JOURNAL OF
AVIAN BIOLOGY

Review

Automated birdsong recognition in complex acoustic environments: a review

Nirosha Priyadarshani, Stephen Marsland and Isabel Castro

N. Priyadarshani (http://orcid.org/0000-0002-4302-5262) (n.priyadarshani@massey.ac.nz), School of Engineering and Advanced Technology, Massey Univ., Palmerston North, New Zealand. – S. Marsland (http://orcid.org/0000-0002-9568-848X), School of Mathematics and Statistics, Victoria Univ. of Wellington, Wellington, New Zealand. – I. Castro (http://orcid.org/0000-0002-3335-2024), Wildlife and Ecology Group, Inst. of Agriculture and Environment, Massey Univ., Palmerston North, New Zealand.

Journal of Avian Biology
2018: e01447
doi: 10.1111/jav.01447

Subject Editor: Simon Verhulst
Editor-in-Chief: Thomas Alerstam
Accepted 8 December 2017

Conservationists are increasingly using autonomous acoustic recorders to determine the presence/absence and the abundance of bird species. Unlike humans, these recorders can be left in the field for extensive periods of time in any habitat. Although data acquisition is automated, manual processing of recordings is labour intensive, tedious, and prone to bias due to observer variations. Hence automated birdsong recognition is an efficient alternative.

However, only few ecologists and conservationists utilise the existing birdsong recognisers to process unattended field recordings because the software calibration time is exceptionally high and requires considerable knowledge in signal processing and underlying systems, making the tools less user-friendly. Even allowing for these difficulties, getting accurate results is exceedingly hard. In this review we examine the state-of-the-art, summarising and discussing the methods currently available for each of the essential parts of a birdsong recogniser, and also available software. The key reasons behind poor automated recognition are that field recordings are very noisy, calls from birds that are a long way from the recorder can be faint or corrupted, and there are overlapping calls from many different birds. In addition, there can be large numbers of different species calling in one recording, and therefore the method has to scale to large numbers of species, or at least avoid misclassifying another species as one of particular interest. We found that these areas of importance, particularly the question of noise reduction, are amongst the least researched. In cases where accurate recognition of individual species is essential, such as in conservation work, we suggest that specialised (species-specific) methods of passive acoustic monitoring are required. We also believe that it is important that comparable measures, and datasets, are used to enable methods to be compared.

Keywords: birdsong recording, passive acoustic monitoring, machine learning, noise, birdsong recognition

NORDIC SOCIETY OIKOS

www.avianbiology.org

Introduction

According to the International Union for the Conservation of Nature Red Data List (IUCN 2014) 1373 (more than 13%) of the total world bird species are vulnerable or in immediate danger of extinction. Effective bird monitoring methods are needed to assess species presence, abundance and evaluate the consequences of current species management-for-conservation practices and provide an indication of overall balance of a given biome (Vielliard 2000, Dawson and Efford 2009, Towsey et al. 2012, Digby et al. 2013). However, methods to accurately estimate bird population sizes require a great deal of time and effort and are costly, thus they are only applied at small scales (Sutherland et al. 2004). Conservation managers need cost-effective tools to monitor the changes in population size of the species they manage, often in difficult terrain and over large areas. Birdsong is often used to detect, monitor, and quantify species because it is effective even when the individuals are out of sight. Humans are capable of identifying birds aurally: the average person can recognise bird calls in their backyard, while experts can identify thousands of bird species by their song alone. The common approach of estimating populations – the call count (point count) survey and similar methods (Barraclough 2000, Taylor and Pollard 2008) – are labour intensive and prone to bias, as they depend on the expertise and hearing capacity of individual observers (Emlen and DeJong 1992, Sauer et al. 1994, McLaren and Cadman 1999, Rosenstock et al. 2002, Brandes 2008). The results of point count call surveys are subjective, have high errors when the call rate is high (Hutto and Stutzman 2009), or those is a lot of ambient noise (Simons et al. 2007), and the presence of observers can affect the vocal activity of the birds (Bye et al. 2001). These surveys are usually carried out during fine weather in easily accessible areas; therefore, they can be biased by weather conditions and location. Further, as call count surveys are short (usually 5–10 min; Dawson and Bull 1975, Angehr et al. 2002), they cannot fully describe temporal patterns (Digby et al. 2013, Potamitis et al. 2014): Loyn (1985) and Vielliard (2000) for example, observed that call counts over periods of less than 20 min underestimated rare species.

Today, high-end weather-proof bio-acoustic recorders with long battery life and high memory capacity are available for affordable prices. These automatic recorders are specially designed for collecting long autonomous field recordings with minimum human intervention. One can schedule the recorders and mount them in the field and return weeks or months later for the data, or set up a sensor network to directly download data to the laboratory (Stattner et al. 2012, Wimmer et al. 2013). The recorders can be operated in 24/7 mode, meaning that they are capable of capturing both the diurnal and nocturnal sonic environment, including any rare or cryptic bird vocalisations, in any habitat, including ecologically sensitive areas or areas that are difficult to access. Practical comparisons between long time recordings by autonomous recorders and human observers have confirmed that the former is capable to detect more species (Cunningham et al. 2004, Acevedo and Villanueva-Rivera 2006, Towsey et al. 2014, Shonfield and Bayne 2017) while sometimes the comparisons are biased when the point counts also consider the visual detection of species (Hutto and Stutzman 2009, Elliot et al. 2016).

Conservation managers are increasingly interested in using these unattended (automated) recorders to infer the presence, abundance and decline of their target species. After collecting the recordings, they are generally processed through spectrogram reading and/or listening by experienced observers, a labour intensive task that means it is not feasible to manually process weeks or months' worth of field recordings (Taylor 1995, Wimmer et al. 2013, Brighten 2015, Colbourne and Digby 2016). As just one example, Figueira et al. (2015) manually analysed more than 2,000 hours of recordings to examine the difference in use of Amazon old forest and secondary forest by parrots.

Thus, automated birdsong recognition could play an important role in environmental monitoring if the recogniser is capable of processing noisy field recordings and producing robust results. While there has been considerable research on automated birdsong recognition to date, the need for further development is evident given that ecologists and conservation managers still spend a great deal of time manually scanning field recordings because none of the available automated birdsong recognition software fulfil their requirements reliably (Swiston and Mennill 2009, Goyette et al. 2011, Potamitis 2014, Potamitis et al. 2014, Ulloa et al. 2016). However, the potential of automated detection combined with a degree of human calibration to cope with large datasets is encouraging (Urazghildiiev and Clark 2007, Digby et al. 2013, Brighten 2015, Rocha et al. 2015).

Evidently, reliable automated recognition of bird species of conservation interest from long duration field recordings is not an easy task: Box 1 summarises the difficulties associated with automated birdsong recognition from automatic recorders. It is important to note that there are many tasks in conservation and wildlife management where automated recorders can provide useful data, and different analysis tools will be needed to successfully achieve many of these tasks. We focus on the underlying requirement for species-level birdsong recognition, which is an important precursor to many of these analysis tools. In addition, machine-based methods are unlikely to be more successful than humans, who can be easily confused by the sounds of juvenile birds and other variations on standard calls, as well as mimics.

In this paper we review the published methods for the automatic processing and recognition of birdsong, primarily from the viewpoint of processing long field recordings (where there is significant noise, and the birds are at a variable distance from the microphone.) We break the problem down into four main areas: preprocessing (particularly noise reduction), call detection/segmentation (while these two terms are often used interchangeably, we differentiate between them by considering call detection as being the detection of putative calls from long recordings, while segmentation includes isolating those calls from the recording), feature choice, and

Box 1. Challenges associated with implementation of automated birdsong recognition to process field recordings.

1. There is a plethora of unavoidable environmental noise overlapped with field recordings.
2. Bird vocalisations are of varying power because birds can be anywhere, some closer and some further away, and at different angles to the recorder's microphones. Accordingly, some songs are louder and some are quieter in the recordings. During song analysis (segmentation), normally the faint songs tend to not be included because noise makes them less visible. The challenge is to maintain accuracy while improving the sensitivity to the target sounds.
3. Birds (of the same or different species) call on top of each other, for example when duetting, during the dawn or dusk chorus, and when they live in flocks.
4. There is large inter- and intra-species song variability. Birds maintain their own song repertoire, with the size of the repertoire and the complexity varying across the species. Some species repeat the same song, while others have a variety of songs and are capable of creating new songs. Many bird species exhibit geographical variations on their songs (Hill et al. 2013). Although this phenomenon is a challenge in species recognition, it could possibly allow individual recognition (Gilbert et al. 1994, Cheng et al. 2010, Baldo and Mennill 2011, Dent and Molles 2016, Ptacek et al. 2016).
5. Similarly to human speech, a bird may generate the same song with short or long duration in different situations. Further, birds are capable of adapting their sounds according to the environment (song plasticity), not just the temporal modulation, some bird species use spectral modulation to successfully deal with increasing anthropogenic noise due to urbanisation (?) Birds also generate incomplete/quick calls in critical situations, especially during the breeding season when they are occupied with incubation and/or chick rearing.
6. During the song learning process, juveniles produce unusual calls, making the recognition more complicated (Williams 2004) and sometimes, even human experts fail to recognise the species from hearing a juvenile.

classification, summarising the literature for each of these areas separately. We use tables to review the research literature in order to allow direct comparisons; the text is used to expand some details of this and highlight particularly interesting examples.

Unfortunately, it is not always easy to present the data in a uniform way. For example, when seeking to compare the amount of data processed, some authors report the number of calls (usually because human-segmented calls were used as input) while other report the number of hours of recordings. One important difference between the results reported in the literature is whether or not the recordings have been processed manually in any way before analysis. We use the label 'Automatic' to mean that data is taken from automatic recorders and used by the algorithms as is, without human modification (except labelling of calls as exemplars for different species). The term 'Manual' covers data that is not collected automatically, either because of the use of humans to perform the recording, or segmentation, or pre-processing.

Within each table papers are ordered by our subjective assessment of how likely their methods are to be successful for the analysis of field recordings. In our experience, it is important to go beyond small-scale, clean and manually processed datasets and deal with real-world data, which exhibits high variability in both signal and noise. Therefore, we have given priority to literature that we believe exhibits the ability to scale to large numbers of long noisy recordings and many species, and thus successfully deal with real-world data. Figure 1 provides a flowchart to assist researchers who wish to identify the most current and appropriate methods to deal with different circumstances according to our review. Following this summary of the methods, we consider the software that is currently available for wildlife managers and ecologists who wish to automatically process field recordings of birdsong.

In addition, the protocols used in data collection directly affect the decisions made based on this data. Therefore, it is critical that standard protocols for data acquisition with automatic recorders are developed. These protocols need to be compatible with the behaviour and the ecology of the birds, their distribution and habitat, and the nature of the vocalisations – hence species-specific. Figure 2 illustrates an abstracted view of how data acquisition, development of protocols, and data analysis (automatic birdsong recognition) can work together in practice even though we do not further discuss the protocols in this paper. There is substantial related work in auditory processing of sound from other animals, see for example Mellinger et al. (2007), Skowronski and Fenton (2008), Marques et al. (2013) for marine mammals, bats, and amphibians. These are all parts of the area known as soundscape ecology (Schafer 1977, Pijanowski et al. 2011, Farina 2014). While there is often substantial crossover between the aims of the methods in these domains, and this is obviously fertile ground for technology transfer, we note that there are also significant differences – not least the huge variety in bird calls between and even within a species – and for reasons of space, do not consider this literature further in this survey.

Preprocessing, call detection and segmentation

Calling birds produce slight fluctuations of air pressure, which the auditory system resolves as sound, as can a microphone; the latter turns these air vibrations into voltages so that it is possible to record them digitally (Catchpole and Slater 2003, Mindlin 2013). To obtain a discrete time digital signal from a continuous time analogue signal, the continuous signal

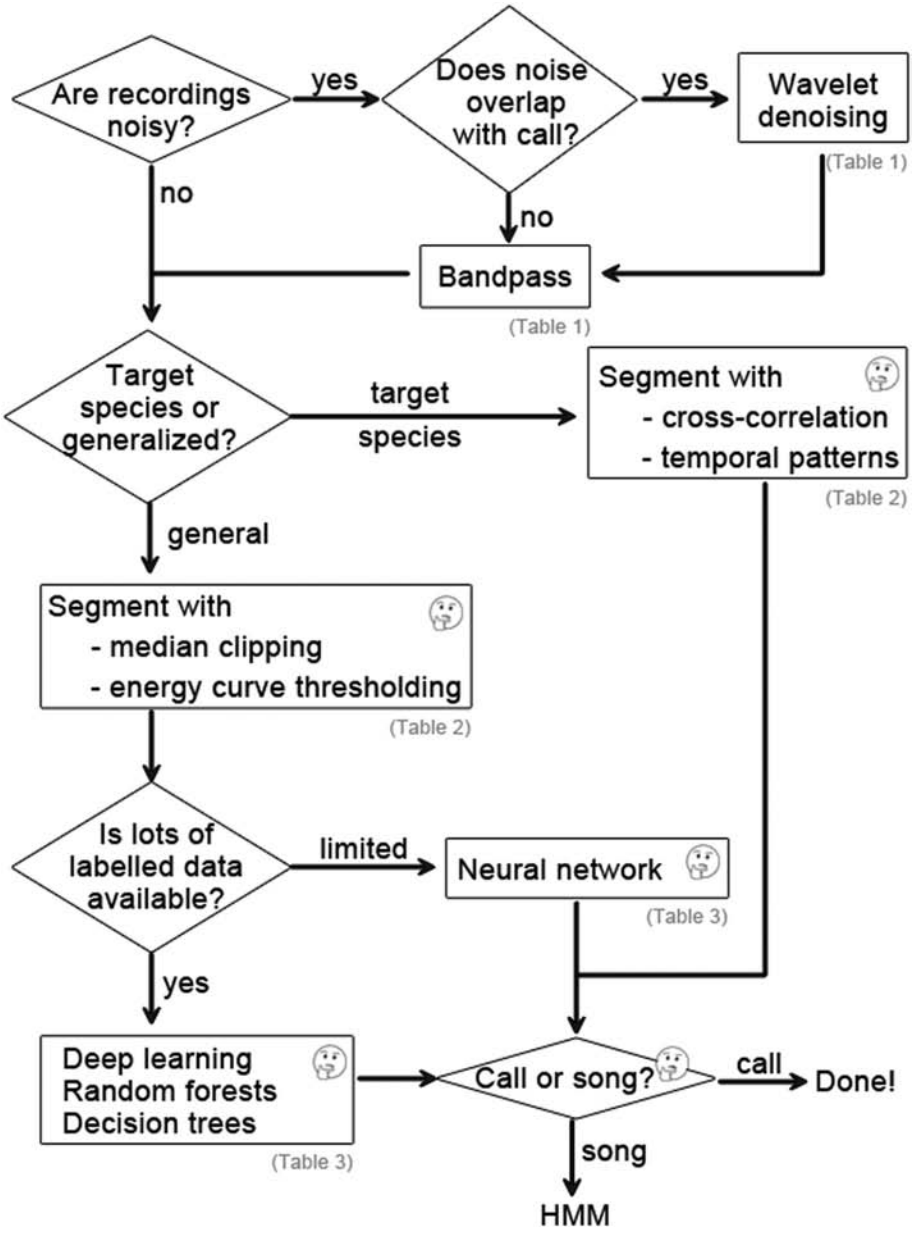

Figure 1. Flowchart representation of the current most suitable approaches to analyse continuous automated recordings and recognise species. For many applications further analysis will be required. The thinking face is used to indicate that further development of methods is necessary.

is sampled at equally spaced intervals. The 'Nyquist sampling theorem' stipulates the minimum sampling frequency required to represent a continuous time signal with a discrete time series (a signal can be recovered from its samples if the sampling frequency is at least twice the highest frequency of the original signal; Landau 1967). The dynamic range used to record the samples is determined by the resolution (the number of bits per sample): larger resolutions provide larger dynamic ranges, but occupy more memory space. Sound files accumulate rapidly when large numbers of automatic recorders are operated continuously, meaning that using an effective method for storing and retrieving sound inventory is crucial in long-term acoustic monitoring. The building blocks that are required for software to reliably recognise calls in a recording are: preprocessing of the recording to remove noise; segmentation of individual calls; extraction of chosen features from the representation of each call; the training of classifiers, generally based on a set of human-labelled training data (Fig. 2).

In order to develop a robust recogniser, it is generally essential to have a rich database that includes clear representations of the possible variations in vocalisations of each species. A well-balanced set of recordings that represents possible variations that can be expected in scheduled autonomous

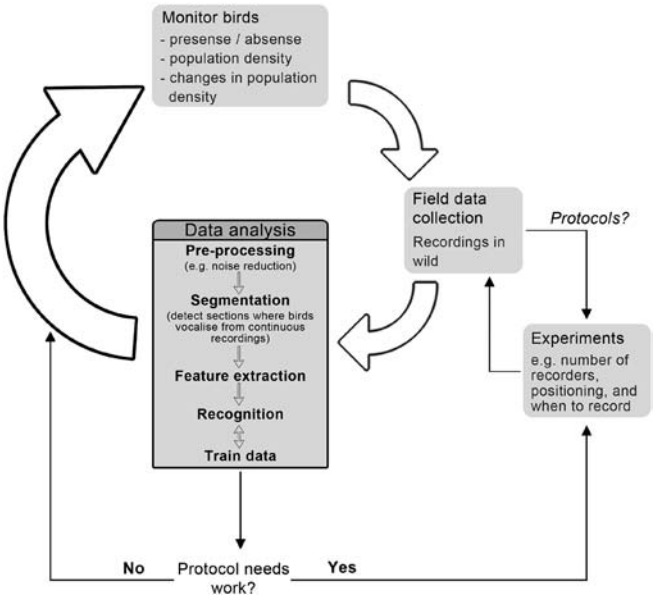

Figure 2. Representation of the full work process required for the use of acoustic recorders in wildlife management, including the development of protocols for the deployment of recorders.

recordings of the target species – not only the regional variations, but also examples from good quality close-range calls through to faint birdsong – could increase the ability of the recogniser to address real-world conservation problems. Unfortunately, while it is easy to obtain unlabelled recordings, it is time-consuming to provide good quality training data for even a small number of species, and we do not know of any large-scale datasets that provide these to the levels needed for conservation tasks. The size of the training data required mainly depends on the target species, with less data required for species with small repertoires and simpler vocalisations. Obtaining this data is a challenging task that requires expertise in handling recorders, and cooperative animals (many birds tend to avoid humans). For rare/nocturnal birds this data can be exceedingly hard to obtain. While data from autonomous recorders can be used to compile a dataset, it requires prior screening of the recordings. This data then has to be manually labelled to provide a training set that can be used to train the classifiers, and an independent test set that can be used to evaluate the success of the trained methods. The labels and annotations need to be generated manually by experts by careful listening to the recordings and/or visual inspection of the spectrograms, a time-consuming but crucial task (currently the main way to process recordings). Only after the recogniser is trained and evaluated on known field recordings, it is ready to process unknown recordings.

In order to validate the algorithms that are used and compare different methods it is necessary to have performance metrics that can be applied to both the call detection and the call recognition. We will briefly summarise a variety of different metrics that have been used in the literature before starting to describe the various approaches to detection and recognition.

Performance measures

There are four possible outcomes when a classifier system makes an prediction for a binary output (such as whether or not there is a birdcall in a given segment of recording): a true positive (TP) is when the detector correctly says that there is a birdcall, a true negative (TN) is when it correctly says that there is not, while false positives (FP) and false negatives (FN) are where the detector incorrectly suggests that there are and are not (respectively) birdcalls. The number of examples of each of the four cases will add up to the number of segments considered. Note that this assumes that the segmentation algorithm took in short time segments and processed each individually. Instead, an identifier may recognise the beginning and end of a bird call.In this case, the same four outcomes can be considered (TP, TN, FP, FN), but the outcome will be the duration rather than the number of calls.

The outcomes can be combined into a few different useful measures:

Recall (also known as sensitivity or the true positive rate)

$$= \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$F_\beta = \frac{\left(1 + \beta^2\right) \times Recall \times Precision}{Recall + \beta^2 \times Precision}$$, where β is a weighting factor (often $\beta = 1$ or 2)

For example, the strength of a call detection algorithm is commonly determined by its recall rate and precision. The recall rate explains how well the segmentation algorithm detects songs, or the percentage of songs retrieved from the total number of songs in the recording. Precision refers to how reliable the detection algorithm is (the percentage of true positives from the positively classified songs). Maintaining a high recall rate while achieving high precision is challenging: recall and precision are often inversely related to each other; therefore, it is possible to gain recall at the cost of losing precision and vice versa. Specificity measures how good a given model is at avoiding false alarms, while accuracy refers to the proportion of the total number of predictions that were correct. The weighted average of recall and precision (F_β) is useful to interpret the combination of recall and precision. For example, while F_1 (harmonic mean) weights recall and precision equally, F_2 emphasises recall over precision.

One way to present some of these measures is to use the Receiver Operating Characteristic (ROC) curve, which

plots the True Positive Rate (TPR = recall) against the False Positive Rate (FPR = 1-specificity) as they change when some parameter(s) of the method are varied. For example, an ROC curve could be generated as a threshold at which sounds are considered as signal rather than noise is adjusted (normally with the effect of increasing TP at the cost of an increased FP or vice versa) producing different (TPR, FPR) pairs to be plotted. The Area Under the (ROC) Curve (AUC) is also used as a measure of accuracy (for further information, see Marsland (2014)).

The same measures can be used to evaluate classification algorithms, except that there are usually more than two species of bird to be identified. The definition of a false positive is then modified to be an incorrect choice of class. For multi-species classification, Mean Average Precision (MAP) is another useful measure. While the majority of researchers have used the aforementioned metrics some have used different measures and there is no consistency. It is not always clear from papers precisely what measures were used, hence Table 1–3 report the performance of each method only to the best of our knowledge and understanding of the relevant materials provided.

Noise reduction

Field recordings capture all the sounds that are present in the geographical area where the recorder is mounted, including birdsong of interest and many other biophony (sounds from other animals), geophony (wind, rain etc.) and anthropophony (man-made sound, such as aeroplanes, and wind turbines). Relatively low frequency sound from abiotic sources such as wind, aeroplanes, running water, and vehicles passing are common, and can either mask species with low frequency songs or generate false positives (Potamitis 2014). Low recognition accuracy is often attributed to noise (Baker and Logue 2003, Fox et al. 2006, Aide et al. 2013), which affects the whole process unless removed initially. Therefore, removing the noise is an important first step in finding the target sounds in a recording.

Noise tends to hide or alter the birdsong: it is superimposed onto the recording, degrading the signal quality and making the sound of the call fainter, leading to poor results by recognisers. Some researchers have found that developing the training data set with clean references (less corrupted with noise) is helpful to avoid false positives (identification of any sound except the target bird sounds; Wildlife Acoustics (2011) and Boucher (2014)). Usually, song examples from the same environment where the test recordings will be collected are preferred (Katz et al. 2016), as it is hoped that these will have similar noise profiles and less variations in calls. While recording clean close-range calls is possible with handheld (manual) recorders (because the recordist can get close to the individual birds and avoid noise by careful screening (Ruse et al. 2016)), it is often more successful if the training data is based on the same type of recorder as will be used in practice. The reason for this is partly because microphones used for manual

Table 1. Noise reduction methods in descending order of effectiveness in regards to automatic analysis of recordings. Effectiveness was decided based on the type of recordings used and the performance of the methods. A = automated; M = manual; L = long (≥ 5 min); S = short; SNR = signal-to-noise ratio; SnNR = modified SNR; AUC = area under the ROC curve; dB = decibels; Y = yes; N = no; N/G = not given.

		Type of recordings				Performance		
Study	Method	A/M	No. of species	L/S	Total length	Direct evaluation	Indirect evaluation	Species-specific?
Priyadarshani et al. (2016)	Wavelet denoising	M	11	S	700 syllables and 55 complete songs	SnNR improved from 0.5 to 5.3 dB	Illustrated visually using examples	N
Ren et al. (2008)	Perceptually scaled wavelet packet decomposition	M	3	S	30 calls – manually added noise	N/G	Similar to uniform band wavelet packet transform	Y
Fox et al. (2008)	Band-pass filtering	M	3	S	N/G	N/G	87% individual identification accuracy	Y
Selin et al. (2007)	Modified filter bank (eight-band)	M	8	S	3132 sound files	N/G	Main focus was species recognition (Table 3)	Y
Chu and Alwan (2009)	Correlation-maximization and band-pass filter	M	5	S	42 min	N/G	Classification error rate improved from 5.4 to 4.6	Y
Lasseck (2013)	Median clipping plus image processing on spectrogram	M	87	S	~2 h	N/G	91.7% AUC – not reversible	N
Wolf (2009)	Wiener filter plus image smoothing (Gaussian blur, median)	A	1	S	single syllable	N/G	Illustrated visually using one spectrogram	N
Briggs et al. (2012)	Whitening filter (using a noise profile)	A	13	S	91 min	N/G	AUC of 0.978 (Table 3)	N
Baker and Logue (2007)	Noise profiling (subtract a sample noise)	M	1	S	1 s × 2 e.g. × 4 manually added noises	SNR improved from 16 to 60 dB	Only useful when the noise is constant	N

Table 2. Call detection and segmentation methods in descending order of their effectiveness to automatic detection of putative calls. Effectiveness was decided based on the performance and the type of recordings used. A = automated; M = manual; L = long (≥ 5 min); S = short; N/G = not given; AUC = area under the ROC curve; SNR = signal-to-noise ratio.

		Recordings					
Study	Method	A/M	No. of species	L/S	Total length	Performance	Comment
Bardeli et al. (2010)	2 species-specific algorithms – based on the temporal patterns in the frequency bands – noise estimation and spectral subtraction	A	2	L	26 h	94% recall, 66% accuracy on Eurasian bittern – 92% detection rate on Savi's warbler	Demonstrated on two very different species
Frommolt and Tauchert (2014)	Template matching by spectrogram correlation using Avisoft-SASLab Pro (Specht 1993)	A	1	L	5 h	84.9% recall, 100% precision (assuming the software output raw detections)	Species-specific – needs testing with noisy field recordings
de Oliveira et al. (2015)	Morphological opening (erosion and dilation) – HMM for recognition	A	1	L	~3 h	56% recall	Required large amounts of training data (> 27 h)
Lasseck (2013)	Median clipping followed by standard image processing techniques	M	87	S	~2 h	91.7% AUC (NIPS4B 2013 competition)	Needs evaluation against long continuous (long) recordings
Jančovič and Köküer (2015)	Sinusoidal detection method	A	30	S	33 h	80% recall with 2 species and 71% recall with 3 species	Assumes the number of species in recordings is known – does not appear to scale well
Potamitis et al. (2014)	Hilbert follower – down sampling and band pass filtering	A	2	S	> 42 h	RMBL-robin: 91% recall and 71% precision – Vouliagmeni (kingfisher): 85% recall and 85% precision	Authors focused on reducing the search space by discarding recordings devoid of target species calls
Ranjard et al. (2015)	Manually defined 5 HMMs to model target species, other species, humans, recorder noise, and silence – MFCC and PLPC features – NN classifier	A	1	L	24 min (test)–~28 h (full dataset)	100% recall and 78% precision on test file	Not scalable
Jinnai et al. (2012)	Time domain energy curve thresholding	A	1	L	N/G	N/G	Its software has reported >95% accuracy on dawn chorus (Boucher et al. 2012)
Schrama et al. (2007)	Frequency band thresholding to extract flight calls	A	N/G	N/G	12 h	27% recall	Emphasised the need of prior noise removal
Neal et al. (2011)	Random forest classifier	M	N/G	S	625 sound files	93.6% recall – 8.6% false positive rate	Original continuous recordings were sampled (15 s) to generate dataset
Zhang and Li (2015)	Adaptive energy detection – SVM classifier – Mel-scale and wavelet features	M	N/G	S	30 bird sounds	85% classification accuracy	Used clean recordings manually polluted with noise
Härmä and Somervuo (2004)	Time domain energy curve thresholding to detect syllables	M	150	S	2000 recordings	Sinusoid model matched with spectral structure of 93% syllables	
Somervuo et al. (2006)	Time domain energy thresholding to detect syllables	M	14	S	792 birdsong	85% recall – 93% precision	High variability in accuracy between species
wa Maina (2015)	Speaker diarisation techniques	M	19	S	179 recordings	42.5% accuracy of estimating the number of species in recordings	Assumed one cluster represents all the non-bird sounds
Härmä (2003)	Sinusoidal modelling to decompose birdsong into syllables and identify them	M	14	S	N/G	38.7% overall recognition rate	Needs estimate of number of syllables
Lee et al. (2006)	Modified Härmä (2003)	M	420	S	420 sound files	87% recall	Classification errors are to poor segmentation and/or background noise
Juang and Chen (2007)	Time domain energy	M	10	N/G	N/G	94.67% recognition rate	Method only useful when the SNR is high

Table 3. Feature extraction and recognition in the descending order of their effectiveness to automatic recognition of bird sounds. Note that the first few rows are allocated to the series of recent birdsong recognition competitions; there is no evidence that these would actually be useful for recognition from automatic recorders. A = automated; M = manual; L = long (≥ 5 min); and S = short; AUC = area under the ROC curve; dB = decibels; GB = gigabytes; CD = compact disk; N/G = not given.

			Type of recordings				
Study	Features used	Recognition method	A/M	No. of species	L/S	Total length	Performance
Murcia and Paniagua (2013)	MFCC – feature reduction by linear discriminant analysis	NN	M	35	S	3.75 h testing – 18 min training	Winning solution of ICML 2013 technical challenge – 0.74 AUC
Dufour et al. (2013)	MFCC	SVM	M	35	S	3.75 h testing – 18 min training	0.64 AUC in ICML 2013 technical challenge
Fodor (2013)	Spectral features were manually selected (30–70) for each species	Random forests	M	19	S	54 m testing – 54 m training	Winning solution of 2013 MLSP – 0.956 AUC (23 random forests) – 0.955 AUC (1 random forest)
Lasseck (2013)	File statistics – segment statistics – segment probabilities	Decision trees	M	87	S	~2 h – 687 train and 1000 test recordings	Winning solution of NIPS4B 2013 – 0.92 AUC – segment probabilities were more useful
Potamitis (2014)	Descriptive, morphological features, and spectrographic cross-correlation	Random forests	M	87	S	~2 h – 687 train and 1000 test recordings	0.91 AUC in NIPS4B 2013 competition
Lasseck (2014)	6669 low level descriptors including MFCC – reduced to 1277 after feature elimination	Decision trees	M	501	S	9688 train and 4339 test recordings	Winning solution in LifeCLEF 2014 Bird Identification Task – 0.92 AUC – 51% MAP
Lasseck (2015)	8541 low level descriptors including MFCC – best results were gained with 500 features	Decision trees – bootstrap aggregating to combine multiple classifiers	M	999	S	33 862 recordings	Wining solution in LifeCLEF 2015 Bird Identification Task – 0.97 AUC – 45% MAP
Sprengel et al. (2016)	Spectrogram as an image with different data augmentation	Convolutional NN	M	999	S	34 128 recordings	Wining solution in LifeCLEF 2016 Bird Identification Task – 69% MAP (56% MAP when background species are used as additional prediction targets)
Ganchev et al. (2015)	MFCC (without Mel-filter bank)	A statistical log-likelihood ratio estimator based on the GMM universal background model and post-processing	A	1	L	2 h 20 min test dataset – > 27 h train dataset	Recall: 97.7% (0–20 dB), 83.8% (0–30 dB), 43.3% (0–40 dB), and 31.2% (0–50 dB) – 100% precision – used to detect southern lapwing *Vanellus chilensis lampronotus* from one month (24/7) recordings
de Oliveira et al. (2015)	Bird sound detection using morphological filtering of the spectrogram	HMM	A	1	L	2 h 20 min test dataset – > 27 h train dataset	Recall (56%) was better than GMM based segmentation (55%; Sahidullah and Saha 2012) and syllable based method (44%; Härmä 2003)

(*Continued*)

Table 3. Continued

Study	Features used	Recognition method	Type of recordings				Performance
			A/M	No. of species	L/S	Total length	
Potamitis et al. (2014)	Perceptual LPCC, a slightly different version of MFCC	HMM	A M	2	S	> 42 h	RMBL-robin: 77% recall and 85% precision; Vouliagmeni: 85% recall and 85% precision – separate models for target and noise
Chu and Blumstein (2011)	–	HMM	M	1	S	~78 min (RMBL-robin)	76% recall and 75% precision – separate HMMs to model target species and background sounds
Acevedo et al. (2009)	Time-frequency features – linear discriminant analysis	Decision trees – SVM	A	12	S (1 min)	>35 h	95% recall and 99% precision – SVM was better than decision trees
Bastas et al. (2012)	DWT, MFCC, Spectrogram-based Image Frequency Statistics (SIFS), and Mixed MFCC and SIFS (MMS)	k nearest neighbours (kNN), MLP, HMM, Evolutionary Neural Network (ENN)	A	5	S	459 test and 128 train bird calls	94% accuracy – Song Scope resulted less accuracy (70%)
Ulloa et al. (2016)	A mean template derived from 10 standardised samples	Spectrogram cross-correlation	A	1	S (1 min)	5 h 36 min test dataset	35% recall and 100% precision – canopy level recorders (20 m) detected more calls (62%) than those placed lower (1.5 m, 38%)
Briggs et al. (2012)	The shape of the binary mask of the segments	Multi Instance Multi Label (MIML) SVM, MIML kNN, MIML Radial Basis Function (RBF)	A	13	S (10 s)	548	MIML RBF reached the highest AUC (0.98)
Andreassen et al. (2014)	Species-specific features (duration, average power, spectral entropy)	SVM	A	1	S	59 GB	precision: 96% (when dry) and 70% (when rain)
Digby et al. (2013)	5 species-specific features	Decision trees (C5.0)	A	1	L	52.2 h test – 66	Automatic: 40% recall, 98% precision – manual scanning: 80% recall – field survey: 94% recall, 96% precision – 3 methods resulted similar annual change of calling activity
Stowell and Plumbley (2014)	Mel spectra and MFCC	Random forest and spherical k-means	M	77	L	7.8 h[1]	Around 0.70 AUC and 35% MAP – row mel spectra (high dimension) yielded better classification than MFCC
Jančovič and Köküer (2015)	Estimation of frequency tracks	HMM	A	30	S	33 h	Accuracy (78%) dropped to 69% when the number of species is unknown

(*Continued*)

Table 3. Continued

Study	Features used	Recognition method	Type of recordings				Performance
			A/M	No. of species	L/S	Total length	
Ventura et al. (2015)	MFCC – noise reduction and frame selection using morphological filter (considering spectrogram image)	HMM	M	40	S (avg. 32 s)	200 test and 400 train recordings	Precision (72%) was higher than MFCC after frame selection with GMM energy detector (70%; Sahidullah and Saha 2012); syllable segmentation (65%; Härmä 2003); region of interest detector (48%; Briggs et al. 2012, Potamitis 2014)
Vilches et al. (2007)	71 pulse features – data mining	HMM, Decision trees (VQ + ID3, J4.8), and Naive Bayes	M	3	S	204 recordings	Best accuracy (98%) from J4.8 (47 features) – incomplete vocalisations and background noise led to low accuracy of HMM (93%)
Selin et al. (2007)	4 features (maximum energy, position, spread, and width) derived from wavelet packet decomposition coefficients	SOM and MLP	M	8	S	2278 train and 854 test sounds	High recognition accuracy with MLP (96%) compared to SOM (78%)
Fagerlund (2007)	MFCC and descriptive parameters	SVM	M	8	S	2278 train and 854 test sounds	98% accuracy – similar performance as (Selin et al. 2007)
Papadopoulos et al. (2015)	Spectral mean, standard deviation, skewness, kurtosis, mode, and spectral flatness (Madhu 2009)	GMM	M	15	S	N/G	Best results when using one feature (mode): > 0.90 AUC for 11 species in 'park' category and 10 species in other cases – lowest AUC from same species (0.70, 0.64, 0.56 for three noise types)
Chen and Maher (2006)	Peak frequency track	Spectral distance (distance between the test frequency track and the reference) and threshold	M	12	S	12 test recordings	Accuracy (99%) decreased to 95% in presence of noise (manually adding white noise) – LPCC and DTW (90% to 71%) – MFCC and HMM (95% to 76%)
Wielgat et al. (2012)	MFCC	HMM	M	30	S	1426 songs	92% recall – not tested on field recordings; separate model for each species
Brandes (2008)	Change in peak frequency and in bandwidth over time	HMM	M	9	S	5871 test and 908 train calls	75–96% recall – cricket and frog species recognition was easier than birds
Taylor (1995)	Peak frequency track (single)	Decision trees (C4.5)	M	9	S	138 flight calls	78%, 4%, and 18% of calls identified correctly, incorrectly, and left unclassified respectively

(*Continued*)

Table 3. Continued

Study	Features used	Recognition method	Type of recordings				Performance
			A/M	No. of species	L/S	Total length	
Tan et al. (2012)	Normalised spectrogram (PCA for dimensionality reduction)	A sparse representation based classifier that represents test feature vector as a sparse linear combination of training data	M	1	S	1022 syllables	90% accuracy – SVM and nearest subspace classifier resulted 88% accuracy
Dong et al. (2015)	Spectral ridge features	Euclidean distance	N/G	19	S	5 h	94% accuracy – when spectral ridges are strong method worked well, but not when the acoustic energy is temporally and spectrally diffused (e.g. parrot shriek)
Kasten et al. (2010)	105 features	Anomaly detection	N/G	10	S	3673 segments	82% accuracy
Lopes et al. (2011a)	MFCC and descriptive parameters (MARSYAS framework; <http://marsyas.info>)	Naive Bayes algorithm, kNN with k = 3, decision tree classifier (J4.8), MLP, SVM, Sequential Minimization Algorithm (SMA)	M	73	S	1619 recordings	2 databases: one with complete audios and one with only pulses, the second was better – MLP and SMA performed better – maximum F_1 when 3, 5, 8, 12, and 20 species considered was 95%, 89%, 86%, 83%, and 78% respectively
Lopes et al. (2011b)	Compared MARSYAS feature set, Inset-Onset Interval Histogram Coefficients feature set, and Sound Ruler feature set	Naive Bayes algorithm, kNN with k = 3, decision tree classifier (J4.8), MLP, and SVM with SMA	M	3	S	101 songs	MARSYAS and Sound Ruler performed equally well – pulses database (99.7% F_1) were better than original audios (79.2%)
Ganchev et al. (2012)	Temporal and spectral features from openSMILE (Eyben et al. 2010)	kNN, Bayes network, MLP, C4.5 decision tree (J4.8), and SVM with SMA	M	N/G	S	150 test recordings – 12 min for training	Recognition accuracy: 86% (MLP), 81% (SVM) – classified each sound frame as bird sound or noise – not addressed species recognition
Ross (2006)	20 frequency, cepstral, and multi frame (global) features	MLP, SVM, Kernel Density estimation of probability (KDE)	M	10	S	403 test and 193 train	79% precision MLP had the highest accuracy (83%) – authors initially discarded noisy examples
Tyagi et al. (2006)	Spectral Ensemble Average Voice Print (SEAV) computed on FFT spectrum by frame wise averaging FFT coefficients	A comparison between SEAV + Euclidean distance, spectrogram + DTW, and MFCC + GMM	N/G	15	N/G	63 recordings	Recognition performance (not defined): SEAV + Euclidean distance 87% – spectrogram + DTW 67% – MFCC + GMM 100%

(*Continued*)

Table 3. Continued

Study	Features used	Recognition method	Type of recordings				Performance
			A/M	No. of species	L/S	Total length	
Graciarena et al. (2010)	MFCC	GMM	M	92	S	≥ 4 recordings per species	Minimum equal error rate (9 on train dataset and 10 on test dataset) was achieved with high number of filters (41) with 100–13 000 Hz
Heller and Pinezich (2008)	Frequency track extraction	Mahalanobis distance	M (high SNR)	4	S	N/G	79% recall
Kogan and Margoliash (1998)	MFCC	DTW and HMM	M (lab recordings)	2 (males)	S	993 songs	90% recall – DTW was better depending on the quality of recordings and complexity of songs
Trifa (2006)	MFCC	HMM using HTK	M	5	S	3368 songs	Accuracy (99.5%) decreased to 90% after introducing low quality recordings
Fox et al. (2006)	MFCC	NN	M	3	S	24 recordings	89% precision – highly restricted experiments (1 recording per individual) – focused individual identification
Jančovič and Köküer (2011)	MFCC and tonal based features (frequency and magnitude of the prominent tonal component per frame)	GMM	M (CD quality)	99	N/G	N/G	83% recall – 1% precision – performance was reported at 10 dB SNR – recognition was in binary format (signal/noise) – tonal features were better than MFCC in the presence of manually added white noise
Mundry and Sommer (2004)	Estimation of fundamental frequency contour	Not addressed	N/G	2	N/G	N/G	No direct evaluation – fundamental contour differed in individuals – looked at the relationship between the structure of begging calls and nutrition need of chicks
Somervuo et al. (2006)	Combination of MFCC and descriptive parameters	DTW, GMM, and HMM	M	14	S	792 birdsong	Precision: 60% (song level); 40% (syllable level) – best results by using MFCC-based syllable trajectory models with DTW – recognition of GMM and HMM were almost similar
Lee et al. (2008)	Two dimensional MFCC	GMM	M (CD)	28	N/G	3789 syllables	84% classification accuracy

(*Continued*)

Table 3. Continued

Study	Features used	Recognition method	Type of recordings				Performance
			A/M	No. of species	L/S	Total length	
Lee et al. (2013)	Angular Radial Transform (ART) on spectrogram	GMM	M (CD)	28	N/G	3789 syllables	95% classification accuracy
Kwan et al. (2006)	MFCC	GMM	N/G	11	N/G	N/G	90% classification accuracy (SNR = 5 dB)
McIlraith and Card (1997)	Time-frequency information	NN	M (CD)	6	S	133 birdsong	> 90% accuracy (not defined)
Cai et al. (2007)	MFCC	MLP	A	14	N/G	N/G	Recognition rate (not defined) stood just below 99% with 4 species, but decreased (87%) after introducing other 10 species
Somervuo and Härmä (2003)	Sinusoidal modelling	SOM and DTW	M	5	S	1000 syllables	Divided the SOM map into five areas (equal to number of species)
Juang and Chen (2007)	LPC	An extension of NN	M (CD)	10	N/G	N/G	96% accuracy – manually segmented data – separate NN for each species
Lee et al. (2006)	LPCC and MFCC	Euclidean distance between test pattern and references	M (CD)	420	S	420 recordings	87% recall – MFCC was better than LPCC – wrong classifications were due to background noise and poor segmentation
Tachibana et al. (2014)	532 features including spectral and cepstral features	SVM	M	1	N/G	N/G	99% accuracy – the research contributes to neuroscience where syllable classification of one species is common

[1] Results for the bldawn dataset.

recording are usually directional, and thus more sensitive to the sound in one particular direction, which is ideal when recording individual animals; in contrast, automated recorders use omni-directional microphones because they aim to record birds from anywhere around the recorder (Brandes 2008). Several authors have stated that the overlap of unwanted sounds with the song of their target species was the biggest obstacle in their automated processing of natural field recordings (Wolf (2009) and Potamitis et al. (2014)). Schrama et al. (2007) claimed that the overall recognition accuracy in automated recognition can only be achieved through an advanced denoising approach. Despite this, there has been relatively little research on it.

The most common approach for denoising found in the literature was to follow standard signal processing techniques and perform noise profiling, followed by filtering. However, noise profiling and noise filtering have their own limitations (Table 1). There are situations where high-pass, low-pass, or band-pass filters cannot be applied successfully. For example, if a particular species produces only low frequency songs, then a low-pass filter with a suitable cut-off frequency removes frequency components beyond the birds' frequency range. However, if there are multiple bird species or a species that produce vocalisations that occupy in different frequency ranges, applying a filter without eliminating some bird vocalisations is impossible. The Wiener filter, an approach that is useful in signal enhancement to remove linear distortions (Vaseghi 2008), is not useful for birdsong because it assumes that the signal and noise are stationary and that spectral information is available. While this can be partially overcome by using the spectrogram window method to split the signal into a series of small timeframes and compute the filter coefficients in each frame, this adaptive Wiener filter (Chen et al. 2006) suffers from signal distortion. So while it has been used for noise reduction in speech signals, it is not useful for birdsong recordings without a priori knowledge of the characteristics of the signal and noise, and this knowledge is rarely available with field recordings.

In recent work Priyadarshani et al. (2016) and Priyadarshani (2017) have shown that wavelets can be effectively used to remove noise from field recordings collected with automatic sound recorders. In a field recording, while the birdsong is transient, a considerable amount of background noise, particularly the geophony, is nearly stationary. Wavelet denoising eliminates this quasi-stationary noise no matter whether it is wideband or narrowband, providing that it is approximately Gaussian.

An alternative to audio analysis is treating the spectrograms as images and applying image analysis methods to clean them from noise (Potamitis 2014). While this method is useful to detect regions of interest (as will be discussed later), it is only approximately reversible, Griffin and Lim (1984). This means that it is of limited utility for recognition in general, as it is common to extract features from the sound file, as well as the spectrogram to achieve recognition.

The result of successful noise reduction is a cleaner recording (although possibly with some artefacts) that is ready to be used as input into segmentation and recognition algorithms.

Call detection and segmentation

In general, automatic recorders turn on and off at set times, and record everything between those times. In order to recognise the birdsong in the recording, potential sections that could contain calls need to be isolated first. This can also be useful to enable human analysts to concentrate only on what is important in a long recording. This is even more relevant when the recordings contain very little birdsong, but lots of noise: useful data is then a very small proportion of the total recording (Andreassen et al. 2014). Selin et al. (2007) concluded that call detection and segmentation is the most complicated and difficult part of the whole automation process; they also highlighted the need for noise reduction.

Following call detection, it may also be desirable to divide a song or a series of calls into syllables, and this is not straightforward (Tchernichovski et al. 2000), especially when the syllables are not followed by a silent interval and not separated clearly. Merging the syllables that are very close to each other is the common practice in syllable detection (Fagerlund 2004). Further, isolation of acoustic units also poses a great challenge due to background noise. Therefore, the majority of the published work has largely avoided automatic segmentation and instead used manually segmented data to test their recognition methods (Anderson et al. 1996, Franzen and Gu 2003, Chen and Maher 2006, Somervuo et al. 2006, Fox et al. 2008). Even when automatic segmentation is used, it is still followed by manual elimination of noisy segments (Selin et al. 2007, Rocha et al. 2015).

Both the waveform and the spectrogram can be used to isolate bird vocalisations (Table 2), based on the assumption that the sections where the birds sing carry more energy than the other parts of the recording (Härmä and Somervuo 2004, Somervuo et al. 2006, Juang and Chen 2007, Jinnai et al. 2012, Towsey et al. 2012, Murcia and Paniagua 2013). This assumption is valid for recordings that are not corrupted by too much noise, but this is not always the case for automatic recordings. Noise causes the bird sounds to be quieter and faded (Briggs et al. 2012), adding to issues with bird proximity to the recorder. Common energy-based song detection coupled with thresholding fails to detect faint songs, but also detect periods of noise that exceed the chosen threshold.

The most common frequency-based method found was to treat the spectrogram as an image and use median clipping, whereby points are identified as birdsong if they are more than some pre-defined multiple of the median of the relevant column and row of the spectrogram (Lasseck 2013, 2014, Potamitis 2014). Image processing techniques such as basic shape morphology methods were used to improve this process (Potamitis 2014). This method works well for calls that are clear, but does not detect calls that are quieter, nor work as well when the noise levels are high. It can, however,

be used to detect the top and bottom frequencies of the call as well, which can be useful for further processing, although these will vary with distance as the harmonics of the signal are attenuated at different rates. Notwithstanding these problems, this form of segmentation has been used as a preprocessing stage successfully: Potamitis (2014) trained a random forest classifier with the features extracted from median clipping to recognise 78 bird species. He reached 91% AUC on the test dataset. The dataset, however, included only short recordings (0.25–5.75 s) and therefore the applicability of the method over long recordings has not yet been evaluated (Table 2).

Morphological opening (erosion and dilation) was employed by de Oliveira et al. (2015) to detect acoustic activity in 14 min long field recordings of southern lapwings *Vanellus chilensis* followed by a species-specific Hidden Markov Model (HMM) based recogniser that required large amount of training data (> 27 h) to ultimately result in 56% recall despite the fact that very faint calls were excluded during the annotation (Table 2).

Very few studies (Bardeli et al. 2010, Frommolt and Tauchert 2014, de Oliveira et al. 2015, Jančovič and Köküer 2015) have evaluated detection methods on natural unattended field recordings (Table 2). Jančovič and Köküer (2015) applied a sinusoidal detection method assuming that the number of species in a given recording was known. The recall decreased significantly (from 80 to 71%) when the number of species increased from two to three. Bardeli et al. (2010) proposed two algorithms that were specifically tailored to two bird species. The methods were based on the temporal patterns in the frequency bands of the target species and noise estimation from each band followed by spectral subtraction to avoid noise, and by evaluating the methods on a large set of automated recordings (26 h). They reported 94% recall and 66% accuracy on the Eurasian bittern *Botaurus stellaris* and 92% overall detection rate on Savi's warbler *Locustella luscinioides*. The problem with these species-specific methods is that they need to be completely redesigned in order to detect other bird species. Frommolt and Tauchert (2014) used template matching in 'Avisoft-SASLab Pro' (Specht 1993) to detect Eurasian bitterns in controlled field recordings (data collected under calm conditions) and reported 85% recall with no false positives (Table 2).

Feature choice and extraction

Turning the denoised and segmented birdcall into something suitable for input into a classification algorithm requires that features of importance are extracted from the call. Possible features can be as simple as the sequence of amplitudes present in the call (or the raw spectrogram values), but generally more success is found with more descriptive features; the overall recognition performance of any pattern recognition task depends heavily on the suitability of the features considered. Regardless of what sounds are being studied, audio feature extraction is a common topic in audio signal processing. There are therefore a large number of toolboxes readily available to easily extract widely used features, particularly for speech and music. Moffat et al. (2015) provide an evaluation of the major feature extraction tools. The result is a huge number of features considered in the literature, often based on those considered in other areas. Despite all of this research there is as yet no clear evidence for and against different representations for birdsong with many different approaches being used (Table 3). In general, if it is unclear which features will be helpful, the tendency is to add more in. Unfortunately, the more features that are included, the more training data is required for learning, something that is known as the curse of dimensionality (Marsland 2014).

There are several ways to categorise the features that are derived from sounds. In general, only local information (based on individual short time windows) is useful for birdsong, particularly since the calls are segmented from the recording. Features can be based on the amplitude plot (such as the bandwidth, number of zero crossings), the energy of the signal within the window, or on the short-time Fourier transform data (such as its statistical moments, fundamental frequency, or spectral variations). Many of these features are related to one another despite being based on different representations. For example, pitch and loudness of a bird call are related to the frequency and energy (cumulative amplitude effect over time) respectively. One set of interesting features use the short-time Fourier transformed data and process it to more closely match how humans process sound. These so-called perceptual features can be based on either a monotonic transformation of the range, such as the Bark scale or the Mel scale. One set of features that are particularly common in the literature are Linear Predictive Coding (LPC) and its extension, the Linear Prediction Cepstrum Coefficients (LPCC), that were initially used for encoding human speech (Zbancioc and Costin 2003), but also seem to be useful to represent birdsong (Table 3). Mel Frequency Cepstral Coefficients (MFCC) provide a relatively low-dimensional representation via the Mel scale filter bank (Graciarena et al. 2010), which consists of linearly (below 1 kHz) and logarithmically (above 1 kHz) spaced Mel scale filters.

MFCC has been useful for human speech recognition (Makhoul and Schwartz 1995, Muda et al. 2010, Priyadarshani et al. 2012), and extended to animal vocalisations (Kogan and Margoliash 1998, Clemins and Johnson 2003, Fox et al. 2006, Lee et al. 2006, Briggs et al. 2009, Stattner et al. 2013). Mostly, MFCC are used with their first and second order derivatives in order to capture dynamic features of the vocal tract. While LPCC is a low cost approach, MFCC has proven to be more accurate for classifying animal sounds (Lévy et al. 2003). However, there are contrasting views regarding the sensitivity of MFCC to noise: Singh et al. (2012) reported that MFCC are less susceptible to additive noise, while Wu and Cao (2005) found the opposite. Given this plethora of possible features and the fact that using more features requires more training data, and can lead to less accurate results, it is necessary to find subsets of the features that are most useful. Accordingly, the goal of feature selection is

to identify redundant or unnecessary features that can be removed in order to reduce the input dimensionality while retaining most of the information, thus enabling more accurate classification. Principal component analysis (PCA) is a useful tool in this regard, and can effectively reduce the data dimensionality. For example, Somervuo and Härmä (2003) significantly reduced the dimension of their birdsong feature vectors from 1000 to 7, with 99% of the variance being explained by those seven components. Another approach, based on data mining of the spectral features was used by Vilches et al. (2006, 2007). An alternative approach is to perform a pre-classification by identifying windows that are similar, so that fewer of them are used. Vector quantization (VQ) can be used for this by clustering together similar windows.

As bird calls are temporal, there are two further processing challenges that have to be dealt with for feature selection: window size selection and temporal alignment. The first of these requirements is needed because many of the features that have been found useful for birdsong recognition are based on short time windows, and establishing a suitable window size is important, in both time and frequency range. For field recordings, where there can be birds with a wide range of different pitches and lengths of calls, both of these choices can be tricky: longer time windows have lower time resolution but higher frequency resolution, and vice versa. Graciarena et al. (2010) investigated the generalising capacity of MFCC over 92 different bird species in conjunction with a Gaussian mixture model (GMM) and found that the optimum frame length is species specific. However, they also found that frequency range optimisation is possible even though the species do not share the same frequency band (100–13 000 Hz was selected experimentally). In their experiment, the best number of filters in the Mel filter bank was 41, which is high compared to human speech (generally 13 filters). Chu and Alwan (2012) proposed an algorithm to optimise the filter bank parameters by using an Expectation Maximisation (EM) algorithm. Using a 42 min recording as test data, with an 85 min recording as training data, Chu and Alwan (2012) improved the identification error rate from 8.7 to 6.2% on the calls of five antbird species (family Thamnophilidae).

The second challenge is to align the calls within the window, a process known as temporal alignment. The most common method for performing temporal alignment is Dynamic Time Warping (DTW), initially proposed by Vintsyuk (1971) for automatic word recognition. It has been successfully used by many researchers to match birdsong (Anderson et al. 1996, Kogan and Margoliash 1998, Somervuo et al. 2006). DTW successfully copes with the different birdsong speeds and lengths by stretching and squashing sections of birdsong so as to find the best matching alignment (Coleman 2005).

While Fourier analysis is the foundation of most of the previously mentioned frequency-based features, the short-time Fourier transform suffers from a lack of time-frequency resolution. In this regard, wavelet analysis can serve as a successful alternative (Priyadarshani et al. 2016). In both the Fourier transform and the wavelet transform, a given signal is converted into frequency domain, but the difference is in the basis functions: the Fourier transform is based on sinusoids, while wavelets are based on self-similar basis functions called mother wavelets. In contrast to sinusoids, wavelet functions are localised in space and are scale-invariant. Therefore, the trade-off in time-frequency resolution can be avoided by using large windows for low frequencies and small windows for high frequencies simultaneously. Bastas et al. (2012) observed that Discrete Wavelet Transformation (DWT)-based features outperformed MFCC. Despite the opportunities that wavelets provide, relatively few publications have so far used them for birdsong analysis (Turunen et al. 2006, Selin et al. 2007, Chou and Liu 2009, Zhang and Li 2015).

Recognition and classification

Reproducing human-level processing of sight and sound is one of the holy grails of machine learning. However, despite recent advances, we are still a long way from this. Machine learning methods generally take vectors of equal length and compute representations of them in order to cluster similar inputs together. The features that comprise the vector are typically based on some subset of the methods described in the previous section. The challenge is to find a representation of the feature vectors that makes the examples of one particular type of call from one species, in all their variation, similar to each other, but dissimilar to other calls, and all calls of any other species.

Once a feature representation has been chosen, feature vectors extracted from the sound file can be fed into a standard machine learning algorithm, which will cluster those that are similar, either in an unsupervised fashion (i.e. without human labels for the calls) or using supervised learning (labels provided by human experts beforehand). Alternatively, exemplars of each call can be treated as templates of a particular type of call, and the distances between vectors representing each call and a new call can be computed, with the closest exemplar being declared a match to the new call.

There are a plethora of machine learning algorithms, and we will only mention those that have been used in the literature for birdsong recognition (Table 3). However, we will first summarise some possible distance metrics between vectors, as this will be important for comparing some of the methods.

One approach is to treat the vector as describing a position in a high dimensional space, for example a feature vector with 16 elements means that is a 16 dimensional space; often the feature vectors are rather longer than that. In this case, the common distance metrics between points can be used; the Euclidean distance, which is a particular example of a Minkowski distance (Marsland 2014), is one example of this type of metric. However, for high dimensional data these methods do not deal well with noise. The alternative is to treat the feature vectors as samples from some unknown probability distribution and seek a match there, for

example by using the Kullback–Liebler (KL) divergence: in general, $KL(A||B) \neq KL(B||A)$.

Alternatively, distance measures can be derived for particular applications, and one that is used for birdsong recognition is the geometric distance (Jinnai et al. 2012), which is the basis for SoundID (see the section Current birdsong recognition related software). This metric aims to be robust to noise by comparing the vectors with a Gaussian distribution and measuring the kurtosis. A more explicit use of the Gaussian distribution is the statistical learning method known as a Gaussian mixture model (GMM). It is assumed that observations (such as recorded birdsong) come from a weighted combination of inputs that can be described by Gaussian random variables. As some of the input can be noise in addition to types of bird call, this method can deal well, at least in theory, with multiple noise sources. One challenge is to know how many sources there are, and training them appropriately. As with the feature selection part of birdsong recognition, many machine learning-based approaches have taken their lead from automated speech recognition, even though the two contexts have as many differences as commonalities (Doupe and Kuhl 1999, Skowronski and Harris 2006, Somervuo et al. 2006, Zhang and Li 2015). In particular, the sounds that birds make are far more variable than those of humans, and the environmental conditions of the recordings are very different.

A variety of different types of neural network (NN) have been used for birdsong recognition, see for example Cai et al. (2007), Lopes et al. (2011a), Sprengel et al. (2016). However, there are two particular problems with this method: it acts as a black box and usually scales badly, so that its ability to discriminate between species degrades as a larger variety of calls or species is introduced (Cai et al. 2007). Hence, while neural network methods work fairly well for limited number of species, this falls off as the number of species increases; for example, Lopes et al. (2011a) started to classify 73 species using NN, but concluded that the performance varied considerably between the species and suggested that the maximum number of species that could be used within the F_1 margin of 80% is 12.

Neural networks that perform unsupervised learning, meaning that they cluster similar calls together rather than using human labels to recognise calls from the same species, are also commonly used, particularly the Self-Organising Map (SOM). Although Stowell and Plumbley (2014) suggest that unsupervised learning is key to successful recognition, Selin et al. (2007) observed higher recognition accuracy with a supervised method (the Multi-Layer Perceptron (MLP)), obtaining 96% test accuracy against 78% using SOM when recognising eight bird species.

The Support Vector Machine, another machine learning approach that maps data into a higher dimensional space where it can be linearly separated, has been used in many of the studies (Table 3). While it often produces very impressive results, as the amount of data increases, so the computational costs increase, and this makes it unsuitable for processing large numbers of calls.

There has also been interest in using decision trees, where at each node of the tree a split is performed using just one of the features from the input vector, and a sequence of these splits enables a decision as to species to be made at the leaves of the tree. A collection of decision trees that are created using random partitions of the data and that independently produce outputs that are combined by majority voting is known as a random forest. They are relatively easy to train and use, and have shown positive results for birdsong recognition at species level on continuous field recordings (Digby et al. 2013), and also in a number of bird species on segmented recordings (Fodor 2013, Potamitis 2014, Stowell and Plumbley 2014, Lasseck 2015). All of the methods that we have considered above use some form of feature vector constructed from a time window of the bird call. However, many birdsongs consist of a sequence of syllables. In order to recognise this, another method that is commonly used in speech recognition has been applied: the Hidden Markov Model (HMM), which creates a time-dependent probability distribution showing how likely certain syllables are to follow from others (Kogan and Margoliash 1998, Kwan et al. 2004, Trifa 2006, Jančovič and Köküer 2015).

The last method that we survey here is of particular interest as it can be part of the feature creation (Lasseck 2015) or a recognition method in its own right. Spectrogram cross-correlation takes a segment of the spectrogram and computes the cross-correlation with a set of template calls. It is simple, yet proven to be successful when scanning a specific species with limited call variations (Frommolt and Tauchert 2014, Ulloa et al. 2016). For example, spectrogram cross-correlation was used to survey screaming piha *Lipaugus vociferans* from autonomous field recordings by Ulloa et al. (2016).

Classification methods have strengths and limitations (Table 3). However, before any of them can be used for automated processing of field recordings there are three main issues that need to be considered: noise, scalability, and the addition of new bird calls. In essence, most of the methods are trained and tested on relatively low numbers of high quality recordings, whereas field recordings are inherently noisy, and can contain bird calls from many different birds. The theory seems to be that by using high quality data for training, the algorithms will deal well with noise in true recordings, although this is not tested, and does not seem particularly likely. Even when noise is included in the investigation, it is sometimes added to clean recordings (Chen and Maher 2006, Ren et al. 2008, Jančovič and Köküer 2011, Zhang and Li 2015), and thus unlikely to be typical of real environmental noise. While denoising methods can help by improving the quality of the calls, this needs to be used for the training data as well as the test data to ensure that the methods deal well with any artefacts that the denoising introduces. In an interesting approach, Papadopoulos et al. (2015) attempted to overcome the non-bird sounds associated with bird recordings by developing individual models for each

target species (in their case, 15 species) using novelty detection, so that the model discriminates the target species from noise. Using audio with a signal-to-noise ratio (SNR) of –3–3 dB (short length/manual), high variability in AUC (0.56 to 0.90) between the species was observed.

It is relatively difficult to overcome the worsening of discrimination with increasing number of species or song types. However, it is possible to improve the quality of recognition markedly by including information about where a call was recorded, since this can reduce the number of possible species that a call could come from, given the limited environmental niches of many birds.

In addition, most machine learning methods are trained off-line, based on the complete dataset, before any recognition occurs. This means that if the user wishes to add new calls or species, or even just further examples of calls that are already in the dataset, a new model has to be trained from scratch, potentially a very computationally expensive operation.

Current birdsong recognition related software

The previous sections have considered the underlying methods that have been applied to automated birdsong recognition. In this section we turn to the software that is available at the current time, and summarise their strengths and weaknesses. A comparison of the various algorithms is given in Table 4, while the text describes the approach taken and references that have used that software.

SoundID is a commercial sound recognition system that is dedicated to bio-acoustic applications such as animal surveys. The main building blocks are LPC for call pattern representation and the geometric distance (which was developed for the software) to perform the recognition (Jinnai et al. 2012). Recordings are segmented using an energy threshold, and the LPC spectral image computed for each extracted segment. Then the LPC pattern (image) is compared with the stored reference patterns using the geometric distance.

The SoundID group reported more than 95% accuracy in analysing the dawn chorus (Jinnai et al. 2012) and they particularly highlight the efficacy of the software for processing large datasets. However, they are clearly expert users, and in our experience it is hard to achieve a reasonable recall rate (Table 4) and optimisation of the parameters is time-consuming and difficult.

Raven Pro is developed by the Cornell Lab of Ornithology for acquisition, visualisation, measurement, and analysis of sounds (Charif et al. 2010) and is popular among ecologists as a spectrogram analysis tool (Bura et al. 2011, Crothers et al. 2011, Kirschel et al. 2011, Vernaleo and Dooling 2011, Sandoval and Barrantes 2012, Aland and Hoskin 2013, Aleixandre et al. 2013, Arévalo and Araya-Salas 2013). The relevant tools provided are band-pass noise filtering and manual or semi-automatic syllable segmentation (Stowell and Plumbley 2011). A couple of syllable-level bird sound detection methods are included that are based on either estimation of background noise or a user-defined SNR, and there are many other user-defined parameters such as duration, or low-pass smoothing (Charif et al. 2010, Duan et al. 2013). A few studies have attempted to apply these automatic detectors to real world data. For example, Sebastián-González et al. (2015) used band-limited energy detector to detect call events of Hawai'i 'amakihi *Hemignathus virens* from field recordings with 93% recall but only 16.8% of them were good selections (precise endpoints). Duan et al. (2013) configured the segmentation module separately for each of five species found in the dawn chorus (based on five hours of data). The overall accuracy is reported as 43%, far below expectation for such a limited number of species.

Song Scope, developed by Wildlife Acoustics, is also equipped with a call detector (Wildlife Acoustics 2011). The feature representation and classification algorithms are MFCC (Agranat 2009, Duan et al. 2011) and HMM (as Song Scope works on clustering multiple syllables). While the developers recommend the software for field biologists to analyse long field recordings made by autonomous recording devices, the main drawback is that their approach is very sensitive to noise (Duan et al. 2013). Therefore, post hoc visual scanning was employed by Buxton et al. (2013) and Cragg et al. (2015) to filter out false positives generated by the Song Scope detectors tailored to detect different call types of nocturnal sea birds. Duan et al. (2013) optimised Song Scope's detector for the same data set used for Raven Pro. Overall accuracy was reported as 37%, which is less than with Raven Pro (43%). Noisy training data decreased the accuracy of Song Scope more significantly than Raven Pro because the HMM treats noise segments as syllables and models them into call structures (Duan et al. 2013). Recently, Wildlife Acoustics introduced Kaleidoscope, an integrated suite of tools for bio-acoustic analysis advancing Song Scope. While the developers claim that the software can generate a set of reports analysing the sound files, the success is yet to be evaluated on third party data.

Sound Analysis Pro 2011 is free open-source software for recording and analysis of animal vocalisation based on less complex features extracted from whole birdsong (Tchernichovski et al. 2000, Tchernichovski 2012). Although they have considered segmentation of songs into syllables and syllable clustering, they report that in this respect the software has limitations. The primary focus is not the analysis of field recordings: the developers recommend the software to be utilised to train animals with playbacks while recording their vocalisation, e.g. throughout the vocal development of a bird, to see the tutor-pupil song relationship (Tchernichovski et al. 2000). Daou et al. (2012) used the features generated from Sound Analysis Pro as the input to their software tool which analyses the syllable transitions within the songs of individual birds.

Avisoft-SASLab Pro (Specht 1993) is a commercial, general purpose sound analysis software that was created by Avisoft Bioacoustics in 1990. The company also provides a

Table 4. A summary of currently available software.

Software	Open access	Main purpose	Pre-processing	Segmentation	Feature representation	Recognition	Performance
SoundID	No	Automatically process field recordings and estimate call counts	Band-pass filter	Time domain energy threshold	LPC	Geometric distance	Developers claim the system can challenge a human expert. Our own best efforts to scan field recordings (morepork) achieved only 40% recall (92% precision) after time consuming parameter tuning (Brighten 2015)
monitoR (R package)	Yes	Automatically process field recordings and estimate call counts	Band-pass filter	Detections are based on user-defined threshold either on the score envelope generated by binary point matching or spectrogram cross-correlation	Two options: mapping anticipated regions of signal within a spectrogram (for binary point matching) or matrix of amplitudes (for spectrogram cross-correlation)	Two options to score each time frame of the test signal: binary point matching or pearson correlation – choosing a score threshold is a user-driven process	The developers evaluated the system with two distinctive species considering only one call type from each species – identification accuracies for two call types were 64% and 72% (binary point matching) and 73% and 72% (using spectrogram cross-correlation) – one template from each call type
Raven Pro	No	Manual review of spectrograms and measure sounds	Band-pass filter	Energy threshold in time (amplitude detector) and frequency (band-limited energy detector)	Facilitates to extract a list of time-frequency features	Not addressed	Developers report that the detectors give false positives even when optimally configured
Song Scope	No	Manual review of field recordings	Noise filtering – requires an estimate of noise	Based on the energy threshold (user-defined frequency band)	MFCC	HMM	Noisy data dramatically decreases the accuracy – recommend to extract training examples from a range of recordings
Kaleidoscope Pro	No	Bat call analysis (also supports non-bat sounds)	Band-pass filter – discard noise only files initially (bat)	A threshold guided method	Zero-crossing – duration	HMM	Uneven accuracy across the species
Avisoft-SASLab Pro	No	Spectrogram analysis	Removes noise below a user defined threshold in frequency	Energy thresholding in time domain	Peak frequency, amplitude at peak frequency, bandwidth, and number of harmonics	Facilitates to cross-correlate sounds	Fail to capture faint calls – encourage to use good quality recordings – developers make no claim about the accuracy

(*Continued*)

Table 4. Continued

Software	Open access	Main purpose	Pre-processing	Segmentation	Feature representation	Recognition	Performance
Sound Analysis Pro	Yes	Assessment of vocal imitation and song development in birds	Focus on noise-free recordings	Complete songs (algorithm does not require the songs to be partitioned into syllables)	Wiener entropy, spectral continuity, pitch, frequency modulation	Based on Euclidean distance	Enough examples to cover call variations
eXtensible BioAcoustic Tool (XBAT)	Yes	Automatically process field recordings, human evaluation and annotation of recordings – support extendability	Not disclosed and could not determine	Not disclosed and could not determine	Not disclosed and could not determine	Not disclosed and could not determine	The developers tested the data template detector on two species, the cerulean warbler *Dendroica cerulean* and the whip-poor-will *Caprimulgus vociferus*. Their best reported results are 100% precision/54% recall and 96% precision/80% recall respectively (Clark and Fristrup 2009)
Arbimon	No	Online storage of short recordings and upload the final analysis	Not disclosed and could not determine	Energy threshold in frequency domain	Frequency range, duration, maximum intensity, and bandwidth	HMM	Developers make no claim about the accuracy – different models for different species (call types) are needed
Praat	Yes	General purpose sound (human speech) analysis	Band-pass filter	Not addressed	Time domain and frequency domain features	Equipped with NN, but not encouraged to use for real-world applications	Useful to annotate and manually segment sound – equipped with programming extension (scripts)
Sonobat	No	Bat call analysis	Filter noisy files and to discard them initially	A threshold guided method	Time-frequency and time-amplitude features (72)	Not disclosed and could not determine	Fails to detect poor quality calls
BatSound	No	Basic bat call analysis	Not addressed	A threshold guided method	Manually measured frequency and temporal features	Not addressed	Far behind the practical requirements
Luscinia	Yes	Archive, measure, and analyse recordings	Band-pass filter	Not relevant or addressed	15 acoustic parameters (contours and hierarchical information)	DTW	> 20 publications, none of them used the software for surveying birds except behavioural studies based on their acoustics
Syrinx	No	Playback of animal sounds, recording, and analysis	Noise profiling – band-pass filter	A threshold guided method	Time-frequency measurements	Not addressed	Useful for ecologists to playback with a minimum of equipment

(*Continued*)

Table 4. Continued

Software	Open access	Main purpose	Pre-processing	Segmentation	Feature representation	Recognition	Performance
PAMGuard	Yes	Passive acoustic monitoring – ocean acoustics	Median filter – average subtraction – thresholding – Gaussian smoothing	Energy threshold	Time-frequency measurements	User-defined classifiers	Not reliable with more species (accuracy dropped from 95% to 59% when the number of species changed from 4 to 12)
SongSeq	Yes	Syllable clustering	Not addressed	Not relevant or addressed	Sound Analysis Pro generated features	Not relevant or addressed	Templates are required
SIGNAL	No	Event detector and analyser modules auto extract and measure sound events from recordings	Band-pass filter	Energy thresholding in time domain	Call duration, call rate, peak frequency, frequency range	Method is not disclosed and could not determine	Set of spectral, temporal, and amplitude parameters to be determined by the user
Ishmael	Yes	Manual review of spectrograms	Not disclosed and could not determine	Energy threshold	Not disclosed and could not determine	Spectrogram correlation	Not reported
SpectraPRO, SpectraLAB	No	Manual review of spectrograms – focus on ocean acoustics	Not relevant or addressed	Not relevant or addressed	Not relevant or addressed	Not relevant or addressed	Not relevant or addressed

bioacoustics recording device and separate recording software. Frommolt and Tauchert (2014) successfully used the software to recognise Eurasian bitterns *Botaurus stellaris*, a species that generates very low frequency and relatively simple vocalisations. The software is useful to automatically measure sound parameters of the spectrogram and waveform. A long list of publications that used this software can be found under the reference list on their web site.

Arbimon is a web-based network for storing, sharing, and analysing acoustic data. Their cyber infrastructure includes a solar-powered remote monitoring station that sends 1-min recordings every 10 min to a base station, which relays the recordings in real-time to the project server, where the recordings are processed and uploaded to the project website (Aide et al. 2013). Recordings to be analysed need be uploaded to Arbimon in order to see the results. However, they do not report the accuracy, and their online recognition facility is expensive and not feasible for the processing of long field recordings.

While we have discussed some of the software tools above, more are given in Table 4.

All the systems that we reviewed have strengths and weaknesses on their own. Overall, the software tools are far behind the practical requirements demonstrating that further developments are essential, particularly, to overcome the problems generated by unwanted sounds (noise) in field recordings and also to deal with faint bird sounds recorded due to the spatial distribution of birds in their natural habitats.

Conclusions

1) While there has been a significant amount of research into computer recognition of bird species based on their calls, this survey has demonstrated that it is not yet sufficient for practical use for data from unattended field recordings. Our review shows that more than two thirds of the work published to date has limited their scope to analyse less noisy and carefully selected recordings. Thus, even though the reported accuracies are impressive, this is because they are largely based on small datasets and relatively clean data.

2) It is important to consider what is wanted from a system: for conservation purposes it is commonly desirable to have extremely accurate (high precision, high recall) recognition of a small number of target species from a set that includes many other species, while for general recognition (such as for a smartphone app) it is sufficient to have reasonably accurate recognition of a large number of species. While there are some similarities between these cases (particularly that noise will be present and the bird will often call a long way from the microphone), they are importantly different in the accuracy requirements.

3) While it is obviously important that the methods be as accurate as possible, the importance of the recall rate for rare species is worth noting, since for these birds confusing their sound with another (more common) species can lead to over-estimate of their abundance, or miss the fact that they

exist at all in the area. Unfortunately, current call detection and segmentation methods generally show poor recall rates and high sensitivity to noise. Our own best efforts using SoundID to scan automatic recordings of morepork *Ninox novaeseelandiae*, the only extant native owl in New Zealand, achieved only 39% recall rate (with 92% precision) after a time-consuming trial and error process for software parameter optimisation (Brighten 2015). This finding highlights another point that is important: overall classification accuracy is not necessarily the most useful measure of utility for a birdsong recogniser; recall and precision provide more useful information.

4) Further research into methods to deal with the noise inherent in field recordings is clearly needed. Many poor results are attributed to high noise levels in at least one of the training and test sets, but conventional noise reduction methods fail to remove noise, particularly in the target frequency range, without losing information.

5) Our review has suggested that the methods that have been successful at screening field recordings are species-specific. Although current species-specific methods cannot be quickly and easily modified to screen other species (mainly because the features are carefully selected for that species, Digby et al. (2013), we believe that for conservation purposes species-specific methods will generally be more successful.

6) It is important that benchmark datasets are available, so that different researchers can compare their methods on the same datasets, and using the same metrics. There are a number of high-quality comparison datasets readily available in the form of bird recognition challenges, such as BirdCLEF (<http://imageclef.org/lifeclef/2016/bird>), NIPS4B (Glotin et al. 2013), and the MLSP 2013 bird classification challenge (Briggs et al. 2013). However, all these datasets provide short recordings rather than the long automated recordings. In addition, they are focused on producing reasonable accuracy on multiple recordings of a large number of bird species. There is the need for shared datasets with annotations of a wide variety of calls for a large number of species if methods that are suitable for conservation work are to be developed.

The RMBL-robin database (<www.seas.ucla.edu/spapl/projects/Bird.html>), an American robin *Turdus migratorius* database, manually recorded with SM1 recorders (but continuous recordings) is readily available with their annotations (song level and syllable level) from UCLA (Chu and Blumstein 2011). For conservation purposes, it is more important to detect individual species extremely accurately, and the detectors can be specialised since the vast majority of the species included in these datasets will not be present in any given conservation area.

There are also datasets such as the Xeno-canto (<www.xeno-canto.org/>) collaborative database, the Cornell Lab of Ornithology's archive of sample bird calls (<http://macaulaylibrary.org/>), and the Tierstimmenarchiv (<www.tierstimmenarchiv.de>) at the Museum für Naturkunde in Berlin; indeed, Ranft (2004) estimated that more than 90% of world's bird species had been represented in major sound archives by 2003. However, for the training of automated recognition systems it is necessary to have a large number of field-noise corrupted raw recordings with accompanying ground truth annotations.

Acknowledgements – Authors would like to thank Amal Punchihewa for comments on the draft of the manuscript.
Funding – This study was partially funded by Higher Education for the Twenty First Century, Sri Lanka (KLN/O-Sci/N6) and School of Engineering and Advanced Technology, Massey Univ., New Zealand (RM1000015982 P-MURF). NP acknowledges funding from WWF Conservation Innovation Awards 2014 – New Ideas for Nature (research innovation) (CIA 14/05) and the support from Univ. of Kelaniya, Sri Lanka.

References

Acevedo, M. A. and Villanueva-Rivera, L. J. 2006. Using automated digital recording systems as effective tools for the monitoring of birds and amphibians. – Wildl. Soc. Bull. 34: 211–214.

Acevedo, M. A., Corrada-Bravo, C. J., Corrada-Bravo, H., Villanueva-Rivera, L. J. and Aide, T. M. 2009. Automated classification of bird and amphibian calls using machine learning: a comparison of methods. – Ecol. Inform. 4: 206–214.

Agranat, I. 2009. Automatically identifying animal species from their vocalizations. – Proceedings of the 5th International Conference on Bio-Acoustics, Holywell Park.

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

Aland, K. V. and Hoskin, C. J. 2013. The advertisement call and clutch size of the golden-capped boulder-frog *Cophixalus pakayakulangun* (Anura: Microhylidae). – ZooTaxa 3718: 299–300.

Aleixandre, P., Montoya, J. H. and Milá, B. 2013. Speciation on Oceanic islands: rapid adaptive divergence vs. cryptic speciation in a Guadalupe Island songbird (Aves: *Junco*) – PLoS One 8: e63242.

Anderson, S. E., Dave, A. S. and Margoliash, D. 1996. Template-based automatic recognition of birdsong syllables from continuous recordings. – J. Acoust. Soc. Am. 100: 1209–1219.

Andreassen, T., Surlykke, A. and Hallam, J. 2014. Semi-automatic long-term acoustic surveying: a case study with bats. – Ecol. Inform. 21: 13–24.

Angehr, G. R., Siegel, J., Aucca, C., Christian, D. G. and Pequeño, T. 2002. An assessment and monitoring program for birds in the Lower Urubamba Region, Peru. – Environ. Monit. Assess. 76: 69–87.

Arévalo, J. E. and Araya-Salas, M. 2013. Collared forest-falcon (*Micrastur semitorquatus*) preying on chestnut-mandibled toucan (*Ramphastos swainsonii*) in Costa Rica. – Wilson. J. Ornithol. 125: 212–216.

Baker, M. C. and Logue, D. M. 2003. Population differentiation in a complex bird sound: a comparison of three bioacoustical analysis procedures. – Ethology 109: 223–242.

Baker, M. C. and Logue, D. M. 2007. A comparison of three noise reduction procedures applied to bird vocal signals. – J. Field Ornithol. 78: 240–253.

Baldo, S. and Mennill, D. J. 2011. Vocal behavior of great curassows, a vulnerable neotropical bird. – J. Field Ornithol. 82: 249–258.
Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K. H. and Frommolt, K. H. 2010. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. – Pattern Recognit. Lett. 31: 1524–1534.
Barraclough, R. K. 2000. Distance sampling: a discussion document produced for the Department of Conservation. – Dept of Conservation Science and Research Internal Report 175.
Bastas, S., Majid, M. W., Mirzaei, G., Ross, J., Jamali, M. M., Gorsevski, P. V., Frizado, J. and Bingman, V. P. 2012. A novel feature extraction algorithm for classification of bird flight calls. – Proc. IEEE Int. Symp. Circuits and Systems, pp. 1676–1679.
Boucher, N., Jinnai, M. and Smolders, A. 2012. A fully automatic wildlife acoustic monitor and survey system. – Proc. Acoustics 2012 Nantes Conf.
Boucher, N. J. 2014. SoundID version 2.0.0 documentation. – SoundID.
Brandes, T. S. 2008. Automated sound recording and analysis techniques for bird surveys and conservation. – Bird Conserv. Int. 18 (Suppl. 1): S163–S173.
Briggs, F., Raich, R. and Fern, X. Z. 2009. Audio classification of bird species: a statistical manifold approach. – Proc. 9th IEEE Int. Conf. Data Mining, pp. 51–60.
Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X. Z., Raich, R., Hadley, S. J., Hadley, A. S. and Betts, M. G. 2012. Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. – J. Acoust. Soc. Am. 131: 4640–4650.
Briggs, F., Huang, Y., Raich, R., Eftaxias, K., Lei, Z., Cukierski, W., Hadley, S. F., Hadley, A., Betts, M., Fern, X. Z. et al. 2013. The 9th annual MLSP competition: new methods for acoustic classification of multiple simultaneous bird species in a noisy environment. – Proc. IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–8.
Brighten, A. 2015. Vocalisations of the New Zealand morepork (*Ninox novaeseelandiae*) on Ponui Island. – Master's thesis, Massey Univ., Palmerston North, New Zealand.
Bura, V. L., Rohwer, V. G., Martin, P. R. and Yack, J. E. 2011. Whistling in caterpillars (*Amorpha juglandis*, Bombycoidea): sound-producing mechanism and function. – J. Exp. Biol. 214: 30–37.
Buxton, R. T., Major, H. L., Jones, I. L. and Williams, J. C. 2013. Examining patterns in nocturnal seabird activity and recovery across the Western Aleutian Islands, Alaska, using automated acoustic recording. – Auk 130: 331–341.
Bye, S. L., Robel, R. J. and Kemp, K. E. 2001. Effects of human presence on vocalizations of grassland birds in Kansas. – Prairie Nat. 33: 249–256.
Cai, J., Ee, D., Pham, B., Roe, P. and Zhang, J. 2007. Sensor network for the monitoring of ecosystem: Bird species recognition. – Proceedings of the 3rd IEEE Int. Conf. Intelligent Sensors, Sensor Networks and Information, pp. 293–298.
Catchpole, C. K. and Slater, P. J. 2003. Bird song: biological themes and variations. – Cambridge Univ. Press.
Charif, R., Strickman, L. and Waack, A. 2010. Raven Pro 1.4 user's manual. Revision 11. – The Cornell Lab of Ornithology, Ithaca, NY.
Chen, J., Benesty, J., Huang, Y. and Doclo, S. 2006. New insights into the noise reduction Wiener filter. – IEEE Trans. Audio Speech Language Process. 14: 1218–1234.
Chen, Z. and Maher, R. C. 2006. Semi-automatic classification of bird vocalizations using spectral peak tracks. – J. Acoust. Soc. Am. 120: 2974–2984.
Cheng, J., Sun, Y. and Ji, L. 2010. A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines. – Pattern Recognit. 43: 3846–3852.
Chou, C.-H. and Liu, P.-H. 2009. Bird species recognition by wavelet transformation of a section of birdsong. – Proc. IEEE Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing. UIC-ATC'09, pp. 189–193.
Chu, W. and Alwan, A. 2009. A correlation-maximization denoising filter used as an enhancement frontend for noise robust bird call classification. – INTERSPEECH, pp. 2831–2834.
Chu, W. and Blumstein, D. T. 2011. Noise robust bird song detection using syllable pattern-based hidden Markov models. – Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp. 345–348.
Chu, W. and Alwan, A. 2012. FBEM: a filter bank EM algorithm for the joint optimization of features and acoustic model parameters in bird call classification. – Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp. 1993–1996.
Clark, C. W. and Fristrup, K. M. 2009. Advanced technologies for acoustic monitoring of bird populations. – Technical report, Cornell Univ., Ithaca, NY.
Clemins, P. J. and Johnson, M. T. 2003. Application of speech recognition to African elephant (*Loxodonta africana*) vocalizations. – IEEE Int. Conf. Acoust. Speech Signal Process. 1: 484–487.
Colbourne, R. and Digby, A. 2016. Call rate behaviour of brown kiwi (*Apteryx mantelli*) and great spotted kiwi (*A. haastii*) in relation to temporal and environmental parameters. – Dept of Conservation Research and Development Series 348.
Coleman, J. 2005. Introducing speech and language processing. – Cambridge Univ. Press.
Cragg, J. L., Burger, A. E. and Piatt, J. F. 2015. Testing the effectiveness of automated acoustic sensors for monitoring vocal activity of marbled murrelets *Brachyramphus marmoratus*. – Mar. Ornithol. 43: 151–160.
Crothers, L., Gering, E. and Cummings, M. 2011. Aposematic signal variation predicts male–male interactions in a polymorphic poison frog. – Evolution 65: 599–605.
Cunningham, R., Lindenmayer, D. and Lindenmayer, B. D. 2004. Sound recording of bird vocalisations in forests. I. Relationships between bird vocalisations and point interval counts of bird numbers–a case study in statistical modeling. – Wildl. Res. 31: 195–207.
Daou, A., Johnson, F., Wu, W. and Bertram, R. 2012. A computational tool for automated large-scale analysis and measurement of bird-song syntax. – J. Neurosci. Methods 210: 147–160.
Dawson, D. and Bull, P. 1975. Counting birds in New Zealand forests. – Notornis 22: 101–109.
Dawson, D. K. and Efford, M. G. 2009. Bird population density estimated from acoustic signals. – J. Appl. Ecol. 46: 1201–1209.
de Oliveira, A. G., Ventura, T. M., Ganchev, T. D., de Figueiredo, J. M., Jahn, O., Marques, M. I. and Schuchmann, K.-L. 2015. Bird acoustic activity detection based on morphological filtering of the spectrogram. – Appl. Acoust. 98: 34–42.
Dent, J. M. and Molles, L. E. 2016. Call-based identification as a potential tool for monitoring great spotted kiwi. – Emu 116: 315–322.

Digby, A., Towsey, M., Bell, B. D. and Teal, P. D. 2013. A practical comparison of manual and autonomous methods for acoustic monitoring. – Methods Ecol. Evol. 4: 675–683.
Dong, X., Towsey, M., Truskinger, A., Cottman-Fields, M., Zhang, J. and Roe, P. 2015. Similarity-based birdcall retrieval from environmental audio. – Ecol. Inform. 29: 66–76.
Doupe, A. J. and Kuhl, P. K. 1999. Birdsong and human speech: common themes and mechanisms. – Annu. Rev. Neurosci. 22: 567–631.
Duan, S., Towsey, M., Zhang, J., Truskinger, A., Wimmer, J. and Roe, P. 2011. Acoustic component detection for automatic species recognition in environmental monitoring. – Proc. 7th Int. Conf. Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), pp. 514–519.
Duan, S., Zhang, J., Roe, P., Wimmer, J., Dong, X., Truskinger, A. and Towsey, M. 2013. Timed probabilistic automaton: a bridge between Raven and Song Scope for automatic species recognition. – Proc. 25th Innovative Applications of Artificial Intelligence Conference, pp. 1519–1524.
Dufour, O., Artieres, T., Glotin, H. and Giraudet, P. 2013. Clusterized mel filter cepstral coefficients and support vector machines for bird song identification. – Proc. 1st workshop on Machine Learning for Bioacoustics 951: 89–93.
Elliot, L., Burwell, C. J., Ashton, L. A., Jones, D. and Kitching, R. L. 2016. Comparison of point counts and automated acoustic monitoring: detecting birds in a rainforest biodiversity survey. – Emu 116: 305–309.
Emlen, J. T. and DeJong, M. J. 1992. Counting birds: the problem of variable hearing abilities. – J. Field Ornithol. 63: 26–31.
Eyben, F., Wöllmer, M. and Schuller, B. 2010. OpenSMILE: the munich versatile and fast open-source audio feature extractor. – Proc. 18th ACM Int. Conf. Multimedia, pp. 1459–1462.
Fagerlund, S. 2004. Automatic recognition of bird species by their sounds. – PhD thesis, Helsinki Univ. Tech.
Fagerlund, S. 2007. Bird species recognition using support vector machines. – EURASIP J. Appl. Signal Process. 2007: 64–64.
Farina, A. 2014. Soundscape ecology: principles, patterns, methods and applications. – Springer.
Figueira, L., Tella, J. L., Camargo, U. M. and Ferraz, G. 2015. Autonomous sound monitoring shows higher use of Amazon old growth than secondary forest by parrots. – Biol. Conserv. 184: 27–35.
Fodor, G. 2013. The ninth annual MLSP competition: first place. – Proc. IEEE Int. Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–2.
Fox, E. J., Roberts, J. D. and Bennamoun, M. 2006. Text-independent speaker identification in birds. – INTERSPEECH ICSLP, pp. 2122–2125.
Fox, E. J., Roberts, J. D. and Bennamoun, M. 2008. Call-independent individual identification in birds. – Bioacoustics 18: 51–67.
Franzen, A. and Gu, I. Y. 2003. Classification of bird species by using key song searching: a comparative study. – Proc. SMC 1: 880–887.
Frommolt, K.-H. and Tauchert, K.-H. 2014. Applying bioacoustic methods for long-term monitoring of a nocturnal wetland bird. – Ecol. Inform. 21: 4–12.
Ganchev, T., Mporas, I., Jahn, O., Riede, K., Schuchmann, K.-L. and Fakotakis, N. 2012. Acoustic bird activity detection on real-field data. – In: Maglogiannis, I., Plagianakos, V. and Vlahavas, I. (eds), Artificial intelligence: theories and applications. Springer, pp. 190–197.
Ganchev, T. D., Jahn, O., Marques, M. I., de Figueiredo, J. M. and Schuchmann, K. L. 2015. Automated acoustic detection of *Vanellus chilensis lampronotus*. – Expert Syst. Appl. 42: 6098–6111.
Gilbert, G., McGregor, P. K. and Tyler, G. 1994. Vocal individuality as a census tool: practical considerations illustrated by a study of two rare species. – J. Field Ornithol. 65: 335–348.
Glotin, H., LeCun, Y., Artiéres, T., Mallat, S., Tchernichovski, O. and Halkias, X. 2013. Neural information processing scaled for bioacoustics, from neurons to big data. – Proc. NIPS4B.
Goyette, J. L., Howe, R. W., Wolf, A. T. and Robinson, W. D. 2011. Detecting tropical nocturnal birds using automated audio recordings. – J. Field Ornithol. 82: 279–287.
Graciarena, M., Delplanche, M., Shriberg, E., Stolcke, A. and Ferrer, L. 2010. Acoustic front-end optimization for bird species recognition. – Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp. 293–296.
Griffin, D. and Lim, J. 1984. Signal estimation from modified short-time Fourier transform. – IEEE Trans. Acoust. Speech Signal Process. 32: 236–243.
Härmä, A. 2003. Automatic identification of bird species based on sinusoidal modeling of syllables. – Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 5: V–545–548.
Härmä, A. and Somervuo, P. 2004. Classification of the harmonic structure in bird vocalization. – Proc. IEEE Int. Conf. Acoust. Speech Signal Process. 5: V–701–704.
Heller, J. R. and Pinezich, J. D. 2008. Automatic recognition of harmonic bird sounds using a frequency track extraction algorithm. – J. Acoust. Soc. Am. 124: 1830–1837.
Hill, S. D., Ji, W., Parker, K. A., Amiot, C. and Wells, S. J. 2013. A comparison of vocalisations between mainland tui (*Prosthemadera novaeseelandiae novaeseelandiae*) and Chatham Island tui (*P. n. chathamensis*). – N. Z. J. Ecol. 37: 214–223.
Hutto, R. L. and Stutzman, R. J. 2009. Humans versus autonomous recording units: a comparison of point-count results. – J. Field Ornithol. 80: 387–398.
IUCN. 2014. The IUCN red list of threatened species. – International Union for Conservation of Nature and Natural Resources.
Jančovič, P. and Köküer, M. 2011. Automatic detection and recognition of tonal bird sounds in noisy environments. – EURASIP J. Adv. Signal Process. 2011: 1–10.
Jančovič, P. and Köküer, M. 2015. Acoustic recognition of multiple bird species based on penalized maximum likelihood. – IEEE Signal Process. Lett. 22: 1585–1589.
Jinnai, M., Boucher, N., Fukumi, M. and Taylor, H. 2012. A new optimization method of the geometric distance in an automatic recognition system for bird vocalisations. – Proceedings of the Acoustics 2012 Nantes Conference, pp. 2439–2445
Juang, C.-F. and Chen, T.-M. 2007. Birdsong recognition using prediction-based recurrent neural fuzzy networks. – Neurocomputing 71: 121–130.
Kasten, E. P., McKinley, P. K. and Gage, S. H. 2010. Ensemble extraction for classification and detection of bird species. – Ecol. Inform. 5: 153–166.
Katz, J., Hafner, S. D. and Donovan, T. 2016. Assessment of error rates in acoustic monitoring with the R package monitoR. – Bioacoustics 25: 1–20.
Kirschel, A. N., Cody, M. L., Harlow, Z. T., Promponas, V. J., Vallejo, E. E. and Taylor, C. E. 2011. Territorial dynamics of Mexican ant-thrushes *Formicarius moniliger* revealed by individual recognition of their songs. – Ibis 153: 255–268.

Kogan, J. A. and Margoliash, D. 1998. Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: a comparative study. – J. Acoust. Soc. Am. 103: 2185–2196.
Kwan, C., Mei, G., Zhao, X., Ren, Z., Xu, R., Stanford, V., Rochet, C., Aube, J. and Ho, K. 2004. Bird classification algorithms: theory and experimental results. – Proc. IEEE Int. Conf. Acoust. Speech Process. 5: 289–292.
Kwan, C., Ho, K., Mei, G., Li, Y., Ren, Z., Xu, R., Zhang, Y., Lao, D., Stevenson, M., Stanford, V. and Rochet, C. 2006. An automated acoustic system to monitor and classify birds. – EURASIP J. Adv. Signal Process. 2006: 1–19.
Landau, H. J. 1967. Sampling, data transmission, and the Nyquist rate. – Proc. IEEE 55: 1701–1706.
Lasseck, M. 2013. Bird song classification in field recordings: winning solution for NIPS4B 2013 competition. – Proc. Int. Symp. Neural Information Scaled for Bioacoustics, joint to NIPS, Nevada, pp. 176–181.
Lasseck, M. 2014. Large-scale identification of birds in audio recordings. – Working Notes of CLEF, pp. 643–653.
Lasseck, M. 2015. Improved automatic bird identification through decision tree based feature selection and bagging. – Working Notes of CLEF.
Lee, C.-H., Lee, Y.-K. and Huang, R.-Z. 2006. Automatic recognition of bird songs using cepstral coefficients. – J. Inform. Technol. Appl. 1: 17–23.
Lee, C.-H., Han, C.-C. and Chuang, C.-C. 2008. Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients. – IEEE Trans. Audio Speech Language Process. 16: 1541–1550.
Lee, C. H., Hsu, S. B., Shih, J. L. and Chou, C. H. 2013. Continuous birdsong recognition using gaussian mixture modeling of image shape features. – IEEE Trans. Multimedia 15: 454–464.
Lévy, C., Linarès, G. and Nocera, P. 2003. Comparison of several acoustic modelling techniques and decoding algorithms for embedded speech recognition systems. – Workshop on DSP in Mobile and Vehicular Systems, Nagoya, Japan, CiteSeer.
Lopes, M. T., Gioppo, L. L., Higushi, T. T., Kaestner, C. A., Silla Jr, C. N. and Koerich, A. L. 2011a. Automatic bird species identification for large number of species. – Proc. IEEE Int. Symp. Multimedia, pp. 117–122.
Lopes, M. T., Junior, C. N. S., Koerich, A. L. and Kaestner, C. A. A. 2011b. Feature set comparison for automatic bird species identification. – Proc. IEEE Int. Conf. Syst. Man Cybern., pp. 965–970.
Loyn, R. H. 1985. The 20-minute search: a simple method for counting forest birds. – Biological Survey Branch, State Forests and Lands Service.
Madhu, N. 2009. Note on measures for spectral flatness. – Electron. Lett. 45: 1195–1196.
Makhoul, J. and Schwartz, R. 1995. State of the art in continuous speech recognition. – Proc. Natl Acad. Sci. USA 92: 9956–9963.
Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., Harris, D. and Tyack, P. L. 2013. Estimating animal population density using passive acoustics. – Biol. Rev 88: 287–309.
Marsland, S. 2014. Machine learning: an algorithmic perspective, 2nd ed. – Chapman and Hall/CRC.
McIlraith, A. L. and Card, H. C. 1997. Birdsong recognition using backpropagation and multivariate statistics. – IEEE Trans. Sig. Process. 45: 2740–2748.
McLaren, M. A. and Cadman, M. D. 1999. Can novice volunteers provide credible data for bird surveys requiring song identification? – J. Field Ornithol. 70: 481–490.
Mellinger, D. K., Stafford, K. M., Moore, S. E., Dziak, R. P. and Matsumoto, H. 2007. An overview of fixed passive acoustic observation methods for cetaceans. – Oceanography 20: 36–45.
Mindlin, G. 2013. The physics of birdsong production. – Contemp. Phys. 54: 91–96.
Moffat, D., Ronan, D. and Reiss, J. D. 2015. An evaluation of audio feature extraction toolboxes. – Proc. 18th Int. Conf. Digital Audio Effects (DAFx-15).
Muda, L., Begam, M. and Elamvazuthi, I. 2010. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. – J. Comput. 2: 138–143
Mundry, R. and Sommer, C. 2004. Tonal vocalizations in a noisy environment: an approach to their semi-automatic analysis and examples of its application. – An. Acad. Bras. Ciênc. 76: 284–288.
Murcia, R. H. and Paniagua, V. S. 2013. Bird identification from continuous audio recordings. – Proc. 1st workshop on Machine Learning for Bioacoustics joint to the 30th ICML 2013: 96–97.
Neal, L., Briggs, F., Raich, R. and Fern, X. Z. 2011. Time-frequency segmentation of bird song in noisy acoustic environments. – Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp. 2012–2015.
Papadopoulos, T., Roberts, S. and Willis, K. 2015. Detecting bird sound in unknown acoustic background using crowdsourced training data. – ArXiv e-prints arXiv: 1505.06443.
Pijanowski, B. C., Farina, A., Gage, S. H., Dumyahn, S. L. and Krause, B. L. 2011. What is soundscape ecology? An introduction and overview of an emerging new science. – Landscape Ecol. 26: 1213–1232.
Potamitis, I. 2014. Automatic classification of a taxon-rich community recorded in the wild. – PLoS One 9: e96936.
Potamitis, I., Ntalampiras, S., Jahn, O. and Riede, K. 2014. Automatic bird sound detection in long real-field recordings: applications and tools. – Appl. Acoust. 80: 1–9.
Priyadarshani, N. 2017. Wavelet-based birdsong recognition for conservation. – PhD thesis, Massey Univ., Palmerston North, New Zealand.
Priyadarshani, N., Marsland, S., Castro, I. and Punchihewa, A. 2016. Birdsong denoising using wavelets. – PLoS One 11: e0146790.
Priyadarshani, P., Dias, N. and Punchihewa, A. 2012. Dynamic time warping based speech recognition for isolated sinhala words. – Proc. IEEE 55th Int. Midwest Symp. Circuits Syst., pp. 892–895.
Ptacek, L., Machlica, L., Linhart, P., Jaska, P. and Muller, L. 2016. Automatic recognition of bird individuals on an open set using as-is recordings. – Bioacoustics 25: 55–73.
Ranft, R. 2004. Natural sound archives: past, present and future. – An. Acad. Bras. Ciênc. 76: 456–460.
Ranjard, L., Withers, S. J., Brunton, D. H., Ross, H. A. and Parsons, S. 2015. Integration over song classification replicates: Song variant analysis in the hihi. – J. Acoust. Soc. Am. 137: 2542–2551.
Ren, Y., Johnson, M. T. and Tao, J. 2008. Perceptually motivated wavelet packet transform for bioacoustic signal enhancement. – J. Acoust. Soc. Am. 124: 316–327.
Rocha, L. H., Ferreira, L. S., Paula, B. C., Rodrigues, F. H. and Sousa-Lima, R. S. 2015. An evaluation of manual and automated

methods for detecting sounds of maned wolves (*Chrysocyon brachyurus illiger* 1815). – Bioacoustics 24: 185–198.
Rosenstock, S. S., Anderson, D. R., Giesen, K. M., Leukering, T., Carter, M. F. and Thompson III, F. 2002. Landbird counting techniques: current practices and an alternative. – Auk 119: 46–53.
Ross, D. J. 2006. Bird call recognition with artificial neural networks, support vector machines, and kernel density estimation. – Master's thesis, Dept of Electrical and Computer Engineering, Univ. of Manitoba, Winnipeg, Canada.
Ruse, M. G., Hasselquist, D., Hansson, B., Tarka, M. and Sandsten, M. 2016. Automated analysis of song structure in complex birdsongs. – Anim. Behav. 112: 39–51.
Sahidullah, M. and Saha, G. 2012. Comparison of speech activity detection techniques for speaker recognition. – CoRR abs/1210.0297.
Sandoval, L. and Barrantes, G. 2012. Characteristics of male spot-bellied bobwhite (*Colinus leucopogon*) song during territory establishment. – J. Ornithol. 153: 547–554.
Sauer, J. R., Peterjohn, B. G. and Link, W. A. 1994. Observer differences in the north american breeding bird survey. – Auk 111: 50–62.
Schafer, R. M. 1977. The tuning of the world. – Alfred A. Knopf.
Schrama, T., Poot, M., Robb, M. and Slabbekoorn, H. 2007. Automated monitoring of avian flight calls during nocturnal migration. – Proc. International Expert meeting on IT-based detection of bioacoustical patterns, Computational bioacoustics for assessing biodiversity, pp. 131–134.
Sebastián-González, E., Pang-Ching, J., Barbosa, J. M. and Hart, P. 2015. Bioacoustics for species management: two case studies with a hawaiian forest bird. – Ecol. Evol. 5: 4696–4705.
Selin, A., Turunen, J. and Tanttu, J. T. 2007. Wavelets in recognition of bird sounds. – EURASIP J. Appl. Signal Process. 2007: 141–141.
Shonfield, J. and Bayne, E. 2017. Autonomous recording units in avian ecological research: current use and future applications. – Avian Conserv. Ecol. 12: 14.
Simons, T. R., Alldredge, M. W., Pollock, K. H. and Wettroth, J. M. 2007. Experimental analysis of the auditory detection process on avian point counts. – Auk 124: 986–999.
Singh, N., Khan, R. and Shree, R. 2012. MFCC and prosodic feature extraction techniques: a comparative study. – Int. J. Comput. Appl. 54: 9–13.
Skowronski, M. D. and Harris, J. G. 2006. Acoustic detection and classification of Microchiroptera using machine learning: lessons learned from automatic speech recognition. – J. Acoust. Soc. Am. 119: 1817–1833.
Skowronski, M. D. and Fenton, M. B. 2008. Model-based automated detection of echolocation calls using the link detector. – J. Acoust. Soc. Am. 124: 328–336.
Somervuo, P. and Härmä, A. 2003. Analyzing bird song syllables on the self-organizing map. – Proc. Workshop on Self-Organizing Maps.
Somervuo, P., Härmä, A. and Fagerlund, S. 2006. Parametric representations of bird sounds for automatic species recognition. – IEEE Trans. Audio Speech Language Process. 14: 2252–2263.
Specht, R. 1993. AVISOFT – sound analysis and synthesis laboratory Pro: a PC program for sonagraphic analysis. – AVISOFT, Berlin, Germany.
Sprengel, E., Jaggi, M., Kilcher, Y. and Hofmann, T. 2016. Audio based bird species identification using deep learning techniques. – CLEF (working notes), pp. 547–559.
Stattner, E., Vidot, N., Hunel, P. and Collard, M. 2012. Wireless sensor network for habitat monitoring: a counting heuristic. – Proc. IEEE 37th Conf. Local Computer Networks Workshops, pp. 753–760.
Stattner, E., Segretier, W., Collard, M., Hunel, P. and Vidot, N. 2013. Song-based classification techniques for endangered bird conservation. – ICML 2013 Workshop on Machine Learning for Bioacoustics, Atlanta, GA, USA.
Stowell, D. and Plumbley, M. D. 2011. Birdsong and C4DM: a survey of UK birdsong and machine recognition for music researchers. – Technical report, Centre for Digital Music, Queen Mary, Univ. of London.
Stowell, D. and Plumbley, M. D. 2014. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. – PeerJ 2: e488.
Sutherland, W. J., Newton, I. and Green, R. 2004. Bird ecology and conservation: a handbook of techniques. – Oxford Univ. Press.
Swiston, K. A. and Mennill, D. J. 2009. Comparison of manual and automated methods for identifying target sounds in audio recordings of pileated, pale-billed, and putative ivory-billed woodpeckers. – J. Field Ornithol. 80: 42–50.
Tachibana, R. O., Oosugi, N. and Okanoya, K. 2014. Semi-automatic classification of birdsong elements using a linear support vector machine. – PLoS One 9: e92584.
Tan, L. N., Kaewtip, K., Cody, M. L., Taylor, C. E. and Alwan, A. 2012. Evaluation of a sparse representation-based classifier for bird phrase classification under limited data conditions. – Proc. Interspeech – 13th Annual Conference of the International Speech Communication Association, pp. 2522–2525.
Taylor, A. 1995. Recognising biological sounds using machine learning. – AI-Conference, CiteSeer, pp. 592–592.
Taylor, S. L. and Pollard, K. S. 2008. Evaluation of two methods to estimate and monitor bird populations. – PLoS One 3: e3047.
Tchernichovski, O. 2012. Sound Analysis Pro 2011 user manual. – Sound Analysis Pro.
Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B. and Mitra, P. P. 2000. A procedure for an automated measurement of song similarity. – Anim. Behav. 59: 1167–1176.
Towsey, M., Planitz, B., Nantes, A., Wimmer, J. and Roe, P. 2012. A toolbox for animal call recognition. – Bioacoustics 21: 107–125.
Towsey, M., Wimmer, J., Williamson, I. and Roe, P. 2014. The use of acoustic indices to determine avian species richness in audio-recordings of the environment. – Ecol. Inform. 21: 110–119.
Trifa, V. 2006. A framework for bird songs detection, recognition and localization using acoustic sensor networks. – Master's thesis, École Polytechnique Fédérale de Lausanne.
Turunen, J., Selin, A., Tanttu, J. T. and Lipping, T. 2006. De-noising aspects in the context of feature extraction in automated bird sound recognition. – In: Gogala, M. and Trilar, T. (eds), Advances in bioacoustics 2, dissertationes classis iv: historia naturalis. Slovenian Academy of Sciences and Arts (Ljubljana), XLVII-3.
Tyagi, H., Hegde, R. M., Murthy, H. A. and Prabhakar, A. 2006. Automatic identification of bird calls using spectral ensemble average voice prints. – Proc. 14th European Signal Processing Conference, pp. 1–5.
Ulloa, J. S., Gasc, A., Gaucher, P., Aubin, T., Réjou-Méchain, M. and Sueur, J. 2016. Screening large audio datasets to determine

the time and space distribution of screaming piha birds in a tropical forest. – Ecol. Inform. 31: 91–99.
Urazghildiiev, I. R. and Clark, C. W. 2007. Detection performances of experienced human operators compared to a likelihood ratio based detector. – J. Acoust. Soc. Am. 122: 200–204.
Vaseghi, S. V. 2008. Noise and distortion. Advanced digital signal processing and noise reduction. – John Wiley and Sons.
Ventura, T. M., de Oliveira, A. G., Ganchev, T. D., de Figueiredo, J. M., Jahn, O., Marques, M. I. and Schuchmann, K.-L. 2015. Audio parameterization with robust frame selection for improved bird identification. – Expert Syst. Appl. 42: 8463–8471.
Vernaleo, B. A. and Dooling, R. J. 2011. Relative salience of envelope and fine structure cues in zebra finch song. – J. Acoust. Soc. Am. 129: 3373–3383.
Vielliard, J. M. 2000. Bird community as an indicator of biodiversity: results from quantitative surveys in Brazil. – An. Acad. Bras. Ciênc. 72: 323–330.
Vilches, E., Escobar, I. A., Vallejo, E. E. and Taylor, C. E. 2006. Data mining applied to acoustic bird species recognition. – Proc. 18th IEEE Int. Conf. Pattern Recogn. 3: 400–403.
Vilches, E., Escobar, I. A., Vallejo, E. E. and Taylor, C. E. 2007. Targeting input data for acoustic bird species recognition using data mining and HMMs. – Proc. 7th IEEE Int. Conf. Data Mining Workshops, pp. 513–518.
Vintsyuk, T. 1971. Element-wise recognition of continuous speech composed of words from a specified dictionary. – Cybern. Syst. Analysis 7: 361–372.
wa Maina, C. 2015. Audio diarization for biodiversity monitoring. – Proc. 12th IEEE Africon International Conference – Green Innovation for African Renaissance, pp. 1–5.
Wielgat, R., Potempa, T., Świętojański, P. and Król, D. 2012. On using prefiltration in HMM-based bird species recognition. – Proc. IEEE Int. Conf. Signals Electron. Syst., pp. 1–5.
Wildlife Acoustics 2011. Song Scope bioacoustics software version 4.0 documentation. – Wildlife Acoustics.
Williams, H. 2004. Birdsong and singing behavior. – Ann. N. Y. Acad. Sci. 1016: 1–30.
Wimmer, J., Towsey, M., Roe, P. and Williamson, I. 2013. Sampling environmental acoustic recordings to determine bird species richness. – Ecol. Appl. 23: 1419–1428.
Wolf, K. 2009. Bird song recognition through spectrogram processing and labelling. – Univ. of Minnesota.
Wu, Z. and Cao, Z. 2005. Improved MFCC-based feature for robust speaker dentification. – Tsinghua Sci. Technol. 10: 158–161.
Zbancioc, M. and Costin, M. 2003. Using neural networks and LPCC to improve speech recognition. – Proc. IEEE Int. Symp. Signals Circuits Syst. 2: 445–448.
Zhang, X. and Li, Y. 2015. Adaptive energy detection for bird sound detection in complex environments. – Neurocomputing 155: 108–116.