This is the peer reviewed version of the following article: Leseberg, N.P., Venables, W.N., Murphy, S.A. and Watson, J.E. (2020) Using intrinsic and contextual information associated with automated signal detections to improve call recogniser performance: a case study using the cryptic and critically endangered Night Parrot (*Pezoporus occidentalis*). *Methods in Ecology and Evolution*, Vol. 11, Iss. 11, Pp 1520-1530; which has been published in final form at **https://doi.org/10.1111/2041-210X.13475**.

Using intrinsic and contextual information associated with automated signal detections to improve call recogniser performance: a case study using the cryptic and critically endangered Night Parrot (*Pezoporus occidentalis*)

Running Title: POST-PROCESSING TO IMPROVE RECOGNISER PERFORMANCE

Nicholas P. Leseberg[A,B], *William N. Venables*[C], *Stephen A. Murphy*[A,B], *James E.M. Watson*[A,B,D]

[A]School of Earth and Environmental Sciences, The University of Queensland, Brisbane 4072, Queensland, Australia

[B]Green Fire Science, The University of Queensland, Brisbane 4072, Queensland, Australia

[C]School of Mathematics and Physics, The University of Queensland, Brisbane 4072, Queensland, Australia

[D]Centre for Biodiversity and Conservation Science, School of Biological Sciences, The University of Queensland, Brisbane 4072, Queensland, Australia

Corresponding author contact details:

Nick Leseberg

School of Earth and Environmental Sciences

University of Queensland

St Lucia QLD 4072

Australia

Email: n.leseberg@uq.edu.au

Phone: 0488 636 010

Abstract

1. Rapid expansion in the collection of large acoustic datasets to answer ecological questions has generated a parallel requirement for techniques that streamline analysis of these datasets. In many cases, automated signal recognition algorithms, often termed 'call recognisers', are the only feasible option for doing this. To date, most research has focused on what types of recognisers perform best, and how to train these recognisers to optimise performance.

2. We demonstrate that once recogniser construction is complete and the data processed, further improvements are possible using intrinsic and contextual information associated with each detection. We initially construct a call recogniser for the Night Parrot (*Pezoporus occidentalis*) using the R package monitoR, and scan a test dataset. We then examine a number of intrinsic variables associated with each detection generated by the recogniser, and several contextual variables, associated with the species' environment and ecology to determine if they might help predict whether a given detection is a true positive (target signal) or false positive (non-target signal). We test several logistic regression models incorporating different combinations of intrinsic and contextual variables, selecting the best-performing model for application. We train the model, using it to calculate the probability each detection is a true or false positive.

3. Substituting this model-derived probability for raw recogniser score improved the recogniser's performance, reducing the number of detections requiring proofing by 60% to achieve recall of 90%, and by 76% to achieve recall of 75%.

4. This technique is applicable to any recogniser output, regardless of the underlying algorithm. Application requires an understanding of how the recogniser algorithm determines matches, and knowledge of a species' ecology and environment. Because advanced programming skills and expertise are not required to apply this technique, it will be particularly relevant to field ecologists for whom building and operating call recognisers is an element of their research toolbox, but not necessarily a focus.

Additional Keywords:

bioacoustics, acoustic monitoring, call recogniser, rare species, night parrot

Introduction

The increasing availability of technology to collect and analyse acoustic data, particularly affordable automated recording units (ARUs), has seen a rapid expansion in this field of research and its applications for ecology and conservation (Shonfield & Bayne, 2017; Teixeira, Maron, & van Rensburg, 2019). The popularity of ARUs is largely due to their efficiency. Particularly for long-term deployments, it is much cheaper to purchase, deploy, and maintain an ARU than a human observer (Digby, Towsey, Bell, & Teal, 2013; Williams, O'Donnell, & Armstrong, 2018). Unlike human observers, ARUs can be left in the field unattended for extended periods, limited only by the availability of power and memory. As solar panels and large capacity memory cards are now also relatively cheap, maintaining permanent acoustic recording stations at remote sites has become feasible.

The easy collection of copious data has advantages and disadvantages. Large acoustic datasets may contain powerful data (Magurran et al., 2010), but extracting that data can be challenging. There are several techniques available to efficiently analyse large acoustic datasets, the most suitable contingent on the nature of the signal of interest (Joshi, Mulder, & Rowe, 2017; Towsey et al., 2018). Increasingly, research has focused on techniques that automate the signal extraction process. This is typically performed using a signal detection algorithm, hereafter termed 'call recogniser' (Potamitis, Ntalampiras, Jahn, & Riede, 2014; Priyadarshani, Marsland, & Castro, 2018). For infrequent signals within large datasets, a call recogniser may be the only feasible solution.

There are several options for researchers wanting to construct a call recogniser. They vary in complexity, from commercial off-the-shelf programs such as Kaleidoscope (Wildlife Acoustics Inc., Concord, Massachusetts, USA), to more recently, advanced machine learning algorithms (Koops, van Balen, & Wiering, 2014; Salamon & Bello, 2017), acoustic indices (Towsey, Wimmer, Williamson, & Roe, 2014), and wavelet based approaches (Priyadarshani, Marsland, Juodakis, Castro, & Listanti, 2020). Although the computational processes behind each differ, the basic premise remains the same; a computer is trained to detect and evaluate acoustic signals by comparing them to a known target signal. Potential signals are classified depending on their similarity to the target signal, with the user controlling the threshold at which a match is declared.

Understanding the impact of this threshold is critical to understanding the performance of a call recogniser. Setting a high threshold increases the precision of the recogniser, meaning a higher proportion of matches will represent actual detections, or true positives. However, this increases the likelihood of false negatives; target signals that do not meet the threshold, for example soft or distant calls. This reduces the recogniser's recall, or ability to identify all target signals within a dataset. Conversely, reducing the threshold ensures that more lower-scoring target signals are returned as matches, but simultaneously returns more lower-scoring non-target signals, or false positives. This increases the recogniser's recall, but also increases the proportion of non-target signals in the resulting dataset, thereby decreasing precision. This false positive / false negative trade-off is a well-known classification problem, with threshold choice driven by the relative cost of false positive or false negative errors.

Besides an obvious focus on which computational techniques create the most successful recognisers, research has also focused on the properties of training data that achieve the best results (Knight & Bayne, 2018; Priyadarshani et al., 2018). Because a call recogniser's output is dependent on how closely the signal of interest compares to the training data, efforts to improve a specific type of recogniser's performance have largely focused on this aspect of their development. However, little research has focused on how post-processing could be used to derive improvements in overall performance. Typically, the output of a recogniser is a list of potential 'detections', each with associated intrinsic information derived from the call recognition process, for example a 'score' reflecting how similar the detection is to the training data. There is also any number of contextual variables associated with each detection, such as time-of-day and geographic location, that are known to affect detectability (Horton, Stepanian, Wainwright, & Tegeler, 2015). Patterns in both intrinsic and contextual data could provide clues to predict whether a detection is actually a signal of interest.

In this paper we outline a novel method to develop a model that uses both intrinsic and contextual information associated with a call recogniser's raw output to generate an improved output. We intentionally present a detailed description of the process, because one of our aims is to demystify the process of automated call recognition for field ecologists, thereby encouraging them to perform their own analyses. Broadly, our process was to first construct a call recogniser for the Night Parrot (*Pezoporus occidentalis*), then investigate relationships between the intrinsic and contextual variables associated with the recogniser's output to establish if any could be incorporated into a model that predicts whether a detection is a true positive or false positive. Following a model development and selection process, we selected the best-performing model and tested whether this model improved recogniser performance.

Methods and Results

Study species and data collection

The Night Parrot is a cryptic and extremely rare bird that formerly occurred throughout arid central Australia (Higgins, 1999), but is now known from only a handful of sites. The species is relatively sedentary, and predictably vocal (Leseberg et al., 2019; Murphy, Silcock, Murphy, Reid, & Austin, 2017). They spend the day roosting in low, dense vegetation, as pairs or small groups. The birds emerge at dusk to engage in a brief period of calling before leaving their roost sites to feed. Birds occasionally return to their roost sites and call during the night, but typically return for another brief period of calling just before dawn. Night Parrot vocalisations are now relatively well known (Leseberg et al., 2019). Given this predictable calling behaviour, acoustic monitoring has proven the most efficient technique for both monitoring the species at known locations, and detecting it at new locations.

Since 2016, Night Parrot calling activity at three long-term stable roost sites in western Queensland has been monitored using Song Meter 3 and Song Meter 4 ARUs (Wildlife Acoustics Inc., Concord, Massachusetts, USA), fitted with standard external omnidirectional microphones. ARUs recorded from sunset to sunrise, using the ARU's default gain settings. Most ARUs recorded at sampling rates of 24000 Hz, or 48000 Hz, although some recorded at 16000 Hz. As required under the Nyquist-Shannon Sampling Theorem (Landau, 1967), these sampling rates are greater than twice the peak frequency of all Night Parrot calls of interest to this study.

Call recogniser development and sound file analysis

We used the R package monitoR (Katz, Hafner, & Donovan, 2016; R Core Team, 2018) to build a call recogniser for the Night Parrot. R is a programming language accessible to users without specialist programming skills, and in a comparison with recognisers using machine learning methods and commercially available packages, monitoR performed well (Knight et al., 2017). We used the technique outlined in Katz et al. (2016) to construct a series of binary point templates. Templates are created by clipping an example call from a sound file and creating a spectrogram (FFT transformation = Hann window, FFT size = 512, overlap = 0). A selection of cells of the resulting spectrogram are then classified as ‘on’ or ‘off’. ‘On’ cells are selected to represent the expected region of strongest signal for the call, while ‘off’ cells are placed strategically where no or little signal is expected (Fig. 1).

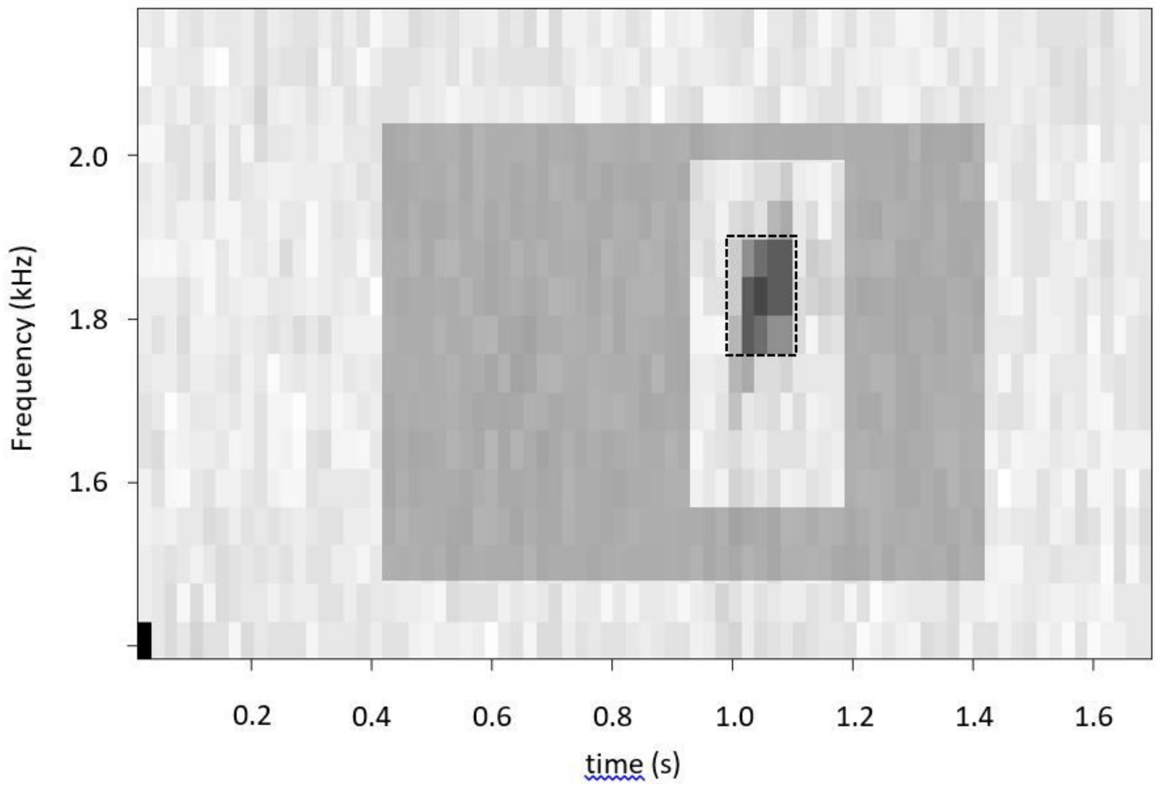

Figure 1. An example of a binary point matching template for the Night Parrot ‘toot’ call, overlaid on the spectrogram of a ‘toot’ call. The central box with dotted outline represents the ‘on’ cells, and ideally contains most of the expected call energy. The shaded area represents the ‘off’ cells.

Although Night Parrots have a variety of different calls, we focused on the bell-like and whistle calls, as these are the calls most likely to be heard in and around roost sites (Leseberg et al., 2019). These broad call types can be broken down further, and we constructed at least one template for each of the ten specific call types known from the study area. We used example calls extracted from the long-term monitoring dataset, adding further templates until testing suggested the recogniser could detect most local variation within these call types. The final recogniser used 31 different templates. Because monitoR requires template files and the sound files that will be scanned to have the same sample rate, these were downsampled or upsampled if required to a sampling rate of 24000 Hz. Qualitative testing confirmed that manipulating the files in this way had no apparent effect on results.

Before analysis, each sound file is converted to a spectrogram using the same parameters as were used to create the templates. Each template is then stepped along that spectrogram, and for every step a similarity score is assigned based on the difference between the amplitude detected in the ‘on’ cells, and the amplitude detected in the ‘off’ cells of the template. When plotted against time this results in a series of peaks; the recogniser returns a list of these peaks with their associated score. As some signals within the sound file are detected by more than one template, a buffer of two seconds was applied so only the highest scoring peak within any two-second period was returned. Because Night Parrot calls are generally short, temporally discrete events, the risk of missing calls due to applying this buffer was low.

Recogniser performance assessment

To evaluate recogniser performance, 90 ten-minute field recordings known to contain Night Parrot calls were extracted from the long-term monitoring dataset. We used field recordings to ensure measured performance reflected what could be achieved on actual field recordings rather than a manufactured test dataset (Potamitis et al., 2014). We used recordings from nights that were either calm or with light winds, as wind noise significantly reduces both ARU and recogniser performance. While this imposes a limitation on the future data the results of this research can be applied to, based on the species' ecology and our experience at the study site, this limitation is not onerous, and is one we are willing to accept to improve efficiency. To avoid overfitting, none of the field recordings contained calls that were used to train the recogniser. The dataset was balanced across the three long-term stable roost sites, and three discrete periods of the night: dusk, night, and dawn. Recordings for the dusk period occurred within one hour of sunset, recordings for the dawn period occurred within one hour of sunrise, and recordings for the night period included any time in between the defined dusk and dawn periods. Using audio-editing public domain software Audacity (version 2.3.0, http://audacity.sourceforge.net/), each clip was viewed in a spectrogram (spectrogram settings: y-axis = 0-4000 Hz, x-axis = 30 secs, FFT transformation = Hann window, FFT size = 256), and listened to at a consistent volume using a set of high-quality noise-cancelling headphones (Sennheiser PXC480). 1850 definite Night Parrot calls were detected, ranging from loud calls made in close proximity to the recorder, to faint, distant calls, that could not be seen on a spectrogram and were only detectable by manual listening.

Each 10-minute recording was then analysed using the call recogniser, with the threshold score set to zero, so all peaks in the similarity score were returned as 'detections'. It is important to note that a 'detection' in this sense is a return from the recogniser representing a prospective detection; it may or may not be an actual detection. The recogniser returned 31437 detections from the 900-minute dataset. These detections were compared to the manually extracted data, and each classified as either a true positive (an actual Night Parrot call) or false positive (not a Night Parrot call). The recogniser did not detect 110 of the 1850 calls in the dataset. These were added to the dataset and classified as false negatives. We assessed baseline performance by producing a precision-recall curve, and calculating the area under the curve (AUC) (Fig. 2).

A precision-recall curve plots recall for each value of precision as the classification threshold is reduced, allowing assessment of the trade-off between the two parameters. AUC of the precision-recall curve is the recommended univariate statistic for comparing call recognisers (Knight et al., 2017).

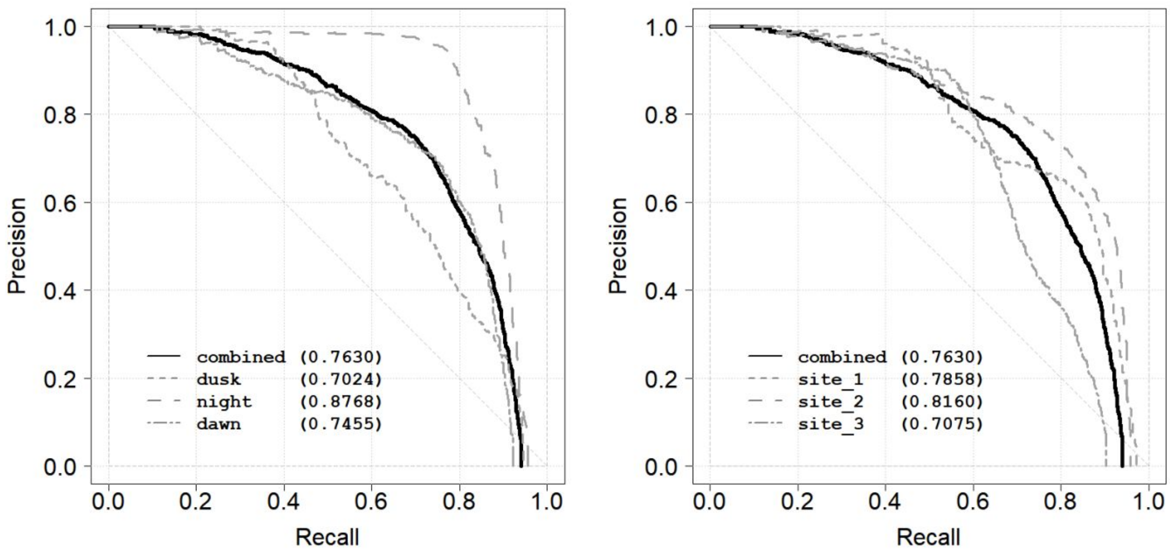

Figure 2. Precision-recall curves calculated using raw recogniser scores, including separate curves for each period (left) and site (right). The figures in brackets give the area under the curve (AUC) for each curve. A higher AUC indicates better recogniser performance.

Identification of potential intrinsic and contextual variables

We next considered what intrinsic and contextual information could be used to assess the likelihood that any given detection was a true positive detection. From the raw recogniser output we extracted the following intrinsic variables for each detection: the score associated with that detection (*score*); which template resulted in the detection (*template*); and, the parent call type of that template (*call_class*). *Score* is the recogniser's most easily interpreted raw output, with obvious predictive value.

A comparison of success rates for different values of *call_class* suggested these could have predictive value. The Night Parrot calls incorporated into this recogniser are generally either short or long. Short single notes are common components of other bird and insect calls occurring in the study area, increasing the probability that templates for short calls will generate false positives. Conversely, longer Night Parrot calls are relatively unique in the study area, meaning their templates are less likely to generate false positives (Table 1).

Table 1. Success rates for different categories of call templates, with recogniser threshold set to zero. Three letter codes represent the different Night Parrot call types incorporated into the recogniser. Short call templates, particularly the '1di' template, generate most false positives. Most of the long call templates perform well.

	Short Calls					**Long Calls**				
	ddt	**too**	**1di**	**2di**	**3nt**	**1tr**	**2tr**	**2wh**	**4wh**	**how**
TRUE POS.	50	287	647	25	5	33	13	567	6	107
FALSE POS.	388	4140	22053	2128	156	46	54	521	138	73

For each detection we clipped a 1.1 second segment of the original file that captured the precise time of that detection, then used R package 'seewave' (Sueur, Aubin, & Simonis, 2008) to calculate the difference between the maximum amplitude and mean amplitude within the frequency range of the template that triggered the detection. Binary point matching compares sound energy within a series of designated 'on' and 'off' cells for each template. Loud sounds within the same frequency range as the binary point template can result in high sound energy flooding both the 'on' and 'off' cells, and if slightly more energy is detected in the 'on' cells this will trigger a detection. Typically though, it will receive a relatively low *score*. We reasoned that if there was a large difference between the maximum and mean amplitude within the template's frequency range, and the detection received only a moderate *score*, this was likely to represent an example of excess sound energy flooding the template, and therefore a false positive. If a large difference in the maximum and mean amplitude within the template's frequency range resulted in a high *score*, the sound energy probably closely matched the 'on' cells of the template, and was more likely to represent a true positive. A plot of amplitude difference (*amp_diff*) against *score* confirmed this relationship (Fig. 3).

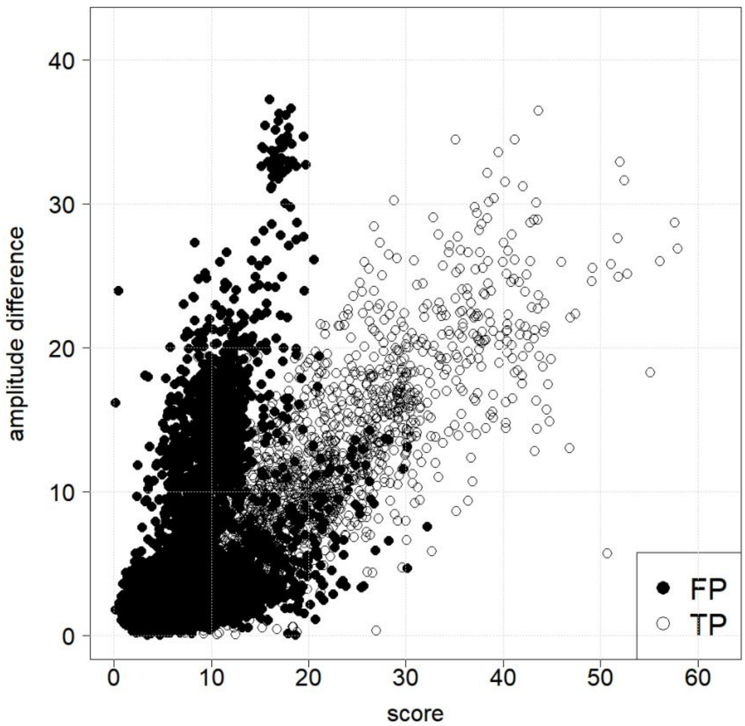

Figure 3. Plot of the relationship between amplitude difference and score for each detection, categorised by detection classification (true positive or false positive). As predicted, detections with a higher amplitude difference but moderate to low score are mostly false positives.

We next considered potential contextual variables. All detections were classified according to which *period* ('dusk', 'night' and 'dawn'), and which *site* they were recorded from ('site_1', 'site_2', 'site_3'). Precision-recall curves were plotted and AUC calculated for each *period* and *site*, then compared to the recogniser's baseline precision-recall curve, to explore their influence on recogniser performance (Fig. 2). Recogniser performance varied between periods, performing best during the night, and most poorly at dusk. This is expected, given the likelihood of false positives is reduced during the night when diurnal birds are not calling. There was no apparent effect of site on recogniser performance. For each detection we also noted which model of ARU (*ARU_type*) and which specific ARU (*machine*) recorded the detection, and in which of the 90 test files (*file*) the detection occurred.

Model development procedure

Our aim was to determine whether a model-derived probability calculated using intrinsic and contextual variables could be substituted for the recogniser's initial *score* value, and achieve better results. We chose a generalised linear mixed-effects model structure, to enable inclusion of both fixed and random effects. As our response variable was binary (true positive or false positive), models were fitted assuming a binomial response distribution, and a logit link function (logistic regression) using the lme4 package (Bates & Sarkar, 2007).

As the practical purpose of this model is to facilitate the process of sifting through recogniser outputs, the process of model building can be more informal than for research purposes that involve *a priori* questions. The approach to selecting the final model was to initially generate a comprehensive set of possible fixed and random effects and compare candidate models containing main effects and interactions for the fixed effect terms, together with the random effects. We then assessed the performance of the candidate models via summary statistics and selected the most promising ones for further development. We determined which variables and variable combinations were critical to those models' performance. Finally, we re-evaluated the refined models before selecting the best performing model as the final model. Model selection was completed using the entire performance dataset.

Fixed and random effects selection

As the aim was to apply the model developed using the performance dataset to any data collected at the study site, we limited fixed effects to those whose complete range of variation was represented in the performance dataset, and which could be determined *a priori* from the resulting raw recogniser output. Factors whose variation was not entirely represented in the performance dataset were included as random effects, and not used in predictions. For example, as *ARU_type* for any data collected at the study site will be either SM3 or SM4, and both were adequately represented in the performance dataset, this could be included as a fixed effect. However, more than 80 individual ARUs have been used at the study site, and only a portion of these were represented in the performance dataset. As this portion represents a random sample from the set of possible ARUs, *machine* (representing the specific ARU used) is included as a random effect. This still allowed the variance associated with this factor to be captured and an allowance made for it in the training phase, but only that level of variance determined during the training phase can be used when the model is applied to future data collected from any machine.

Data exploration revealed interactions were needed between *score* and both *period* and *amp_diff*, so these were initially included as a three-way interaction fixed effect. Because the relationship between a detection's *score* and the probability that the detection is a true positive is curved in the logistic scale, *score* was fitted as a quadratic term. Also included as fixed effects were *call_class* and *ARU_type*. As factors whose level will very likely be new for future datasets, *site*, *file*, and *machine* were all included as random effects. The factor *template* can be established *a priori* from the raw results, but as it contains 31 levels and is nested within *call_class* its predictive power is likely to be limited. However, understanding its impact on model performance may still be important, so it was included as a random effect.

We initially tested a series of 16 models. Each model included all fixed effects, but varied in the combination of random effects. All possible combinations of the four random effects were tested, including a model with no random effects. Models were compared using both Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC). AIC and BIC are statistics for comparing relative model performance, with the primary difference being that BIC penalises more heavily for model complexity (Burnham & Anderson, 2004). Four models stood out as having much lower AIC than the other 12 (Table 2). These four models also had a much lower BIC than the other 12 models. Examining the variance components for each random effect revealed that *file* and *template* were the source of most variation in each of the four best-ranked models, with the contribution of both *machine* and *site* limited (Table 3). Therefore, we retained *file* and *template* as random effects.

We next ran the model including all fixed effects and our chosen random effects, before examining the significance of resulting individual fixed effect coefficients (Table 4). These suggest that the three-way interaction between *period*, *score* and *amp_diff* is not substantially influencing model performance, but that each of the two way interactions between these variables should be retained. *Call_class* has an effect on model performance, but not consistently across classes. Calls that are short have less influence on the model than calls which are long. To investigate this, we created two new variables based on call length. The variable *call_length_1* categorised detections based on the template that detects the call as either short or long, while *call_length_2* categorised all detections based on the template that detects the call as either short, medium, or long. The influence of *ARU_type* is significant, but marginally so.

Table 2. Summary statistics for all random effects models, ranked by AIC. There is strong support for the top four models, warranting further inspection of each component's variation within these models.

Random effects	AIC	BIC	Deviance	log lik.	Resid. df
file + template + site	2520.79	2779.82	2174.11	-1229.40	31406
machine + file + template	2520.96	2779.98	2174.03	-1229.48	31406
machine + file + template + site	2522.75	2790.14	2174.13	-1229.38	31405
file + template	2528.47	2779.14	2172.61	-1234.23	31407
file + site	2716.34	2967.01	2436.64	-1328.17	31407
machine + file	2716.37	2967.04	2436.45	-1328.18	31407
machine + file + site	2718.31	2977.34	2436.58	-1328.16	31406
file	2722.61	2964.93	2434.75	-1332.31	31408
machine + template	2730.59	2981.26	2561.90	-1335.30	31407
machine + template + site	2732.03	2991.06	2561.98	-1335.01	31406
template + site	2740.30	2990.97	2581.41	-1340.15	31407
template	2840.21	3082.53	2699.26	-1391.10	31408
machine	2955.80	3198.11	2873.55	-1448.90	31408
machine + site	2957.47	3208.15	2873.65	-1448.74	31407
site	2965.63	3207.95	2892.61	-1453.82	31408
fixed effects only	3066.66	3300.62	3010.66	-1505.33	31409

We tested a series of nine models, including all possible combinations of the following fixed effects: *score*, *period* and *amp_diff* as either a three-way, or three separate two-way interactions; template category as either *call_class*, *call_length_1* or *call_length_2*; and, with or without *ARU_type*. The random effects for *file* and *template* were retained for all models. The three best models had an AIC value no larger than one unit above the model with the minimum AIC (Table 5). However, the third ranked of these models had a much lower BIC than the other two, with ΔBIC > 30 between this model and the next ranked model by BIC.

Given there was not clear support for one of these three models using AIC, we contend that the best-ranked model using BIC could be considered preferable. We selected this model for use in practice.

Table 3. Variance of each random effects component within each of the top four models used for random effects testing. The contribution of both *machine* and *site* are limited in each case, supporting the decision to retain only *file* and *template* for model simplicity.

file + template + site	
Component	Std dev.
file	1.2177
template	1.2545
site	0.6554

machine + file + template	
Component	Std dev.
file	1.2113
template	1.2492
machine	0.5789

machine + file + template + site	
Component	Std dev.
file	1.2127
template	1.2538
machine	0.2847
site	0.5536

file + template	
Component	Std dev.
file	1.3584
template	1.2222

Model testing

To test the model, we partitioned the performance dataset, using one third of the files, balanced by site and period, to train the model. The remaining files were set aside to test the model. After training, the model was used to predict whether each detection in the test dataset was a true positive. Because we would not know file in advance for a future dataset, this random effect was predicted using the estimate from model training. The predicted probability for each detection was then then substituted for raw recogniser score, and the precision-recall curves replotted (Fig. 4).

Table 4. Significance of the fixed effect coefficients for the model incorporating all fixed effects. Of particular note are the consistent differences between short calls (‘ddt’, ‘1di’, ‘2di’, ‘3nt’, ‘too’) and long calls (‘1tr’, ‘2tr’, ‘2wh’, ‘4wh’, ‘how’).

Fixed effect	Estimate	Std. Error	*z* value	Pr(>|*z*|)
(Intercept)	-7.271	0.702	-10.356	0.000
period ‘dusk’	1.623	0.543	2.986	0.003
period ‘night’	0.970	0.588	1.650	0.099
score^2(1)	133.608	56.883	2.349	0.019
score^2(2)	-494.164	59.590	-8.293	0.000
amp_diff	0.801	0.086	9.258	0.000
ARU_type ‘SM4’	-1.367	0.416	-3.288	0.001
call_class ‘1tr’	5.001	1.469	3.404	0.001
call_class ‘2di’	0.054	0.787	0.068	0.945
call_class ‘2tr’	5.767	1.868	3.088	0.002
call_class ‘2wh’	3.352	0.763	4.391	0.000
call_class ‘3nt’	1.972	1.198	1.645	0.100
call_class ‘4wh’	3.824	1.449	2.638	0.008
call_class ‘ddt’	2.254	1.348	1.673	0.094
call_class ‘how’	5.211	1.323	3.938	0.000
call_class ‘too’	1.597	0.984	1.623	0.105
period ‘dusk’: *score*^{2}(1)	173.151	71.366	2.426	0.015
period ‘night’: *score*^{2}(1)	-63.386	107.565	-0.589	0.556
period ‘dusk’: *score*^{2}(2)	198.939	73.734	2.698	0.007
period ‘night’: *score*^{2}(2)	-214.750	102.485	-2.095	0.036
period ‘dusk’:*amp_diff*	-0.588	0.094	-6.288	0.000
period ‘night’:*amp_diff*	-0.428	0.106	-4.024	0.000
score^2(1):*amp_diff*	-7.443	5.904	-1.261	0.207
score^2(2):amp_diff	34.896	5.888	5.926	0.000
period ‘dusk’: *score*^{2}(1):*amp_diff*	9.761	7.708	1.266	0.205
period ‘night’: *score*^{2}(1):*amp_diff*	27.259	12.111	2.251	0.024
period ‘dusk’: *score*^{2}(2):*amp_diff*	-8.322	8.106	-1.027	0.305
period ‘night’: *score*^{2}(2):*amp_diff*	5.248	11.050	0.475	0.635

Table 5. Summary statistics for the final set of nine models. Only fixed effects for each model are shown; the random effects for each model were file and template. There is strong support for each of the top three models by AIC, but the third of these (in bold) has much stronger support by BIC and was selected as the final model.

Fixed effects	AIC	BIC	Deviance	log lik.	Resid. df
period * score^2 * amp_diff + call_length_1 + ARU_type	2522.12	2705.94	2171.67	-1239.06	31415
period * score^2 * amp_diff + call_length_2 + ARU_type	2522.60	2714.78	2169.74	-1238.30	31414
period * score^2 + score^2 * amp_diff + period * amp_diff + call_length_1 + ARU_type	2522.84	2673.25	2180.75	-1243.42	31419
period * score^2 + score^2 * amp_diff + period * amp_diff + call_length_2 + ARU_type	2524.54	2683.30	2179.03	-1243.27	31418
period * score^2 * amp_diff + call_class + ARU_type	2528.47	2779.14	2172.61	-1234.23	31407
period * score^2 + score^2 * amp_diff + period * amp_diff + call_class + ARU_type	2529.69	2746.94	2182.30	-1238.84	31411
period * score^2 + score^2 * amp_diff + period * amp_diff + call_length_1	2532.23	2674.28	2179.28	-1249.12	31420
period * score^2 + score^2 * amp_diff + period * amp_diff + call_length_2	2533.67	2684.07	2177.83	-1248.83	31419
period * score^2 + score^2 * amp_diff + period * amp_diff + call_class	2538.85	2747.74	2181.04	-1244.42	31412

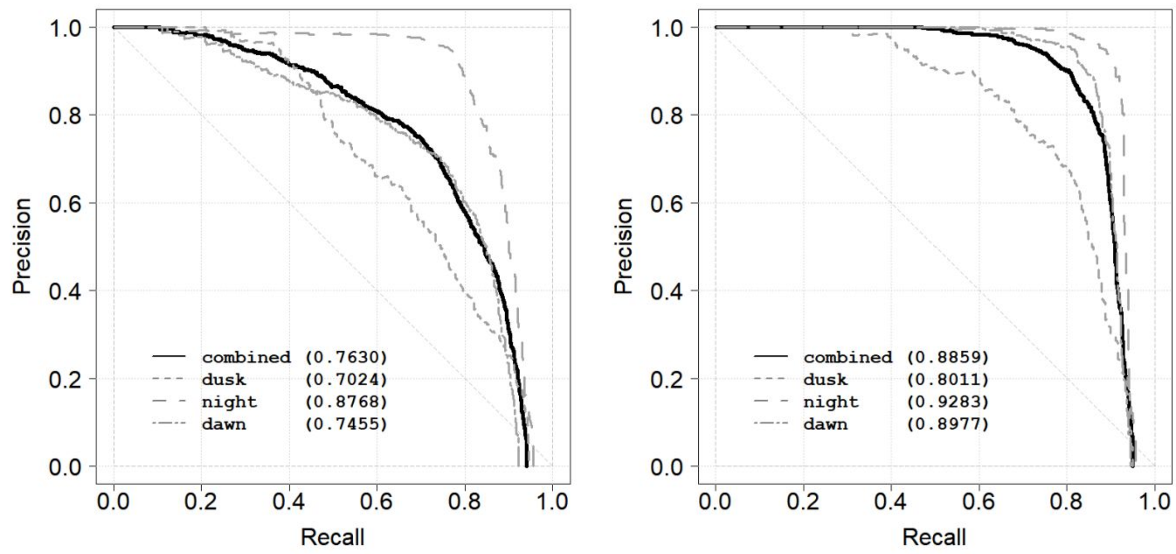

Figure 4. Precision recall curves calculated for each period using raw recogniser scores (left), and model-derived probabilities (right). When using model-derived probabilities, the increase in AUC is evident overall, and across all periods, meaning this approach improves recogniser performance.

The precision-recall curves for the combined data, and for each *period*, demonstrate that substituting model-derived probability for raw score results in an increased AUC overall (AUC = 0.89 for model-derived probability, and AUC = 0.76 for raw score), meaning overall recogniser performance is improved. As expected, this improvement is modest for the night *period*, but marked for both the dusk and dawn *period*, with AUC improving by 0.10 and 0.15 respectively.

To quantify the practical improvements resulting from this modelling procedure, we investigated the number of detections requiring proofing to achieve a specific level of recall. Recall is of particular importance because the recall of a recogniser equals the probability that a species will be detected if it is available for detection, an important component of the overall probability of detection (Pollock et al., 2004). Furthermore, it is important for rare species research because prioritising recall maximises the likelihood of detecting the species if it is available in the acoustic dataset. This emphasis on recall manifests itself in the increased number of detections that require proofing to achieve the increased level of recall.

We calculated the mean number of false positive detections requiring proofing per 10-minute file in the test dataset to achieve a specific recall; a proxy for the amount of time an analyst needs to spend proofing recogniser output. We initially calculated the score cut-off that achieved a specified recall for both raw score, and for the model-derived probability. Because model-derived probability incorporates *period* as a fixed effect in the calculation, cut-off scores for a specific value of recall under the model-derived probability may vary between periods. Accordingly, the model-derived probability cut-off for each recall threshold was calculated separately for each *period* using only the test dataset to avoid overfitting. Using these data, we also simulated for both raw score and model-derived probability, how many false positive detections would need to be checked during a complete 12 hour night of acoustic data, with one hour of 'dusk', ten hours of 'night' and one hour of 'dawn' recordings to be assessed.

The model-derived probability markedly reduced the number of false positives that needed checking to achieve each level of recall tested (Table 6). This improvement is most pronounced during the night period, and at lower levels of recall. However, even at 90% recall, if using the model-derived probability as a substitute for score, the number of false positives that would need checking during an entire night of acoustic data is 40% of what would need to be checked if using the raw score.

Table 6. The mean number of false positives requiring proofing in a 10-minute recording for a set level of recall, using either raw recogniser score (Score), or the model-derived probability (MDP). The final three columns present the number of false positives that would need proofing if analysing a 12-hour night of recordings, with the '%' column representing the percentage of proofing, and therefore time required when using model-derived probability compared to raw score.

	Dusk		Night		Dawn		12-hour Night		
Recall	Score	MDP	Score	MDP	Score	MDP	Score	MDP	%
0.50	2.80	0.85	0.20	0.00	0.70	0.00	30.6	5.1	17
0.55	4.30	1.10	0.25	0.05	1.00	0.00	43.8	9.0	21
0.60	6.20	1.30	0.25	0.05	1.55	0.00	58.5	10.2	17
0.65	7.55	2.15	0.25	0.05	2.00	0.10	69.3	15.9	23
0.70	9.80	3.30	0.40	0.05	2.60	0.35	93.6	24.3	26
0.75	15.30	4.45	0.50	0.05	3.65	0.55	137.7	32.4	24
0.80	22.25	6.55	0.70	0.25	5.80	0.85	201.9	56.4	28
0.85	29.30	13.70	1.85	0.55	9.70	2.35	322.8	122.7	38
0.90	45.35	34.95	9.05	1.35	22.25	10.10	840.0	335.1	40

Discussion

The method we have outlined demonstrates that intrinsic and contextual information associated with a call recogniser's output can be used to improve the performance of that recogniser. This approach is compatible with any signal detection algorithm, not just binary point matching as is the case here. While the improvements are revealed through the AUC of the precision-recall curve, this representation is somewhat abstract. The practical benefits of this approach are more clearly demonstrated in the reduced effort required to achieve a specific recall. For practitioners using call recognisers to analyse large quantities of field recordings, the limiting factor is typically time, which manifests itself as the number of detections that can be manually proofed. However, while this technique does result in efficiencies, there are limitations.

Raw recogniser performance and improvement

These improvements will only apply to detections within the recogniser's output; it does not change the recogniser's ability to detect false negatives. False negatives occur for two reasons. The recogniser may detect some other signal that occurs concurrently with the call of interest and achieves a higher score, meaning the call of interest is missed. Such events are difficult to overcome. Alternatively, a call of interest may not match the training data. Post-processing techniques, as outlined here, will not improve recogniser performance in that respect. This can only be overcome by updating the recogniser's training dataset to improve the probability the recogniser will detect that missed call. If new templates are added to the recogniser, the model selection process will need to be rerun, with sufficient training and test files added to model the impact of the new templates.

Model application for different species and new sites

Even though calls used to create this recogniser's templates were excluded from the training and test datasets, because the Night Parrot population at the study site is very small, it is likely calls from the same individuals were incorporated into the training and test datasets. There is a resultant risk of model overfitting. Additionally, the repertoire of this population is well-known (Leseberg et al., 2019), and the recogniser templates featured most of the variation that occurs at the study site. It is possible this combination of factors has exaggerated the success of our model. In scenarios where the subject species does not have such a consistent repertoire, because it has a larger number of individuals, a more dynamic population, or greater variation in its calls, this technique will still be applicable provided this variation is incorporated into the training and test datasets.

The properties of the general soundscape, including likely non-target calls that occur in the dataset will also influence model applicability. For example, the model developed here could be reasonably applied to other datasets from western Queensland, where Night Parrots are known to have similar calls to those in this dataset (NL pers. obs.), and where the suite of likely non-target species will also be similar. However, the model may not be as effective if applied to a dataset from Western Australia, where the suite of Night Parrot and likely non-target species are slightly different to western Queensland. Testing on an annotated dataset would determine if the model does improve recogniser performance and by how much. Otherwise, the model selection and training process would need to be rerun using a performance dataset compiled from the new region of interest.

Impact of model treatment of different call types

The fixed effect *call_length_1* boosts the model-derived probability for longer calls, when compared to shorter calls. In a scenario where shorter calls predominate at a site, this may affect the recogniser's ability to detect birds at that site. It is likely that faint short calls are most affected. Because an ARU established at a prospective long-term stable roost site will record a variety of short calls over time, the probability of at least some calls being detected by the recogniser is high. Additionally, over long periods at long-term stable roost sites, there is typically a mix of long and short calls (SM, NL unpub. data), ensuring the recogniser will detect birds if they are present. This may still be an issue if a short deployment limits the variety of calls that occur within the dataset.

An additional consequence of the differing treatment of call types by the model will be the distortion of potential distance effects. Researchers can extract distance information from acoustic data, using signal strength, or variables closely related to signal strength such as the call recogniser's raw score, as a proxy for distance from the recorder (Knight & Bayne, 2018; Lambert & McDonald, 2014). This information is then used in distance-sampling procedures, or for establishing survey effort parameters (Yip, Leston, Bayne, Solymos, & Grover, 2017). The mechanics of this modelling technique will confound any attempts to use the model-derived probabilities as a proxy for distance, because they are influenced by factors other than signal strength, whereas raw score is typically heavily dependent on signal strength (Knight & Bayne, 2018). For example, if ranked by model-derived probability, a faint long call is likely to rank higher than if it were ranked by raw score alone. If model-derived probability is being used as a proxy for distance from the recorder, this would be equivalent to the call being made closer to the recorder, an incorrect assumption that could distort conclusions around that call's likely distance from the recorder.

Depending on the aim of the distance-sampling approach, this issue could be overcome in several ways, although each has limitations. Research could assess the relationship between model outputs and distance, although this is likely to vary across call types, and for a species like the Night Parrot would require a test dataset that would be almost impossible to collect. Alternatively, signal strength or raw score for a given detection could be extracted after model application to determine distance data, although this will mean the calls extracted will be influenced by the model. Again, long calls are more likely to be extracted than short calls, possibly interfering with subsequent conclusions. A final option could be to first sort data by raw score, before applying the model to the subset of data whose raw score satisfies the distance sampling criteria.

Other parameters with potential predictive power

The modelling approach applied here was successful using a relatively limited number of parameters, some that were particular to the subject species' biology, such as *call_length_1* and *period*, while others were generic, such as *amp_diff*, *ARU_type* and the random effects *template* and *file*. It is likely that a number of other parameters could be incorporated to further improve results. As Night Parrots call more frequently in response to local rain events (Murphy, Austin, et al., 2017), a variable quantifying antecedent rainfall could be an obvious inclusion. An emerging question in Night Parrot research is the merit of acoustic surveys at water points and likely feeding sites, compared to current protocols that focus solely on roosting habitat. If autecological research determines a consistent pattern of nocturnal activity, *site resource* (i.e. water point, feeding site, roosting site) could be included as a fixed effect in the model.

The predictable calling behaviour and site fidelity of the Night Parrot make it particularly suited to the approach we have outlined here, but with careful consideration, it will be applicable in other scenarios. Intrinsic variables related to raw recogniser output can be developed that are species specific, as call type was here, or recogniser specific, as *amp_diff* was in this case, being relevant specifically to the binary point matching technique used in this recogniser. There are likely to be similar variables that could be developed for the numerous other recogniser algorithms. Improvements to the raw output for more advanced algorithms may not be as significant as for the relatively basic binary point matching, but for field ecologists, any reduction in the time required to proof recogniser returns will be beneficial. The contextual variables that could be trialled will relate to a species' biology and might include long-term seasonal and short-term weather effects, habitat or other environmental parameters at both the local and landscape scale, and calling biology. The number of contextual parameters that could be tested is limited only by a researcher's ability to compile a performance testing dataset that satisfactorily represents the variation in each parameter.

This technique's biggest advantages are its simplicity, and compatibility with any recognition algorithm. For the ecologist or practitioner, call recogniser development is daunting, with high performing recognisers generally built using state-of-the-art techniques that in many cases require advanced programming skills and research time. The foundation of the post-processing technique we outline in this paper is a relatively straightforward procedure that can be completed using graduate level statistics. For that reason, it will be of particular use to practicing field ecologists looking to improve a simple recogniser, which may only be one part of a broader research project. It may also be applied to any state-of-the-art recognition algorithm to further improve results.

Acknowledgements

The authors would like to acknowledge the Maiawali people, on whose land the research for this paper was conducted. Bush Heritage Australia, Fortescue Metals Group, and the Australian Government's National Environmental Science Program through the Threatened Species Recovery Hub provided support for this research. NL received an Australian Government Research Training Program (RTP) Scholarship, and additional support through the Max Day Environmental Fellowship, University of Queensland strategic funding and Birds Queensland.

Authors' contributions

NL, SM, WV, and JW conceived the ideas and designed the methodology; NL, SM and JW collected the data; NL and WV analysed the data; NL led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data availability

We intend to make the outputs of the recogniser, and the code used to create and apply our model available via Github.

References

Bates, D. M., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference:Understanding AIC and BIC in Model Selection. *Sociological Methods & Research, 33*(2), 261-304. doi:10.1177/0049124104268644

Digby, A., Towsey, M., Bell, B. D., & Teal, P. D. (2013). A practical comparison of manual and autonomous methods for acoustic monitoring. *Methods in Ecology and Evolution, 4*, 675-683. doi:10.1111/2041-210X.12060

Higgins, P. J. (1999). *Night Parrot (Pezoporus occidentalis)* (Vol. 4, Parrots to Dollarbird). South Melbourne: Oxford University Press.

Horton, K. G., Stepanian, P. M., Wainwright, C. E., & Tegeler, A. K. (2015). Influence of atmospheric properties on detection of wood-warbler nocturnal flight calls. *International Journal of Biometeorology, 59*, 1385-1394. doi:10.1007/s00484-014-0948-8

Joshi, K. A., Mulder, R. A., & Rowe, K. M. (2017). Comparing manual and automated species recognition in the detection of four common south-east Australian forest birds from digital field recordings. *Emu, 117*(3), 233-246. doi:10.1371/journal.pone.0199396

Katz, J., Hafner, S. D., & Donovan, T. (2016). Tools for automated acoustic monitoring within the R package monitoR. *Bioacoustics, 25*(2), 197-210. doi:10.1080/09524622.2016.1138415

Knight, E. C., & Bayne, E. M. (2018). Classification threshold and training data affect the quality and utility of focal species data processed with automated audio-recognition software. *Bioacoustics*. doi:10.1080/09524622.2018.1503971

Knight, E. C., Hannah, K. C., Foley, G. J., Scott, C. D., Brigham, R. M., & Bayne, E. (2017). Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. *Avian Conservation and Ecology, 12*(2). doi:10.5751/ACE-01114-120214

Koops, H. V., van Balen, J., & Wiering, F. (2014). A deep Neural Network Approach to the LifeCLEF 2014 Bird task. *CEUR Workshop Proceedings, 1180*, 634-642.

Lambert, K. T. A., & McDonald, P. G. (2014). A low-cost, yet simple and highly repeatable system for acoustically surveying cryptic species. *Austral Ecology, 39*, 779-785. doi:10.1111/aec.12143

Landau, H. J. (1967). Sampling, data transmission, and the Nyquist rate. *Proceedings of the IEEE, 55*(10), 1701-1706. doi:10.1109/PROC.1967.5962

Leseberg, N. P., Murphy, S. A., Jackett, N. A., Greatwich, B. R., Brown, J., Hamilton, N., . . . Watson, J. E. M. (2019). Descriptions of known vocalisations of the Night Parrot *Pezoporus occidentalis*. *Australian Field Ornithology, 36*, 79-88. doi:10.20938/afo36079088

Magurran, A. E., Baillie, S. R., Buckland, S. T., Dick, J. P., Elston, D. A., Scott, E. M., . . . Watt, A. D. (2010). Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends in Ecology and Evolution, 25*(10), 574-582. doi:10.1016/j.tree.2010.06.016

Murphy, S. A., Austin, J. J., Murphy, R. K., Silcock, J., Joseph, L., Garnett, S. T., . . . Burbidge, A. H. (2017). Observations on breeding Night Parrots (*Pezoporus occidentalis*) in western Queensland. *Emu, 117*(2), 107-113. doi:10.1080/01584197.2017.1292404

Murphy, S. A., Silcock, J., Murphy, R. K., Reid, J. R. W., & Austin, J. J. (2017). Movements and habitat use of the night parrot *Pezoporus occidentalis* in south - western Queensland. *Austral Ecology, 42*(7), 858-868. doi:10.1111/aec.12508

Pollock, K. H., Marsh, H., Bailey, L. L., Farnsworth, G. L., Simons, T. R., & Alldredge, M. W. (2004). Separating Components of Detection Probability in Abundance Estimation: An Overview with Diverse Examples. In W. L. Thompson (Ed.), *Sampling rare or elusive species: concepts, designs, and techniques for estimating population parameters*. Washington, USA: Island Press.

Potamitis, I., Ntalampiras, S., Jahn, O., & Riede, K. (2014). Automatic bird sound detection in long real-field recordings: Applications and tools. *Applied Acoustics, 80*, 1-9. doi:10.1016/j.apacoust.2014.01.001

Priyadarshani, N., Marsland, S., & Castro, I. (2018). Automated birdsong recognition in complex acoustic environments: a review. *Journal of Avian Biology, 49*(5). doi:10.1111/jav.01447

Priyadarshani, N., Marsland, S., Juodakis, J., Castro, I., & Listanti, V. (2020). Wavelet filters for automated recognition of birdsong in long-time field recordings. *Methods in Ecology and Evolution, 11*, 403-417. doi:10.1111/2041-210X.13357

R Core Team. (2018). R: A Language and Environment for Statistical Computing: R Foundation for Statistical Computing, Vienna.

Salamon, J., & Bello, J. P. (2017). Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters, 24*(3), 279-283. doi:10.1109/LSP.2017.2657381

Shonfield, J., & Bayne, E. M. (2017). Autonomous recording units in avian ecological research: current use and future applications. *Avian Conservation and Ecology, 12*(1). doi:10.5751/ACE-00974-120114

Sueur, J., Aubin, T., & Simonis, C. (2008). Seewave: A free modular tool for sound analysis and synthesis. *Bioacoustics, 18*, 213-226. doi:10.1080/09524622.2008.9753600

Teixeira, D., Maron, M., & van Rensburg, B. J. (2019). Bioacoustic monitoring of animal vocal behaviour for conservation. *Conservation Science and Practice*. doi:10.1111/csp2.72

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. *Ecological Informatics, 21*, 110-119. doi:https://doi.org/10.1016/j.ecoinf.2013.11.007

Towsey, M., Znidersic, E., Broken-Brow, J., Indraswari, K., Watson, D. M., Phillips, Y., . . . Roe, P. (2018). Long-duration, false-colour spectrograms for detecting species in large audio data-sets. *Journal of Ecoacoustics, 2*. doi:10.22261/JEA.IUSWUI

Williams, E. M., O'Donnell, C. F. J., & Armstrong, D. P. (2018). Cost-benefit analysis of acoustic recorders as a solution to sampling challenges experienced monitoring cryptic species. *Ecology and Evolution, 8*, 6839-6848. doi:10.1002/ece3.4199

Yip, D. A., Leston, L., Bayne, E. M., Solymos, P., & Grover, A. (2017). Experimentally derived detection distances from audio recordings and human observers enable integrated analysis of point count data. *Avian Conservation and Ecology, 12*(1). doi:10.5751/ACE-00997-120111