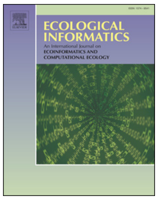
Ecological Informatics 34 (2016) 108–116

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Classifying and ranking audio clips to support bird species richness surveys

Liang Zhang *, Michael Towsey, Jinglan Zhang, Paul Roe

Eco-acoustics group, Science and Engineering Faculty, Queensland University of Technology, Australia

ARTICLE INFO

Article history:
Received 24 February 2016
Received in revised form 25 May 2016
Accepted 25 May 2016
Available online 27 May 2016

Keywords:
Acoustic sampling
Classification
Acoustic monitoring
Ranking
Bird species richness

ABSTRACT

Advances in programmable field acoustic sensors provide immense data for bird species study. Manually searching for bird species present in these acoustic data is time-consuming. Although automated techniques have been used for species recognition in many studies, currently these techniques are prone to error due to the complexity of natural acoustics.

In this paper we propose a smart sampling approach to help identify the maximum number of bird species while listening to the minimum amount of acoustic data. This approach samples audio clips in a manner that can direct bird species surveys more efficiently. First, a classifier is built to remove audio clips that are unlikely to contain birds; second, the remaining audio clips are ranked by a proxy for the number of species. This technique enables a more efficient determination of species richness.

The experimental results show that the use of a classifier enables to remove redundant acoustic data and make our approach resilient to various weather conditions. By ranking audio clips classified as "Birds", our method outperforms the currently best published strategy for finding bird species after 30 one-minute audio clip samples. Particularly after 60 samples, our method achieves 10 percentage points more species. Despite our focus on bird species, the proposed sampling approach is applicable to the search of other vocal species.

1. Introduction

Bird species are good indicators of environmental health and have been used to monitor the dynamic change of the natural environment (Carignan and Villard 2002; Catchpole and Slater 2003). The use of acoustics to monitor birds confers several advantages (Bardeli et al. 2010). First, it allows for covering a large spatial area; second, it provides continuous recordings for a long period of time; third, it functions well even with poor lighting or visual impediment; and fourth, audio signals are cheaper to store and compute than visual signals. With acoustic recordings, the cost of in-the-field observation is translated into an analysis problem. This paper focuses on the study of bird species richness (Spellerberg and Fedor 2003), which aims to determine the number of unique bird species in a specific area within a specific period of time using post-analysis of acoustic recordings.

1.1. Bird species richness survey

Due to spatiotemporal limitations, conducting an in-the-field bird species richness survey requires effective sampling protocols (Schneider 1994). Point count is one of the most popular sampling protocols where skilled bird observers document bird species they encounter in a specific site at fixed period of time (Huff et al. 2000). Normally, these recorded bird species can be subjective and hard to verify.

Acoustic sensor offers an effective approach to collect data at large spatiotemporal scales (Acevedo and Villanueva-Rivera 2006). The recorded acoustic data can be stored permanently and provide a convenient way to verify bird species. However, the increased dataset also necessitates the development of efficient techniques.

The difficulty of conducting the bird species richness survey by acoustics lies in the diversity of bird vocalizations. Competition for the acoustic space and environmental constraints, such as temperature and vegetation structure, may lead to significant variations within and between species vocalizations (Farina 2014). Additionally, simultaneous vocalizations could make the acoustic recognition of bird species even more difficult.

Manually listening to audios and inspecting the corresponding spectrograms for bird species identification is reliable, if experienced persons are involved, but time-consuming. For example, a one-minute audio clip often requires twice the time to investigate because people frequently replay the audio to identify which species are vocalizing (Wimmer et al. 2013). Although automated techniques offer computational power to alleviate this problem, their development is still in infancy. Unlike human speech and music, bird vocalizations are less structured and their repetition is unpredictable. These problems hinder the use of automated techniques for bird species identification.

Several studies have contributed to developing automated techniques for bird species recognition (Fagerlund 2007; Kasten et al.

[*] Corresponding author at: Room 1002, Level 10, S Block, Gardens Point Campus, 2 George Street, Brisbane, QLD 4000, Australia.
E-mail address: l68.zhang@hdr.qut.edu.au (L. Zhang).

http://dx.doi.org/10.1016/j.ecoinf.2016.05.005
1574-9541/© 2016 Elsevier B.V. All rights reserved.

2010; Somervuo et al. 2006). These methods work well when vocalization structures are simple. For example, only a single species that has multiple vocal types is calling or several different species sharing a common type of vocal structure. Automated call recognition is a promising alternative for acoustic data analysis, but accuracy is still far from perfect, especially on detecting vocal species recorded from the natural environment. Recently, a multi-label classification method has been introduced to detect bird vocalizations (Briggs et al. 2012). Unlike prior work on single-label classification where there is only one label associated with a study object, multi-label classification is possible to associate the study object with multiple labels, providing a potential solution to recognize simultaneous bird vocalizations. However, this method is a supervised machine learning that can only predict predefined labels in the training data; consequently, it cannot handle any unexpected vocalizations that may appear.

Confronted with a large volume of data, Wimmer et al. (2013) first introduced sampling methods to assist bird species richness survey. They compared five temporal sampling strategies on a one-day recording, pointing out that the most efficient strategy to find bird species is dawn sampling. Here dawn sampling is referred to randomly selecting audio clips 3 h after dawn. However, this is an intuitive approach based on the fact that many bird species vocalize during dawn. There are no further instructions on how to effectively investigate these 3-hour acoustic data. They also suggested that using automated techniques to locate periods that are likely to contain unique species might improve the efficiency of bird species surveys. The use of automated techniques to direct the sampling of acoustic data for bird species surveys is called "*smart sampling*". Prior works have shown the ability of using the linear regression (Towsey et al. 2013) or the clustering technique (Eichinski et al. 2015) for smart sampling, but they did not take into consideration of various weather conditions such as heavy rain and strong wind gusts that can affect the efficiency of these methods.

1.2. Acoustic characteristics of audio clips

Direct use of automated techniques may not be effective in the case of non-targeted and multi-species inventories; nevertheless, we can still utilize these techniques to make bird species recognition more efficient than manual analysis. Fig. 1 shows five commonly encountered examples of one-minute audio clips. We categorize them as "Birds", "Insects", "Low activity", "Rain", and "Wind". In temperate woodland ecosystems in spring, a prior study describes the daily acoustic activity to be birds vocalizing during the day and insects chirping from sunset to the sunrise of the next day (Wimmer et al. 2013). Occasional heavy rain and strong wind gusts are important acoustic information because they may interrupt bioacoustics activities. We also define low activity as the time when little amount of acoustic energy is recorded. Apparently, removing data that do not contain bird species can improve the efficiency of species finding. Since these five acoustic patterns have discriminative time-frequency characteristics, it is possible to filter the other four patterns from "Birds" using a classification method.

1.3. Indicators of acoustic diversity

Biodiversity assessment is one of the most challenging problems that ecologists are confronted with. Indices have been used to

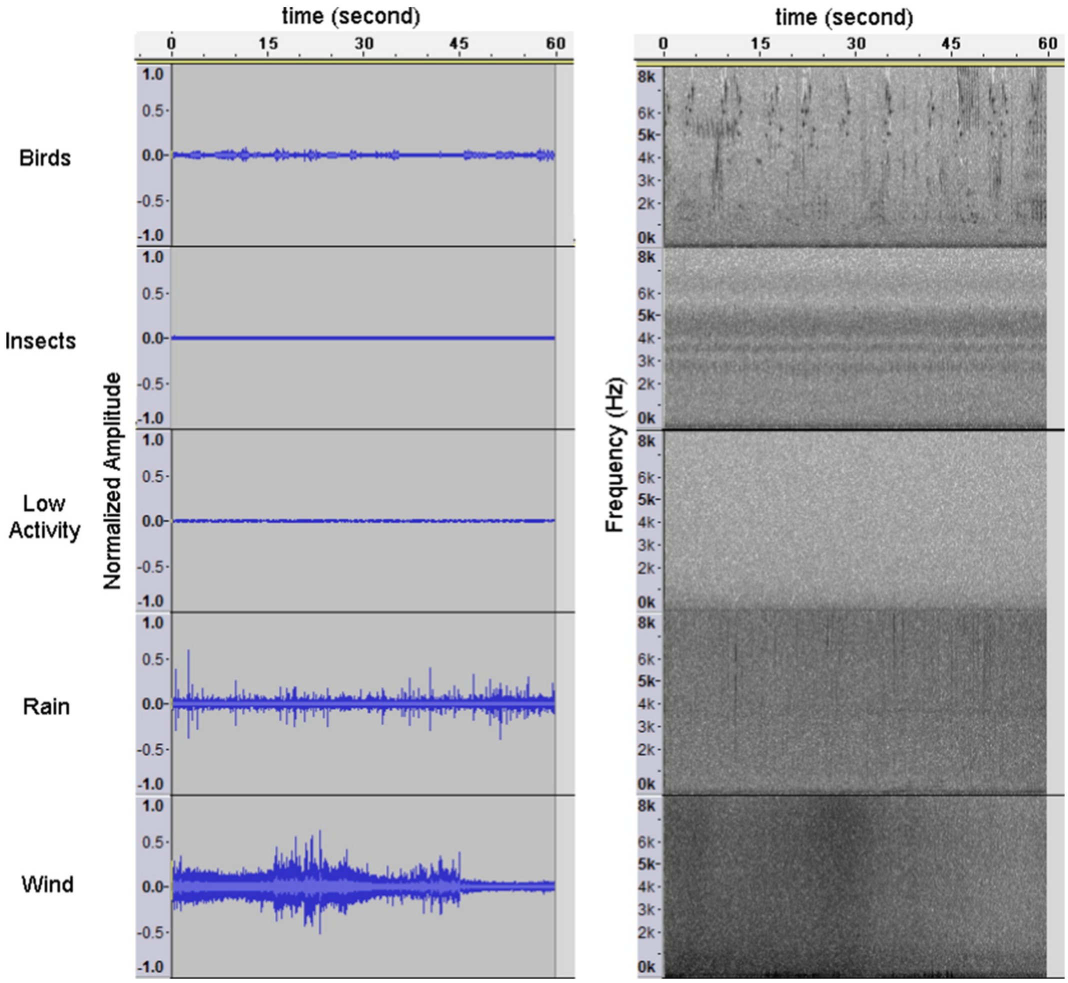


Fig. 1. Characteristics of five acoustic patterns in one-minute audio clips (left) and their corresponding spectrograms (right).

characterize different facets of animal and plant communities (Magurran and McGill 2011; Pavoine and Bonsall 2011). Traditional indices are derived from species inventory where lists of species are recorded on-site. Recently, acoustic indices have been introduced to estimate acoustic diversity in the natural environment (Sueur et al. 2014). Acoustic indices are summarized values of acoustic energy in an audio clip. They can be exemplified by acoustic entropy index (Sueur et al. 2008) and acoustic complexity index (Pieretti et al. 2011). The former is calculated based on the Shannon index, which aims at characterizing acoustic energy dispersion; whereas the latter is highlighted to capture rapid change of acoustic frequency information.

In this paper, we propose a new smart sampling approach to direct bird species richness surveys. Our approach utilizes the inherent time-frequency characteristics of various acoustic patterns to classify audio clips, removing 58% of one-day acoustic data that are likely to be redundant for bird species richness surveys. Using the presence/absence data of bird annotations, we show that the proposed ranking method can assist in finding 10 percentage points more bird species than the dawn sampling when given more than 30 one-minute audio clip samples. The primary contributions of this work are as follows.

- We build a supervised machine learning model to classify five common acoustic patterns including “Birds”, “Insects”, “Low activity”, “Rain”, and “Wind”;
- We propose to sample audio clips with spectral information as a more efficient way for directing bird species richness surveys.

The remainder of this paper is divided as follows. We propose a classification and ranking method for efficient determination of bird species in Section 2. Section 3 demonstrates the experimental results and compares our method with some benchmarks. We discuss new acoustic features and some limitations of our method in Section 4. Section 5 concludes and gives a brief description of our future work.

2. Method

In this section, we describe our datasets and acoustic features used to characterize audio recordings. To improve the efficiency of finding bird species, we propose a two-step procedure for assistive determination of bird species in a one-day recording. First, a classifier is built to filter the redundant non-bird one-minute audio clips; a spectrum index is then used to rank the one-minute audio clips classified as “Birds” in step one.

2.1. Datasets

The raw acoustic data are collected from the Samford Ecological Research Facility, Brisbane, Australia (27.39°S, 152.88°E). The recording sites are mainly covered by inland open-forest and woodland comprised of *Eucalyptus tereticornis*, *Eucalyptus crebra* and *Melaleuca quinquenervia* in moist drainage. The Samford Creek flows to the west of the study area. There are small areas of gallery rainforest and areas of open pasture along the southern boundary. A frog pond can be found in the southern open pasture.

Table 1
Basic information of collected acoustic data.

Data	Site	Dates	Formats
Training	Northeast	13th and 14th October 2010	MP3
	Northeast	16th and 17th October 2010	MP3
	Northeast	13th April 2013	WAV
	Southeast	16th October 2010	MP3
Test	Northeast	15th October 2010	MP3

All recordings are recorded at a 22,050 Hz sampling rate of a two-channel signal, 16 bits (Table 1). They are later downmixed to a 17,640 Hz mono signal and cut into one-minute audio clips for computational convenience. The training data are selected from 2 sites over 6 days by listening to the recording and visually inspecting the corresponding spectrograms. A total of 150 one-minute audio clips are selected across five acoustic patterns, each of which contains 30 one-minute samples. Our test data are collected on 15th October 2010 when there is intermittent rain and wind. There are 62 bird species annotated as presence or absence at a one-minute resolution by three experienced bird observers for the test data. More details about the dataset can be found in this paper (Wimmer et al. 2013).

Note that the pre-defined five acoustic patterns are not always exclusive. For example, some birds may vocalize in the rainy and windy days; and there might be one or two remote and faint bird vocalizations being recorded during “Low activity”. To ameliorate this problem, we conservatively labeled the audio clips that possibly contain bird species as “Birds”. This bias is reasonable because we aim to remove irrelevant acoustic data and retain as many bird species as possible.

2.2. Acoustic features

As with most pattern recognition problems, selecting proper features is crucial for successful classification. Here our study objects are one-minute audio clips. For each audio clip, acoustic features can be derived from its waveform envelope or spectrogram amplitude. In this paper, a waveform envelope is a smoothed waveform by using a 512-point rectangular window with 50% overlap. A spectrogram is calculated by applying a Fourier transform to small segments of an audio clip. To obtain a comparable temporal resolution as the waveform envelope, the small audio segments are cut by a 512-point hamming window. Depending on whether acoustic features are derived from waveform envelopes or spectrograms, they can be categorized into temporal features and spectral features respectively.

2.2.1. Temporal features

Given a time series $x(n)$, $1 \le n \le N$, temporal features are calculated as follows:

1. *AveSignalAmplitude* (Towsey et al. 2013): It is the average amplitude of the waveform envelope; values are converted to decibel:

$$aveSignalAmplitude = 10log_{10}\left(\sum^{N} x(n)/N\right)^2$$

2. *Signal-to-noise ratio*: It is derived from dividing waveform envelope energy by background noise energy, values are in decibel. Background noise (BN) is estimated as the mode of the waveform envelope (Towsey 2013).

$$signal\text{-}to\text{-}noiseratio = 10log_{10}\left(\sum^{N} |x(n)|/BN\right)^2$$

3. Temporal Entropy (H[t]) (Sueur et al. 2008): It is the entropy of waveform envelope.

4–10. *Matching pursuit* (*MP*) *features*: matching pursuit algorithm enables to decompose local waveform structures into a sparse time-frequency representation using an over-complete dictionary of basis functions (Chu et al. 2009). The decomposition process can be depicted as following steps:

a. Generate an over-complete set of basis functions using a dictionary of Gabor function;
b. Compute the correlations between the signal and all basis functions by using inner product;

c. Find the highest correlation and subtract the weighted basis function from the signal. The weight is the inner product of this basis function and the signal;
d. Go to step b and c until a certain number of iterations or a predefined signal-to-residual energy ratio has been reached.

We used seven matching pursuit features in this study, including signal-to-residual ratio (*MP_SRR*), the means and standard deviations of chirp, frequency, and position. Among them, frequency and position (also called as 'scale') are used in this paper (Chu et al. 2009). The rest features are useful for this specific application. For example, chirp is referred to the frequency change rate that can be found in bird vocalizations and the signal-to-residual ratio is used to estimate the complexity of an audio clip. An efficient implementation of matching pursuit algorithm – Matching Pursuit Toolkit (Krstulovic and Gribonval 2006) was used. A user-specified iteration number is set to determine how many basis functions are used to approximate the original signal. Here, we ran a grid search on optimizing the classification accuracy of using MP features at different iterations of 5, 10, 100, 500, 1000 and other features proposed in this study. It was found that 500 iterations provided a good compromise between MP features used alone and with other features.

2.2.2. Spectral features

After applying a short-time Fourier transform to the original audio signal, we obtain a spectrogram. It is a matrix of *N* time frames by *M* frequency bins. Here **s** denotes a short-time Fourier transformed spectrum, and a denotes each amplitude value.

$$\mathbf{s}_\mathrm{i} = (\mathrm{a}_1,\ \mathrm{a}_2, \cdots, \mathrm{a}_\mathrm{M})^{\mathrm{transpose}} \quad \mathrm{i} = 1, 2, \cdots, \mathrm{N}$$

Spectral features are calculated from the spectrogram of a one-minute audio clip.

11. *Acoustic Complexity Index* (*ACI*) (Pieretti et al. 2011): It is the average absolute amplitude differences between adjacent time frames.

12–14. *FrequencyCover* (Towsey et al. 2014): It refers to the count of amplitude values divided by the number of time frames and the number of frequency bins. A threshold is used to set low values to zeros and high values to ones prior to the calculation. *Frequency cover* can be divided into low (a_j, $1 \le j \le 7$), mid (a_j, $7 < j \le 51$), and high (a_j, $51 < j \le 256$) -frequency cover respectively. Here *j* is an integer implying a frequency bin and K in the following equation is the number of frequency bins, depending on the frequency range it calculates.

$$\mathbf{FrequencyCoverSpectrum} = \sum\nolimits^{\mathrm{N}} \mathbf{s}_\mathrm{i}/\mathrm{N}$$

$$\mathrm{FrequencyCover} = \sum\nolimits^{\mathrm{K}} \mathbf{FrequencyCoverSpectrum}/\mathrm{K}$$

15. *Entropy of the average spectrum* (H[*s*]) (Towsey et al. 2013): It is entropy of average amplitude values in frequency bins (a_j, $7 < j \le 256$). The vector of average amplitude values is:

$$\mathbf{v} = \sum\nolimits^{\mathrm{N}} \mathbf{s}_\mathrm{i}/\mathrm{N}$$

So the entropy of the average spectrum is calculated as:

$$\mathrm{H[s]} = -\sum\nolimits^{\mathrm{K}} \mathbf{v} \log_2 \mathbf{v} / \log_2 \mathrm{K}$$

16. *Entropy of spectral maxima* (H[*m*]) (Towsey et al. 2013): It is the entropy of amplitude values that has maximum counts in frequency bins (a_j, $7 < j \le 256$). Here, **c** is referred to a vector of frequency bins having maximum values.

$$\mathrm{H[m]} = -\sum\nolimits^{\mathrm{K}} \mathbf{c} \log_2 \mathbf{c} / \log_2 \mathrm{K}$$

17. *Mel-frequency cepstral coefficients* (*MFCCs*) (Molau et al. 2001): The spectra of a spectrogram are mapped to mel-frequency banks and the energy is summed in each bank. The mel scale relates physical frequency to perceived frequency by humans. The equation for converting frequency to mel-frequency is:

$$\textit{Mel-frequency} = 1125 \ln(1 + \textit{frequency}/700)$$

Typically 12 mel-frequency cepstral coefficients are obtained by using discrete cosine transform on the logarithmic mel-frequency. In *MFCCs*, low frequency has a higher resolution than high frequency. Since traditional *MFCCs* have detailed frequency information, to compare it with other acoustic features which have average frequency

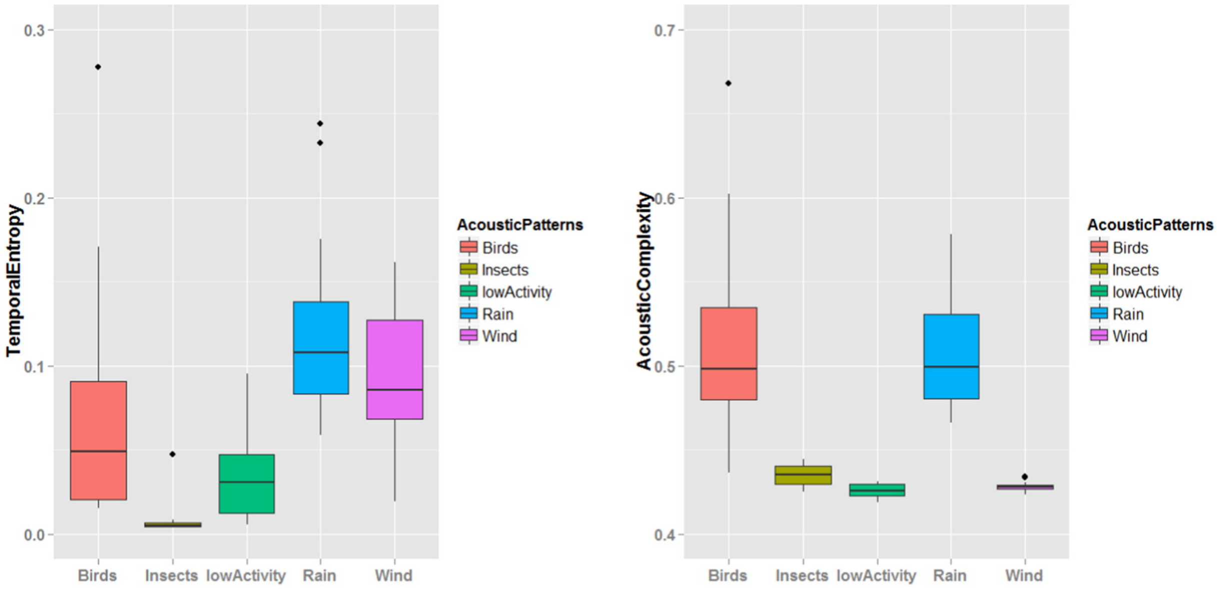


Fig. 2. Temporal entropy H[t] and acoustic complexity index (*ACI*) of 150 one-minute training audio clips. In practical use, the values of acoustic indices will be rescaled between 0 and 1 inclusive.

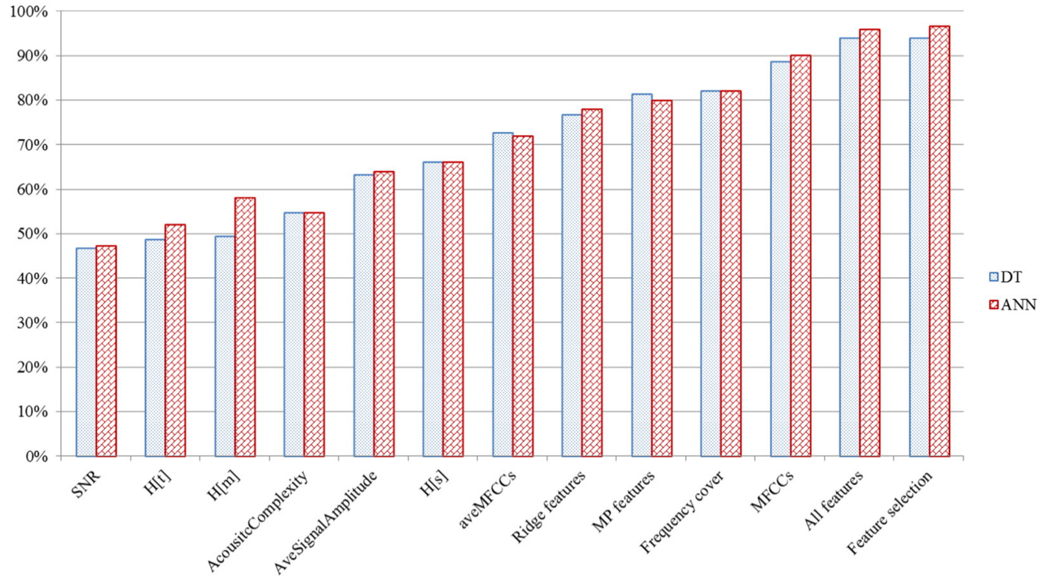


Fig. 3. Classification accuracy using different feature sets, decision tree (DT) and multilayer perceptron (ANN) are used as classifiers. Forward feature selection includes 7 features: *AveSignalAmplitude*, *AcousticComplexity* (*ACI*), H[*s*], H[*m*], *verRidge*, *horRidge*, and *MP_SRR*.

information, we also average 12 mel-frequency cepstral coefficients to obtain *aveMFCCs*.

18–19. *Ridge features* (*verRidge* and *horRidge*): If a spectrogram is considered as an image comprised of pixels, ridges are local maxima in different directions of a spectrogram. These features were first introduced for bird vocalization retrieval (Dong et al. 2013). We use the same approach to calculate the ridges, and average the count of vertical and horizontal ridges of a spectrogram as two separate ridge features.

2.3. Classification

Decision tree and multilayer perceptron (Duda et al. 2012) are used to investigate classification performance. The classifiers are trained in a 10-fold stratified cross validation method implemented in Weka 3.7.11 (Hall et al. 2009), where the number of instances for each acoustic pattern is assigned proportionally to each fold. Decision tree is implemented by C4.5 algorithm and the number of nodes in the hidden layer of multilayer perceptron is set to 6. We also utilize a forward stepwise method to select features by optimizing classification accuracy (Hall 1999).

We use precision and recall to measure the performance of the classification results. Precision measures "how useful the classified results are", and recall measures "how complete the classified results are". In other words, high precision means that a classifier returns substantially more correct results than incorrect, while high recall means that a classifier returns most of the correct results. Their equations are depicted as follows:

$$\text{precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

In this study, *true positives* is the number of instances that are correctly classified, *false positives* is the number of instances that are incorrectly classified, and *false negatives* is the number of instances that are misclassified as one of the other four acoustic patterns.

2.4. Ranking

Some one-minute audio clips naturally contain more bird species than others. Consider, for example, the dawn and dusk choruses. Listening to audio clips which contain more birds has high probability to find new bird species. Although the exact number of bird species in each audio clip is difficult to identify simply by their vocalizations, an audio clip that contains more species normally has high acoustic activity. Since acoustic features should reflect acoustic activity, we can use these features as a proxy to assign priority to each audio clip.

In this paper, we use correlation coefficients (Edwards 1976) between acoustic feature and the number of bird species to select a feature which best indicates the acoustic activity of audio clips. Spearman's, instead of Pearson's, correlation coefficient is used. The former measures the monotonic relationships without any assumption about statistical distribution of either variable; whereas the latter measures the linear relationships of two variables that have normal distributions. Since correlations between acoustic features and the number of bird species do not have to be linear in this case, Spearman's correlation coefficient is more appropriate than Pearson's correlation. Then the best correlated acoustic feature is utilized to rank audio clips, giving each of them a priority to listen to.

It is also reported that spectral information might be important for bird species to avoid inter-specific acoustic competition (Farina 2014); therefore, the complexity of spectral information could be a good proxy of the number of bird species. Since all aforementioned acoustic features are summarized acoustic features; that is, their temporal and spectral information has been averaged, we need to recalculate spectral features by averaging only the time frames of a spectrogram other than frequency bins. In this study, we will select one acoustic index that has the maximum correlation coefficient with the number of bird species to calculate its spectral feature.

Table 2
Classification accuracy of the test dataset.

	Birds	Insects	Low activity	Rain	Wind
Precision (%)	96.7	49.9	91.9	90.9	56.1
Recall (%)	88.4	87.1	78.0	70.3	65.3

Table 3
Confusion matrix of testing dataset using acoustic indices from feature selection.

Classified as →	Birds	Insects	Low activity	Rain	Wind	Actual total
Birds	**585**	67	3	2	5	662
Insects	7	**169**	9	9	0	194
Low activity	4	53	**248**	2	11	318
Rain	6	46	**2**	**149**	9	212
Wind	3	4	8	2	**32**	49
Classified total	605	339	270	164	57	1435

Values in bold are referred to the number of correctly classified one-minute audio clips for different classes.

The matching pursuit features are calculated by the MPTK 0.7.0 package and *MFCCs* are calculated by the signal processing toolbox in MATLAB 2014b. The software used to generate the rest of the indices was a proprietary C# application developed by the QUT Ecoacoustics Research Group (Truskinger et al. 2014). The program, named AnalysisPrograms.exe, was compiled on August 15th 2014 with the version number 14.08.0.0. Statistical analysis is conducted with R 3.0.2 and RStudio 0.98.994.

3. Results

3.1. Analysis of acoustic indices

Different indices emphasize different aspects of acoustic information. Fig. 2 illustrates the values of two indices calculated from 150 one-minute audio clips, grouped by five acoustic patterns. Take *temporal entropy* H[t] for example. It can separate “Insects” from other four acoustic patterns ($p < 0.001$). This is mainly because “Insects” have flat waveforms while others have rapid waveform changes. *ACI* captures rapid changes of spectral energy, but it fails to differentiate narrow band acoustic energy (bird vocalizations) from wide band energy (rain) ($p > 0.1$). This also confirms that *ACI* is preferably used in high signal-to-noise ratio situations (Pieretti et al. 2011). Using a single index is hardly sufficient to describe the complexity of natural acoustics. It is recommended that combinations of several indices can complement each other and provide more efficient representations for audio clips (Towsey et al. 2013). In this experiment we use acoustic indices as features for classification and ranking, so the terms ‘acoustic indices’ and ‘acoustic features’ are used interchangeably.

3.2. Classification accuracy of training data

The overall classification accuracy of five acoustic patterns is compared in Fig. 3 using decision tree and multi-layer perceptron. As shown in the graph, acoustic features that average frequency information (*aveMFCCs* and those on the left) have lower classification accuracy than those considering frequency information (*Ridge features* and those on the right). As reported in other research (Chu et al. 2006), adding all features may not build the best classifier due to the redundant information of different features. This is why features selected by a forward stepwise method provide slightly better classification accuracy than simply using all features. In our experiment, forward feature selection provides classification accuracy of 94% and 96.7% for decision tree and multilayer perceptron respectively. The multi-layer perceptron slightly outperforms decision tree in most cases. Consequently, we use multilayer perceptron and the selected features for the rest of our experiment.

3.3. Confusion matrix of test data

To verify the reliability of the generated multilayer perceptron classifier, we applied it to a separate test dataset. The overall classification accuracy on test dataset is 82.4%. Table 2 shows the precision and recall of each acoustic pattern of the test dataset. Generally, “Birds” has the highest precision and recall; whereas “Insects” has the lowest precision and “Wind” has the lowest recall.

A confusion matrix (Table 3) is created for a better understanding of misclassification instances among different acoustic patterns. “Birds” is the most common class in the test data, which also has the highest classification accuracy. “Wind” is the least common class. Moreover, “Birds”, “Low activity” and “Rain” are often misclassified as “Insects”, but not vice versa. There are 605 of 1435 (42%) minutes classified as “Birds”, which means we successfully remove 58% of acoustic data that are likely to be redundant for bird species surveys. After classification, the remaining 605 one-minute audio clips will further be used for determining the number of bird species.

The classification model aims to filter minutes that are less likely to contain birds so that determination of bird species can be conducted more efficiently. Fig. 4 demonstrates how the classification model affects the distribution of minutes in terms of their content of bird species. The proposed classification method enables to discard a large volume of audio clips without birds and retain the majority of those with more than 5 species (including 5). By contrast, our method misclassified audio clips with 1 to 4 bird species possibly due to lack of acoustic energy originated from bird vocalizations. The impact of these misclassified audio clips on bird species richness surveys depends on the behavior of specific species. From the presence/absence bird annotations in our study, we find that 59 out of 62 (95.2%) bird species remain in 605 min classified as “Birds”; that is, all but three bird species are retained after filtering irrelevant acoustic data. It also guarantees that our ranking strategy has the potential to obtain a more efficient way to find bird species.

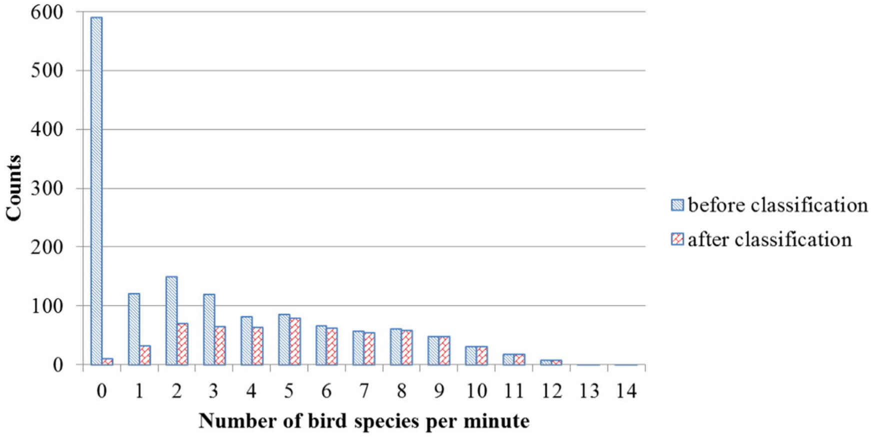


Fig. 4. Histograms of the number of bird species per minute (15th October 2010, southeast site). After classification, redundant minutes that contain no birds have been removed.

Table 4
Spearman's correlations ($p < 0.01$, $n = 605$) between acoustic features and the number of bird species per minute, and the percent of bird species found in first 60 one-minute samples ranked by the corresponding features.

	verRidge	*MP_SRR*	*ACI*	H[*s*]	*horRidge*
Correlation	0.14	0.16	0.36	0.47	0.62
Percent of bird species found in first 60 min (%)	58.1	48.4	64.5	66.1	71.0

3.4. Audio clip ranking by acoustic features

Listening to audio clips in a chronological order does not necessarily provide satisfactory efficiency of finding bird species. Indeed, clips before dawn contain only a few bird vocalizations. These minutes are correctly classified as “Birds” but normally have few of them. To resolve this issue, we propose a ranking method to re-order the sequence of audio clips so that those containing more acoustic activities will have higher priority to be listened to.

We calculate Spearman's correlation coefficient to determine which acoustic feature has the highest correlation with the number of bird species (Table 4, first row). The experimental results show that *horRidge* has the highest correlation with the number of bird species per minute. This particular acoustic feature is considered as a good indicator of acoustic activity of one-minute audio clips. To exemplify its utility, we rank one-minute audio clips with the corresponding acoustic features and estimate the percent of bird species found at the 60th minute (Table 4, second row). When one-minute audio clips are ranked by the *horRidge*, we can find the most bird species (71%) in the first 60 min compared to that ranked by other features.

To investigate whether spectral information can provide more information on assisting bird species surveys, we further calculate horizontal ridge spectrum because this feature has the highest correlation with the number of bird species in each audio clip.

Fig. 5 shows four examples of horizontal ridge spectra calculated from different one-minute audio clips with different number of bird species. We can observe that the more bird species in an audio clip, the larger the number of local maxima in the spectrum. Although the mapping between them is not perfect in a low signal-to-noise ratio case (bottom right), the number of local maxima in horizontal ridge spectrum seems to be a good proxy of the number of bird species in a one-minute audio clip.

We estimate the number of local maxima for all one-minute audio clips by calculating the second derivative of the smoothed ridge spectrum. Audio clips are later ranked by the number of local maxima in a descending order to determine its sampling sequence.

3.5. Species accumulation curve

Species accumulation curves are plotted to examine the efficiency in searching for birds by different methods. We compare the result ranked by local maxima of ridge spectra with three benchmark curves (Fig. 6). There are two benchmark curves in this graph. The top one denotes the theoretical best which maximizes the bird species found at each sampled audio clip using the bird annotations; the solid circle at the bottom is derived from the baseline method which is the mean of sampling 1435 one-minute audio clips 1000 times at random. Any useful species accumulation curve should reside between these two curves. The triangular curve is the dawn sampling that currently performs the best to our knowledge (Wimmer et al. 2013). This method assumes that there are more bird species in the morning chorus; therefore, it recommends to inspecting audio clips during 180 min after dawn. In this paper, we simulate the process by averaging 1000 times of dawn sampling at random.

By taking a pairwise t-test ($p < 0.001$), we find that smart sampling (shown as squares in Fig. 6) outperforms dawn sampling after 30 one-minute samples. Specifically in 60 min, our new method can find 82.2% of bird species, which is 10 percentage points higher than the mean of dawn sampling (72.6%).

4. Discussions

Our smart sampling approach is tested on a one-day recording of a temperate area. The robustness of our approach needs to be tested with data collected from other types of habitats. One might wonder if high or low bird species richness could affect the efficiency of our approach. This problem of different levels of species richness, we believe, might not be directly associated with the sampling efficiency. Instead it is the vocal activities of birds that could make a difference since our approach exploits the temporal and spectral characteristics of acoustics for bird species surveys. Provided that there are high densities of bird vocalizations in one-minute audio clips and different bird species are separated well, our approach can offer promising results regardless of different levels of bird species richness.

A dawn chorus of bird species mixed with other vocal taxonomies such as insects and amphibians could affect the dawn sampling strategy.

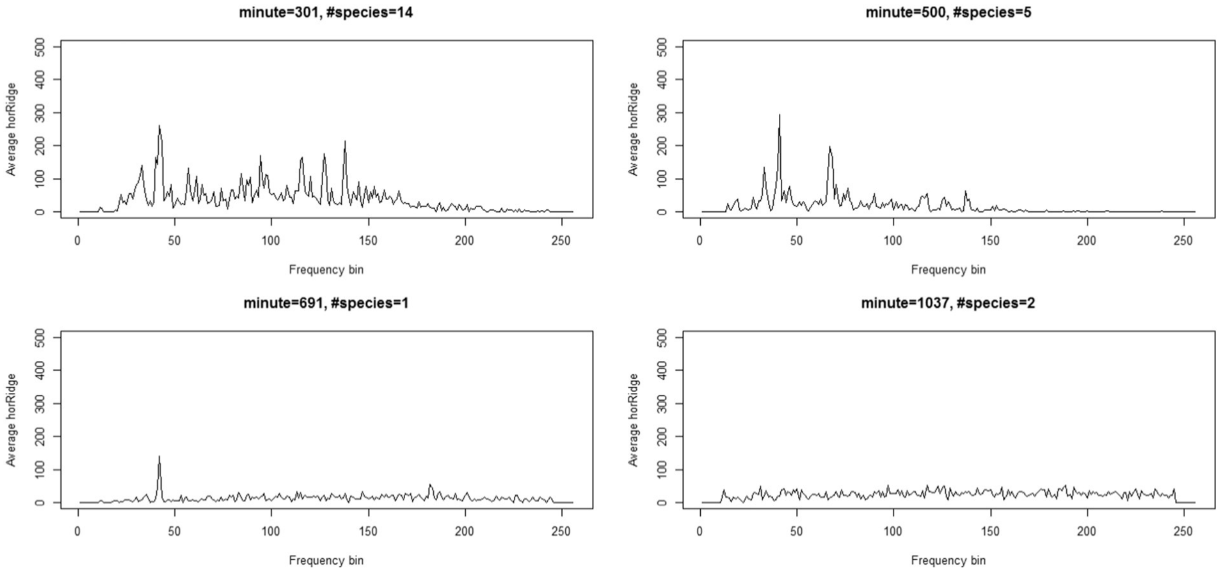


Fig. 5. Examples of ridge spectra calculated from four different one-minute audio clips with different number of species.

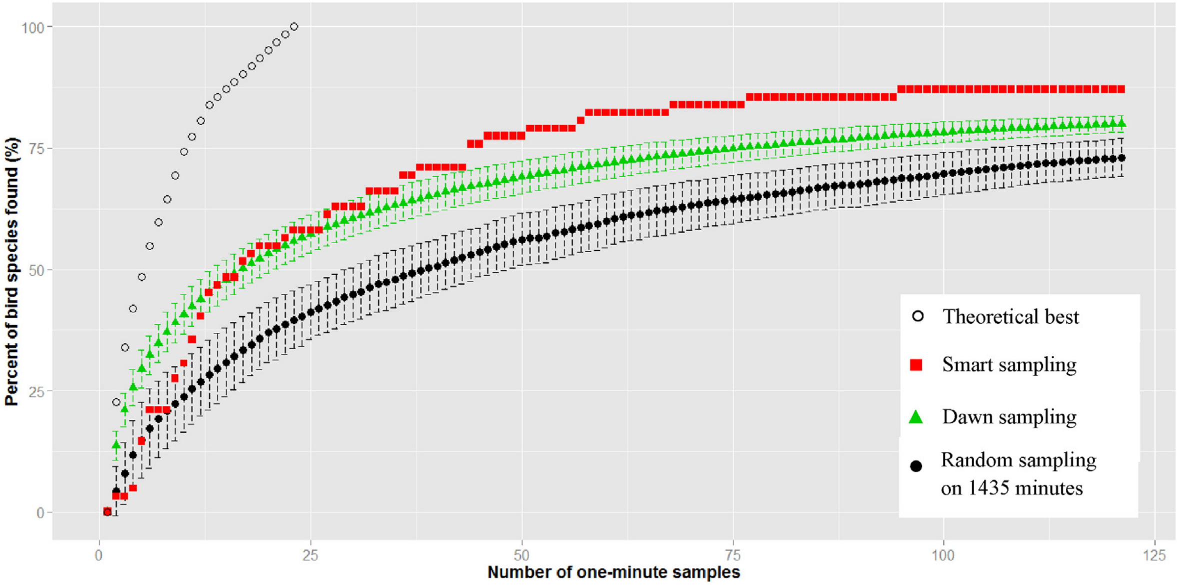


Fig. 6. The percent of bird species found using different sampling methods on 15th October 2010. The error bars indicate one standard deviation of the corresponding mean.

As for the proposed approach, we pre-define a class of insects in our classifier model, aiming to identify audio clips that contain strong insects chirping. One might also wish to filter amphibians from the recording. We should note that indices/features used in this study capture salient acoustic information rather than species information. Although vocalizations of amphibians have not been included in our training dataset, some amphibians such as frogs having dominant frequency of vocal structures could be classified as “Insects”; whereas others could be randomly classified into any the other four acoustic patterns depending on its acoustic characteristics. Such a bias is acceptable in the current study as long as our focus is on bird species. Consequently, the proposed approach is relatively more robust than dawn sampling when confronted with mixed vocalizations of various taxonomies. Further work is needed if it is necessary to build a classifier to distinguish amphibians.

This work is a step towards using automated techniques to assist bird species surveys. The exact time that ornithologists can spend on listening to recordings has received little study. A reference time for comparisons is 60 one-minute samples, which is comparable effort to traditional 120 minutes in-field surveys because audios take twice as much time to listen to (Wimmer et al. 2013). The proposed approach outperforms dawn sampling at 60 one-minute samples.

The proposed approach is beneficial to the study of soundscape ecology which deals with large numbers of audio recordings. It has the potential to detect rain and wind events over large spatiotemporal scales with little manual effort. Moreover, acoustic indices are developed to assess biodiversity and investigate landscape under the framework of soundscape ecology; whereas bird species richness surveys are closely related to bioacoustics which mainly focuses on studying individual species. Therefore, our work of using acoustic indices to assist bird species surveys signifies the link between soundscape ecology and bioacoustics.

There are some limitations in our approach from a technical perspective. The ranking of one-minute audio clips only considers the number of bird species in a descending order. Although audio clips with more bird species have high priority to be listened to, it is possible that they might share the same species. An ideal solution to this problem is to give low priority to audio clips that may contain the same bird species as those being audited. We investigate two potential approaches. The first one is to consider temporal redundancy. The assumption is that consecutive audio clips are more likely to share the same bird species, so once an audio clip is inspected, its temporally adjacent audio clips should be given low priority. The disadvantage of such a method is that the number of adjacent audio clips which should be given low priority to is completely subject to empirical decision. The other solution is to identify and remove ‘acoustically similar’ audio clips. Acoustic similarity is measured by calculating Euclidean distances or correlations between audio clips. A low priority is given to audio clips that are similar to audited ones. However, features used in this study cannot reflect detailed acoustic information because they are summarized from one-minute audio clips.

5. Conclusions and future work

Acoustics offers a wealth of ecologically meaningful information on top of what the visual cue can provide. The use of acoustic sensors increases the spatiotemporal scale of field observations, but the escalating data makes the manual analysis difficult. In practice, manually listening to a one-day recording for bird species richness survey is expensive in terms of time and effort required.

In this paper, we propose a computer-assisted approach to support bird species richness survey. The proposed approach consists of two stages: classification and ranking. Our classification results demonstrate that 58% of the total acoustic data can be removed while retaining the 95.2% of bird species in that day. At the ranking stage, the number of local maxima of ridge spectra is used to rank one-minute audio clips, providing a final sequence of audio clips for people to listen to. The final results show that the proposed approach achieves higher efficiency than the dawn sampling in determining bird species of a one-day recording. Moreover, this new approach is adaptive to various weather conditions such as rain and wind, which are the weakness of using dawn sampling.

Our future work will focus on developing more descriptive features to differentiate species compositions in audio clips. Segmenting audio clips into shorter durations provides one potential way to characterize more detailed bird vocalizations.

Acknowledgments

This research was conducted with the support of the QUT Samford Ecological Research Facilities. The authors would like to acknowledge the contribution of the experts in annotating the bird species and Anthony Truskinger in managing annotation datasets and generating acoustic features. The primary author was supported with a joint scholarship provided by QUT (201306150009) and China Scholarship Council.

References

Acevedo, M.A., 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.

Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K.H., Frommolt, K.H., 2010. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. Pattern Recogn. Lett. 31, 1524–1534.

Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X.Z., Raich, R., Hadley, S.J.K., Hadley, A.S., Betts, M.G., 2012. Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. J. Acoust. Soc. Am. 131, 4640–4650.

Carignan, V., Villard, M.-A., 2002. Selecting indicator species to monitor ecological integrity: a review. Environ. Monit. Assess. 78, 45–61.

Catchpole, C.K., Slater, P.J., 2003. Bird Song: Biological Themes and Variations. Cambridge University Press.

Chu, S., Narayanan, S., Kuo, C.C.J., Mataric, M.J., 2006. Where Am I? Scene Recognition for Mobile Robots Using Audio Features, IEEE International Conference on Multimedia and Expo, 885–888.

Chu, S., Narayanan, S., Kuo, C.C.J., 2009. Environmental sound recognition with time-frequency audio features. IEEE Trans. Audio Speech Lang. Process. 17, 1142–1158.

Dong, X., Towsey, M., Zhang, J., Banks, J., Roe, P., 2013. A Novel Representation of Bioacoustic Events for Content-Based Search in Field Audio Data. Digital Image Computing, Techniques and Applications (DICTA).

Duda, R.O., Hart, P.E., Stork, D.G., 2012. Pattern Recognition. John Wiley & Sons.

Edwards, A.L., 1976. An Introduction to Linear Regression and Correlation. W. H. Freeman, San Francisco, CA.

Eichinski, P., Sitbon, L., Roe, P., 2015. Clustering acoustic events in environmental recordings for species richness surveys. Procedia Comput. Sci. 51, 640–649.

Fagerlund, S., 2007. Bird species recognition using support vector machines. EURASIP J. Adv. Sign. Process. 1, 64.

Farina, A., 2014. Soundscape Ecology: Principles, Patterns. Springer, Methods and Applications.

Hall, M.A., 1999. Correlation-Based Feature Selection for Machine Learning. The University of Waikato.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. 11, 10–18.

Huff, M.H., Bettinger, K.A., Ferguson, H.L., Brown, M.J., Altman, B., 2000. A Habitat-Based Point-Count Protocol for Terrestrial Birds, Emphasizing Washington and Oregon.

Kasten, E.P., McKinley, P.K., Gage, S.H., 2010. Ensemble extraction for classification and detection of bird species. Ecol. Inform. 5, 153–166.

Krstulovic, S., Gribonval, R., 2006. MPTK: Matching pursuit made tractable. IEEE International Conference on Acoustics, Speech and Signal Processing.

Magurran, A.E., McGill, B.J., 2011. Biological Diversity: Frontiers in Measurement and Assessment. Oxford University Press Oxford.

Molau, S., Pitz, M., Schluter, R., Ney, H., 2001. Computing mel-frequency cepstral coefficients on the power spectrum. 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001 vol. 1, pp. 73–76.

Pavoine, S., Bonsall, M.B., 2011. Measuring biodiversity to explain community assembly: a unified approach. Biol. Rev. 86, 792–812.

Pieretti, N., Farina, A., Morri, D., 2011. A new methodology to infer the singing activity of an avian community: the acoustic complexity index (ACI). Ecol. Indic. 11, 868–873.

Schneider, D.C., 1994. Quantitative Ecology: Spatial and Temporal Scaling. Elsevier.

Somervuo, P., Harma, A., Fagerlund, S., 2006. Parametric representations of bird sounds for automatic species recognition. IEEE Trans. Audio Speech Lang. Process. 14, 2252–2263.

Spellerberg, I.F., Fedor, P.J., 2003. A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener' index. Glob. Ecol. Biogeogr. 12, 177–179.

Sueur, J., Pavoine, S., Hamerlynck, O., Duvail, S., 2008. Rapid acoustic survey for biodiversity appraisal. PLoS One 3, e4065.

Sueur, J., Farina, A., Gasc, A., Pieretti, N., Pavoine, S., 2014. Acoustic indices for biodiversity assessment and landscape investigation. Acta Acust. United Acust. 100, 772–781.

Towsey, M., 2013. Noise removal from wave-forms and spectrograms derived from natural recordings of the environment, QUT ePrints, http://eprints.qut.edu.au/61399/, Brisbane, Australia.

Towsey, M., Wimmer, J., Williamson, I., Roe, P., 2013. The use of acoustic indices to determine avian species richness in audio-recordings of the environment. Ecol. Inform. (Special issue).

Towsey, M., Zhang, L., Cottman-Fields, M., Wimmer, J., Zhang, J., Roe, P., 2014. Visualization of long-duration acoustic recordings of the environment. Procedia Comput. Sci. 29, 703–712.

Truskinger, A., Cottman-Fields, M., Eichinski, P., Towsey, M., Roe, P., 2014. Practical analysis of big acoustic sensor data for environmental monitoring. Paper Presented at the 2014 IEEE Fourth International Conference on Big Data and Cloud Computing (BdCloud), Sydney, Australia.

Wimmer, J., Towsey, M., Roe, P., Williamson, I., 2013. Sampling environmental acoustic recordings to determine bird species richness. Ecol. Appl. 23, 1419–1428.