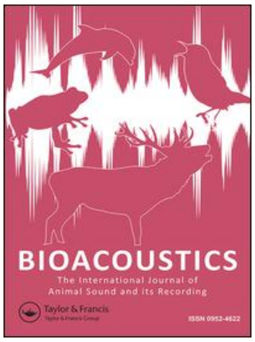

Taylor & Francis
Taylor & Francis Group

Bioacoustics
The International Journal of Animal Sound and its Recording

ISSN: 0952-4622 (Print) 2165-0586 (Online) Journal homepage: https://www.tandfonline.com/loi/tbio20

Automatic bird sound detection: logistic regression based acoustic occupancy model

Yi-Chin Tseng, Bianca N. I. Eskelson, Kathy Martin & Valerie LeMay

To cite this article: Yi-Chin Tseng, Bianca N. I. Eskelson, Kathy Martin & Valerie LeMay (2020): Automatic bird sound detection: logistic regression based acoustic occupancy model, Bioacoustics, DOI: 10.1080/09524622.2020.1730241

To link to this article: https://doi.org/10.1080/09524622.2020.1730241

Published online: 24 Feb 2020.

BIOACOUSTICS
https://doi.org/10.1080/09524622.2020.1730241

Taylor & Francis
Taylor & Francis Group

Automatic bird sound detection: logistic regression based acoustic occupancy model

Yi-Chin Tseng[a], Bianca N. I. Eskelson[a], Kathy Martin[b,c] and Valerie LeMay[a]

[a]Department of Forest Resources Management, University of British Columbia, Vancouver, Canada; [b]Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, Canada; [c]Environment and Climate Change Canada, Pacific Wildlife Research Centre, Vancouver, BC, Canada

ABSTRACT
Avian bioacoustics research was greatly assisted by the introduction of autonomous recording units, which not only allow remote monitoring but also make large-scale studies possible. However, manual inspection of acoustic recordings becomes more challenging with increasingly larger datasets. In this study, we developed a logistic model to predict the probability of bird presence in audio recordings using sound frequency percentiles. The acoustic recordings covered bird songs and calls in a wide range of environments (e.g. grassland, forest, urban areas) along with the presence of noise due to weather, traffic, insects, and human speech. Based on leave-one-out cross-validation, our final logistic model resulted in a 75% overall accuracy and a 16% false negative rate using the optimal cut-off of 0.35 (i.e. probability ≥ 0.35 indicates the presence of birds). Compared with a convolutional neural network model using the same dataset, the logistic model was about seven times faster in terms of the processing time, but achieved slightly lower overall accuracy. This bird sound detection model using sound frequency percentiles in a logistic model opens up promising approaches to aid in automatic, accurate, and efficient analyses of large audio datasets for monitoring wildlife communities.

ARTICLE HISTORY
Received 5 December 2019
Accepted 11 February 2020

KEYWORDS
Autonomous recording units (ARUs); bird song recognition; biodiversity monitoring; convolutional neural networks; IEEE bird sound detection challenge

Introduction

Ecoacoustics, a field of research that focuses on the investigation and interpretation of environmental sounds (Sueur and Farina 2015), benefits from the development of autonomous recording units (ARUs), which enable remote and autonomous data collection of soundscapes at large scales. ARUs have become increasingly popular in recent years, especially for avian biodiversity assessments and wildlife community investigations across landscapes (Sueur et al. 2014; Shonfield and Bayne 2017). Through aural detection, field observers have been able to reliably detect birds during point count surveys (Simons et al. 2007). Acoustic surveys with ARUs similarly allow reliable, aural detection of birds (Shonfield and Bayne 2017), and have been used to derive ecological indices such as avian species occupancy (Lambert and McDonald 2014; Sovern et al. 2014; Drake et al. 2016), and abundance (Hedley et al. 2017). Large-scale monitoring using ARUs has been widely applied to assess

CONTACT Yi-Chin Tseng sunnyyctseng@gmail.com

avian community composition (Klingbeil and Willig 2015), monitor endangered birds (Garnett et al. 2011), and conduct bird counting (Rosenstock et al. 2002).

Despite increasing interest in using ARUs, analysing audio recordings continues to be a challenge given the terabytes of data generated by large-scale monitoring projects (Joly et al. 2019). Accordingly, commercial and open source software have been developed to assist bird species identification from audio recordings (Shonfield and Bayne 2017), including Song Scope (Wildlife Acoustic, Maynard, Massachusetts, USA), Kaleidoscope Pro (Wildlife Acoustic, Maynard, Massachusetts, USA), Raven Pro (Cornell Laboratory of Ornithology, Ithaca, New York, USA), Sound Analysis Pro (Tchernichovski et al. 2000), and the R package 'monitoR' (Hafner and Katz 2017). However, using species identification software generally requires substantial computational processing time and memory resources (Stowell et al. 2016). One way to minimise the processing time and memory resources is to filter the recording periods without bird sounds before using the species identification software (Stowell et al. 2016). A pre-filtering step is especially useful for audio recordings with low bird sound activity, such as acoustic recordings collected for monitoring nocturnal birds (Rognan et al. 2012). Specifically, employing an automatic bird sound detection algorithm as a pre-filtering step for bird sounds can greatly increase the efficiency of audio analyses.

Convolutional neural networks (CNN) are the state-of-the art models that have been widely applied for detecting sound events in audio recordings, and have demonstrated high accuracy in distinguishing recordings with versus without bird sounds (Adavanne et al. 2017; Cakir et al. 2017; Kong et al. 2017; Pellegrini 2017). A CNN contains a series of convolutional layers that include filters to be applied on the input image and enable a model to recognise underlying relationships in a set of images (Wyse 2017). The input 'images' in a CNN bird sound detection model are spectrograms derived from audio recordings, which contain information of sound frequency in the vertical axis and time in the horizontal axis (Byers and Kroodsma 2016). Although being the state-of-the art model form for the bird sound detection task, a CNN model generally requires considerable computational resources for training and making predictions (Dreiseitl and Ohno-Machado 2002). Further, a CNN typically contains thousands of parameters in order to define kernels in convolutional layers, making the interpretation of model parameters a difficult task (Dreiseitl and Ohno-Machado 2002).

Due to their simplicity, effectiveness, and relative ease of interpretation, logistic models (i.e. generalised linear models using a logit link and a binomial distribution) are also often used for predicting the presence of wildlife (Mladenoff et al. 1999). A previous study showed that logistic detection models can achieve similar accuracy as neural network detection models but can greatly reduce the computational resources and enable drawing of statistical inferences (Dreiseitl and Ohno-Machado 2002). Logistic prediction models can relate a binary response variable, bird presence or absence, with predictor variables extracted from audio recordings. Several low-level descriptive parametric representations (i.e. dimension reduction approach) have been proven to be simple and powerful predictor variables for bird vocalisation detection (Fagerlund 2014). For example, peak spectral components of recordings have been used to detect flight calls (Tanttu et al. 2006); highest frequency and the loudest frequency have also been used to recognise migratory birds (Schrama et al. 2007). Given that bird sound frequencies are distinctively high with most bird species having sound frequencies ranging from 1 to 8 kHz (Bonney 2007), sound frequency-related predictor variables were commonly used for detecting bird sounds in audio recordings (Fagerlund 2014).

In this study, we propose a logistic bird sound detection model using sound frequency percentiles as an alternative to CNN for automatic bird sound detection. The logistic model and sound frequency percentiles, descriptive parametric representations of sound frequency distribution of an audio recording, were chosen aiming to reduce the computational resources needed for training a bird sound detection model, and to use relatively interpretable predictor variables. Specific objectives were: 1) to test the effectiveness of using sound frequency percentiles in a logistic model to predict the presence of bird sounds in audio recordings, 2) to compare the performance of this approach to results from a CNN model, and 3) to provide an algorithm to implement this method. This study provides an efficient and accurate algorithm for detecting birds from audio recordings that could potentially be used in a wide variety of environments.

Materials and methods

Data

In 2016, the Institute of Electrical and Electronics Engineers (IEEE) Signal Processing Society initiated a Bird Audio Detection challenge (Stowell et al. 2016), offering a dataset in real live bioacoustics monitoring projects. The audio recordings dataset included 15,690 10-second recordings (Table 1). About half of the recordings were from a United Kingdom (UK) bird-sound crowd-sourced research spinout called Warblr (Warblr 2015). Warblr included smartphone-derived recordings covering a wide distribution of UK locations and environments. The other half of the recordings were from a world-wide field recording project called FreeSound (Stowell and Plumbley 2014). The recordings from FreeSound were diverse in location and environment from all around the world with a relatively high proportion taken in Europe. Each 10-second recording was categorised into bird presence or absence by a network of volunteers, including a revalidation process that minimised mislabelled recordings (Stowell et al. 2016). Audio recordings covered a wide range of bird species, but the species information was not provided in the dataset. The recordings further included noise due to weather, traffic, large mammals, insects, human speech, and even human bird imitations. All recordings were formatted into a 44.1 kHz sampling rate and the mono pulse code modulation WAV (Stowell et al. 2016).

Table 1. Number of audio recordings by data source and bird presence/absence.

	Presence of birds	Absence of birds	Total
Warblr	6,045 (39%)	1,955 (13%)	8,000 (52%)
FreeSound	1,935 (12%)	5,755 (37%)	7,690 (48%)
Total	7,980 (51%)	7,710 (49%)	15,690 (100%)

Audio processing and variable extraction

A band-pass filter with a lower limit sound frequency of 400 Hz and an upper limit sound frequency of 10,000 Hz was applied to each recording, in order to reduce the low-frequency noise caused by wind or mechanical operation (de Oliveira et al. 2015; Adavanne et al. 2017; Kong et al. 2017). A spectrogram (Figure 1(a)) was generated for each audio recording with sound frames being 12 milliseconds long with a 6 milliseconds overlap (i.e. the first sound frame contained 1–12 milliseconds, the second sound frame contained 7–18 milliseconds, etc.). For an audio recording sampled at 44.1 kHz, the frame length of 12 milliseconds corresponds to 512 samples and the overlap of 6 milliseconds corresponds to a frame shift in 256 samples. The frame shift determined the time resolution. The discrete Fourier transformation was then applied to each sound frame (Boll 1979) with a frequency resolution of 40 Hz, which was the ratio between the frequency range in Hz (400–10,000 Hz) and the frame size (512 samples). The dominant sound frequency (i.e. the sound frequency with the highest intensity) was selected from each sound frame (Figure 1(b)) and these were used to obtain an empirical cumulative distribution (Figure 1(c)). Since bird sounds typically range from 1 to 8 kHz (Bonney 2007), it was hypothesised that the higher sound frequencies would be more powerful in detecting bird presence. Therefore, 10 sound frequency percentiles (i.e. 30th, 40th, 50th, 60th, 70th, 80th, 90th, 95th, 97.5th, and 99th), with focus on the higher percentiles, were extracted from the empirical cumulative distribution as candidate predictor variables. Furthermore, the source of the recording (i.e. Warblr or FreeSound), SOURCE, was added as an additional candidate predictor variable. This indicator variable was to account for the difference between the two sound projects (e.g. crowdsourcing, Warblr, using uncontrolled equipment versus remote monitoring project, FreeSound, using fixed and known recording equipment).

Logistic detection model formulation

Logistic models (McCullagh and Nelder 1989) were developed, specifically:

$$\widehat{Prob}(Y=1) = \frac{exp\left(\beta_0 + \beta_1 x_{1i} + \quad \ldots \quad + \beta_m x_{mi}\right)}{1 + exp\left(\beta_0 + \beta_1 x_{1i} + \quad \ldots \quad + \beta_m x_{mi}\right)}$$

where $\widehat{Prob}(Y=1)$ is the predicted probability of a bird presence, $x_{1,\ldots,}x_m$ are predictor variables, and $\beta_1, \ldots, \beta_m$ are the parameters corresponding to each of the predictor variables. Model parameters were estimated using maximum likelihood (McCullagh and Nelder 1989) and the glm() function in R (R Core Team 2019), specifying a generalised linear model with a logit link and a binomial distribution.

A 'base model' was first developed. Likelihood ratio tests (α = 0.05) were used to test whether the candidate predictor variables (i.e. 10 frequency percentiles, SOURCE, and interactions between percentiles and SOURCE) contributed significantly to the detection of bird sound in audio recordings. Specifically, likelihood ratio tests were used to test: 1) whether to include any of the candidate variables in the model, 2) whether to include SOURCE in the model, and 3) whether to include interactions between percentiles and SOURCE in the model. This resulted in the selection of the 'base model', and the possibility of further variable reductions was examined by applying backwards elimination. Specifically, percentiles with their

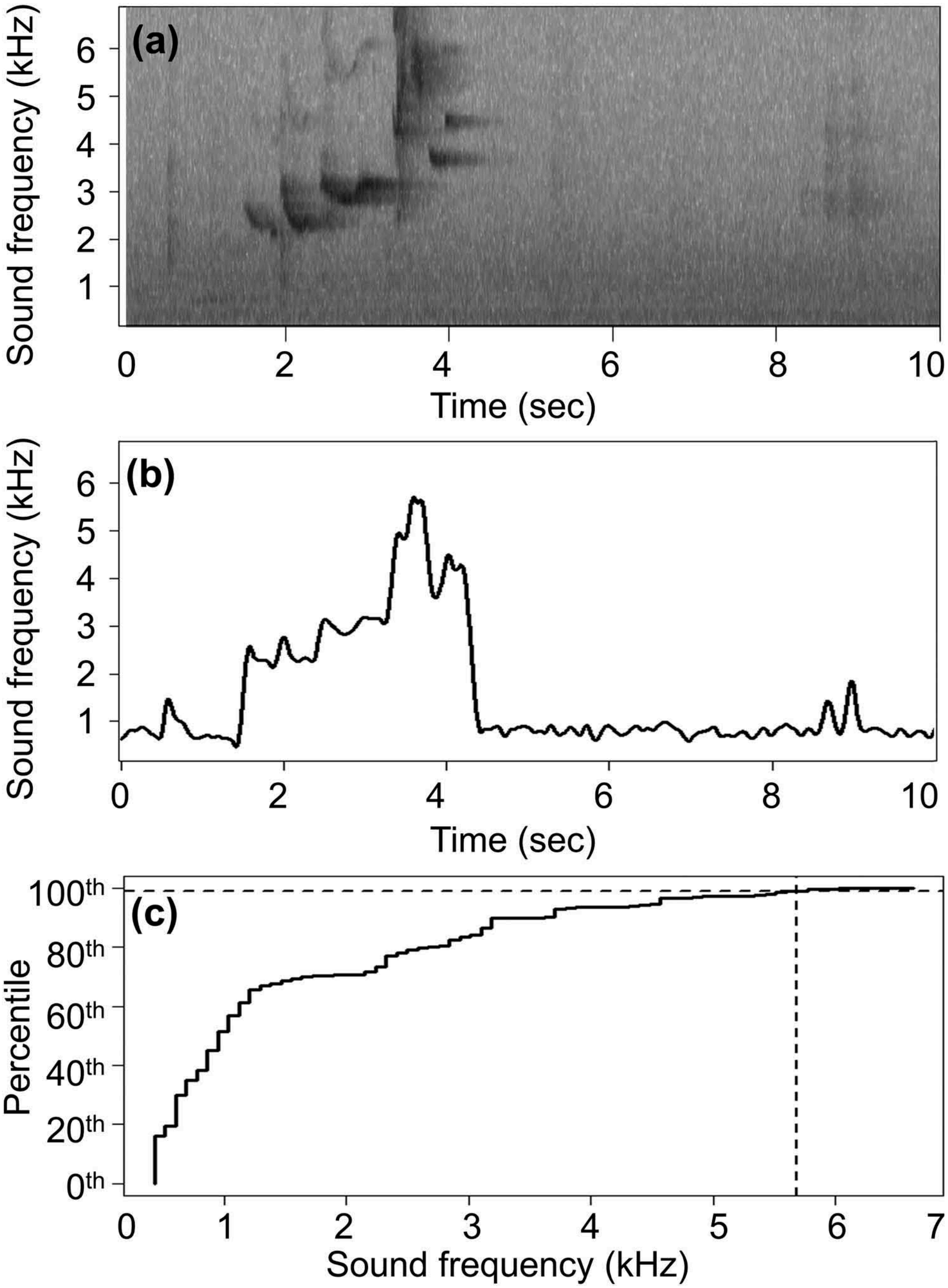

Figure 1. Process of extracting sound frequency percentiles from a recording. (a) Spectrogram of a ten-second long recording. (b) Dominant sound frequencies selected from each of the sound frames in the spectrogram. (c) Empirical cumulative distribution of dominant sound frequencies. The 99th percentile was marked by dashed lines for demonstration.

interactions with SOURCE were dropped until all remaining percentiles with SOURCE interactions were significant ($\alpha = 0.05$).

Logistic detection model evaluation

Since the accuracy of a logistic model is often overestimated when using the same observations for model building and testing (Hosmer et al. 2013), the leave-one-out cross-validation method (Snee 1977) was applied. Specifically, the predicted probability for each recording was based on the model fitted without that particular observation. Several metrics were selected to evaluate the logistic detection model based on recommendations for habitat models provided by Pearce and Ferrier (2000), and for acoustic recognisers provided by Knight et al. (2017).

First, the Hosmer-Lemeshow goodness-of-fit test (α = 0.05) (Hosmer et al. 2013) was used to test the agreement between observed bird presence (i.e. true bird presence) and the predicted probabilities. The chi-square statistic was derived with a group number of 20, where the first group included audio recordings with 0 to 0.05 (excluding 0.05) predicted probabilities, and the second group included audio recordings with 0.05 to 0.1 (excluding 0.1) predicted probabilities, and so on. A significant p-value indicates model lack-of-fit.

Then, the predicted presence or absence of each audio recording was obtained by comparing the predicted probability to a cut-off probability: an observation was categorised as predicted presence if the predicted probability ≥ cut-off probability. In order to assess how the cut-off probability influences the logistic detection model performance, overall accuracy, false negative rate, and false positive rate corresponding to each cut-off probability ranging from 0.2 to 0.9 in increments of 0.05 were calculated (Pearce and Ferrier 2000).

$$Overall\ accuracy = 1 - \frac{\#\ of\ false\ negative\ predictions + \#\ of\ false\ positive\ predictions}{Total\ \#\ of\ audio\ recordings}$$

$$False\ negative\ rate = \frac{\#\ of\ false\ negative\ predictions}{\#\ of\ audio\ recordings\ with\ bird\ presence}$$

$$False\ positive\ rate = \frac{\#\ of\ false\ positive\ predictions}{\#\ of\ audio\ recordings\ with\ bird\ absence}$$

where *false negative predictions* were audio recordings with bird presence but the model predicted as absence of birds, and *false positive predictions* were audio recordings with bird absence but the model predicted as presence of birds. The false negative rate indicates the chances of the model missing the bird sound in the recording, while the false positive rate indicates the chances of the model failing to identify the recordings without bird sound. An optimal cut-off probability was selected based on the highest overall accuracy.

Finally, the relative operating characteristic (ROC) curve was derived by plotting the true positive rate (i.e. 1 – false negative rate) against the false positive rate across a gradient of cut-off probabilities (DeLong et al. 1988). Accordingly, the area under the ROC curve (AUC) was calculated as an aggregate measure of performance of a detection model: a model with an AUC value above 0.7 and 0.8 provides satisfactory and excellent discrimination, respectively (Hosmer et al. 2013).

Comparison to a convolutional neural network detection model

A convolutional neural network (CNN) detection model was developed as a comparison to the logistic detection model. The CNN architecture was based on an eight-layer CNN from the Canadian Institute for Advanced Research (CIFAR-10) in Keras API (Chollet 2015). This CNN architecture was selected given its reported high accuracy in the CIFAR-10 image classification task. The source code was provided by Lima (2018) and was further modified to process the audio recording dataset in this study. The input layer of the CNN included sound spectrograms derived from the 10-second recordings with each spectrogram down-sized to a single channel image of a 50 Hz-by-50 second pixel to reduce the computer processing time. This CNN architecture included four convolutional layers, two pooling layers, and two dense layers (Figure 2). For regularisation, dropout (Srivastava et al. 2014) with a rate of 0.25 was employed in convolutional layers. The network was trained with 30 epochs (i.e. 30 full training cycles) and a batch size of 100 using the Adam optimiser (Kingma and Ba 2014) and binary cross-entropy as the loss function.

In order to compare the logistic model to the CNN detection model, the dataset was randomly partitioned into a training dataset (80%) and a test dataset (20%). Both the logistic and CNN models were fit using the training dataset and evaluated using the test dataset. The processing time to fit a model based on the training dataset and to make predictions on the test dataset was recorded for both types of models using a computer with an Intel Core i7-4790 CPU operating at 3.6 GHz. The overall accuracy, false negative, and false positive rates were calculated based on the predictions made using the models on the test dataset.

All audio processing and statistical analyses were performed in the computing environment R (R Core Team 2019), version 3.3.1. Packages ‘tuneR’ (Ligges et al. 2018) and ‘seewave’ (Sueur et al. 2008a) were used for audio processing and dominant frequency extraction. The glm() function in the package ‘nlme’ (Pinheiro et al. 2019) was used for developing logistic detection models. Finally, the CNN detection model was developed with package ‘keras’ (Allaire and Chollet 2019).

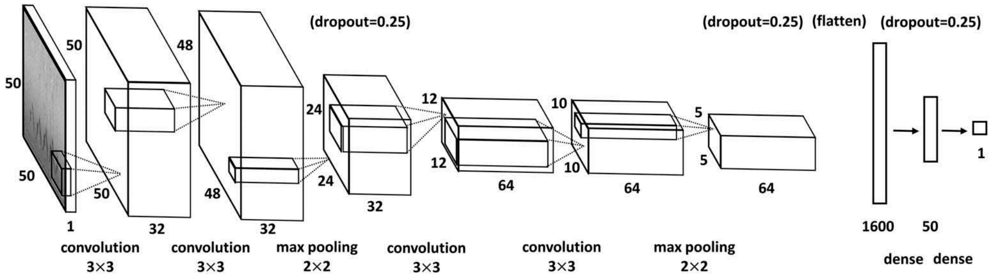

Figure 2. The complete convolutional neural network (CNN) model architecture used in this study. The numbers indicate the dimension of each object. The input layer of the model is the spectrogram of the recording and the output is the predicted probability of the presence of bird sounds.

Results

Characteristics of sound frequency percentiles

Two recordings were selected from the dataset to visually demonstrate the difference between audio recordings with and without bird sounds. The spectrogram of an audio recording with bird sounds had higher dominant sound frequencies compared to the one with human speech and traffic noise (Figure 3(a)). This difference in sound frequency distribution resulted in distinct patterns in the empirical cumulative distribution of the dominant sound frequencies (Figure 3(b)). The recording with bird sounds had a curve that rose steadily and achieved the 100th percentile with high sound frequency at around 6 kHz, while the recording without bird sounds had a curve that rose sharply and achieved the 100th percentile in low sound frequency at around 3 kHz.

Furthermore, 10 sound frequency percentiles were summarised using all recordings with and without bird sounds (Figure 4). The recordings without bird sounds had around 60% of dominant sound frequencies lower than 1 kHz and 90% of dominant sound frequencies lower than 2 kHz. In general, recordings with bird sounds had higher mean dominant sound frequency in each percentile than those without bird sounds (Figure 4).

Logistic detection models

To test the contribution of candidate predictor variables, four logistic models, the null model (GLM.0), two reduced models (GLM.R1, GLM.R2), and the full model (GLM.F), were fitted as possible base models (upper part of Table 2). All three models (GLM.R1, GLM.R2, and GLM.F) were significantly different from the GLM.0 based on likelihood ratio tests (α= 0.05). Of these, GLM.F had the highest log likelihood and the lowest AIC (Table 2). Therefore, GLM.F was selected as the base model (i.e. a full model with all percentiles, SOURCE, and interactions between SOURCE and all percentiles).

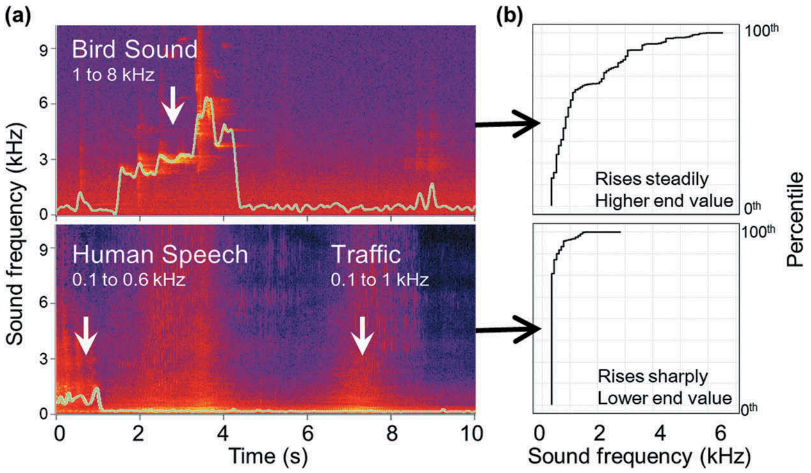

Figure 3. Illustration of audio recordings with and without bird sounds, specifically: (a) spectrograms of a recording with bird sounds (top), and a recording with human speech and traffic noise (bottom); and (b) associated empirical cumulative distributions of dominant sound frequencies.

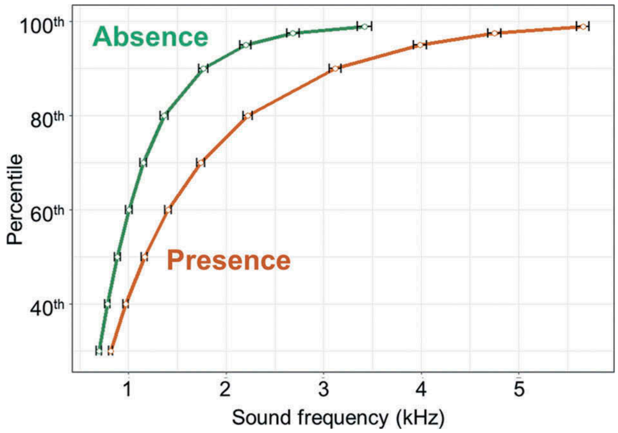

Figure 4. Sound frequency percentiles versus mean dominant sound frequencies with 99% confidence interval marked as horizontal error bars. Audio recordings were grouped by bird presence (orange) and absence (green).

Table 2. Logistic detection model comparison for selecting base model (upper part) and conducting variable selection based on the selected base model (lower part).

	Variables	Log likelihood	AIC	ΔAIC relative to GLM.0	p-value
GLM.0	None	−10,873.2	21,748	–	–
GLM.R1	10 percentiles	−9314.8	18,652	3096	$< 2.2*10^{-6}$
GLM.R2	10 percentiles + SOURCE	−8105.8	16,236	5512	$< 2.2*10^{-6}$
GLM.F	10 percentiles * SOURCE	−8043.9	16,132	5616	$< 2.2*10^{-6}$
GLM.R3	9 percentiles * SOURCE (excluding 40th)	−8044.2	16,128	5620	$< 2.2*10^{-6}$
GLM.R4	8 percentiles * SOURCE (excluding 40th and 80th)	−8044.8	16,126	5622	$< 2.2*10^{-6}$
GLM.R5	7 percentiles * SOURCE (excluding 40th, 80th, and 60th)	−8046.8	16,126	5622	$< 2.2*10^{-6}$

GLM.F was then reduced using backwards elimination (lower part of Table 2). The 40th, 80th, and 60th percentiles and their interactions with SOURCE were dropped sequentially (e.g. GLM.R3 was derived from GLM.F by dropping 40th percentile and its interaction with SOURCE). The final model, GLM.R5, included 15 predictor variables (i.e. 30th, 50th, 70th, 90th, 95th, 97.5th, and 99th percentiles, SOURCE, and the interactions of these percentiles with SOURCE).

Logistic detection model accuracy

The distribution of predicted probabilities using the leave-one-out cross-validation process was clearly left-skewed for recordings with bird sounds versus right-skewed for recordings without bird sounds (Figure 5). The recordings with bird sounds had a higher median predicted probability (0.72) compared to that without bird sounds (0.20). Furthermore, potential false predictions could be noticed on these distributions. Specifically, the small peak on the left side of the orange frequency distribution are recordings with bird sounds but having low predicted probability (i.e. potential false negative). Alternatively, the small

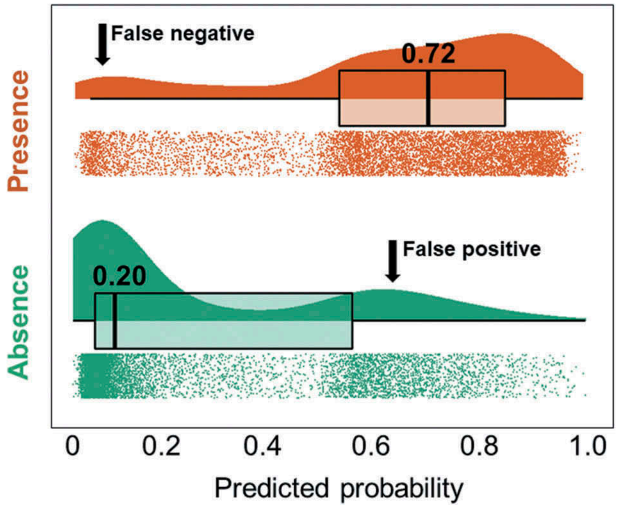

Figure 5. The distributions of predicted probabilities for recordings with (orange) versus without (green) bird sounds. The left and right side of the boxes are the first quantile and the third quantile of the predicted probabilities, respectively. The recordings with bird sounds had a higher median predicted probability (0.72) than that of recordings without bird sounds (0.20).

peak on the right side of the green frequency distribution are recordings without bird sounds but having high predicted probability (i.e. potential false positive).

The Hosmer-Lemeshow test for the final logistic detection model indicated a lack-of-fit (p-value <0.001). However, it should be noted that the power of this test is very high given the large number of audio recordings; thus, even a small lack-of-fit would be detected (Neter et al. 1996). To illustrate this, the observed presence proportions were plotted against predicted probabilities for the 20 equal-interval probability groups (Figure 6), as recommended by Pearce and Ferrier (2000). Generally, the observed presence proportions were higher than the predicted probabilities in the mid-range of predicted probabilities, indicating some lack-of-fit. This generally corresponded to the predicted probability groups with less data (i.e. shown as smaller blue bubbles). Further, the slope and the intercept of the fitted line were 0.96 (parameter standard error 0.05) and 0.03 (parameter standard error 0.03), respectively. The 1:1 correspondence fell in the region of the 95% confidence band, indicating the model was well-calibrated.

The optimal cut-off probability was 0.35 based on the highest overall accuracy of 75% (31% true positives and 44% true negatives; Figure 7). The corresponding false negative rate was 16%, with a false positive rate of 35%. The model achieved an AUC of 82%, showing that the model provided excellent discrimination between recordings with and without bird sounds.

Logistic vs CNN models

The logistic and CNN models were compared by partitioning the dataset into a training dataset (80%) and a test dataset (20%). The processing time for the logistic model was shorter

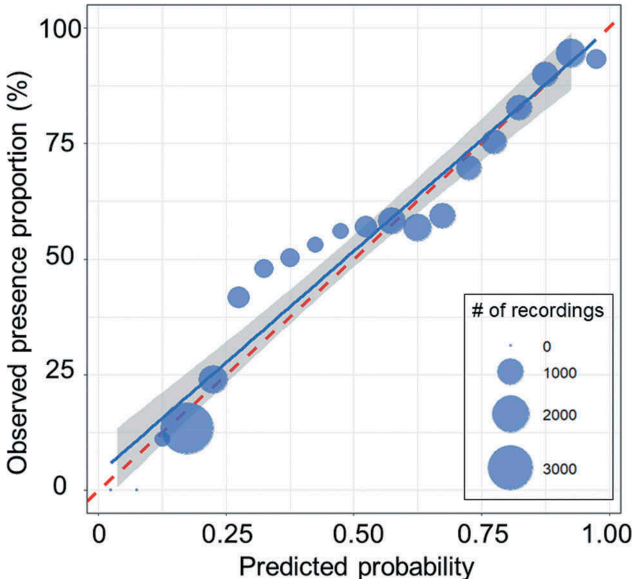

Figure 6. Observed presence proportions versus predicted probabilities for 20 equal-interval predicted probability groups (blue bubbles). The size of the bubbles represents the number of recordings in each group. The fitted line (solid blue line) is shown along with a 95% confidence band (grey area), compared with a 1:1 correspondence (dashed red line).

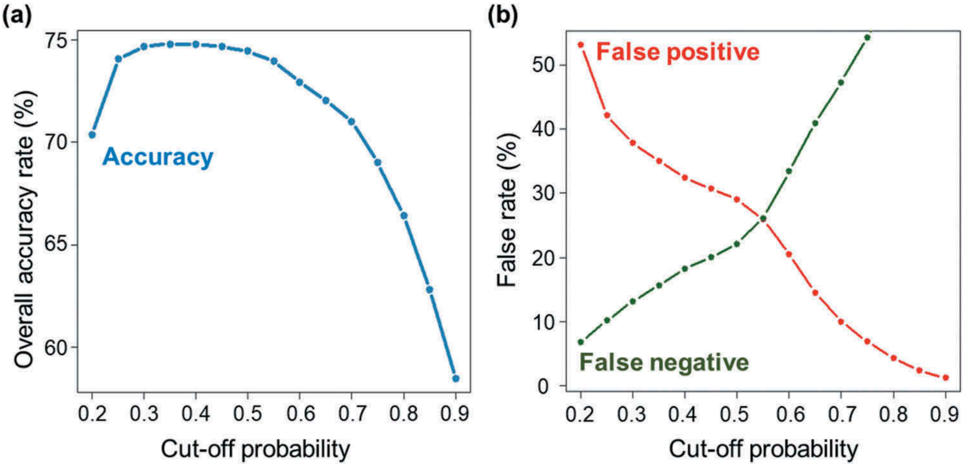

Figure 7. (a) Overall accuracy (blue), (b) false negative rate (green), and false positive rate (red) across a gradient of cut-off probabilities.

than that of the CNN model for both training and making predictions (Table 3). In particular, each 10-second recording required around 0.3 seconds to make predictions using the logistic model versus 2 seconds using the CNN model. However, the logistic model had a lower overall accuracy than CNN model with a higher false positive rate and a similar false negative rate.

Table 3. Processing times and model performance comparison between the logistic model and the CNN. Processing times were tested on a computer with an Intel Core i7-4790 CPU operating at 3.6 GHz.

Model	Processing time for training (hr)	Processing time for making predictions (hr)	Overall accuracy	False negative	False positive
Logistic	1:09	0:17	76%	14%	34%
CNN	6:48	1:40	84%	19%	13%

Discussion

Using sound frequency percentiles as low-level descriptive parametric representations

It was hypothesised that sound frequency percentiles would be powerful low-level descriptive parametric representations for detecting bird presence in audio recordings based on the fact that bird sound has distinctively high sound frequencies (Bonney 2007). This hypothesis was supported by our results, where a logistic detection model with sound frequency percentiles achieved an overall accuracy of 75% in detecting bird sounds in audio recordings. Frequency percentiles not only allow dimension reduction but also provide relatively interpretable predictor variables.

Challenges of CNN detection models

Despite achieving high overall accuracy, CNN models have several potential challenges when used to detect bird sound in audio recordings. First, the simplification of spectrograms as images might result in the loss of information (Wyse 2017). Specifically, the input of the CNN model were the spectrograms of the audio recordings, which have fundamentally different meanings in the axes: one being sound frequency (Hz) and the other being time (second). Although treating spectrograms as images (i.e. with same meaning of axes) in CNN models has become a common approach for sound detection and proved to be extremely useful (Stowell et al. 2016), some information, such as harmonics of sounds, might be missed with the strong localisation ability of CNN models (Wyse 2017). Second, the interpretability of the CNN model was limited. Although the large number of parameters in a CNN model enable it to model nonlinear relationships in the data, they also limit the interpretability of the CNN model since the contribution of the kernels in each convolutional layer is difficult to determine (Dreiseitl and Ohno-Machado 2002). Finally, the CNN model in this study required seven times more computational resources than the logistic model, even with a relatively simple CNN architecture. A ‘deeper’ CNN model (i.e. a CNN model with more convolutional layers) might provide even better model performance, but would require greater processing resources (Dreiseitl and Ohno-Machado 2002). This would be of particular concern where the processing resources are limited. In contrast, the logistic detection model is relatively simple, easier to modify, and can be applied readily to large datasets.

Challenges of the logistic detection model

Although having several advantages over the CNN model, the logistic detection model developed in this study still had some potential limitations. First, the logistic detection model might miss bird species with extremely low sound frequency (e.g. owl, grouse, ptarmigan). This was due to the application of the high pass filter during the audio processing and given that more high sound frequency percentiles were selected in the model. The application of a high pass filter with a cut-off frequency of 400 Hz will efficiently reduce the noise appearing in the spectrogram, but it might also filter out bird sounds coming from species with lower sound frequency such as doves. Further, more high sound frequency percentiles were selected in the model, which means that the logistic detection model was more sensitive to bird species having higher sound frequency. Second, the logistic detection model also resulted in high false positive (35%) and false negative (16%) rates. For these false predictions, false positives would increase the examination effort after applying the detection model, but would not influence the truthfulness of the following analysis as the recordings of interest would be retained in the sample (Pearce and Ferrier 2000). In contrast, recordings with false negatives would be identified as not including bird sounds, while they might include important information. This could cause issues particularly if 'filtered out' recordings contain species of interest (Pearce and Ferrier 2000). This approach might still work if these 'filtered out' recordings do not include the target group of species being monitored.

False positives likely resulted from the inclusion of non-bird high frequency sounds occurring in this diverse dataset, such as insect vocalisations. For example, cicadas have their sound frequency ranges from 4 kHz to 16 kHz (Bennet-Clark and Young 1994), which overlaps with the range for birds of 1 to 8 kHz (Bonney 2007). As a result, using upper frequency percentiles in the logistic model may not have recognised the cicadas as non-bird sounds. This issue with cicada sounds affecting the performance of a bird sound detection models has been noted previously in other research (Towsey et al. 2014). A solution for this issue was proposed by Brown et al. (2019), where nine acoustic indices were used to filter cicada chorus in a recording. Towsey et al. (2014) also used spectral entropy and background noise to develop a simple classifier to detect cicadas. Instead of manually checking and identifying the false positives from the logistic model results, these insect sound detection algorithms could be applied after the bird sound detection model to reduce the false positives and save time examining all recordings predicted as including bird sounds.

False negatives likely resulted from the inclusion of unconstrained content (e.g. wind, rain, traffic noise), which masks bird sounds (Rumsey and Mccormick 2012). In the logistic detection model, the dimension reduction approach for calculating predictor variables involved the selection of the dominant sound frequency in a specific sound frame. This approach is robust even with bird sound and non-bird sound overlapping in an audio recording, as long as the intensity of the bird sound is higher than other noises. However, the use of dominant frequencies might be more challenging for audio recordings collected from locations with constant loud noises, where bird sounds might be easily covered by noise. For this, one solution is to integrate other predictor variables into the logistic detection model that are more robust when multiple sound sources exist in the same sound frame (Fagerlund 2014). Another solution is to filter the unconstrained content in the recording before applying the detection model. Applying a high pass filter is the easiest way to remove such noises and, in particular, several indices have been proposed for this filtering purpose, such as the acoustic complexity index (Pieretti et al. 2011), spectral and temporal entropy (Sueur et al. 2008b), background noise (Towsey 2013), and spectral cover (Towsey et al. 2014).

Future work

Covering a wide range of bird species and ecosystems, the audio recordings used in this study provided an opportunity to build a model with broad potential applications. However, the lack of information about bird species in the dataset limited the possibility of examining the model performance across different bird species groups. Thus, future studies are needed to test this logistic detection model on audio recordings focusing on finding different target species. For this, a better model performance is expected given decreased variation in the data.

Further, the audio recordings used in this study did not include environments with complex soundscapes, such as tropical environments (Stowell et al. 2016). Tropical environments have a high diversity of sounds that lead to heterogeneous soundscapes (Rodriguez et al. 2014), which may result in including more than one bird species or increasing the possibility of overlapping sounds in an audio recording. In this case, the results of the presented logistic detection model may not hold. Therefore, future studies could test the performance of the developed model with audio recordings collected across a range of environments that differ in their complexity of soundscapes.

Acknowledgements

Great gratitude for the Machine Listening Lab in Queen Mary University of London for providing and summarizing the audio recordings used in this study. Special thanks to Min-Zhi Liao for his recommendations in developing the convolutional neural network models. Thanks to Kristina Hick and Woongsoon Jang for their reviews, which greatly assisted the research.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Soundscape Association of Taiwan.

ORCID

Yi-Chin Tseng http://orcid.org/0000-0002-8621-2244

Data availability statement

Data used in this study are available from the Bird Audio Detection Challenge (http://dcase.community/challenge2018/task-bird-audio-detection, accessed on 23 October 2019). The R code for the logistic model, the CNN model, and example recordings are available on the GitHub repository (https://github.com/SunnyTseng/Bird-Sound-Detection-2019.git, accessed on 8 November 2019).

References

Adavanne S, Drossos K, Cakir E, Virtanen T. 2017. Stacked convolutional and recurrent neural networks for bird audio detection. Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO); Kos Island. p. 1729–1733.

Allaire JJ, Chollet F. 2019. keras: R interface to 'Keras'. [online]; [Accessed 2020 Jan 24]. https://CRAN.R-project.org/package=keras

Bennet-Clark H, Young D. 1994. The scaling of song frequency in cicadas. J Exp Biol. 191 (1):291–294.

Boll S. 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Acoust. 27(2):113–120.

Bonney R. 2007. Citizen science at the Cornell lab of ornithology. In: Yager R, Falk J, editors. Exemplary science in informal education settings: standards-based success stories. Arlington (VA): National Science Teachers Association. p. 213–229.

Brown A, Garg S, Montgomery J. 2019. Automatic rain and cicada chorus filtering of bird acoustic data. Appl Soft Comput. 81:105501.

Byers BE, Kroodsma DE. 2016. Avian vocal behavior. In: Lovette IJ, Fitzpatrick JW, editors. The Cornell Lab of Ornithology handbook of bird biology. 3rd ed. Hoboken (NJ): John Wiley & Sons. p. 355–405.

Cakir E, Adavanne S, Parascandolo G, Drossos K, Virtanen T. 2017. Convolutional recurrent neural networks for bird audio detection. Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO); Kos Island. 1744–1748.

Chollet F. 2015. Keras. Github repository. [online]. [Accessed 2019 Nov 8]. https://github.com/fchollet/keras

R Core Team. 2019. R: a language and environment for statistical computing.

de Oliveira AG, Ventura TM, Ganchev TD, de Figueiredo JM, Jahn O, Marques MI, Schuchmann KL. 2015. Bird acoustic activity detection based on morphological filtering of the spectrogram. Appl Acoust. 98:34–42.

DeLong ER, DeLong DM, Clarke-Pearson DL. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 44 (3):837–845.

Drake KL, Frey M, Hogan D, Hedley R. 2016. Using digital recordings and sonogram analysis to obtain counts of yellow rails. Wildlife Soc B. 40(2):346–354.

Dreiseitl S, Ohno-Machado L. 2002. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform. 35(5–6):352–359.

Fagerlund S. 2014. Studies on bird vocalization detection and classification of species [dissertation]. Espoo (Finland): Aalto University.

Garnett S, Szabo J, Dutson G. 2011. The action plan for Australian birds 2010. Victoria (Australia): CSIRO Publishing.

Hafner S, Katz J. 2017. MonitoR: acoustic template detection in R. [online]. [Accessed 2019 Nov 8]. https://cran.r-project.org/web/packages/monitoR/index.html.

Hedley RW, Huang YW, Yao K. 2017. Direction-of-arrival estimation of animal vocalizations for monitoring animal behavior and improving estimates of abundance. Avian Conserv Ecol. 12(1): art6.

Hosmer DW, Lemeshow S, Sturdivant RX. 2013. Applied logistic regression. Hoboken (NJ): Wiley.

Joly A, Goëau H, Botella C, Kahl S, Servajean M, Glotin H, Bonnet P, Planqué R, Robert-Stöter F, Vellinga WP, et al. 2019. Overview of LifeCLEF 2019: identification of Amazonian plants, South & North American birds, and biche prediction. Cham (Switzerland): Springer.

Kingma DP, Ba J. 2014. Adam: a method for stochastic optimization. arXiv Preprint. arXiv: 1462.6908:1–15.

Klingbeil BT, Willig MR. 2015. Bird biodiversity assessments in temperate forest: the value of point count versus acoustic monitoring protocols. PeerJ. 3:e973.

Knight EC, Hannah KC, Foley GJ, Scott CD, Brigham RM, Bayne E. 2017. Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. Avian Conserv Ecol. 12(2):art14.

Kong Q, Xu Y, Plumbley MD. 2017. Joint detection and classification convolutional neural network on weakly labelled bird audio detection. Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO); Kos island. 1749–1753.

Lambert KTA, McDonald PG. 2014. A low-cost, yet simple and highly repeatable system for acoustically surveying cryptic species. Austral Ecol. 39(7):779–785.

Ligges U, Krey S, Mersmann O, Schnackenberg S. 2018. tuneR: analysis of music and speech. [online]. [Accessed 2020 Jan 24]. https://CRAN.R-project.org/package=tuneR.

Lima F. 2018. Convolutional neural networks in R. [online]. [Accessed 2019 Nov 8]. https://www.r-bloggers.com/convolutional-neural-networks-in-r/.

McCullagh P, Nelder JA. 1989. Generalized linear models. London (UK): Routledge.

Mladenoff DJ, Sickley TA, Wydeven AP. 1999. Predicting gray wolf landscape recolonization: logistic regression models vs. New Field Data. Ecol Appl. 9(1):37–44.

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. 1996. Applied linear statistical models. New York (NY): McGraw-Hill/Irwin.

Pearce J, Ferrier S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol Model. 133(3):225–245.

Pellegrini T. 2017. Densely connected CNNs for bird audio detection. Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO); Kos Island. p. 1734–1738.

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(3):868–873.

Pinheiro J, Bates D, DebRoy S, Sarker D. 2019. nlme: linear and nonlinear mixed effects models. [online]. [Accessed 2020 Jan 24]. https://CRAN.R-project.org/package=nlme.

Rodriguez A, Gasc A, Pavoine S, Grandcolas P, Gaucher P, Sueur J. 2014. Temporal and spatial variability of animal sound within a neotropical forest. Ecol Inform. 21:133–143.

Rognan CB, Szewczak JM, Morrison ML. 2012. Autonomous recording of great gray owls in the Sierra Nevada. NW Nat. 93(2):138–144.

Rosenstock SS, Anderson DR, Giesen KM, Leukering T, Carter MF, Thompson III F. 2002. Landbird counting techniques: current practices and an alternative. Auk. 119(1):46–53.

Rumsey F, Mccormick T. 2012. Sound and recording. Abingdon (UK): Routledge.

Schrama T, Poot M, Robb M, Slabbekoorn H. 2007. Automated monitoring of avian flight calls during nocturnal migration. Proceedings of the International Expert Meeting on IT-Based Detection of Bioacoustical Patterns; Isle of Vilm.

Shonfield J, Bayne EM. 2017. Autonomous recording units in avian ecological research: current use and future applications. Avian Conserv Ecol. 12(1):art14.

Simons TR, Alldredge MW, Pollock KH, Wettroth JM, Dufty AM. 2007. Experimental analysis of the auditory detection process on avian point counts. Auk. 124(3):986–999.

Snee RD. 1977. Validation of regression models: methods and examples. Technometrics. 19 (4):415–428.

Sovern SG, Forsman ED, Olson GS, Biswell BL, Taylor M, Anthony RG. 2014. Barred owls and landscape attributes influence territory occupancy of northern spotted owls. J Wildl Manag. 78 (8):1436–1443.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. 2014. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Rec. 15:1929–1958.

Stowell D, Plumbley MD. 2014. An open dataset for research on audio field recording archives: freefield1010. Proceedings of the 53rd International Conference, Semantic Audio; London, UK.

Stowell D, Wood M, Stylianou Y, Glotin H. 2016. Bird detection in audio: a survey and a challenge. Proceedings of the 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP); Salerno.

Sueur J, Aubin T, Simonis C. 2008a. Seewave, a free modular tool for sound analysis and synthesis. Bioacoustics. 18:213–226.

Sueur J, Farina A. 2015. Ecoacoustics: the ecological investigation and interpretation of environmental sound. Biosemiotics. 8:493–502.

Sueur J, Farina A, Gasc A, Pieretti N, Pavoine S. 2014. Acoustic indices for biodiversity assessment and landscape investigation. Acta Acust United Ac. 100(4):772–781.

Sueur J, Pavoine S, Hamerlynck O, Duvail S, Reby D. 2008b. Rapid acoustic survey for biodiversity appraisal. PLoS One. 3(12):e4065.

Tanttu JT, Turunen J, Selin A, Ojanen M. 2006. Automatic feature extraction and classification of crossbill (loxia spp.) flight calls. Bioacoustics. 15(3):251–269.

Tchernichovski O, Nottebohm F, Ho CE, Pesaran B, Mitra PP. 2000. A procedure for an automated measurement of song similarity. Anim Behav. 59(6):1167–1176.

Towsey M, Wimmer J, Williamson I, Roe P. 2014. The use of acoustic indices to determine avian species richness in audio-recordings of the environment. Ecol Inform. 21:110–119.

Towsey MW. 2013. Noise removal from wave-forms and spectrograms derived from natural recordings of the environment. Brisbane: Queensland University of Technology.

Warblr. 2015. Warblr: identify UK bird songs and calls. [online]. [Accessed 2019 Nov 8]. https://www.warblr.co.uk/.

Wyse L. 2017. Audio spectrogram representations for processing with convolutional neural networks. Proceedings of the First International Workshop on Deep Learning and Music joint with IJCNN; Anchorage. 37–41.