markov models (HMMs) to detect calls that are modulated in frequency and/or time (Duan *et al.*, 2013; Oswald *et al.*, 2022), by identifying properties of the target sound beyond simply frequency range. Such models can provide more robust and sensitive detections. More recently, there has been a strong push towards the use of data-driven machine learning, exemplified by deep learning (DL), using techniques such as convolutional neural networks (CNNs) (LeCun, Bengio & Hinton, 2015), recurrent neural networks (RNNs) (Yu *et al.*, 2019) and more recently transformers (Lin *et al.*, 2022). Transformers have been shown to obtain impressive detection accuracies, e.g. BirdNET (Kahl *et al.*, 2021), and the BTO Acoustic Pipeline (Anon., 2023b).

There is, however, a highly fragmented landscape in the field of automatic detection – in particular between the fields of computer science/machine learning, and bioacoustics/acoustic ecology – and it can be very challenging for practitioners to know where to get started. Should one build their own classifier from scratch, fine-tune an existing model, or simply use an off-the-shelf pretrained model (Stowell, 2022a; Dufourq *et al.*, 2022b)? Good quality detectors already exist in a relatively user-friendly format for birds, e.g. BirdNet (Kahl *et al.*, 2021); bats, e.g. BTO Acoustic Pipeline (Anon., 2023b), Kaleidoscope (Anon., 2023a); cetaceans, e.g. PAMGuard (Gillespie *et al.*, 2009); rodents, e.g. DeepSqueak (Coffey, Marx & Neumaier, 2019), MUPET (Van Segbroeck *et al.*, 2017). However, these detectors tend to be known only by those using them in the field and are not straightforward to generalise to other taxa without retraining or altering the model architecture or assumptions. There is also an imbalance with some taxa being better represented than others in terms of the availability of detectors. The process of building or fine-tuning a new deep learning model for a practitioner's particular habitat and species of interest is non-trivial and involves several tasks such as cloning repositories from Github, designing data-loaders, and training models on specialised computing hardware such as Graphical Processing Unit (GPU) clusters. This serves as a major barrier to widespread adoption of these new techniques unless a tame computer scientist can be persuaded to assist in the process. In contrast, the more mature field of automatic detection in camera trapping, e.g. WildlifeInsights, CameraTrapDetectorR (Hendry & Mann, 2018); Camelot (Hendry & Mann, 2018); Agouti (Casaer *et al.*, 2019); MegaDetector, can serve as an exemplar for deriving best practices, as existing tools are easy to use for non-programmers, and easily generalised to different taxa.

2.3. What do we aspire for from automatic detection?

Despite the challenges associated with the automatic detection of acoustic signals, rapid advances in machine learning are starting to bring this concept into reality. Although the context under which acoustic data is recorded and its end use will differ, the common requirement is for algorithms that take acoustic data as an input, and then detect and return extracted sounds as the output. Some users may only require outputs of particular target sounds, such as a specific species or anthropogenic sound, whereas others may require all sounds to be classified. Ideally, the ultimate end goal of automatic detection for biologists would be a universal, off-the-shelf algorithm capable of detecting and classifying all animal vocalisations such that anybody, including those without any training in computational methods, could process their acoustic data more efficiently and flexibly tailor it to their particular use-case (Romero-Mujalli *et al.*, 2021). Where an off-the-shelf detector for a sound of interest is not readily available, algorithms that are easy to train with a relatively small amount of data and minimal annotation effort should be the aim.

3. PERSPECTIVES FROM BIOLOGICAL SCIENCES

In this section, we give, largely for the benefit of the reader from a computer science or other non-biological background, an overview of the possible roles for bioacoustics in addressing several important evolutionary, ecological, and conservation questions, highlighting the potential benefit that automatic detection can provide.

3.1. Overview of uses of automatic detection in the biological sciences

Detecting acoustically active animals through their acoustic signals can provide a wealth of information that is important to conservation biology, ecology, evolutionary biology, animal behaviour, and welfare (Mcloughlin, Stewart & McElligott, 2019; Odom *et al.*, 2021; Erbe & Thomas, 2022). Often, these areas of study can overlap: animals can produce sounds to influence the behaviour of others in a wide range of contexts, e.g., to attract a mate or warn off an intruder, or as a by-product of other behaviours, e.g., the sound of wings flapping or footsteps.

Historically, conservation efforts and biodiversity surveys have been skewed towards species that are easy to trap or track across the landscape, often depending on direct observation or finding physical traces like scat or hair (Boakes *et al.*, 2010). However, the field of bioacoustics allows us to survey remote or otherwise inaccessible areas, e.g., deep sea environments, arctic and antarctic regions, and rainforests (Staaterman *et al.*, 2017), with research often focusing on the loud and persistent calls of target species to detect their presence. Like camera trapping, bioacoustics generates large datasets which challenge analysis, but, unlike camera traps, the same event can be recorded in multiple places, multiplying the data to be assessed and analysed.

Below, we provide a broad review of the use of acoustic data in the biological and ecological sciences, from measures of biodiversity at geographic scales to tracking the movements and behaviours of individual animals, and highlight how automatic detection can increase the efficiency and efficacy of monitoring.

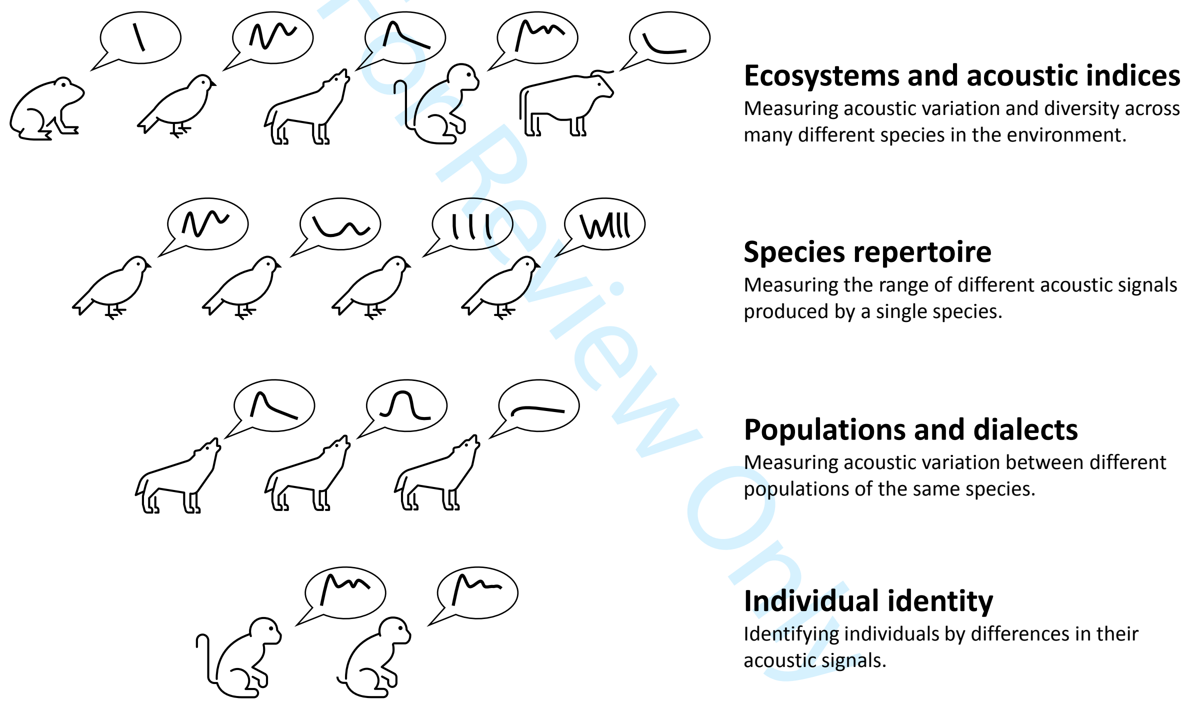

Figure 1: Hierarchy of acoustic signal specificity

3.1.1. Ecosystems and acoustic indices

Any multi-species soundscape will consist of a wide range of frequencies being used by different species in the same environment (Krause, 1993). To maximise the chance that their signal will be detected, animals usually avoid acoustic signal interference by vocalising in different frequency ranges or at different times, as described by the acoustic adaptation hypothesis (Hansen, 1979; Rothstein & Fleischer, 1987). This ecological phenomenon makes it possible to detect particular clades or species. It also means that estimates of biodiversity can be made based on the number of different acoustic signals being produced at different times/frequencies.

Acoustic indices provide a quantitative measure of acoustic complexity by analysing variation in the frequency and timing of acoustic signals, rather than identifying individual sounds. Such indices offer metrics for wildlife monitoring and assessment, characterising biological communities through sound (Sueur *et al.*, 2014; Buxton *et al.*, 2018). While acoustic indices are informative about the acoustic complexity or general biodiversity of a landscape, they are less useful for deriving specific information about species or the individuals vocalising.

Acoustic indices typically do not use automatic detection and classification of acoustic signals, as, by their nature, they characterise the soundscape as a whole. However, automatic detection of sound classes, for example distinguishing acoustic signals of anthropogenic origin from those of biological origin, can improve the ability of acoustic indices to provide useful indications of biological activity (Narasimhan, Fern & Raich, 2017; Fairbrass *et al.*, 2019; Clark *et al.*, 2023). Thus, effective automatic classification of acoustic signals may become an important element of improving acoustic indices in future research.

3.1.2. Species occupancy and density

Occupancy modelling is the statistical analysis of the patterns and dynamics of a species in a given space over time (MacKenzie *et al.*, 2003), which can be informed by acoustic signals (Wood & Peery, 2022; Cole *et al.*, 2022). Bioacoustic occupancy monitoring can provide critical information on the presence and absence of species and the dynamics of the ecosystem, particularly for cryptic or elusive species.

Population density estimates model a species' abundance within a defined space. Density estimates are an extremely important tool for assessing spatiotemporal population changes that can be the result of declining prey numbers, land-use change, human-wildlife conflict (Wolf & Ripple, 2016; Ogutu *et al.*, 2016; Rostro-García *et al.*, 2023), or other factors, and bioacoustics data can provide an important tool for estimating the densities of animal populations.

3.1.3. Spatial analyses

Population surveys and behavioural investigations often need to be able to determine the location and/or movement patterns of animals. Bioacoustic surveys have been used in more recent years to supplement or replace previous tracking methods (Frommolt & Tauchert, 2014). For example, the tracking of migratory species across their extensive ranges, where radio/satellite telemetry is only useful if the individuals tagged with a transmitter survive what may be a high mortality journey, can benefit from the application of bioacoustic techniques. While telemetry is an effective method for learning about a species' movement, it can also be highly invasive, can affect the behaviour of individuals being trapped, and is not always suitable for all species/age groups, e.g. species that are too small to carry the weight of a transmitter, or species in remote areas (Sharpe *et al.*, 2009).

Depending on the intended research goals, it may be sufficient simply to detect the presence/absence of an animal within a recorder's range (macro-localisation), or one may need to infer the exact position of an individual (micro-localisation). There are benefits and limitations to each: macro-localisation can inform on occupancy, habitat suitability, territory use, and migratory patterns. On the other hand, with a significant increase in the complexity of use, advanced tools also allow a more targeted approach such as multilateration, where the exact individual's location is calculated based on the time difference of arrival (TDOA) of an acoustic signal to multiple recording devices (Mennill *et al.*, 2012). Such micro-localisation allows avoiding double counting for density estimates, can inform on animal movement speed and direction, as well as providing fine grained territory boundaries, but requires additional downstream processing to carry out the localisation analysis.

Estimation of a focal animal's home range and territory provides wildlife managers with a boundary for their activity (Powell, 2000), permitting the study of intraspecific dynamics and spatial distribution of individuals across a landscape (Burgos & Zuberogoitia, 2020), which can be important for conservation action. Real-time automatic detection combined with localisation reduces the research effort required for follow-up visual observation and can obviate the need for visual observation entirely.

3.1.4. Species characteristics

Automatic detection of acoustic signals is complicated by the fact that there are relatively few species that, like the American toad (*Anaryxus americanus*), produce a single call (Bee, 2012), while many species produce multiple call-types, e.g., the northern mockingbird (*Mimus polyglottos*) produces hundreds of different song types (Derrickson, 1988). Thus, while it is relatively easy to link a croak to the presence of a toad, it can be more challenging to capture all the potential acoustic signals of the mockingbird. This is further complicated if the species' different call-types need to be classified beyond simple detection (e.g., as contact calls vs alarm calls).

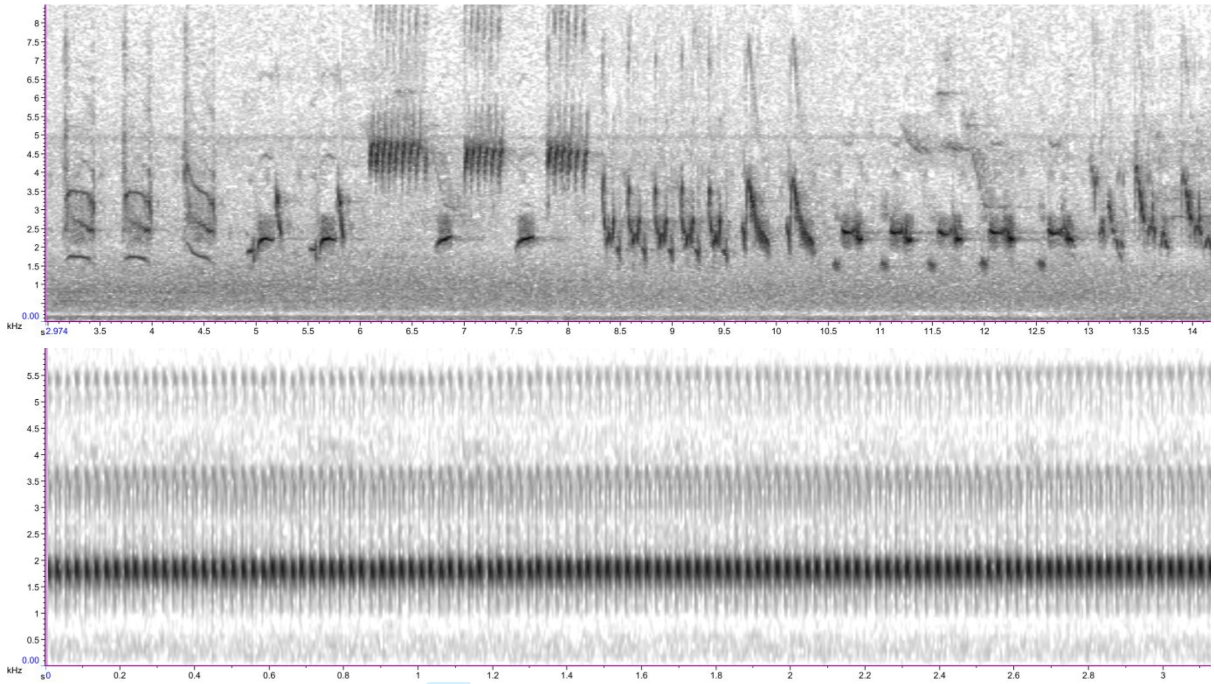

Figure 2. Top: the varied mimicry of the northern mockingbird (*Mimus polyglottus*), composed of varied songs of other species, which would be difficult to detect in a general way. Bottom: the call of the American toad (*Anaxyrus americanus*), which only produces a single call, repeated for long periods. The spectrograms show time on the x-axis and frequency on the y-axis.

Collectively, all the distinct call types a species produces can be defined as the vocal repertoire. The size of the repertoire may be thought of as a simple proxy for vocal complexity (Bouchet, Blois-Heulin & Lemasson, 2013; Manser *et al.*, 2014), and the structure of the repertoire (e.g., how often call-types are used and interpretations of the potential uses) are important for describing a species' behavioural ecology. Therefore, both general acoustic signal detection ("the target species made a sound in some way") and specific call-type detection ("the leopard-specific alarm call has been produced") are useful to different studies and these analyses can be nested. Comparisons of vocal complexity between species, species groups, and taxa (Kershenbaum *et al.*, 2021; Leighton & Birmingham, 2021) may enable research into broad evolutionary or ecological questions, such as cognitive abilities, adaptive advantages of cognitive skills, or the evolution of language (McComb & Semple, 2005; Dunn & Smaers, 2018).

The more varied and less stereotyped calls are, the larger the challenge to automatic detection. However, the implications of variability within a single call type on the performance of automatic detection and classification have not been adequately investigated.

3.1.5. Populations and social groups

The same species can show variation in their vocalisations among social groups and or across geographic regions. Research into these differences can offer unique insight into either phylogenetic patterns, including speciation (Meyer *et al.*, 2012; Riesch *et al.*, 2012; Heaphy & Cain, 2021), historic geographical patterns (Laiolo *et al.*, 2001; Kershenbaum *et al.*, 2012; Hebets *et al.*, 2021), or differences between social groups (Ford, 1991; Velásquez *et al.*, 2013; Garland, Castellote & Berchok, 2015; Kershenbaum *et al.*, 2016b). Automatic detection can scan through long-term recordings to unveil temporal and cultural variations of vocal behaviours, for example in whales (McDonald, Hildebrand & Mesnick, 2009; Garland *et al.*, 2011; Best *et al.*, 2022).

3.1.6. Individual characteristics

It may be important to identify individual animals and/or characterise the traits or states of individuals of a target species, such as age, sex, body size, emotional valence/arousal, and physiology. Acoustic signals can potentially encode all of this information. Examples of the benefits of individual identification include gaining insights into the evolution and ecology of a species, such as life history stages and social structure (Clutton-Brock & Sheldon, 2010); facilitating conservation efforts, for example tracking movement of critically endangered species in the landscape (Mcloughlin *et al.*, 2019); and improving management in captivity, for example measuring vocal activity as an indicator of welfare in zoo housed animals

(Castellote & Fossa, 2006). A diverse range of species' calls have been found to encode individual identity from birds (Fox, Roberts & Bennamoun, 2008; Martin *et al.*, 2022) to cattle (Green *et al.*, 2019), cetaceans (Kershenbaum, Sayigh & Janik, 2013; Bøttcher *et al.*, 2018), and frogs (Qian *et al.*, 2023).

Individual identification provides an open scope for spatiotemporal monitoring of species without tagging (Aide *et al.*, 2013), while also offering the opportunity for population estimation using mark-capture recapture methods, which rely on individual identification (Marques *et al.*, 2013b; Buxton *et al.*, 2018).

Acoustic signals can be used in a wide range of species to assess the intensity (high to low) and valence (positive to negative) of emotional arousal of animals, which in turn can be used as an estimate of welfare in animals in captivity (Volodina & Volodin, 1999; Clark & Dunn, 2022) and farms (Manteuffel, Puppe & Schön, 2004). Inferring emotional arousal from acoustic signals also allows for the assessment of 'positive welfare' in animals (Laurijs *et al.*, 2021), and it is possible to monitor farm animals for the onset of disease, e.g., pigs (*Sus domesticus*) (Exadaktylos *et al.*, 2008; Mcloughlin *et al.*, 2019) and chickens (*Gallus domesticus*) (Mao *et al.*, 2022).

3.2. Key challenges

As outlined above, many studies in ecology and evolution require relatively precise identification of the type of acoustic signal, for example different call types, the source of the sound, individual identification, or the localisation of the source of the sound in space. Despite the huge potential of automatic detection to answer these challenges, the field is still facing significant barriers during implementation in biological studies, ranging from limitation in infrastructure, lack of training, inaccessibility of methods, and practical limitations in the field. For example, field recordings are often not of optimal recording quality and have a low signal to noise ratio. Even under ideal conditions, acoustic signals themselves may be highly varied and irregular, with low stereotypy and a high degree of variability between individuals and groups, or geographical dialects (Nelson, 2000), all of which can present a challenge for automatic detection. The broad implementation of automatic detection requires that the model is robust to the variation presented.

The training of models requires data to be robustly identified and correctly attributed to the study species or individuals, often produced by visual observation of the callers. Collecting these data can be challenging as, for instance, individuals may remain visually cryptic, or call only at certain times. Thus, ground-truthing data requires high quality, reliably identified call datasets which can be difficult to obtain but are essential. Furthermore, generalising data from captive animals or in unique circumstances might give rise to misleading results. Thus, robust identification of large datasets is rare but essential and should be a focus for future research.

4. TECHNICAL PERSPECTIVES

4.1. Perspectives from computer science

4.1.1. The role of computation in automatic detection

Advanced computational methods can provide solutions to a wide range of bioacoustics problems. For example, acoustic signals of interest can be merely detected (i.e., the start and end times identified), or additional information can be extracted, such as classification of signal type, or location of the sound source. If different types of acoustic signal are present, they can be grouped into multiple classes, which might represent different species, or different call types within a single species. Even when a single type of acoustic signal is present, the task of counting the number of such events or sub-elements of the events is often non-trivial (e.g., the different notes in a birdsong, or the individual barks of a dog). Therefore, the role of automatic detection and automatic processing of bioacoustic data is a broad field, with many possible applications.

Computational methods can help with any task which can be clearly defined. One way to define the task is through explicit rules (an engineering approach), for example, to specify that a target acoustic signal occurs solely and uniquely in a certain range of frequencies. Alternatively, a set of examples can be provided to the algorithm (a machine-learning approach), and the algorithm is trained to generalise those examples to detect successfully when presented with novel examples. In the case of automatic detection, some tasks are simple enough that a good method can be designed directly using the engineering approach: this typically happens with situations of highly-stereotyped sounds, where template-matching often works well (Barker, Herrera & West, 2014), or low-noise environments with few interfering sounds, where energy-detection may work well (Hood, Flogeras & Theriault, 2016).

When the target sounds, or the background, are more complex—such as with recordings of elaborate bird song or soundscapes with high levels of anthropogenic noise—then machine learning (ML) is of benefit. As noisy problems can rarely be defined in a clear-cut "engineering" way, ML attempts to reach a solution by generalising from a set of examples instead. Although ML has been investigated for many years (Towsey *et al.*, 2012), it is the era of deep learning that now makes many bioacoustic detection tasks achievable (Stowell, 2022b). It is still important to define the task to be solved clearly – by curating good datasets for training and evaluating systems, and by specifying the input and output data formats. Input data format, in bioacoustic applications, is generally some representation of the sounds recorded, whereas the output format is defined by the nature of the "answer" that the system is trained to supply, e.g., species presence, individual, call type, etc.

Data curation aside, the power of ML comes from having techniques that can "train" (optimise) the system to achieve a particular goal, and so the output data format matters because it is closely tied to this procedure of optimisation. If the output format is a yes/no answer about species presence, this is the same format as a *binary classification* task in ML and can be addressed directly by training a classifier (Stowell, 2022b), which takes sound as input, and outputs a corresponding indicator: present/absent. Very often, however, the output format wanted is more complex; for instance, given a long audio recording as input, we may want to output a list of (predicted) events giving each event's start and end time, and optionally its frequency range as well. Note that this is quite similar to "object detection" in image recognition, and indeed, most bioacoustic research uses spectrograms as a visual representation of a sound, rather than working with the sound directly. In this case, we may typically be looking for a list of "bounding boxes" along the time axis or in time-frequency, leading some to directly adapt image object detection algorithms to spectrograms (Kershenbaum & Roch, 2013; Venkatesh, Moffat & Miranda, 2022; Wu *et al.*, 2022). When a ML model has been trained, better results may be obtained if the model is applied in the same conditions as the training data, i.e., "in-domain" as opposed to "out-of-domain" data (Best *et al.*, 2020). For example, conditions might be "in-domain" if they have the same background conditions, microphone type, and sampling protocol as in the training data.

4.1.2. State of the art in automatic detection methods

No algorithm will generalise perfectly to all situations: the choice of training data represents the choice of intended domain. Classic machine learning advice would be to avoid "out-of-domain" situations. Yet many taxa do not benefit from such a large amount of prior work as on birds. Could we nevertheless make use of off-the-shelf models from similar tasks, or must we start building a large new dataset?

Happily, a recent widespread trend is "transfer learning", in which one or more pretrained models are used that have been trained on tasks that are different from (but usually related to) the original domain: for example, we could consider models trained on human speech recognition. The models are then re-used for the current application (i.e., animal vocalisations), and it is often found that the original learning makes training the model on the current data more effective (Zhuang *et al.*, 2021).

A common approach to transfer learning, known as fine-tuning, consists of modifying only a small subset of parameters and adapting the inputs and/or outputs format. The modification requires training the model on a new set of examples, made up of audio recordings and corresponding annotations. This procedure is computationally much lighter than performing the process from scratch. It also requires fewer labels since it exploits many of the regularities in the initial data set. As a rule of thumb, one may try to choose a base model which has been trained on similar target sounds or background noise, e.g. BirdNet (Kahl *et al.*, 2021). Yet we have observed successful attempts in adapting models from significantly different acoustic data, even from different frequency ranges (Çoban *et al.*, 2020; Sethi *et al.*, 2020; Leroux *et al.*, 2021; Sarkar & -Doss, 2023).

When using transfer learning (also known as "pretrained" models), special care must be taken. The model must be applied to acoustic data that closely resemble the data on which it has been trained. The user must reflect on details such as matching sampling rates, normalisations, SNR-levels, and duration of the input audio segments. Usually, the producers of such models will have trained models on diverse data to ensure generalisation. However, optimal performance is achieved when staying within the region of operation for which the model was designed.

In this paradigm, the algorithm trained on a different system can be considered to perform a role similar to the role of the spectrographic representation in aiding human interpretation of sounds. In the same way that a spectrogram or filterbank takes a sound waveform and presents it in a different format (and one where the important features are easy to detect by eye), so a model trained on a different species, for example, cannot detect the target species well, but may nonetheless produce an output (known as extracted acoustic features) that can be used as the input to train another model, which will then be more successful in finding the focal species. In the ML literature the resulting features are often referred to as embeddings or latent representations. Unlike traditional acoustic features like a spectrogram these embeddings are often difficult to interpret on their own. They are the result of a large composition of complex functions whose parameters have been optimised to solve a particular task such as classifying an acoustic scene or discriminating from a given set of videos the one that matches a particular sound.

Despite this, fine-tuning alone may not be sufficient to obtain the desirable level of accuracy. We may then further adapt the model to our specific needs by retraining all of its parameters on the acoustic data of interest. One must take into consideration that these models have been designed with a large number of parameters (317 million parameters for the large version of HuBERT for instance; (Hsu *et al.*, 2021), to be optimised on thousands of hours of audio. When trained on a small number of examples this may quickly lead to overfitting, where the model will work as expected on the data presented during training but will fail to produce satisfactory predictions for unseen audio examples.

Even when many hours of field recordings are available, it is not clear if the acoustic data will be sufficiently diverse to produce acoustic features that will be performant enough for downstream tasks such as the detection of vocalisations. In other words, if a bioacoustic dataset does not contain any useful (or additional) information which could be reemployed in the downstream detection tasks, then retraining the pre-trained model might not improve performance. Furthermore, re-training these models on large amounts of data is usually a tedious task which calls for the expertise of trained computer scientists and access to costly computational resources such as GPU clusters.

The approach of adapting transfer learning models to automatic bioacoustic detection, can still be carried out by pretraining models on bioacoustic data directly, instead of human speech or generic audio. It has been shown to yield interesting results in the downstream detection performances for a variety of species (Hagiwara, 2022), but much work still needs to be done in this area. The success of this method relies on the availability of large datasets which could allow for the pretraining of a single, large-scale, multispecies foundation model. As is the case in the speech processing and image recognition domains, making such a model available to the bioacoustic community could then allow for efficient user-friendly classifiers to be trained in few-shot learning contexts within a unified pipeline.

An alternative to the transfer learning approaches is to use smaller models, with fewer parameters that may be trained entirely on the target audio data. For example, an algorithm called TweetyNet is designed for detecting/segmenting bird vocalisations in a laboratory context, based on a CNN to be trained specifically for each target bird; the package includes an interface to simplify that training process (Cohen *et al.*, 2022); DeepSqueak can do the same for rodent vocalisations (Coffey *et al.*, 2019). Those algorithms directly train the CNN as a classifier/detector. Another approach used by many in the bioacoustics community is to train a so-called 'auto-encoder' on the dataset of interest to extract deep feature representations from unlabelled data. This unsupervised approach consists in optimising a neural network to compress an audio snippet into a numerical vector which is decompressed to reconstruct the original sound. This technique has been applied to call categorisation in a variety of species (Sainburg, Thielk & Gentner, 2020; Best *et al.*, 2023).

Even using such methods, it is common that bioacoustic datasets are not large enough to train an ML detector well, or that some categories/contexts are underrepresented in the training

595 data. It is thus common (and recommended) to use ‘data augmentation’ to assist with this:
596 ‘new’ training examples can be created by small modifications of existing ones. This has
597 been widely investigated and found to improve performance, to a similar extent as the use of
598 pretrained networks (Lostanlen *et al.*, 2018).

599 The bioacoustics community often faces complex scenarios with sound events overlapping
600 both in time and frequency (e.g., dawn chorus) or with highly non-stationary background
601 noise (e.g., urban scenes). These require more advanced and specific solutions that tackle the
602 problem of working with mixtures of sounds. Data-augmentation techniques serve this
603 purpose by artificially constructing similar data for which annotations can be created by
604 design (Jansson *et al.*, 2017; Zhang *et al.*, 2018; Wisdom *et al.*, 2020). These approaches
605 have been applied to improve performance on up to ten simultaneously-calling bird species in
606 a simulation study (Parrilla & Stowell, 2022) and in real recordings with significantly fewer
607 simultaneous calls (Denton, Wisdom & Hershey, 2021; Bermant, 2021).

608 4.1.3. Assessing pre-existing models

609 The fast pace at which the ML community publishes new pretrained models renders them
610 outdated quickly. The availability of accessible learning resources for some models makes
611 them a go-to solution for many practitioners, despite having been superseded by other
612 options. Model publishers should document their work in a way approachable by non-experts
613 if they aspire to have an important impact on the bioacoustic community. On the other hand,
614 users of these models may consult the latest benchmarks and challenges that target diverse
615 applications of audio ML representations. For instance, HEAR (Turian *et al.*, 2022)
616 benchmarked multiple state-of-the-art methods on a varied set of tasks in speech, music and
617 environmental sounds. More recently BEANS (Hagiwara *et al.*, 2022) proposes a benchmark
618 specific to bioacoustics where representations are tested on detection and classification tasks
619 of several species.

4.2. Conclusions of the technical constraints on the current uses, limitations and expectations of automatic detection

Automatic detection has been used for density estimation (McDonald & Fox, 1999; Marques *et al.*, 2013a), occupancy (Dawson & Efford, 2009), species presence (Obrist *et al.*, 2010), trends (Abrahams & Geary, 2020), and phenology, e.g., the start of breeding, or daily onset of song (Willacy, Mahony & Newell, 2015; Oliver *et al.*, 2018). This technology can be used in conjunction with other non-invasive monitoring methods such as camera traps, scat surveys, hair collection, and human observation (Long, 2008), providing additional information and allowing monitoring of otherwise cryptic species that might elude detection. There should be ongoing conversations between biologists and computer scientists, bidirectional and iterative, improving the survey quality, accuracy, and algorithm usability over time. Biologists can provide the ground-truthing and validation of the use of automatic detection, while computer scientists can develop the system and work with them to iteratively improve the automatic detection system.

While we have argued for the widespread use of automatic detection systems, there are limitations, and these should be considered at the start of a project. Some of these are self-evident: signals that do not rise above background noise will be lost as undetectable. Also, signals can be difficult to separate if they overlap with either intraspecific, interspecific, or unrelated sounds, as in the dawn chorus when birds sing with many overlapping, very similar elements, making extraction/detection of a single unit difficult. Dataset sizes (for both training and deployment) may be a limiting factor. We have referred to data augmentation and denoising to synthetically account for data limitations. These and other tools (e.g., data imputation, generative deep learning) are often helpful, but the results are unlikely to be as reliable or unbiased as they would be with a large representative dataset. They should not be relied upon as a silver bullet when recordings are rarely observed, noisy or otherwise hard to analyse. Just like with human annotation, automatic detection will always be subject to some level of bias and inaccuracy; one advantage of automatic systems is that these factors can be numerically evaluated. Automatic detection model predictions are only ever as good as the input training data. Annotations which are not accurate or have not been conducted appropriately for the intended application may worsen the efficacy of the model. Furthermore, density estimation relies on the choice of robust thresholds for confidence in attribution of sounds. There can be an accumulation of errors over time if the thresholds are chosen either to be too low or too high, discarding weak identifications wrongly, or placing too much confidence in others. Finally, all acoustic detection relies on the sound event occurring, and often species may choose to not vocalise or create a sound and thus can be missed. What is not heard cannot be counted. However, despite these caveats, we believe that automatic detection and PAM offer the opportunity to collect and analyse data that cannot be processed by other means, providing an exciting and valuable new tool for the biological sciences.

5. A PRACTICAL GUIDE TO AUTOMATIC DETECTION

We now present a practical guide for using automatic detection. There are many decisions that we must make when designing a study that uses automatic detection, and our goal is to help practitioners optimise these decisions. We realise that some of these decisions may be constrained by access to financial resources, lack of training in bioacoustics, limited technical skills in coding and machine learning, and/or lack of access to high-speed internet for cloud storage and computing. These limitations may be particularly pronounced for researchers in the Global South. We acknowledge that there is still much to be done to make these tools and approaches accessible for all.

668 This guide is developed to help users implement an 'off-the-shelf' automatic detection approach, or for developing or adapting their own approach. We strongly advocate that practitioners implement a pilot study to ensure the approach they plan to use is feasible before embarking on a large-scale endeavour. Importantly, even with the most sophisticated automated approach, a substantial amount of human investment is needed to create training datasets, evaluate detector performance, and verify the detections.

5.1. Define research questions

The most important thing to consider when using automatic detection is the specific research question. For example, if you are interested in detecting the presence or absence of a rare signal (e.g., a gunshot or the presence of an endangered species) then you will want to use an approach that will ensure high recall (i.e., high probability of detection) and you may tolerate a relatively high number of false positives. Alternatively, if you are interested in subsequently classifying individuals from the detections, you may prefer to focus on retaining high signal to noise ratio (SNR) calls and will tolerate lower recall with higher precision. Your research question will influence every decision you make in the automatic detection workflow, including study design, data collection and the analytical approach. For guidance on study design, we point readers to (Sugai *et al.*, 2020).

5.2. Study design

Depending on the nature of the research question, researchers will need to determine their study design, including hardware needs, recording schedule and whether the processing of data will be carried out in real time or at a later date (or, 'offline'). For instance, for the detection of a single species, researchers may deploy ARUs over the landscape for a period of time and then download the data onto a hard drive to be processed offline. The recording schedule also needs to be determined according to the research goal. We refer the readers to more extensive discussions of this issue for further details (e.g., (Browning *et al.*, 2017; Metcalf *et al.*, 2023a). Real-time processing is an emerging area, but due to the limitations of placing power-efficient computation in the field, real-time automatic detection typically is more bespoke and less accurate than offline processing.

5.3. Start with a pilot study (if possible)

Given the costs (both financial and in human labour) of implementing projects that use automatic detection, we strongly advocate that researchers start with a small scale setup to test out their planned approach. For a large-scale PAM study, deploying a few recorders over a smaller spatial scale and a shorter time period may provide enough acoustic data to get started with automatic detection. If the signals are relatively rare (e.g., gunshots) perhaps finding online repositories or datasets of samples would be necessary. A well-designed pilot study will help researchers make informed decisions about annotations, choosing an automated detector, and reporting and interpreting their results.

5.4. Data collection and archiving

Data storage and archiving remains challenging, since the large data volume of the raw audio in many projects often goes beyond the limits of free or easily-available services. Furthermore, (Metcalf *et al.*, 2023a) recommend backing up audio data in multiple copies, and also making use of cloud storage. Simply storing the audio is typically only part of the issue: you and your collaborators will also need to access it, for example to visualise or to apply an algorithm to the dataset, which means that speed of upload and download (bandwidth) may be an equal or greater concern. Cost of storage and bandwidth are often significant questions. Arbimon (Ganchev, 2020) is one project that aims to store and share large volumes of wildlife audio on behalf of others.

Reducing data sizes can be achieved in many ways, including audio file compression and data subsampling. Lossless compression (such as FLAC) can reduce file size without losing information; lossy compression (such as MP3 or AAC) will discard at least some information from the signal, but might still support reliable analysis (Heath *et al.*, 2021), depending on the research question. An alternative strategy very relevant in automatic detection is to keep only the audio corresponding to the positive detections: for rarely occurring sounds this will greatly reduce the storage requirements, while keeping the detected audio clips available for inspection or re-analysis. However, any missed (false-negative) sound events will be irretrievably lost. This would prohibit future interrogation of the raw data for other potential uses.

Good-quality metadata including time, date, location and more, is crucial for the success and reproducibility of any project. This can be stored in the audio files (as "RIFF tags") or separately (Metcalf *et al.*, 2023a). Research and other publicly-shared data should be "FAIR"- findable, accessible, interpretable, reusable (Wilkinson *et al.*, 2016) – and publishing metadata in standardised formats is key to this. The Biodiversity Information Standards (TDWG) group maintains the metadata standards Darwin Core (Darwin Core Task Group, 2009) and Audiovisual Core (GBIF/TDWG Multimedia Resources Task Group, 2013) which help with this through a lightweight approach of specifying common field names and their definitions (such as "Capture Device", "Taxon Coverage", "Locality", "Start Timestamp"). By using such standards, researchers can ensure that their metadata will be understood by others and be findable. It also enables a next generation of methods that could automatically generalise across multiple available datasets, since the metadata are compatible.

5.5. Data annotation

A well-annotated dataset is critical to the performance of a ML-based automated detector. When creating annotations, many decisions must be made, including which program will be used, the specific approach, as well as (often subjective) decisions regarding specifics about the granularity, or what ‘counts’ as an annotation, for example individual vocalisation bouts or whole sequences. There have been calls to standardise annotation approaches in bioacoustics (Nicholson, 2023), similar to what has been done for human speech (Gibbon, Moore & Winski, 1998) and music (Humphrey *et al.*, 2014). However, to our knowledge a standardised protocol does not yet exist, perhaps due to the diversity of signal types and research questions across bioacoustics and/or a lack of communication among fields. Here, we aim to provide some guidance for annotating a dataset for automatic detection (Figure 4).

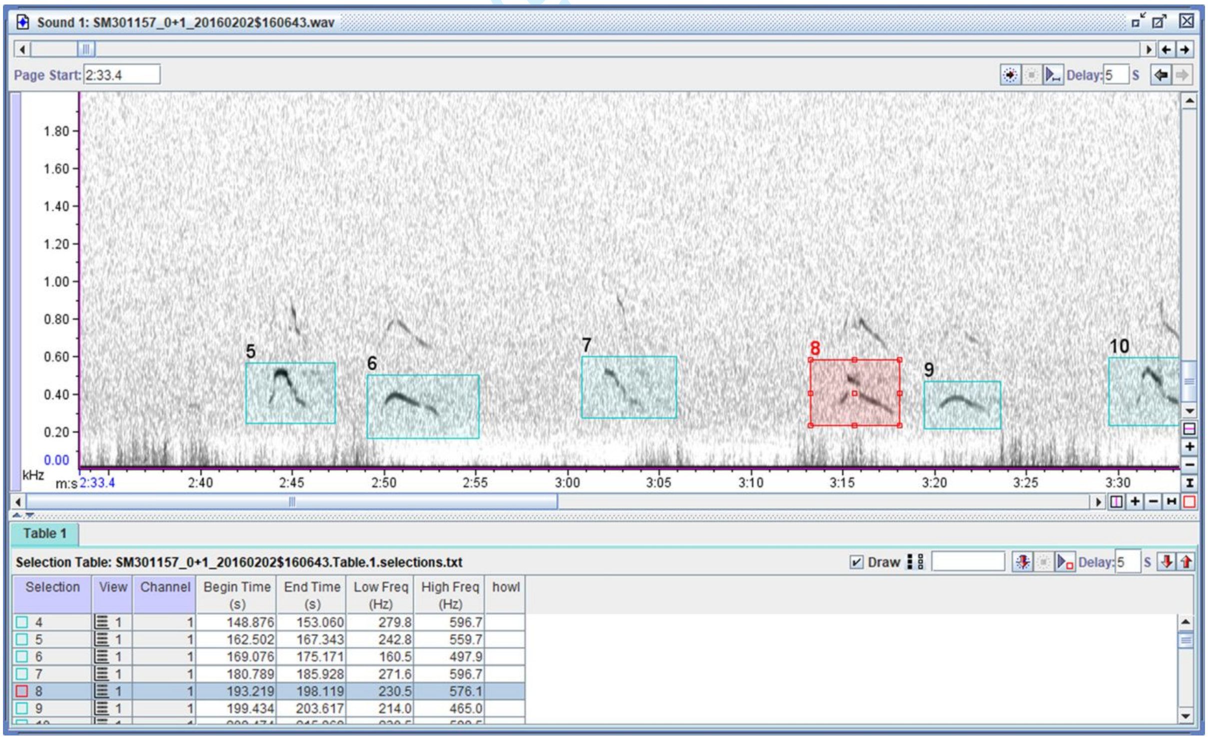

Figure 4. Example annotation of acoustic signals, in this case, wolf howls. Taken from (Kershenbaum *et al.*, 2019), showing a spectrogram generated using Raven Pro.

Due to the relatively large amount of human investment required to get high-quality annotations, researchers often ask themselves how many annotations are needed. This generally depends on the research question, and it is often recommended to annotate as many signals as possible, however there are more specific questions that can help guide these decisions. The first concerns the classes or discrete types of signals in your dataset. For example, will you annotate every bird species in a long-term recording? Will you annotate a single call type from a single species? Or will you annotate all the notes or elements in a sequence from a single individual? In addition, one must decide whether to annotate the "negative class" (oftentimes the "noise" or "absence" category). If doing exhaustive annotation where all the signals of interest are annotated, then it can be assumed that anything that is not annotated is the "negative class". However, strategically annotating other "distractor/noise" sound events may improve detector performance, especially sounds occurring within the target frequency range which are loud or easily confused with the target signal. These "noise" labels can help with error analysis and with the training of an algorithm.

Decisions about the temporal scale of the annotations must also be made. A common approach is to annotate the smallest acoustic unit, e.g., note or syllable (Kershenbaum *et al.*, 2016a); however this method can be very time-consuming for large datasets. For vocal sequences that are comprised of multiple acoustic units (e.g., gibbon vocalisations) another approach is to annotate particular call types or phrases within the longer sequence, e.g., annotate only the female gibbon contribution to the duet (Clink *et al.*, 2023).

The number of annotations needed will be influenced by the research question and the choice of the automatic detection approach (see below) but may also be limited by external factors such as funding support for analysts. It is important to consider the diversity of signal types as well as background noise, and to work to include a distribution of annotations or samples across sites, times of day, groups, individuals, etc. A higher number of annotations (and therefore more available samples for training data) will likely improve detector performance and may be necessary in cases where the signals of interest are highly variable. In some cases, such as the use of transfer learning, a smaller number of training samples (~ 25) may be sufficient (Dufourq *et al.*, 2022a), but even in these cases, only a test set in the order of one hundred examples would enable a reliable evaluation of the model. Researchers also need to make decisions about which target signals to include in their annotations, such as whether to include low SNR acoustic signals, signals that substantially overlap with non-target signals, or signals that are abnormal in structure.

A common way to do annotations is by visualising spectrograms in a graphical user interface (GUI) such as Raven Pro (K. Lisa Yang Center for Conservation Bioacoustics, 2014), SonicVisualizer (Cannam, Landone & Sandler, 2010) or Praat (Boersma & Weenink, 2007) and creating bounding boxes around the signal(s) of interest. Other possibilities include the use of an energy or coherence (Wijers *et al.*, 2021) detector to identify all signals above a certain threshold in a given frequency range and then labelling these detections, applying an unsupervised clustering algorithm and labelling the batches of samples that have been grouped together, or the use of DL approaches to identify the start and stop times of signals of interest automatically, e.g., TweetyNet (Cohen *et al.*, 2022). However, one must be cautious about mass semi-automated annotations, since these may introduce non-obvious bias that can affect the conclusions of the study. We recommend including random sampled manual inspection steps in the procedure. It is important to document your annotation protocol, including the decisions you made and why you made them, in a way that can be reproduced by others. We suggest including these protocols as online supporting material in publications. In addition, it is crucial to check both intra- and inter-observer reliability for creating annotations (Nguyen Hong Duc *et al.*, 2021).

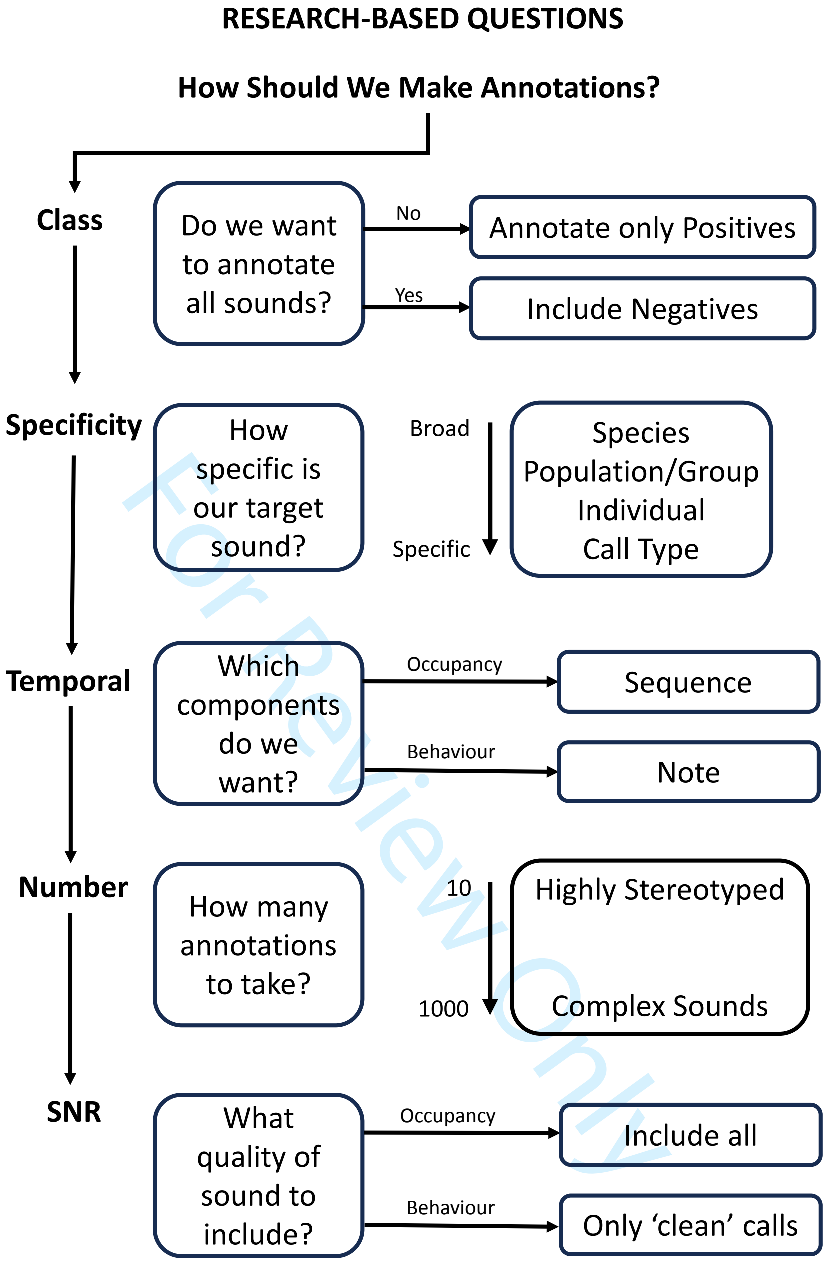

Figure 5a. A flowchart for designing research questions in relation to automatic detection.

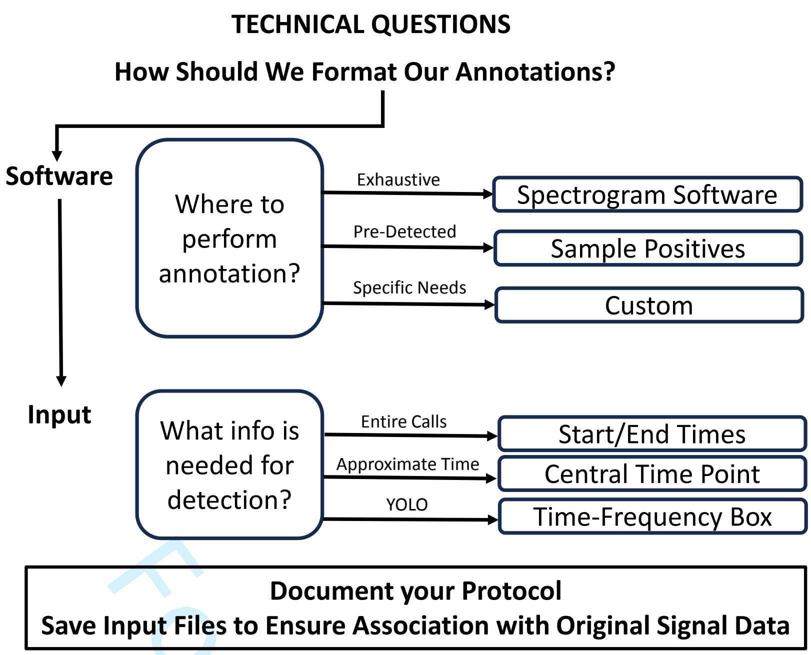

Figure 5b. A flowchart showing annotation decisions for automatic detection.

5.6. Choose your Detection Pipeline

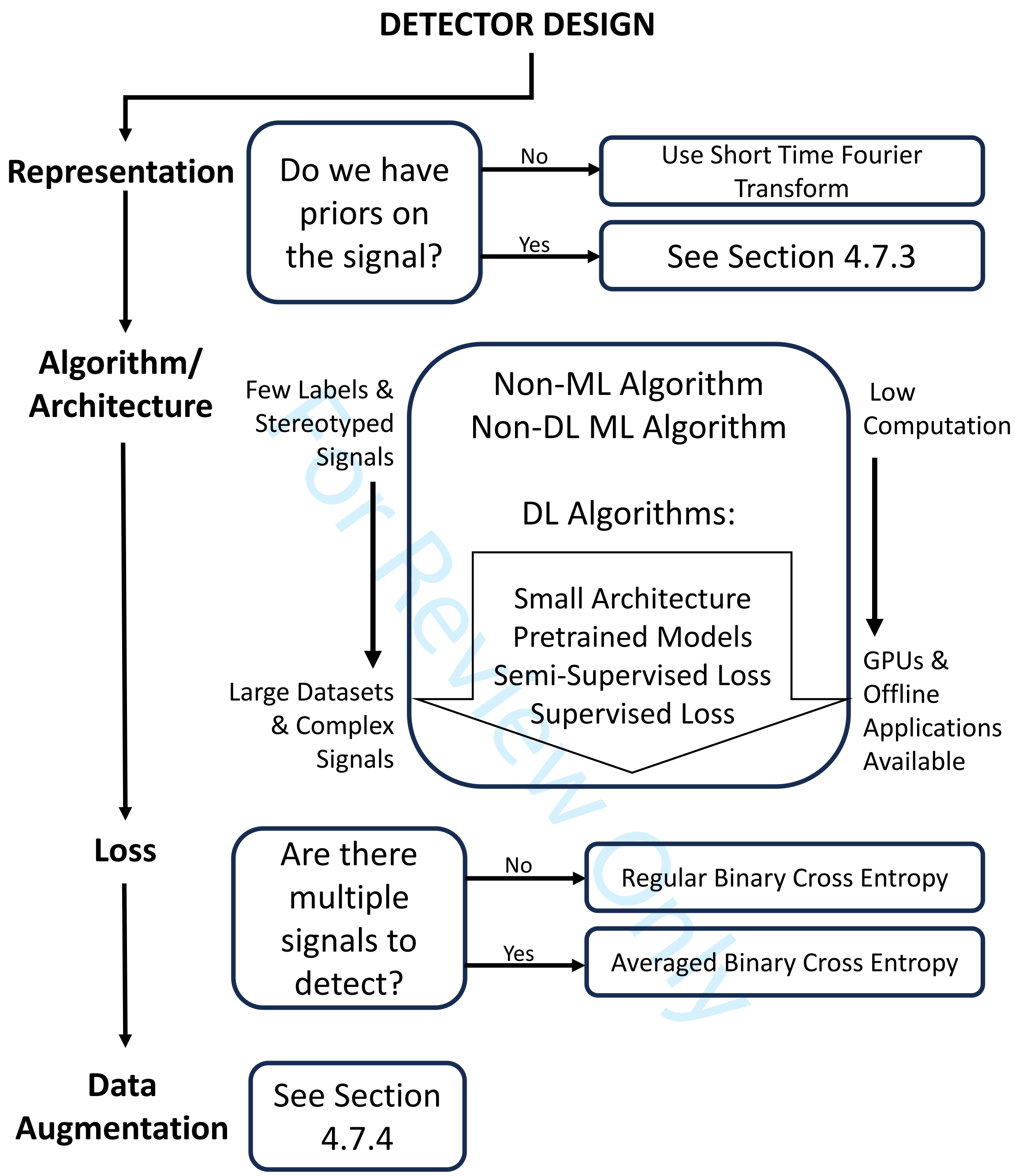

Figure 5c. A flowchart showing the decisions necessary in automated detector design.

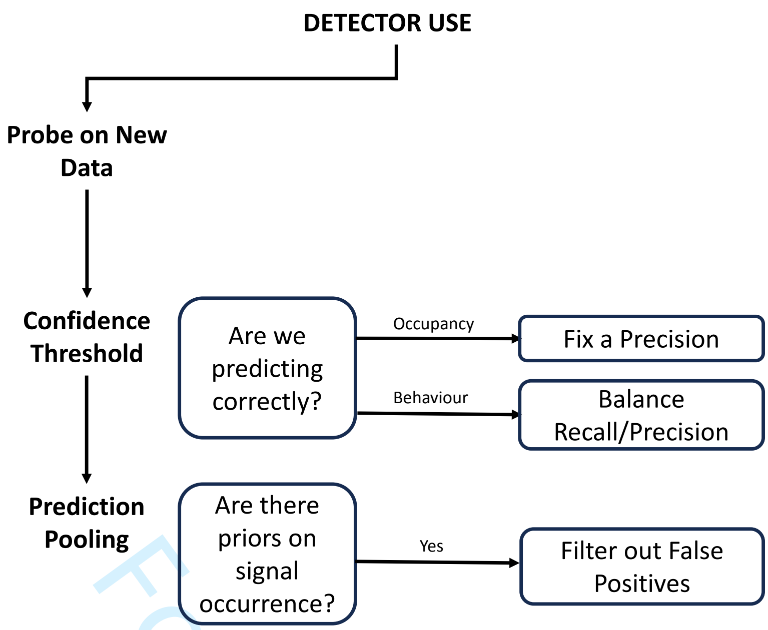

Figure 5d. A flowchart showing the constraints of end use on automated detector design.

5.6.1. Interfacing with your pipeline

Selecting an automatic detection approach depends on factors such as technical familiarity, desired granularity, and budgetary constraints. Products such as Kaleidoscope (https://www.wildlifeacoustics.com/), PAMGuard (https://www.pamguard.org), and Arbimon (https://rfcx.org/ecoacoustics) provide easy to use interfaces for systems that can perform automatic detection on audio samples originating from a wide variety of environmental samples. These tools come equipped with traditional approaches rooted in standard signal processing techniques but are limited in their ability to utilise modern advances in machine learning. Conversely, modern DL frameworks, such as TensorFlow, PyTorch, and Keras (Stowell, 2022b), as well as the models built with them, rarely come with an easy to use interface which makes them less accessible. Commercial approaches offering cloud-based machine learning as a service (MLaaS) solutions, such as those from Amazon, IBM, or Microsoft, allow easier access to these advanced methods, but can be prohibitively expensive.

Practitioners must decide whether easy-to-use tools are sufficient for the problem at hand, or whether it would be advantageous to exploit the often superior performance of DL methods, which require more investment of time, money or both. The complexity of the research question has a significant influence on the selection but may be outweighed by the need to invest further in expertise or funding.

In the case of any automatic detection approach, the pipeline must be evaluated in the context of the research questions which necessitates dividing the data to properly evaluate performance and generalisability, the choice of an appropriate detection mechanism, and the selection of relevant, comparable, and appropriate metrics.

5.6.2. Split your data

As for most machine learning tasks, datasets should be split into train, test, and validation subsets to ensure the true generalisability and comparability of a model's performance. This means that an amount of data (usually around 10 to 20% of the total dataset) need to be kept unseen during training and validation of the model, for which the remaining 80-90% of the data are used. This helps to avoid model overfitting, which would cause the model to learn only the characteristics of the training data, without the ability to generalise to new data, and would bias performance scores (Gareth James *et al.*, 2013, p176).

The validation (or development) set is used for hyperparameter tuning. This is especially useful in the case of DL models which involve empirical testing of optimal values and setups for elements such as optimisers, learning-rates, or early stopping. The best performing model, as determined using the validation set is then applied to the test set. Finally, the best performing model on the validation set is applied to the test set. The test set should not be used to fit the values of such hyperparameters or to compare model architectures since it would no longer serve for generalisation assessment; it is kept for final performance evaluation. Creating an effective test dataset may include the selection of a separate microphone entry, specific time frames, separate recording locations, or subsets of vocalisations from an individual which were not included in the training set, amongst others. The general idea here is to separate the prediction capabilities of the computer model from recording specificities and data related biases. We always want to ensure that an automatic detection model is generalisable rather than specifically trained for a single recording setup, location, or individual.

To provide an example, in the case of creating a presence/absence detection model, one should not use annotations from the same file for training and testing. But instead, certain audio files should be used to create the train dataset, and independent files should be used to test the detection model. Furthermore, the model should be applied to entire testing audio files and not only to parts of the test file that have been annotated, as this might result in an overly optimistic evaluation of the model and potential false positives would be missed.

5.6.3. Pick your feature representation

Depending on the automatic detection approach, acoustic data may be transformed through feature extraction to ease the automatic detection process. In the computational bioacoustics literature, an array of such feature extraction methods can be found, each presenting their own advantages and limitations.

In bioacoustics, the dominant approach is undoubtedly spectral representations such as spectrograms or mel-spectrograms. This type of representation usually allows for interpretable visualisation of acoustic data and provides an easy route to use popular vision-based models such as CNNs for object detection and image classification. Despite this, some information from the raw waveform may get lost when computing these representations. This is especially the case for transient signals such as odontocetes' clicks which are poorly represented by Fourier transforms (Jiang *et al.*, 2018). CNNs developed for spectrograms cannot be used directly for waveforms, because the data is of different dimensionality;

however there have been a lot of recent developments in DL methods applied directly to waveforms and so this is increasingly feasible (Baevski *et al.*, 2020).

DL methods now allow for high-dimensional inputs such as whole spectrograms, with the succession of layers extracting higher level features and information. However, historically users were the ones responsible for selecting relevant features to represent signals. In this context, MFCCs were often used, and given to a classification algorithm such as a support vector machine (Mitrovic, Zeppelzauer & Breiteneder, 2006). For relatively simple use cases e.g., stereotyped signals and low background noise, this approach might suffice in bringing satisfactory performances.

Recently, as stated in Section 0, extracting pretrained latent representations as features is also being adopted as a promising solution. This approach may imply additional effort on the part of the user and raises an array of questions on pretraining datasets, selected model architectures or the need for higher computational power. It can also prove successful in easing the downstream learning process or allowing for smaller annotated datasets in few-shot learning perspectives.

Despite the advantage of using such abstract representations, using traditional engineered features such as fundamental frequency, call duration or number of notes may still prove to be effective depending on the task at hand. These can also be combined with features extracted from the time domain such as energy and zero-crossing rates. These can then allow for the use of simpler algorithms which may be easier to implement and require little computational power and training time.

Overall, there is no such thing as the perfect feature extraction method for bioacoustics. Comparing different feature representations should always be the preferred approach and can be carried out on the previously mentioned validation set, ideally in a pilot study.

904 5.6.4. Decide on feature transformation

905 Prior to feature extraction, specifically in the case of noisy recordings characterised by low
906 SNR, some detectors may benefit from denoising, i.e., the automatic removal of background
907 noise from the acoustic signal of interest. An extensive overview of recent approaches can be
908 found in (Xie, Colonna & Zhang, 2021) with accessible open-source solutions. Some of these
909 methods are built on light-weight algorithms such as spectral-gating (Sainburg, 2019), others
910 involve the use of DL with CNNs, Noise-2-Noise-based approaches (Bergler *et al.*, 2020), or
911 denoising-autoencoder models (Vickers *et al.*, 2021; Yang *et al.*, 2021).

912 Although it is useful in some applications, this pre-processing step is not always
913 recommended and must be used with caution as it may result in a loss of information. In
914 some cases, noise can also be directly handled by the detector itself, especially when using
915 noise-resilient DL architectures or when stationary noise is not overlapping the target signals.
916 In cases where noise reduction is applied prior to training, the evaluation and test datasets
917 will need to be put through the same process, to ensure that training and testing data have
918 comparable characteristics and contain similar acoustic information. When building a noise
919 resilient model, one may also resort to multi-condition training approaches. This method can
920 imply adding noisy corrupted versions of the data to the training set or including both the
921 original and the denoised versions of the data during training to help with model robustness
922 to noisy acoustic contexts. This approach is fairly common in speech processing (Yin *et al.*,
923 2015) but needs further exploration in bioacoustics.

924 Depending on the amount of training data available, data augmentation techniques may be
925 used to artificially increase the variability of the data on which models are optimised. The
926 choice of which augmentation technique to use depends on the application. One should aim
927 to apply transformations that cover the range of variations found in real signals. However,
928 care must be taken to avoid transformations that could invalidate the annotations. For

instance, in a bird call detector reversing sounds could be a tempting simple transformation, yet this could result in artificially making a bird call more similar to that of another species. Simple transformations may also create artefacts that can complicate the modelling, for example a pitch shift of a howl may also unrealistically shift the background noise. Commonly used techniques include stretching or compressing the duration of acoustic signals, shifting their pitch, making small volume modifications, or adding a variety of noise or mixing with other audio events via some linear or non-linear combination (e.g., taking one presence event and mixing it with one or more absence events). These transformations may also be combined to produce more variation.

Recently generative deep-learning methods, such as Generative Adversarial Networks (GANs) have been proposed in order to generate synthetic examples (Wang, She & Ward, 2022; Bergler *et al.*, 2022a).

5.6.5. Decide on a method

5.6.5.1. Deep learning or not

As mentioned above, the choice of a detection mechanism is dependent at least partially on the complexity of the problem. If the signals are well defined, have high SNR, are highly stereotyped, and the research question involves simple segmentation and can be done offline, a package such as Kaleidoscope or Arbimon may be more than adequate.

Using machine (deep) learning may be advantageous in situations requiring a more complex analysis, such as call type classification, or where robustness to environmental noise is necessary (Aodha *et al.*, 2018; Stowell, 2022a). However, in situations where access to either a large amount of computing resources or the training / expertise to use them effectively is limited, the use of DL may not be possible. Additionally, it must be considered where the detection mechanism will be deployed. If access to a large computing cluster is readily available but the end result must function on a small device for field deployment, then a large and complex model may not work. Conversely, if the final model will only be used offline using minimal computing resources (budget GPU), then the model choice becomes somewhat more flexible. Different machine learning approaches are given in Table 1, together with their requirements and example studies.

959 Table 1. Different types of machine learning techniques

Learning Type	**Labelled Data Requirements**	**Metrics**	**Visualisations**	**Examples**
Supervised (Segmentation, Classification)	Large amount of labelled data	Accuracy, Precision, Recall, F-Score, AUROC, mAP, UAR	Confusion Matrix ROC-Curve, PR-Curve	(Bergler *et al.*, 2022b)
Unsupervised or self-supervised (Clustering)	Labelled Data Not Necessary	Reconstruction Loss (MAE, MSE), homogeneity, completeness	Reconstructions, Dim-Reduction (t-SNE, UMAP)	(Cuevas *et al.*, 2017)
Semi-Supervised Learning	Some Labelled Data - Large amount of Unlabelled data (optional)	Both Supervised & Unsupervised	Both Supervised & Unsupervised	(Bermant *et al.*, 2019; Saeed, Grangier & Zeghidour, 2021; Leroux *et al.*, 2021; Hagiwara *et al.*, 2022)

5.6.5.2. Choose your evaluation metrics

The evaluation of the automatic detection mechanism depends primarily on the type of task to be performed. A fully supervised detection / classification task is typically evaluated using metrics such as accuracy, precision, recall, F-score, or area under the receiver operating characteristic curve (AUROC) (Lever, Krzywinski & Altman, 2016). These all provide different insights and can help evaluate how the model is performing. For example, precision indicates the fraction of relevant results (true positives) that are found among all detected events, whereas recall indicates the fraction of signals in the dataset which were effectively found. Typically, a balance must be decided which metrics are most important for a particular task. For example, recall may be an important score to consider when detecting rare phenomena where missing a single detection of an underrepresented class may prove costly. Wrong choice of metrics may bias the results, for example, in the case of highly unbalanced datasets, i.e., when the acoustic object to be detected is rather underrepresented in the dataset compared to negative labels, accuracy may be very high despite low performances on the small number of positive test samples.

Visualising results from supervised training methods can involve a confusion matrix, which is a table that shows the ground truth values on one axis and predicted values on the other, allowing visual analysis of model performance which is easy to digest. Another option is the receiver operating characteristic curve (ROC curve), which plots the trade-off between true positive rate (TPR) and false positive rate (FPR) at all confidence thresholds, enabling the analyst to more easily choose a prediction threshold which suits their needs. The area under the ROC curve (AUROC) gives a summary of the model’s performance across threshold and is agnostic of threshold choice.

A similar visualisation to the ROC curve is the Precision-Recall (PR) curve, which also highlights the balance between missing out events (false negative) and making false alarms (false positive). The area under the PR curve is commonly referred to as the mean average precision (mAP). The important difference between PR and ROC curves is that the precision gives the proportion of correct detection among all detections and the FPR indicates the proportion of wrong detections among all negative examples. In the case of highly unbalanced datasets (e.g., 1% of positive examples), the FPR can be rather optimistic as compared to the precision, and thus the mAP might come out to be significantly lower than the AUROC. Detailed discussions on possible performance metrics can be found in (Davis & Goadrich, 2006; Hildebrand *et al.*, 2022).

Useful metrics for unsupervised learning are harder to identify, as it depends on the research question. If labelled data are available, they can be used to assess the quality of a clustering attempt by measuring completeness (across how many clusters are samples with the same label) or homogeneity (the proportion of samples in a cluster with the same label). Visualisation for unsupervised clustering results are often done by reducing the dimensionality of the reductions to either two or three dimensions using t-Stochastic Neighbour Embedding (t-SNE) (Maaten & Hinton, 2008), Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy & Melville, 2020), or a similar method.

5.7. Verifications - check your results

The verification of model performance on the test data should involve quantitative and qualitative evaluations. Quantitative metrics give the performance in terms of comparable values like the F1-score, accuracy, precision, or recall. Whilst the qualitative metrics would help to understand the practical implications of the model. Qualitative analysis involves manually checking or visualising the predictions. This may involve plotting automatic segmentation results on spectrograms to visually account for the precision of detected time frames. It may also be carried out through a simple manual inspection of a subset of results.

Careful manual analysis of the signals with missed detections or false alarms could help to identify the characteristics that trigger the models and help to improve the models further by adding the specific variations needed in the training data or clean up train data (especially wrong annotations or mislabelled data).

5.7.1. When is a model good enough? Performance thresholds

Understanding the performance thresholds and being realistic about the task is a pragmatic way of approaching the problem. It is important to understand that machine learning models are statistical in nature and may never provide 100% performance even with perfect data or models. Understanding the limitations of the model and the desirable performance in the real world scenario can help set the thresholds for performance, for example, trade-off between false positives and missed detections (Karnan, Akila & Krishnaraj, 2011). In some scenarios it may not be even practically feasible to achieve a desirable performance due to factors like overlapping sounds, environment noise or very low SNR. But understanding and defining the problem based on a trade-off between what is feasible with the acoustic data and what is desirable could help define performance thresholds and build practical models. For example, defining the range of distance within which the target species needs to be detected.

5.7.2. How harmful are mistakes (false positives vs false negatives)?

The use case for automatic detection will influence how much (section 4.8.1) and what kind of errors are acceptable. For instance, if doing an analysis on vocal behaviour, missing a call in a sequence might strongly distort results. Conversely, if occupancy trends are aimed for, missing one call in a sequence is insignificant, and imperfect detection can be incorporated into occupancy models (Bailey, MacKenzie & Nichols, 2014). Recall is thus more or less important depending on the type of study being conducted.

In general, false positives are undesirable, but a certain amount might be acceptable (Shiu *et al.*, 2020). In any case, converting the precision into the number of false positives per hour allows an unambiguous interpretation by the user and the planning of how to deal with false alarms.

Additionally, prior knowledge on vocal behaviour such as the sequence regularities might allow filtering out of false positives. Such priors can be used to reduce confidence thresholds and increase the recall, but with the risk of imposing too strong priors and missing out on uncommon sequences.

5.7.3. Reproducibility and accessibility

We also expect automated vocalisation detection systems to be made available to other users, thus broadening the contribution to the field of bioacoustics (especially to users without a strong computer science background). For this purpose, code for detection systems should be shared in comprehensive and accessible ways, such as version control repositories, and should be well documented with detailed user manuals (Braga *et al.*, 2023). An easy way to make a detection model available to the community is also to follow common APIs that will allow their integration into pre-existing interfaces, such as ARISE (Hogeweg & Stowell, 2023) or Raven Pro (K. Lisa Yang Center for Conservation Bioacoustics, 2014).

Besides publishing code for experiments to be reproducible, datasets used for training and testing should be made available to the community for building new systems and comparing them using standard annotation protocols (see Section 5.5). Indeed, public benchmarking datasets exist (Joly *et al.*, 2015; Politis *et al.*, 2020) but cover only a relatively small set of species targeted by bioacoustic studies.

1055 5.7.4. Access to raw recordings

1056 Ideally, additional to labelled training dataset, raw recordings (as opposed to cut-out
1057 snapshots) are of potential value to the machine learning community (to train self-supervised
1058 models for instance) and to the research community in general to reuse the data for other
1059 tasks or to create new annotated datasets from previously recorded data. But it might not be
1060 always feasible to make this readily accessible in public repositories due to storage and other
1061 constraints. We encourage researchers to store the raw recordings locally and share them on
1062 demand with the community or with interested parties.

1063 6. WAYS FORWARD

1064 We now consider some important ways forward for automatic detection for bioacoustics,
1065 including best practices which should be implemented now, the challenges still to be
1066 overcome, and the future direction of the field.

1067 6.1. Challenges

1068 6.1.1. Bioacoustic challenges

1069 Although automatic detection has already brought large improvements to the field of
1070 bioacoustics, challenges remain which are closely related to the nature of animal sound
1071 and/or the desired uses of such data. For instance, population density estimates rely on
1072 detections being reliable, without double-counting individuals' vocalisations when they are
1073 picked up by multiple devices (Kimura *et al.*, 2010; Marin-Cudraz *et al.*, 2019), and are
1074 further improved if calls can be localised and attributed to an identified individual (Nijman,
1075 2007; Knight & Bayne, 2019; Hedley *et al.*, 2021; Law *et al.*, 2021). Moreover, in most cases
1076 population density cannot be estimated without knowing the detection range of the system
1077 (Metcalf *et al.*, 2023a). The detection range of the acoustic signal will depend on multiple

factors including source level and frequency range of the signal, characteristics of the habitat including ambient noise levels, vegetation and topography, along with specifications of the ARU (Haupert, Sèbe & Sueur, 2022). However, detection range is often difficult to estimate, especially in forest environments or areas with extreme topography, and in many cases is ignored or assumed to be consistent across studies when this may not be the case. When species of interest are near the limit of the detection range of the device, recordings of vocal signals may become attenuated or missed. This might cause problems in some tasks which try to capture specific aspects of the vocalisation, for example to infer behaviour, caller identity or communication patterns, rather than generic tasks which look at occupancy (Spillmann *et al.*, 2017).

Even when accurately focusing on our target species' vocal signals, animals might engage in simultaneous vocalisations or choruses (Torti *et al.*, 2018), which makes a simple timestamped detection system insufficient for acoustic behaviour analysis. Also, it can be difficult to distinguish vocalisations of similar species if they share characteristics, e.g., dog barks and coyote barks share a number of similarities which make it difficult to determine which species produced the rapid-fire sequence of noisy barks, though there are some quantitative differences (Feddersen-Petersen, 2000).

6.1.2. Computational challenges

Computational challenges in this field include questions of algorithms, datasets, computational efficiency, computing platforms and more.

One overarching challenge within machine learning in the broad, and with particular relevance to automatic detection, is the ability to generalise. For example, a model well-trained for a particular species can perform poorly with even slight variations in recording devices, ambient noise, or operating environments. This could lead to low accuracy without further testing and adjustment. Creating scalable models that have the flexibility to add new species to the training dataset, to increase the number of vocally active species which can be detected, is still a challenging task. Transferring knowledge from models built with data from one species to a new species without further training data is even more desirable. We also note that many models are highly task specific - the data specification, annotations, model architectures, and systems are highly optimised for best performance. For example, a system used to determine the occupancy of a species may not be suitable for individual identification, understanding communication, or behaviour patterns which superficially appear to be related but are subtly different tasks. It is not immediately clear to a user how far to trust in the generalisation of a detector.

Acquiring generic datasets that can address multiple tasks, such as population density estimation and behavioural characteristics, poses a significant challenge due to the limitations in data collection strategies. Typically, data collection is initially planned to address specific tasks, which makes it difficult to acquire datasets that can be scaled to any given task. This is a challenge as it is essential to streamline and optimise the recordings to collect only data of interest to a particular task to increase storage and computational efficiency. But, at the same time, the data collected might not include the context or information that was needed to use it for a new task. A lack of generic, benchmark datasets has significant implications for the standardisation of methods in the field and the appropriate evaluation of research.

In bioacoustics as in other fields, DL comes with very limited interpretability, an issue known as the 'black box problem'. This amplifies the problem that conclusions drawn about DL models will be specific to the dataset they were tested on, which significantly hinders the process of finding a consensus for the best architecture or training procedure to be used. In certain cases, it is also unclear as to how different research studies split their datasets and conduct model evaluation. As it stands, little to no standards on the best approaches exist and without these best practices put in place, authors will implement their own approaches within their research. The best opportunity to overcome issues such as these is firstly to encourage further development of public access or benchmark datasets, and secondly to probe models on their detailed behaviour regarding these datasets (Alain & Bengio, 2018). Within the current literature, the approach that authors have taken to implement their machine learning testing methodologies and model evaluation differs drastically. In most cases, comparisons are not made to existing results on datasets that are publicly available, instead, most studies present their findings related to their proposed method on the dataset that was collected for the study. These observations are quite different to what has been observed within the computer vision and natural language processing literature whereby most studies will compare their proposed method to various baselines and existing state-of-the-art methods on the same datasets. Consequently, a comparison between research studies within bioacoustics is not feasible and determining the state-of-the-art is non-trivial. Various initiatives exist that provide bioacoustic benchmark datasets and standardised public evaluations, including automatic detection in particular, though these are neither as large nor as widely-used as in mainstream ML application domains (Stowell *et al.*, 2019; Ferrari *et al.*, 2020; Hagiwara *et al.*, 2022).

Training machine learning models, particularly deep neural networks, is computationally intensive. Specifically, computers, workstations, or servers with a large amount of CPU (central processing unit) and GPU may be needed, to speed up the training or just to make it achievable in reasonable time. Furthermore, certain deep neural networks require a large amount of GPU RAM to load the model into memory given the large number of trainable neural network parameters that need optimisation. The issue of access to computational power can exacerbate inequalities between people, institutions, and countries. However, the good news is that the widespread use of pretrained models can massively decrease the amount of computation needed: most researchers should not need to train a model from scratch. This helps to reduce inequalities as well as the carbon footprint incurred through a move to ML methods.

In conjunction with computation, data storage requirements have skyrocketed with the amount of data being collected from PAM and necessities to store, share and create backups of these very large datasets. In certain cases, practitioners have had to ship hard drives physically across the world to share acoustic datasets, and in other cases practitioners share large datasets via cloud-based solutions. It is unlikely that storing all audio for all projects is feasible, and yet discarding audio takes away the possibility of reanalysis or new uses. Bioacoustics will benefit from the development of mixed schemes with well-designed heuristics to store some audio in full resolution (e.g., detected audio clips) and the remainder in highly compressed formats which are still reusable (e.g., embeddings or low-bitrate lossy compression).

There are other considerations that arise from the large data volumes that are required both for training automatic detection systems, and for investigating biological questions using bioacoustics. Logistical challenges in maintaining the data collection devices include changing batteries, calibration of microphones, and general wear and tear. Sometimes the devices need to be deployed in remote, difficult-to-access, or even dangerous locations, which makes the maintenance even more challenging. Therefore, the effort required to gather the volume of data needed for training automatic detection models needs to be considered carefully. However, artificial intelligence being a rapidly evolving field means that new techniques and models may ease (or indeed exacerbate) the problems of providing enough data.

6.2. Future directions

6.2.1. Accessibility

The extent to which automatic detection for bioacoustics is accessible to a wide range of researchers across different fields and geographical regions is patchy and insufficient. Future developments in the field must include increasing the ease with which researchers can implement and customise the technology. Usable, stable, and open-source tool kits with an associated GUI, and potentially a cloud-based solution, can aid the entry of practitioners from a non-machine-learning background and reduce the learning curve. Standards-based interoperability and component-based approaches will help ensure that solutions remain well-maintained and usable.

To move to the next generation of automatic detection, we look forward to further work developing the scale, reliability, and generality of machine learning methods in bioacoustics. But even considering the current state of the art, the barrier to entry for practitioners, students and researchers who are new to the field of machine learning is high (Broll & Whitaker, 2017; Schultze, Gruenefeld & Boll, 2020). This barrier is potentially even higher for newcomers in machine learning for bioacoustics than those entering the field of machine learning for computer vision or natural language processing. For the latter two, there are large quantities of educational material, including blog posts, online tutorials, books, videos, and software repositories. The number of research laboratories, and researchers from tertiary educational institutions working on automatic detection for PAM or bioacoustics in general is not evenly distributed between the Global North and South, and thus, the ability to train students may differ between regions. There is a pressing need for more educational material to become available so that those entering the field can rapidly learn the necessary skills to facilitate progress, and as such, we encourage researchers and practitioners to create and share open-access educational material.

Complementary to educational materials is of course that systems themselves should be more accessible and user-friendly. The required use of Python or R (let alone libraries such as Tensorflow, and repositories such as Github, etc) acts as a barrier to many potential users, and so projects that develop good interfaces are to be celebrated. However, the pace of change in ML methods is fast, as well as the diversity of platforms (e.g., mobile devices), so it is risky to advocate a single graphical interface. The solution is to rely on component-based approaches and well-documented standards; as long as user interfaces can use standards-based methods to “talk to” algorithms and datasets, and each of these components can be replaced, substituted and improved, we provide a good substrate that makes it easy for interface developers to add value to the work (Darwin Core Task Group, 2009; GBIF/TDWG Multimedia Resources Task Group, 2013). For all these components, the community needs to consider their maintenance models (open source or commercial, free or subscription-based) and the ongoing maintenance of core components should not be left to chance.

6.2.2. Foundation models

As with the maturation of machine learning in fields such as image or speech recognition, we expect animal vocalisation detection models to progressively standardise, not only in terms of model architectures but also in data representation. Indeed, pretrained models created from large datasets with a variety of species or taxa can yield rather generic embeddings, allowing good performances when fine-tuning for a specific task, even when relatively few labels are available (see Section 0). Fields such as text processing and image recognition are beginning to move to a scale where “foundation models” emerge, meaning DL models which are trained across massive and highly varied datasets, whose scales lead to emergent generalisation behaviour and which can be reused for a wide range of downstream tasks (Bommasani *et al.*, 2022). The same could happen for bioacoustics and automatic detection: although the size of the benefit is hard to foresee, large-scale highly generalised models could indeed overcome the significant limitation in bioacoustics that many custom tasks do not come with strong training datasets. An alternative approach is few-shot learning, recently explored to generalise robustly from as few as five examples (Nolasco *et al.*, 2023). Such methods indicate that "one big dataset" is not necessarily the main objective for the field. These trends may converge, with the many public bioacoustic datasets forming a richly structured pretraining curriculum for systems to generalise well from simple examples.

6.2.3. Multi-modal detection

Some challenges posed by automatic bioacoustic detection, including difficulties in separating individual emitters, precisely assessing population density, double counting, or missing detections, could potentially be eased by multi-modal approaches. In fact, incorporating additional modalities such as images, video or GPS data, may result in complementary information missing from the acoustic data and enhance the detector's performance, which can then enable uses such as abundance estimation (Akamatsu *et al.*, 2013) and activity tracking (Li *et al.*, 2020; Morrison & Novikova, 2023). Automatic multimodal approaches can also allow tackling complex and innovative behavioural questions for species known to communicate in multimodal ways, such as primates (Slocombe, Waller & Liebal, 2011; Liebal & Oña, 2018) and spiders (Uetz & Roberts, 2002; Hebets, 2005). Multimodal data thus presents many advantages for automatic bioacoustic detection, all the while raising an array of limitations and adding a certain degree of complexity to machine learning solutions. Recording multimodal data is a first important challenge which can be partly addressed through the increasing availability of new efficient hardware solutions, such as lightweight, inexpensive camera traps and drones. The automatic processing of non-acoustic data is also being investigated and numerous machine learning models exist as promising solutions (Akamatsu *et al.*, 2013). Yet, the simplicity, diversity and quantity of information contained in bioacoustic data seem to make it a superior solution in most detection tasks (Enari *et al.*, 2019), at least as long as vision-based machine learning and visual recording hardware / large data storage and processing don't show significant improvements.

6.2.4. Keeping a biologist in the loop

Some of the ML models and systems are designed without the full domain knowledge or context of the problem being addressed. There needs to be close collaboration between the ML engineer designing the systems and training models, and biological scientists, as domain experts, who can validate the solutions and performance of the models. The process pipeline needs to be designed such that domain experts should closely monitor every stage from the methodology for data collection, design of data collection devices, data annotation techniques or methodology, data splits, model architecture (including inputs and outputs), and performance metrics and performance threshold values. It is also worth noting that the very same biologists may also be the ideal audience for the commercialisation of foundational models once they become available and the technologies and methods are easily accessible. The system should be iteratively improved with the active feedback from experts in the field or through the knowledge of the domain expert. This in turn maps to the process flow standardisation discussed in earlier sections.

Since bioacoustic tasks deal with big datasets, demanding high computational power, there needs to be considerations on the environmental impacts of data storage, data transfer, computation power in terms of model training, validation or deployment in the real world. Training machine learning models is computationally very expensive and the use of GPUs results in large amounts of energy consumption. This raises the question of sustainability with respect to the research being conducted. Various independent researchers training similar models on the same datasets would result in a suboptimal use of resources. Energy consumption may be reduced by training smaller models (from model pruning, or "distillation") or by sharing models. There are options of cloud storage or cloud computations (Aide *et al.*, 2013) which could benefit from the usage of green data centres in remote locations (Ministry of Local Government and Modernisation, 2021) that have green infrastructure for energy production (through renewable energy sources) and are perhaps less harmful to the environment rather than local GPUs or server solutions.

It is also important to think of low footprint, low power usage models and systems in real world deployment for data collection or final deployment. Currently, many research studies are applying automatic detection algorithms on data that were collected in the past. We, however, anticipate that the field will move towards real-time algorithms which require systems that consume less energy in comparison to modern GPUs. To achieve this, more efforts are required within model compression, for these models to be embedded into small devices during data collection or deployment in the field.

Automatic detection holds large opportunities for advances in the field of conservation and welfare, and drawing on the domain knowledge of biologists not currently involved in bioacoustics can open up new research directions. The advantages of processing large amounts of acoustic data seem clear to those currently involved in the field, but the wider biological community should be involved to find new fundamental research questions in the field of ecology and evolution (Clutton-Brock & Sheldon, 2010; De Frenne *et al.*, 2018), for example around species occurrence (Sebastián-González *et al.*, 2015; Rice *et al.*, 2021; Sattar, 2023), population density (Marques *et al.*, 2013b) and diversity (Kotera & Phillott, 2022), habitat use (Brookes, Bailey & Thompson, 2013; Kotila *et al.*, 2023), phenology (Dede *et al.*, 2014; Monczak *et al.*, 2017), and the early detection of invasive species (Juanes, 2018). Such questions offer opportunities for research into major conservation challenges like biodiversity loss and the effects of climate change (Sugai & Llusia, 2019; Ross *et al.*, 2023). Presently, studies driven by existing bioacoustics practitioners mostly focus on occurrence, or spatial or temporal distribution of a single species, whereas the advancement of automatic detection potentially allows for a focus on multiple species and to map biodiversity and potentially the functioning of whole ecosystems (Ross *et al.*, 2018).

Another example of how biologists and ecologists can steer the direction in which automatic detection may be developed in the future is to identify research questions without current technological solutions. For example, although detecting signs of poor animal welfare in captivity has been the subject of many studies (Zhang *et al.*, 2022; Mao *et al.*, 2022), there are comparably very few studies investigating the of wild animals (Mcloughlin *et al.*, 2019). This is surprising given the great potential acoustic monitoring of threatened species could provide, for example on species' reproduction, or social behaviour (Teixeira, Maron & Rensburg, 2019; Greggor *et al.*, 2021).

7. CONCLUSIONS

7.1. Need for AD

Automatic detection is no longer an optional capability in bioacoustics. Increasing data volumes, the need for near real-time analysis, and the expanding range of questions that biologists want to answer using passive acoustics mean that opening up the capabilities of this promising technology require parallel new developments in the field of machine learning.

7.2. Cooperation between disciplines

Mature fields in machine learning, such as image or voice recognition, are not immediately transferrable to automatic detection in bioacoustics. Close cooperation between biologist practitioners and machine learning developers will help advance solution creation by (a) enabling developers with an understanding of the problems facing bioacoustics practitioners, and (b) inform biologists what can and cannot be provided by the state of the art in machine learning.

7.3. Deep neural networks

Despite this, impressive advances in machine learning, particularly deep neural networks, hold out the potential for very significant developments that would cut processing time and enable a new wave of bioacoustics applications.

7.4. Development pipelines

Application development pipelines are of necessity problem-specific, however, certain guidelines and workflows (Section 4) should smooth the integration of solutions constrained both by the biological features of the problem, and by the available machine learning capabilities.

8. ACKNOWLEDGEMENTS

This paper arose from an investigative workshop, “Automatic detection for bioacoustics”, organised by JD and AK, supported with funding from the Cambridge Centre for Data-Driven Discovery and Accelerate Programme for Scientific Discovery, made possible by a donation from Schmidt Futures.

ED is supported by a research chairship from the African Institute for Mathematical Sciences South Africa. ED's work was carried out with the aid of a grant from the International Development Research Centre, Ottawa, Canada, www.idrc.ca, and with financial support from the Government of Canada, provided through Global Affairs Canada (GAC), www.international.gc.ca.

AM is supported by NERC NE/W005468/1. We thank the Agence Nationale de la Recherche (ANR) for funding RM grant ANR-20-CE23-0012-01 MIM, and PB's postdoc grant ANR-

20-CHIA-0014. PB and JC's participation in the workshop were funded by the Institute of Convergence ILCB (ANR-16-CONV-0002), which has benefited from the support of the French government (France 2030) and is managed by the Excellence Initiative of Aix-Marseille University (A*MIDEX).

9. BIBLIOGRAPHY

Abrahams, C. & Geary, M. (2020) Combining bioacoustics and occupancy modelling for improved monitoring of rare breeding bird populations. *Ecological Indicators* **112**, 106131.

Acevedo, M.A., Corrada-Bravo, C.J., Corrada-Bravo, H., Villanueva-Rivera, L.J. & Aide, T.M. (2009) Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics* **4**, 206–214. Elsevier B.V.

Aide, T.M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G. & Alvarez, R. (2013) Real-time bioacoustics monitoring and automated species identification. *PeerJ* **1**, e103.

Akamatsu, T., Ura, T., Sugimatsu, H., Bahl, R., Behera, S., Panda, S., Khan, M., Kar, S.K., Kar, C.S., Kimura, S. & Sasaki-Yamamoto, Y. (2013) A multimodal detection model of dolphins to estimate abundance validated by field experiments. *The Journal of the Acoustical Society of America* **134**, 2418–2426.

Alain, G. & Bengio, Y. (2018) Understanding intermediate layers using linear classifier probes. arXiv. http://arxiv.org/abs/1610.01644 [accessed 5 July 2023].

ANON. (2023a) Kaleidoscope Pro Analysis Software. *Wildlife Acoustics*. https://www.wildlifeacoustics.com/products/kaleidoscope-pro [accessed 12 August 2023].

ANON. (2023b) BTO Acoustic Pipeline. *BTO - British Trust for Ornithology*. https://www.bto.org/our-science/products-and-technologies/bto-acoustic-pipeline [accessed 12 August 2023].

AODHA, O.M., GIBB, R., BARLOW, K.E., BROWNING, E., FIRMAN, M., FREEMAN, R., HARDER, B., KINSEY, L., MEAD, G.R., NEWSON, S.E., PANDOURSKI, I., PARSONS, S., RUSS, J., SZODORAY-PARADI, A., SZODORAY-PARADI, F., ET AL. (2018) Bat detective—Deep learning tools for bat acoustic signal detection. *PLOS Computational Biology* **14**, e1005995. Public Library of Science.

BAEVSKI, A., ZHOU, Y., MOHAMED, A. & AULI, M. (2020) wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems (NeurIPS)* **33**, 12449–12460.

BAILEY, L.L., MACKENZIE, D.I. & NICHOLS, J.D. (2014) Advances and applications of occupancy models. *Methods in Ecology and Evolution* **5**, 1269–1279.

BARKER, D.J., HERRERA, C. & WEST, M.O. (2014) Automated detection of 50-kHz ultrasonic vocalizations using template matching in XBAT. *Journal of Neuroscience Methods* **236**, 68–75.

BEE, M.A. (2012) Sound source perception in anuran amphibians. *Current Opinion in Neurobiology* **22**, 301–310.

Bergler, C., Barnhill, A., Perrin, D., Schmitt, M., Maier, A. & Nöth, E. (2022a) ORCA-WHISPER: An Automatic Killer Whale Sound Type Generation Toolkit Using Deep Learning. In *Interspeech 2022* pp. 2413–2417. ISCA.

Bergler, C., Schmitt, M., Maier, A., Smeele, S., Barth, V. & Nöth, E. (2020) ORCA-CLEAN: A Deep Denoising Toolkit for Killer Whale Communication. In *Interspeech 2020* pp. 1136–1140. ISCA.

Bergler, C., Smeele, S.Q., Tyndel, S.A., Barnhill, A., Ortiz, S.T., Kalan, A.K., Cheng, R.X., Brinkløv, S., Osiecka, A.N., Tougaard, J., Jakobsen, F., Wahlberg, M., Nöth, E., Maier, A. & Klump, B.C. (2022b) ANIMAL-SPOT enables animal-independent signal detection and classification using deep learning. *Scientific Reports* **12**, 21966. Nature Publishing Group.

Bermant, P.C. (2021) BioCPPNet: automatic bioacoustic source separation with deep neural networks. *Scientific Reports* **11**, 23502. Nature Publishing Group.

Bermant, P.C., Bronstein, M.M., Wood, R.J., Gero, S. & Gruber, D.F. (2019) Deep Machine Learning Techniques for the Detection and Classification of Sperm Whale Bioacoustics. *Scientific Reports* **9**, 12588. Nature Publishing Group.

Best, P., Ferrari, M., Poupard, M., Paris, S., Marxer, R., Symonds, H., Spong, P. & Glotin, H. (2020) Deep Learning and Domain Transfer for Orca Vocalization Detection. In *2020 International Joint Conference on Neural Networks (IJCNN)* pp. 1–7.

Best, P., Marxer, R., Paris, S. & Glotin, H. (2022) Temporal evolution of the Mediterranean fin whale song. *Scientific Reports* **12**, 13565. Nature Publishing Group.

Best, P., Paris, S., Glotin, H. & Marxer, R. (2023) Deep audio embeddings for vocalisation clustering. *PLOS ONE* **18**, e0283396. Public Library of Science.

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Chang-qing, D., Clark, N.E., O'Connor, K. & Mace, G.M. (2010) Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. *PLOS Biology* **8**, e1000385. Public Library of Science.

Boersma, P. & Weenink, D. (2007) PRAAT: Doing phonetics by computer (Version 5.3.51).

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., et al. (2022) On the Opportunities and Risks of Foundation Models. arXiv. http://arxiv.org/abs/2108.07258 [accessed 7 August 2023].

Bøttcher, A., Gero, S., Beedholm, K., Whitehead, H. & Madsen, P.T. (2018) Variability of the inter-pulse interval in sperm whale clicks with implications for size estimation and individual identification. *The Journal of the Acoustical Society of America* **144**, 365–374.

Bouchet, H., Blois-Heulin, C. & Lemasson, A. (2013) Social complexity parallels vocal complexity: a comparison of three non-human primate species. *Frontiers in Psychology* **4**.

Braga, P.H.P., Hébert, K., Hudgins, E.J., Scott, E.R., Edwards, B.P.M., Sánchez Reyes, L.L., Grainger, M.J., Foroughirad, V., Hillemann, F., Binley, A.D., Brookson, C.B., Gaynor, K.M., Shafiei Sabet, S., Güncan, A., Weierbach, H., et al. (2023) Not just for programmers: How GitHub can accelerate collaborative and reproducible research in ecology and evolution. *Methods in Ecology and Evolution* **14**, 1364–1380.

BROLL, B. & WHITAKER, J. (2017) DeepForge: An open source, collaborative environment for reproducible deep learning.

BROOKES, K.L., BAILEY, H. & THOMPSON, P.M. (2013) Predictions from harbor porpoise habitat association models are confirmed by long-term passive acoustic monitoring. *The Journal of the Acoustical Society of America* **134**, 2523–2533.

BROWNING, E., GIBB, R., GLOVER-KAPFER, P. & JONES, K.E. (2017) Passive acoustic monitoring in ecology and conservation. Report, WWF-UK.

BURGOS, G. & ZUBEROGOITIA, I. (2020) A telemetry study to discriminate between home range and territory size in Tawny Owls. *Bioacoustics* **29**, 109–121. Taylor & Francis.

BUXTON, R.T., MCKENNA, M.F., CLAPP, M., MEYER, E., STABENAU, E., ANGELONI, L.M., CROOKS, K. & WITTEMYER, G. (2018) Efficacy of extracting indices from large-scale acoustic recordings to monitor biodiversity. *Conservation Biology* **32**, 1174–1184.

CANNAM, C., LANDONE, C. & SANDLER, M. (2010) Sonic visualiser: an open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the 18th ACM international conference on Multimedia* pp. 1467–1468. Association for Computing Machinery, New York, NY, USA.

CASAER, J., MILOTIC, T., LIEFTING, Y., DESMET, P. & JANSEN, P. (2019) Agouti: A platform for processing and archiving of camera trap images. *Biodiversity Information Science and Standards* **3**, e46690.

CASTELLOTE, M. & FOSSA, F. (2006) Measuring acoustic activity as a method to evaluate welfare in captive beluga whales (Delphinapterus leucas). *Aquatic Mammals* **32**, 325–333.

Clark, F.E. & Dunn, J.C. (2022) From Soundwave to Soundscape: A Guide to Acoustic Research in Captive Animal Environments. *Frontiers in Veterinary Science* **9**.

Clark, M.L., Salas, L., Baligar, S., Quinn, C.A., Snyder, R.L., Leland, D., Schackwitz, W., Goetz, S.J. & Newsam, S. (2023) The effect of soundscape composition on bird vocalization classification in a citizen science biodiversity monitoring project. *Ecological Informatics* **75**, 102065.

Clink, D.J., Kier, I., Ahmad, A.H. & Klinck, H. (2023) A workflow for the automated detection and classification of female gibbon calls from long-term acoustic recordings. *Frontiers in Ecology and Evolution* **11**.

Clutton-Brock, T. & Sheldon, B.C. (2010) Individuals and populations: the role of long-term, individual-based studies of animals in ecology and evolutionary biology. *Trends in Ecology & Evolution* **25**, 562–573.

Çoban, E.B., Pir, D., So, R. & Mandel, M.I. (2020) Transfer Learning from Youtube Soundtracks to Tag Arctic Ecoacoustic Recordings. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 726–730.

Coffey, K.R., Marx, R.E. & Neumaier, J.F. (2019) DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. *Neuropsychopharmacology* **44**, 859–868.

Cohen, Y., Nicholson, D.A., Sanchioni, A., Mallaber, E.K., Skidanova, V. & Gardner, T.J. (2022) Automated annotation of birdsong with a neural network that segments spectrograms. *eLife* **11**, e63853. eLife Sciences Publications, Ltd.

Cole, J.S., Michel, N.L., Emerson, S.A. & Siegel, R.B. (2022) Automated bird sound classifications of long-duration recordings produce occupancy model outputs similar to manually annotated data. *Ornithological Applications* **124**, duac003.

Cuevas, A., Veragua, A., Español-Jiménez, S., Chiang, G. & Tobar, F. (2017) Unsupervised blue whale call detection using multiple time-frequency features. In *2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)* pp. 1–6.

Darwin Core Task Group (2009) Darwin Core. https://www.tdwg.org/standards/dwc/ [accessed 5 July 2023].

Davis, J. & Goadrich, M. (2006) The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* pp. 233–240. Association for Computing Machinery, New York, NY, USA.

Dawson, D.K. & Efford, M.G. (2009) Bird population density estimated from acoustic signals. *Journal of Applied Ecology* **46**, 1201–1209. WILEY-BLACKWELL.

De Frenne, P., Van Langenhove, L., Van Driessche, A., Bertrand, C., Verheyen, K. & Vangansbeke, P. (2018) Using archived television video footage to quantify phenology responses to climate change. *Methods in Ecology and Evolution* **9**, 1874–1882.

Dede, A., Öztürk, A.A., Akamatsu, T., Tonay, A.M. & Öztürk, B. (2014) Long-term passive acoustic monitoring revealed seasonal and diel patterns of cetacean presence in the Istanbul Strait. *Journal of the Marine Biological Association of the United Kingdom* **94**, 1195–1202.

DENTON, T., WISDOM, S. & HERSHEY, J.R. (2021) Improving Bird Classification with Unsupervised Sound Separation. arXiv. http://arxiv.org/abs/2110.03209 [accessed 6 July 2023].

DERRICKSON, K.C. (1988) Variation in Repertoire Presentation in Northern Mockingbirds. *The Condor* **90**, 592–606.

DUAN, S., ZHANG, J., ROE, P., WIMMER, J., DONG, X., TRUSKINGER, A. & TOWSEY, M. (2013) Timed Probabilistic Automaton: A Bridge between Raven and Song Scope for Automatic Species Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence* **27**, 1519–1524.

DUFOURQ, E., BATIST, C., FOQUET, R. & DURBACH, I. (2022a) Passive acoustic monitoring of animal populations with transfer learning. *Ecological Informatics* **70**, 101688.

DUFOURQ, E., BATIST, C., FOQUET, R. & DURBACH, I. (2022b) Passive acoustic monitoring of animal populations with transfer learning. *Ecological Informatics* **70**, 101688.

DUFOURQ, E., DURBACH, I., HANSFORD, J.P., HOEPFNER, A., MA, H., BRYANT, J.V., STENDER, C.S., LI, W., LIU, Z., CHEN, Q., ZHOU, Z. & TURVEY, S.T. (2021) Automated detection of Hainan gibbon calls for passive acoustic monitoring. *Remote Sensing in Ecology and Conservation* **7**, 475–487.

DUNN, J.C. & SMAERS, J.B. (2018) Neural Correlates of Vocal Repertoire in Primates. *Frontiers in Neuroscience* **12**.

ENARI, H., ENARI, H.S., OKUDA, K., MARUYAMA, T. & OKUDA, K.N. (2019) An evaluation of the efficiency of passive acoustic monitoring in detecting deer and primates in comparison with camera traps. *Ecological Indicators* **98**, 753–762.

Erbe, C. & Thomas, J.A. (eds) (2022) *Exploring Animal Behavior Through Sound: Volume 1: Methods*. Springer Nature.

Exadaktylos, V., Silva, M., Aerts, J.-M., Taylor, C.J. & Berckmans, D. (2008) Real-time recognition of sick pig cough sounds. *Computers and Electronics in Agriculture* **63**, 207–214.

Fairbrass, A.J., Firman, M., Williams, C., Brostow, G.J., Titheridge, H. & Jones, K.E. (2019) CityNet—Deep learning tools for urban ecoacoustic assessment. *Methods in Ecology and Evolution* **10**, 186–197.

Feddersen-Petersen, D.U. (2000) Vocalization of European wolves (<i>Canis lupus lupus</i> L.) and various dog breeds (<i>Canis lupus</i> f. fam.). *Archives Animal Breeding* **43**, 387–398.

Ferrari, M., Glotin, H., Marxer, R. & Asch, M. (2020) DOCC10: Open access dataset of marine mammal transient studies and end-to-end CNN classification. In *2020 International Joint Conference on Neural Networks (IJCNN)* pp. 1–8.

Ford, J.K.B. (1991) Vocal traditions among resident killer whales (Orcinus orca) in coastal waters of British Columbia. *Canadian Journal of Zoology* **69**, 1454–1483. NRC Research Press.

Fox, E.J.S., Roberts, J.D. & Bennamoun, M. (2008) Call-Independent Individual Identification in Birds. *Bioacoustics* **18**, 51–67. Taylor & Francis.

Frick, W.F. (2013) Acoustic monitoring of bats, considerations of options for long-term monitoring. *Therya* **4**, 69–70. Centro de Investigaciones Biológicas del Noroeste.

Frommolt, K.-H. & Tauchert, K.-H. (2014) Applying bioacoustic methods for long-term monitoring of a nocturnal wetland bird. *Ecological Informatics* **21**, 4–12.

Ganchev, T.D. (2020) Chapter 8 - Ubiquitous computing and biodiversity monitoring. In *Advances in Ubiquitous Computing* (ed A. Neustein), pp. 239–259. Academic Press.

Gareth James, Daniela Witten, Trevor Hastie, & Robert Tibshirani (2013) *An Introduction to Statistical Learning : With Applications in R*. Springer, New York, NY.

Garland, E.C., Castellote, M. & Berchok, C.L. (2015) Beluga whale (Delphinapterus leucas) vocalizations and call classification from the eastern Beaufort Sea population. *The Journal of the Acoustical Society of America* **137**, 3054–3067.

Garland, E.C., Goldizen, A.W., Rekdahl, M.L., Constantine, R., Garrigue, C., Hauser, N.D., Poole, M.M., Robbins, J. & Noad, M.J. (2011) Dynamic Horizontal Cultural Transmission of Humpback Whale Song at the Ocean Basin Scale. *Current Biology* **21**, 687–691. Elsevier.

GBIF/TDWG Multimedia Resources Task Group (2013) Audiovisual Core Multimedia Resources Metadata Schema. https://www.tdwg.org/standards/ac/ [accessed 5 July 2023].

Gibbon, D., Moore, R. & Winski, R. (eds) (1998) *Vol 1 Spoken Language System and Corpus Design*. De Gruyter Mouton, Berlin, Boston.

Gillespie, D., Gordon, J., McHugh, R., McLaren, D., Mellinger, D.K., Redmond, P., Thode, A., Trinder, P. & Deng, X.Y. (2009) PAMGUARD: Semiautomated, open source software for real-time acoustic detection and localisation of cetaceans. In *The*

Journal of the Acoustical Society of America pp. 2547–2547. Acoustical Society of AmericaASA.

Green, A., Clark, C., Favaro, L., Lomax, S. & Reby, D. (2019) Vocal individuality of Holstein-Friesian cattle is maintained across putatively positive and negative farming contexts. *Scientific Reports* **9**. Nature Research.

Greggor, A.L., Masuda, B., Gaudioso-Levita, J.M., Nelson, J.T., White, T.H., Shier, D.M., Farabaugh, S.M. & Swaisgood, R.R. (2021) Pre-release training, predator interactions and evidence for persistence of anti-predator behavior in reintroduced `alalā, Hawaiian crow. *Global Ecology and Conservation* **28**, e01658.

Hagiwara, M. (2022) AVES: Animal Vocalization Encoder based on Self-Supervision. arXiv. http://arxiv.org/abs/2210.14493 [accessed 6 July 2023].

Hagiwara, M., Hoffman, B., Liu, J.-Y., Cusimano, M., Effenberger, F. & Zacarian, K. (2022) BEANS: The Benchmark of Animal Sounds. arXiv. http://arxiv.org/abs/2210.12300 [accessed 6 July 2023].

Hansen, P. (1979) Vocal learning: Its role in adapting sound structures to long-distance propagation, and a hypothesis on its evolution. *Animal Behaviour* **27**, 1270–1271. Elsevier Science, Netherlands.

Harrington, F.H. & Mech, L.D. (1982) An Analysis of Howling Response Parameters Useful for Wolf Pack Censusing. *The Journal of Wildlife Management* **46**, 686–693. Allen Press.

Haupert, S., Sèbe, F. & Sueur, J. (2022) Physics-based model to predict the acoustic detection distance of terrestrial autonomous recording units over the diel cycle and across seasons: Insights from an Alpine and a Neotropical forest. *Methods in Ecology and Evolution*. Wiley.

Heaphy, K. & Cain, K. (2021) Song variation between sexes and among subspecies of New Zealand Fantail (Rhipidura fuliginosa). *Emu - Austral Ornithology* **121**, 198–210. Taylor & Francis.

Heath, B.E., Sethi, S.S., Orme, C.D.L., Ewers, R.M. & Picinali, L. (2021) How index selection, compression, and recording schedule impact the description of ecological soundscapes. *Ecology and Evolution* **11**, 13206–13217.

Hebets, E.A. (2005) Attention-altering signal interactions in the multimodal courtship display of the wolf spider Schizocosa uetzi. *Behavioral Ecology* **16**, 75–82.

Hebets, E.A., Bern, M., McGinley, R.H., Roberts, A., Kershenbaum, A., Starrett, J. & Bond, J.E. (2021) Sister species diverge in modality-specific courtship signal form and function. *Ecology and Evolution* **11**, 852–871.

Hedley, R.W., Wilson, S.J., Yip, D.A., Li, K. & Bayne, E.M. (2021) Distance truncation via sound level for bioacoustic surveys in patchy habitat. *Bioacoustics* **30**, 303–323.

Hendry, H. & Mann, C. (2018) Camelot—intuitive software for camera-trap data management. *Oryx* **52**, 15–15.

Hildebrand, J.A., Frasier, K.E., Helble, T.A. & Roch, M.A. (2022) Performance metrics for marine mammal signal detection and classification. *The Journal of the Acoustical Society of America* **151**, 414–427.

Hill, A.P., Prince, P., Snaddon, J.L., Doncaster, C.P. & Rogers, A. (2019) AudioMoth: A low-cost acoustic device for monitoring biodiversity and the environment. *HardwareX* **6**, e00073.

Hogeweg, L. & Stowell, D. (2023) An API for AI species recognition. *Arise*. https://www.arise-biodiversity.nl/post/an-api-for-ai-species-recognition [accessed 4 July 2023].

Hood, J.D., Flogeras, D.G. & Theriault, J.A. (2016) Improved passive acoustic band-limited energy detection for cetaceans. *Applied Acoustics* **106**, 36–41.

Hsu, W.-N., Bolte, B., Tsai, Y.-H.H., Lakhotia, K., Salakhutdinov, R. & Mohamed, A. (2021) HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. arXiv. http://arxiv.org/abs/2106.07447 [accessed 6 July 2023].

Humphrey, E.J., Salamon, J., Nieto, O., Forsyth, J., Bittner, R.M. & Bello, J.P. (2014) JAMS: A JSON ANNOTATED MUSIC SPECIFICATION FOR REPRODUCIBLE MIR RESEARCH.

Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A. & Weyde, T. (2017) Singing voice separation with deep U-Net convolutional networks. conference, Suzhou, China. https://ismir2017.smcnus.org/ [accessed 6 July 2023].

Jiang, J., Bu, L., Wang, X., Li, C., Sun, Z., Yan, H., Hua, B., Duan, F. & Yang, J. (2018) Clicks classification of sperm whale and long-finned pilot whale based on continuous wavelet transform and artificial neural network. *Applied Acoustics* **141**, 26–34.

Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.-P., Planqué, R., Rauber, A., Palazzo, S., Fisher, B. & Müller, H. (2015) LifeCLEF 2015: Multimedia Life Species Identification Challenges. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (eds J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. Jones, E. San Juan, L. Capellato & N. Ferro), pp. 462–483. Springer International Publishing, Cham.

Juanes, F. (2018) Visual and acoustic sensors for early detection of biological invasions: Current uses and future potential. *Journal for Nature Conservation* **42**, 7–11.

K. Lisa Yang Center for Conservation Bioacoustics (2014) Bioacoustics Research Program. https://ravensoundsoftware.com/ [accessed 5 July 2023].

Kahl, S., Wood, C.M., Eibl, M. & Klinck, H. (2021) BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics* **61**, 101236.

Karnan, M., Akila, M. & Krishnaraj, N. (2011) Biometric personal authentication using keystroke dynamics: A review. *Applied Soft Computing* **11**, 1565–1573.

Kershenbaum, A., Blumstein, D.T., Roch, M.A., Akçay, Ç., Backus, G., Bee, M.A., Bohn, K., Cao, Y., Carter, G., Cäsar, C., Coen, M., DeRuiter, S.L., Doyle, L., Edelman, S., Ferrer-i-Cancho, R., et al. (2016a) Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biological Reviews* **91**, 13–52.

Kershenbaum, A., Demartsev, V., Gammon, D.E., Geffen, E., Gustison, M.L., Ilany, A. & Lameira, A.R. (2021) Shannon entropy as a robust estimator of Zipf's Law in animal vocal communication repertoires. *Methods in Ecology and Evolution* **12**, 553–564.

Kershenbaum, A., Ilany, A., Blaustein, L. & Geffen, E. (2012) Syntactic structure and geographical dialects in the songs of male rock hyraxes. *Proceedings of the Royal Society B: Biological Sciences* **279**, 2974–2981. Royal Society.

Kershenbaum, A., Owens, J.L. & Waller, S. (2019) Tracking cryptic animals using acoustic multilateration: A system for long-range wolf detection. *The Journal of the Acoustical Society of America* **145**, 1619–1628.

Kershenbaum, A. & Roch, M.A. (2013) An image processing based paradigm for the extraction of tonal sounds in cetacean communications. *The Journal of the Acoustical Society of America* **134**, 4435–4445.

Kershenbaum, A., Root-Gutteridge, H., Habib, B., Koler-Matznick, J., Mitchell, B., Palacios, V. & Waller, S. (2016b) Disentangling canid howls across multiple species and subspecies: Structure in a complex communication channel. *Behavioural Processes* **124**, 149–157.

Kershenbaum, A., Sayigh, L.S. & Janik, V.M. (2013) The Encoding of Individual Identity in Dolphin Signature Whistles: How Much Information Is Needed? *PLOS ONE* **8**, e77671. Public Library of Science.

Kimura, S., Akamatsu, T., Li, S., Dong, S., Dong, L., Wang, K., Wang, D. & Arai, N. (2010) Density estimation of Yangtze finless porpoises using passive acoustic sensors and automated click train detection. *The Journal of the Acoustical Society of America* **128**, 1435.

Knight, E.C. & Bayne, E.M. (2019) Classification threshold and training data affect the quality and utility of focal species data processed with automated audio-recognition software. *Bioacoustics* **28**, 539–554.

Kotera, M.M. & Phillott, A.D. (2022) Calls for conservation: A review of bioacoustics monitoring with case studies from India. *Asian Journal of Environment & Ecology*, 142–150.

Kotila, M., Suominen, K.M., Vasko, V.V., Blomberg, A.S., Lehikoinen, A., Andersson, T., Aspi, J., Cederberg, T., Hänninen, J., Inkinen, J., Koskinen, J., Lundberg, G., Mäkinen, K., Rontti, M., Snickars, M., et al. (2023) Large-scale long-term passive-acoustic monitoring reveals spatio-temporal activity patterns of boreal bats. *Ecography* **2023**, e06617.

Krause, B.L. (1993) The niche hypothesis: a virtual symphony of animal sounds, the origins of musical expression and the health of habitats. *The Soundscape Newsletter* **6**, 6–10.

Laiolo, P., Rolando, A., Delestrade, A. & Sanctis, A. de (2001) GEOGRAPHICAL VARIATION IN THE CALLS OF THE CHOUGHS. *The Condor* **103**, 287–297. American Ornithological Society.

Laurijs, K.A., Briefer, E.F., Reimert, I. & Webb, L.E. (2021) Vocalisations in farm animals: A step towards positive welfare assessment. *Applied Animal Behaviour Science* **236**, 105264.

Law, B., Gonsalves, L., Burgar, J., Brassil, T., Kerr, I., Wilmott, L., Madden, K., Smith, M., Mella, V., Crowther, M., Krockenberger, M., Rus, A., Pietsch, R., Truskinger, A., Eichinski, P., et al. (2021) Estimating and validating koala. *Wildlife Research* **49**, 438–448.

LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *Nature* **521**, 436–444.

Leighton, G.M. & Birmingham, T. (2021) Multiple factors affect the evolution of repertoire size across birds. *Behavioral Ecology* **32**, 380–385.

Leroux, M., Al-Khudhairy, O.G., Perony, N. & Townsend, S.W. (2021) Chimpanzee voice prints? Insights from transfer learning experiments from human voices. arXiv. http://arxiv.org/abs/2112.08165 [accessed 6 July 2023].

Lever, J., Krzywinski, M. & Altman, N. (2016) Classification evaluation. *Nature Methods* **13**, 603–604. Nature Publishing Group.

Li, N., Ren, Z., Li, D. & Zeng, L. (2020) Automated techniques for monitoring the behaviour and welfare of broilers and laying hens: towards the goal of precision livestock farming. *Animal* **14**, 617–625.

Liebal, K. & Oña, L. (2018) Different Approaches to Meaning in Primate Gestural and Vocal Communication. *Frontiers in Psychology* **9**.

Lin, T., Wang, Y., Liu, X. & Qiu, X. (2022) A survey of transformers. *AI Open* **3**, 111–132.

Long, R.A. (2008) *Noninvasive survey methods for carnivores*. In p. 385. Island Press.

Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S. & Bello, J.P. (2018) Birdvox-Full-Night: A Dataset and Benchmark for Avian Flight Call Detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 266–270.

Maaten, L. van der & Hinton, G. (2008) Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605.

MacKenzie, D.I., Nichols, J.D., Hines, J.E., Knutson, M.G. & Franklin, A.B. (2003) Estimating Site Occupancy, Colonization, and Local Extinction When a Species Is Detected Imperfectly. *Ecology* **84**, 2200–2207.

Manser, M.B., Jansen, D.A.W.A.M., Graw, B., Hollén, L.I., Bousquet, C.A.H., Furrer, R.D. & le Roux, A. (2014) Chapter Six - Vocal Complexity in Meerkats and Other Mongoose Species. In *Advances in the Study of Behavior* (eds M. Naguib, L. Barrett, H.J. Brockmann, S. Healy, J.C. Mitani, T.J. Roper & L.W. Simmons), pp. 281–310. Academic Press.

Manteuffel, G., Puppe, B. & Schön, P.C. (2004) Vocalization of farm animals as a measure of welfare. *Applied Animal Behaviour Science* **88**, 163–182.

Mao, A., Giraudet, C.S.E., Liu, K., De Almeida Nolasco, I., Xie, Z., Xie, Z., Gao, Y., Theobald, J., Bhatta, D., Stewart, R. & McElligott, A.G. (2022) Automated identification of chicken distress vocalizations using deep learning models. *Journal of The Royal Society Interface* **19**, 20210921.

Marin-Cudraz, T., Muffat-Joly, B., Novoa, C., Aubry, P., Desmet, J.-F., Mahamoud-Issa, M., Nicolè, F., Van Niekerk, M.H., Mathevon, N. & Sèbe, F. (2019) Acoustic monitoring of rock ptarmigan: A multi-year comparison with point-count protocol. *Ecological Indicators* **101**, 710–719.

Marques, T.A., Thomas, L., Martin, S.W., Mellinger, D.K., Ward, J.A., Moretti, D.J., Harris, D. & Tyack, P.L. (2013a) Estimating animal population density using passive acoustics. *Biological reviews of the Cambridge Philosophical Society* **88**, 287–309.

MARQUES, T.A., THOMAS, L., MARTIN, S.W., MELLINGER, D.K., WARD, J.A., MORETTI, D.J., HARRIS, D. & TYACK, P.L. (2013b) Estimating animal population density using passive acoustics. *Biological Reviews* **88**, 287–309.

MARTIN, K., ADAM, O., OBIN, N. & DUFOUR, V. (2022) Rookognise: Acoustic detection and identification of individual rooks in field recordings using multi-task neural networks. *Ecological Informatics* **72**, 101818.

MCCOMB, K. & SEMPLE, S. (2005) Coevolution of vocal communication and sociality in primates. *Biology Letters* **1**, 381–385. Royal Society.

MCDONALD, M., HILDEBRAND, J. & MESNICK, S. (2009) Worldwide decline in tonal frequencies of blue whale songs. *Endangered Species Research* **9**, 13–21.

MCDONALD, M.A. & FOX, C.G. (1999) Passive acoustic methods applied to fin whale population density estimation. *Journal of the Acoustical Society of America* **105**, 2643–2651. ACOUSTICAL SOC AMER AMER INST PHYSICS.

MCINNES, L., HEALY, J. & MELVILLE, J. (2020) UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv. http://arxiv.org/abs/1802.03426 [accessed 7 August 2023].

MCLOUGHLIN, M.P., STEWART, R. & MCELLIGOTT, A.G. (2019) Automated bioacoustics: methods in ecology and conservation and their potential for animal welfare monitoring. *Journal of The Royal Society Interface* **16**, 20190225.

MENNILL, D.J., BATTISTON, M., WILSON, D.R., FOOTE, J.R. & DOUCET, S.M. (2012) Field test of an affordable, portable, wireless microphone array for spatial monitoring of animal ecology and behaviour. *Methods in Ecology and Evolution* **3**, 704–712.

Metcalf, O., Abrahams, C., Ashington, B., Baker, E., Bradfer-Lawrence, T., Browning, E., Carruthers-Jones, J., Darby, J., Dick, J., Eldridge, A., Elliott, D., Heath, B., Howden-Leach, P., Johnston, A., Lees, A., et al. (2023a) Good practice guidelines for long-term ecoacoustic monitoring in the UK. In pp. 1–82. Report, The UK Acoustics Network.

Metcalf, O., Abrahams, C., Ashington, B., Baker, E., Bradfer-Lawrence, T., Browning, E., Carruthers-Jones, J., Darby, J., Dick, J., Eldridge, A., & others (2023b) Good practice guidelines for long-term ecoacoustic monitoring in the UK. The UK Acoustics Network.

Meyer, D., Hodges, J.K., Rinaldi, D., Wijaya, A., Roos, C. & Hammerschmidt, K. (2012) Acoustic structure of male loud-calls support molecular phylogeny of Sumatran and Javanese leaf monkeys (genus Presbytis). *Bmc Evolutionary Biology* **12**, 16–16. BIOMED CENTRAL LTD.

Miller, B.S., Madhusudhana, S., Aulich, M.G. & Kelly, N. (2023) Deep learning algorithm outperforms experienced human observer at detection of blue whale D-calls: a double-observer analysis. *Remote Sensing in Ecology and Conservation* **9**, 104–116.

Ministry of Local Government and Modernisation (2021) Norwegian data centres - sustainable, digital powerhouses. Plan, regjeringen.no. *Government.no*. https://www.regjeringen.no/en/dokumenter/norwegian-data-centres-sustainable-digital-powerhouses/id2867155/ [accessed 7 August 2023].

Mitrovic, D., Zeppelzauer, M. & Breiteneder, C. (2006) Discrimination and retrieval of animal sounds. In *2006 12th International Multi-Media Modelling Conference* p. 5 pp.-.

Monczak, A., Berry, A., Kehrer, C. & Montie, E. (2017) Long-term acoustic monitoring of fish calling provides baseline estimates of reproductive timelines in the May River estuary, southeastern USA. *Marine Ecology Progress Series* **581**, 1–19.

Morrison, A. & Novikova, A. (2023) Monitoring technologies for animal welfare: A review of aspirations and deployments in zoos. In *Proceedings of the Future Technologies Conference (FTC) 2022, Volume 3* (ed K. Arai), pp. 155–178. Springer International Publishing, Cham.

Morrissey, R.P., Ward, J., DiMarzio, N., Jarvis, S. & Moretti, D.J. (2006) Passive acoustic detection and localization of sperm whales (Physeter macrocephalus) in the tongue of the ocean. *Applied Acoustics* **67**, 1091–1105.

Narasimhan, R., Fern, X.Z. & Raich, R. (2017) Simultaneous segmentation and classification of bird song using CNN. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 146–150.

Nelson, D.A. (2000) Song overproduction, selective attrition and song dialects in the white-crowned sparrow. *Animal Behaviour* **60**, 887–898.

Nguyen Hong Duc, P., Torterotot, M., Samaran, F., White, P.R., Gérard, O., Adam, O. & Cazau, D. (2021) Assessing inter-annotator agreement from collaborative annotation campaign in marine bioacoustics. *Ecological Informatics* **61**, 101185.

Nicholson, D. (2023) Crowsetta: A Python tool to work with any format for annotating animal vocalizations and bioacoustics data. *Journal of Open Source Software* **8**, 5338.

Nijman, V. (2007) Effects of vocal behaviour on abundance estimates of rainforest Galliforms. *Acta Ornithologica* **42**, 186–190.

Nolasco, I., Singh, S., Morfi, V., Lostanlen, V., Strandburg-Peshkin, A., Vidaña-Vila, E., Gill, L., Pamuła, H., Whitehead, H., Kiskin, I., Jensen, F.H., Morford, J., Emmerson, M.G., Versace, E., Grout, E., et al. (2023) Learning to detect an animal sound from five examples. arXiv. http://arxiv.org/abs/2305.13210 [accessed 5 July 2023].

Obrist, M.K., Pavan, G., Sueur, J., Riede, K., Llusia, D. & Márquez, R. (2010) Bioacoustics approaches in biodiversity inventories. *Manual on field recording techniques and protocols for all taxa biodiversity inventories* **8**, 68–99.

Odom, K.J., Araya-Salas, M., Morano, J.L., Ligon, R.A., Leighton, G.M., Taff, C.C., Dalziell, A.H., Billings, A.C., Germain, R.R., Pardo, M., de Andrade, L.G., Hedwig, D., Keen, S.C., Shiu, Y., Charif, R.A., et al. (2021) Comparative bioacoustics: a roadmap for quantifying and comparing animal sounds across diverse taxa. *Biological Reviews* **96**, 1135–1159.

Ogutu, J.O., Piepho, H.-P., Said, M.Y., Ojwang, G.O., Njino, L.W., Kifugo, S.C. & Wargute, P.W. (2016) Extreme Wildlife Declines and Concurrent Increase in Livestock Numbers in Kenya: What Are the Causes? *PLOS ONE* **11**, e0163249. Public Library of Science.

Oliver, R.Y., Ellis, D.P.W., Chmura, H.E., Krause, J.S., Pérez, J.H., Sweet, S.K., Gough, L., Wingfield, J.C. & Boelman, N.T. (2018) Eavesdropping on the Arctic:

Automated bioacoustics reveal dynamics in songbird breeding phenology. *Science Advances* **4**, eaaq1084.

Oswald, J.N., Erbe, C., Gannon, W.L., Madhusudhana, S. & Thomas, J.A. (2022) Detection and Classification Methods for Animal Sounds. In *Exploring Animal Behavior Through Sound: Volume 1: Methods* (eds C. Erbe & J.A. Thomas), pp. 269–317. Springer International Publishing, Cham.

Parrilla, A.G.A. & Stowell, D. (2022) Polyphonic sound event detection for highly dense birdsong scenes. arXiv. http://arxiv.org/abs/2207.06349 [accessed 6 July 2023].

Pérez-Granados, C. & Schuchmann, K.-L. (2021) Passive acoustic monitoring of the diel and annual vocal behavior of the Black and Gold Howler Monkey. *American Journal of Primatology* **83**, e23241.

Petso, T., Jamisola, R.S. & Mpoeleng, D. (2021) Review on methods used for wildlife species and individual identification. *European Journal of Wildlife Research* **68**, 3.

Politis, A., Mesaros, A., Adavanne, S., Heittola, T. & Virtanen, T. (2020) Overview and Evaluation of Sound Event Localization and Detection in DCASE 2019. *arXiv:2009.02792 [cs, eess]*.

Powell, R. (2000) Animal home ranges and territories and home range estimators. *Research Techniques in Animal Ecology: Controversies and Consequences*, 65–110.

Qian, T., Deng, G., Li, Y. & Yang, D. (2023) Description of the advertisement call of Boulenophrys nanlingensis (Anura, Megophryidae), with a case of individual identification using its dorsum pattern. *Herpetozoa* **36**, 123–128. Pensoft Publishers.

Rice, A., Širović, A., Trickey, J.S., Debich, A.J., Gottlieb, R.S., Wiggins, S.M., Hildebrand, J.A. & Baumann-Pickering, S. (2021) Cetacean occurrence in the Gulf of Alaska from long-term passive acoustic monitoring. *Marine Biology* **168**, 72.

Riesch, R., Barrett-Lennard, L.G., Ellis, G.M., Ford, J.K.B. & Deecke, V.B. (2012) Cultural traditions and the evolution of reproductive isolation: ecological speciation in killer whales? *Biological Journal of the Linnean Society* **106**, 1–17. WILEY-BLACKWELL.

Romero-Mujalli, D., Bergmann, T., Zimmermann, A. & Scheumann, M. (2021) Utilizing DeepSqueak for automatic detection and classification of mammalian vocalizations: a case study on primate vocalizations. *Scientific Reports* **11**, 24463. Nature Publishing Group.

Ross, S.R.P. -J., O'Connell, D.P., Deichmann, J.L., Desjonquères, C., Gasc, A., Phillips, J.N., Sethi, S.S., Wood, C.M. & Burivalova, Z. (2023) Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. *Functional Ecology* **37**, 959–975.

Ross, S.R.P.-J., Friedman, N.R., Dudley, K.L., Yoshimura, M., Yoshida, T. & Economo, E.P. (2018) Listening to ecosystems: data-rich acoustic monitoring through landscape-scale sensor networks. *Ecological Research* **33**, 135–147.

Rostro-García, S., Kamler, J.F., Sollmann, R., Balme, G., Augustine, B.C., Kéry, M., Crouthers, R., Gray, T.N.E., Groenenberg, M., Prum, S. & Macdonald, D.W. (2023) Population dynamics of the last leopard population of eastern Indochina in the context of improved law enforcement. *Biological Conservation* **283**, 110080.

Rothstein, S.I. & Fleischer, R.C. (1987) Vocal Dialects and Their Possible Relation to Honest Status Signalling in the Brown-Headed Cowbird. *The Condor* **89**, 1–23.

Saeed, A., Grangier, D. & Zeghidour, N. (2021) Contrastive Learning of General-Purpose Audio Representations. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 3875–3879.

Sainburg, T. (2019) timsainb/noisereduce: v1.0. Zenodo. https://zenodo.org/record/3243139 [accessed 5 July 2023].

Sainburg, T., Thielk, M. & Gentner, T.Q. (2020) Latent space visualization, characterization, and generation of diverse vocal communication signals. bioRxiv. https://www.biorxiv.org/content/10.1101/870311v2 [accessed 6 July 2023].

Sarkar, E. & -Doss, M.M. (2023) Can Self-Supervised Neural Representations Pre-Trained on Human Speech distinguish Animal Callers? arXiv. http://arxiv.org/abs/2305.14035 [accessed 6 July 2023].

Sattar, F. (2023) A new acoustical autonomous method for identifying endangered whale calls: A case study of blue whale and fin whale. *Sensors* **23**, 3048.

Schultze, S., Gruenefeld, U. & Boll, S. (2020) Demystifying deep learning: Developing a learning app for beginners to gain practical experience. *Proceedings of the Mensch und Computer 2020Workshop*.

Sebastián-González, E., Pang-Ching, J., Barbosa, J.M. & Hart, P. (2015) Bioacoustics for species management: two case studies with a Hawaiian forest bird. *Ecology and Evolution* **5**, 4696–4705.

Sethi, S.S., Jones, N.S., Fulcher, B.D., Picinali, L., Clink, D.J., Klinck, H., Orme, C.D.L., Wrege, P.H. & Ewers, R.M. (2020) Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set. *Proceedings of the National Academy of Sciences* **117**, 17049–17055. Proceedings of the National Academy of Sciences.

Sharpe, F., Bolton, M., Sheldon, R. & Ratcliffe, N. (2009) Effects of color banding, radio tagging, and repeated handling on the condition and survival of Lapwing chicks and consequences for estimates of breeding productivity. *Journal of Field Ornithology* **80**, 101–110.

Shiu, Y., Palmer, K.J., Roch, M.A., Fleishman, E., Liu, X., Nosal, E.-M., Helble, T., Cholewiak, D., Gillespie, D. & Klinck, H. (2020) Deep neural networks for automated detection of marine mammal species. *Scientific Reports* **10**, 607. Nature Publishing Group.

Slocombe, K.E., Waller, B.M. & Liebal, K. (2011) The language void: the need for multimodality in primate communication research. *Animal Behaviour* **81**, 919–924.

Smith, B.R., Root-Gutteridge, H., Butkiewicz, H., Dassow, A., Fontaine, A.C., Markham, A., Owens, J., Schindler, L., Wijers, M., Kershenbaum, A., Smith, B.R., Root-Gutteridge, H., Butkiewicz, H., Dassow, A., Fontaine, A.C., et al. (2021) Acoustic localisation of wildlife with low-cost equipment: lower sensitivity, but no loss of precision. *Wildlife Research* **49**, 372–381. CSIRO PUBLISHING.

Soulsbury, C.D., Gray, H.E., Smith, L.M., Braithwaite, V., Cotter, S.C., Elwood, R.W., Wilkinson, A. & Collins, L.M. (2020) The welfare and ethics of research involving wild animals: A primer. *Methods in Ecology and Evolution* **11**, 1164–1181.

Spillmann, B., van Schaik, C.P., Setia, T.M. & Sadjadi, S.O. (2017) Who shall I say is calling? Validation of a caller recognition procedure in Bornean flanged male orangutan (*Pongo pygmaeus wurmbii*) long calls. *Bioacoustics* **26**, 109–120.

Staaterman, E., Ogburn, M.B., Altieri, A.H., Brandl, S.J., Whippo, R., Seemann, J., Goodison, M. & Duffy, J.E. (2017) Bioacoustic measurements complement visual biodiversity surveys: preliminary evidence from four shallow marine habitats. *Marine Ecology Progress Series* **575**, 207–215.

Stowell, D. (2022a) Computational bioacoustics with deep learning: a review and roadmap. *PeerJ* **10**, e13152. PeerJ Inc.

Stowell, D. (2022b) Computational bioacoustics with deep learning: a review and roadmap. *PeerJ* **10**, e13152. PeerJ Inc.

Stowell, D., Wood, M.D., Pamuła, H., Stylianou, Y. & Glotin, H. (2019) Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge. *Methods in Ecology and Evolution* **10**, 368–380.

Sueur, J., Farina, A., Gasc, A., Pieretti, N. & Pavoine, S. (2014) Acoustic Indices for Biodiversity Assessment and Landscape Investigation. *Acta Acustica united with Acustica* **100**, 772–781.

Sugai, L.S.M., Desjonquères, C., Silva, T.S.F. & Llusia, D. (2020) A roadmap for survey designs in terrestrial acoustic monitoring. *Remote Sensing in Ecology and Conservation* **6**, 220–235.

Sugai, L.S.M. & Llusia, D. (2019) Bioacoustic time capsules: Using acoustic monitoring to document biodiversity. *Ecological Indicators* **99**, 149–152.

Teixeira, D., Maron, M. & Rensburg, B.J. (2019) Bioacoustic monitoring of animal vocal behavior for conservation. *Conservation Science and Practice* **1**.

Torti, V., Valente, D., De Gregorio, C., Comazzi, C., Miaretsoa, L., Ratsimbazafy, J., Giacoma, C. & Gamba, M. (2018) Call and be counted! Can we reliably estimate the number of callers in the indri's (Indri indri) song? *PLOS ONE* **13**, e0201664.

Towsey, M., Planitz, B., Nantes, A., Wimmer, J. & Roe, P. (2012) A toolbox for animal call recognition. *Bioacoustics* **21**, 107–125. Taylor & Francis.

Turian, J., Shier, J., Khan, H.R., Raj, B., Schuller, B.W., Steinmetz, C.J., Malloy, C., Tzanetakis, G., Velarde, G., McNally, K., Henry, M., Pinto, N., Noufi, C., Clough, C., Herremans, D., et al. (2022) HEAR 2021: Holistic Evaluation of Audio Representations. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track* (eds D. Kiela, M. Ciccone & B. Caputo), pp. 125–145. PMLR.

Uetz, G.W. & Roberts, J.A. (2002) Multisensory Cues and Multimodal Communication in Spiders: Insights from Video/Audio Playback Studies. *Brain Behavior and Evolution* **59**, 222–230.

Van Segbroeck, M., Knoll, A.T., Levitt, P. & Narayanan, S. (2017) MUPET—Mouse Ultrasonic Profile ExTraction: A Signal Processing Tool for Rapid and Unsupervised Analysis of Ultrasonic Vocalizations. *Neuron* **94**, 465-485.e5.

Velásquez, N.A., Marambio, J., Brunetti, E., Méndez, M.A., Vásquez, R.A. & Penna, M. (2013) Bioacoustic and genetic divergence in a frog with a wide geographical distribution. *Biological Journal of the Linnean Society* **110**, 142–155.

Venkatesh, S., Moffat, D. & Miranda, E.R. (2022) You Only Hear Once: A YOLO-like Algorithm for Audio Segmentation and Sound Event Detection. *Applied Sciences* **12**, 3293. Multidisciplinary Digital Publishing Institute.

Vickers, W., Milner, B., Risch, D. & Lee, R. (2021) Robust North Atlantic right whale detection using deep learning models for denoisinga). *The Journal of the Acoustical Society of America* **149**, 3797–3812.

Volodina, E.V. & Volodin, I.A. (1999) Bioacoustics in zoos: A review of applications and perspectives. *International Zoo News* **46**, 208–213.

Vu, T.T., Chi, T.N., Doherty Jr, P.F., Nguyen, H.T., Clink, D.J., Dac, M.N., Thanh, H.D. & Trong, T.G. (2023) Using mobile smartphones and bioacoustics to monitor endangered bird species. *Ibis*. Wiley Online Library.

Vu, T.T. & Tran, L.M. (2019) An Application of Autonomous Recorders for Gibbon Monitoring. *International Journal of Primatology* **40**, 169–186. Springer US.

Wang, Z., She, Q. & Ward, T.E. (2022) Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy. *ACM Computing Surveys* **54**, 1–38.

Wijers, M., Loveridge, A., Macdonald, D.W. & Markham, A. (2021) CARACAL: a versatile passive acoustic monitoring tool for wildlife research and conservation. *Bioacoustics* **30**, 41–57. Taylor & Francis.

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., et al. (2016) The FAIR Guiding

Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018. Nature Publishing Group.

Willacy, R.J., Mahony, M. & Newell, D.A. (2015) If a frog calls in the forest: Bioacoustic monitoring reveals the breeding phenology of the endangered Richmond Range mountain frog (Philoria richmondensis). *Austral Ecology* **40**, 625–633.

Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R.J., Wilson, K. & Hershey, J.R. (2020) Unsupervised Sound Separation Using Mixture Invariant Training. arXiv. http://arxiv.org/abs/2006.12701 [accessed 6 July 2023].

Wolf, C. & Ripple, W.J. (2016) Prey depletion as a threat to the world's large carnivores. *Royal Society Open Science* **3**, 160252. Royal Society.

Wood, C.M. & Peery, M.Z. (2022) What does 'occupancy' mean in passive acoustic surveys? *Ibis* **164**, 1295–1300.

Wu, S.-H., Chang, H.-W., Lin, R.-S. & Tuanmu, M.-N. (2022) SILIC: A cross database framework for automatically extracting robust biodiversity information from soundscape recordings based on object detection and a tiny training dataset. *Ecological Informatics* **68**, 101534.

Xie, J., Colonna, J.G. & Zhang, J. (2021) Bioacoustic signal denoising: a review. *Artificial Intelligence Review* **54**, 3575–3597. Springer.

Yang, W., Chang, W., Song, Z., Zhang, Y. & Wang, X. (2021) Transfer learning for denoising the echolocation clicks of finless porpoise (Neophocaena phocaenoides sunameri) using deep convolutional autoencoders. *The Journal of the Acoustical Society of America* **150**, 1243–1250.

Yin, S., Liu, C., Zhang, Z., Lin, Y., Wang, D., Tejedor, J., Zheng, T.F. & Li, Y. (2015) Noisy training for deep neural networks in speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing* **2015**, 2.

Yu, Y., Si, X., Hu, C. & Zhang, J. (2019) A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation* **31**, 1235–1270.

Zhang, H., Cisse, M., Dauphin, Y.N. & Lopez-Paz, D. (2018) mixup: Beyond Empirical Risk Minimization. arXiv. http://arxiv.org/abs/1710.09412 [accessed 6 July 2023].

Zhang, Z., Zhang, H., He, Y. & Liu, T. (2022) A review in the automatic detection of pigs behavior with sensors. *Journal of Sensors* **2022**, 1–17.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. & He, Q. (2021) A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE* **109**, 43–76.

Zimmer, W.M.X. (2011) *Passive Acoustic Monitoring of Cetaceans*. Cambridge University Press.

Zwerts, J.A., Stephenson, P.J., Maisels, F., Rowcliffe, M., Astaras, C., Jansen, P.A., van der Waarde, J., Sterck, L.E.H.M., Verweij, P.A., Bruce, T., Brittain, S. & van Kuijk, M. (2021) Methods for wildlife monitoring in tropical forests: Comparing human observations, camera traps, and passive acoustic sensors. *Conservation Science and Practice* **3**, e568.