Evaluating and optimising performance of multispecies call recognisers for ecoacoustic restoration monitoring

Simon Linke[1], Daniella Teixeira[2], and Katie Turlington[3]

[1]CSIRO Environment
[2]QUT
[3]Griffith University

April 12, 2023

Abstract

Monitoring the effect of ecosystem restoration can be difficult and time consuming. Autonomous sensors, such as acoustic recorders, can aid monitoring across long time scales. This project successfully developed, tested and implemented call recognisers for eight species of frog in the Murray-Darling Basin. Recognisers for all but one species performed well and substantially better than many species recognisers reported in the literature. We achieved this through a comprehensive development phase, which carefully considered and refined the representativeness of training data, as well as the construction (amplitude cut-off) and the similarity thresholds (score cut-offs) of each call template used. Recogniser performance was high for almost all species examined. Recognisers for C. signifera, L. fletcherii, L. dumerilii, L. peronii, and C. parinsignifera all performed well, with most templates having ROC values (the proportion of true positive and true negatives) over 0.7, and some much higher. Recognisers for L. peronii, L. fletcherii and L. dumerilii performed particularly well in the training dataset, which allowed for responses to environmental watering events, a restoration activity, to be clearly observed. While slightly more involved than building recognisers using commercial packages, the workflows ensure that a high quality recogniser can be built and the performance fine-tuned using multiple parameters. Using the same framework, recognisers can be improved in future iterations. We believe that multi-species recognisers are a highly effective and precise way to detect the effects of ecosystem restoration.

Introduction

Bioacoustics – the study of sound production, dispersion and reception in animals – has been practiced for millennia. Even in underwater systems, Aristotle had described communication between animals in great anatomic and behavioural detail (Aristotle, 1910; Linke et al., 2020). Bioacoustics can be used to study animal ecology – for example, reproductive behaviour and success (Teixeira et al., 2019) - to monitor population dynamics of native or invasive species (Brodie et al., 2020b), and to detect rare and endangered soniferous animals (Dema et al., 2020; Dutilleux and Curé, 2020; Znidersic et al., 2020). The sister discipline ecoacoustics is a new field that is not restricted to biotic organisms, but – like ecology to biology – investigates biodiversity, its relation to habitats as well as populations and ecological communities (Sueur and Farina, 2015).

Ecoacoustics has been used to quantify ecological responses to environmental restoration or improvement in condition, providing a rapid and continuous monitoring framework that can detect both degradation and restoration success (Greenhalgh et al., 2021; Linke and Deretic, 2020; Znidersic and Watson, 2022). Often, acoustic indices are used in assessments. These indices are analogous to measurements of diversity or richness in classical ecology – they summarise the acoustic properties of an overall soundscape, for example its spatial, temporal or combined complexity, its overall volume or the relation between natural and human-influenced frequency bands (Sueur et al., 2014). However, given inherent variations in soundscapes between places,

ecoacoustic indices must be calibrated by ecosystem. While some authors have described clear variation along landscape gradients (Ng et al., 2018), others have found little relation of acoustic indices to human disturbance (Mitchell et al., 2020). Other studies have found that acoustic indices can be dominated by single acoustic events, for example river flows (Linke and Deretic, 2020) or single species that dominate the soundscape, such as snapping shrimp (Bohnenstiehl et al., 2018).

To examine restoration of wetlands in Australia's most highly regulated river system, the Murray-Darling, Linke and Deretic (2020) used both manual annotation and ecoacoustic indices to track recovery of amphibian and waterbird populations. The Murray-Darling system currently flows at only ˜40% of its natural capacity, with the bulk of the extracted water used for irrigated agriculture (Grafton et al., 2014). Under a federal government initiative - the "Murray-Darling Basin Plan" - water is being returned to rivers and wetlands via water buybacks from irrigators, however quantifying ecological recovery over the long term is difficult (King et al., 2015; Souchon et al., 2008). Linke and Deretic (2020) pioneered the use of ecoacoustic analysis as a tool to continuously monitor populations after restorative water returns to wetlands. When manually listening to recordings of frog and bird calls, they found highly significant responses in richness of water-dependent biota to environmental watering. However, the response of acoustic indices was much weaker, and in some cases non-significant, partially obfuscated by ambient noises, and also subject to high diurnal variation. This led the authors to conclude that a logical next step was to trial multi-species call recognisers that would combine the advantage of species specificity with the automated processing of acoustic indices (Linke and Deretic, 2020).

Call recognisers usually function to detect single species, since bioacoustics is often used to detect cryptic or rare animals. However, as the application of acoustics to environmental monitoring increases, multi-species recognisers are likely to become more important. Multi-species recognisers detect sympatric species simultaneously (Wright et al., 2020; Zhong et al., 2020), and outputs can be analysed for species separately or combined. This is useful where groups of species (e.g. mixed species frog choruses) represent environmental change or other ecological values. Like single-species recognisers, multi-species recognisers can use acoustic indices to detect soundscapes in which target species are likely to occur (Brodie et al., 2020a), or they can implement several single-species algorithms to detect discrete calls (Ruff et al., 2020). There are many challenges to creating reliable multi-species recognisers, however methods for reducing the increased risk of false detections are beginning to be examined (Campos et al., 2019; Wright et al., 2020).

Performance metrics used to evaluate and report on call recogniser performance are highly variable in the literature (Knight et al., 2017). Terminology is inconsistent and studies may report only a small number of possible performance metrics. This makes comparisons and repeatability difficult. Perhaps more importantly, there are major inconsistencies in type and amount of training data used and the test datasets upon which recognisers are evaluated. While strictly standardised methods are unlikely be feasible (e.g. for rare species, datasets can be extremely difficult to acquire), studies should, as a minimum, report the representativeness of the training data, how these were chosen or tested, and any limitations or assumptions. Decisions relating to, for example, geographical coverage may have important consequences for recogniser performance and transferability among regions. Moreover, the extent to which training and test data include real-world ambient noise should be explained, because factors like wind, noise and other species' calls can significantly impact false detections (Brandes, 2008; Cragg et al., 2015; Crump and Houlahan, 2017; Kahl et al., 2021; Knight et al., 2017; Priyadarshani et al., 2018; Salamon et al., 2016; Towsey et al., 2012). To standardise the reporting of performance metrics, Knight et al. (2017) recommended all studies report precision, recall, F-score and area under the precision-recall curve (AUC) or, for comparison with the broader classifier literature, receiver operating characteristics (ROC) curve.

Using a template-matching algorithm (binary point matching, Towsey et al., 2012) in the R package monitor (Katz et al., 2016b) , we aimed to establish a free and open source protocol to optimise multi-species call recogniser construction and evaluation using three levers: template selection, amplitude cut-off and score-cut-off.

- First, we tested the performance of geographically-representative candidate call templates (training

data) and, from this, selected a small number of high-performing templates from which to construct call recognisers.
- Second, we examined call templates at a range of amplitude cut-offs, which alters their detection sensitivity.
- Third, we tested templates across a wide range of score cut-offs, which defines the threshold of similarity between templates and sound data at which a detection is returned.

As a case study, we tested this protocol on the calls of eight sympatric frog species from the Koondrook-Pericoota wetland complex in the Murray-Darling Basin.

Methods

Overview and computational strategy

To build and evaluate the recognisers for eight target frog species, we used a large database of annotated calls from the study area in Koondrook-Perricoota (KP) forest. The Forestry Corporation of NSW– the body commissioning the study – had previously annotated 831 files from 20 sites and found varying levels of presence for the different species. From these files, we extracted between 100 to 200 reference calls per species, from all sites where the species was detected. Following manual inspections for call clarity and variation, we used these reference calls to build approximately 20 candidate recognisers per species (i.e., one template equals one recogniser per species). Some recognisers were based on templates of the same reference call, but were created using different amplification settings, which is modifiable in monitoR. We then ran the recognisers on the pre-annotated files to calibrate the score cut-off (similarity threshold between reference call templates and the sound files) and estimate omission and commission errors. We then chose the final recognisers based on the best ROC (Receiver Operator Criterion, Zou et al., 2007), thus minimising both false positive and false negative detections.

Study area and data collection

The Forestry Corporation of NSW provided two years of acoustic data files recorded between February 2018 and February 2020. A SongMeter 3 or SongMeter 4 (Wildlife Acoustics Inc.) sound recorder was deployed at each of the 20 study sites (see Appendix 1) in the KP Forest. Prior to January 2019, each recorder recorded five minutes of audio per day at dusk. The start time of each recording changed according to the time of sunset that day. From the acoustic data provided by the Forestry Corporation of NSW, eight frog species were identified as potential indicators of ecological health. A list of previously annotated detections (i.e. times, dates, and locations that these frogs had been detected via manual listening) was also provided by the Forestry Corporation of NSW (see Table 1 for the number of files where the candidate frog species were present in annotated files).

Table 1 Count of annotated evaluation files where calls of candidate frog species were present

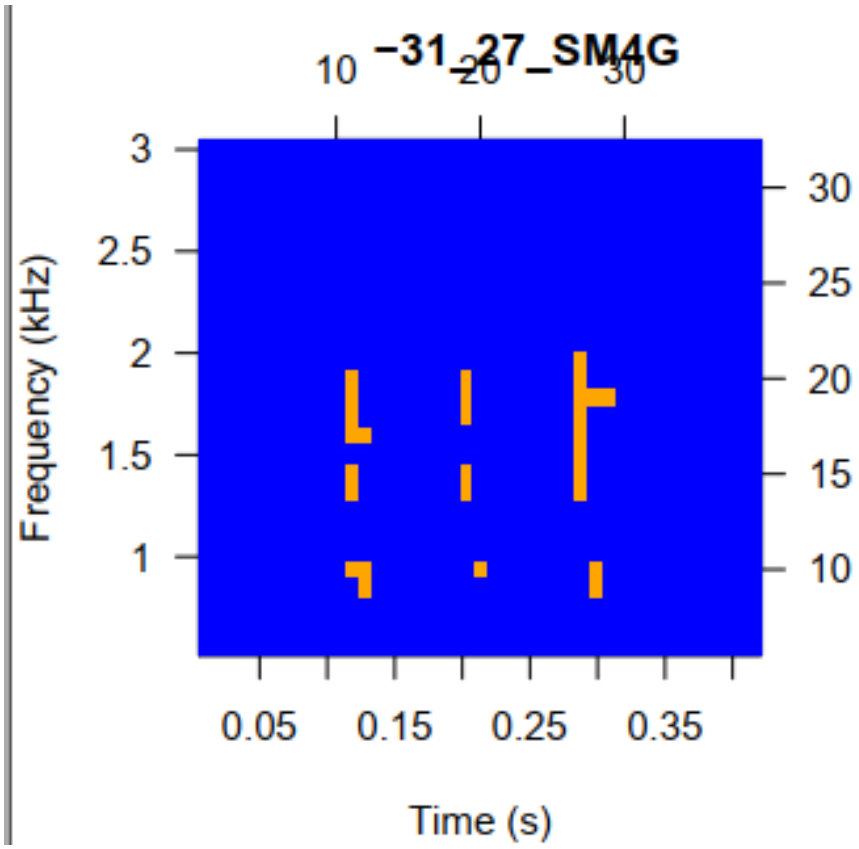
Species	**Evaluation files (n)**
Crinia signifera	300
Limnodynastes tasmaniensis	146
Limnodynastes fletcherii	78
Limnodynastes dumerilii	74
Litoria peronii	118
Crinia parinsignifera	298
Neobatrachus sudelli	62
Litoria raniformis	24

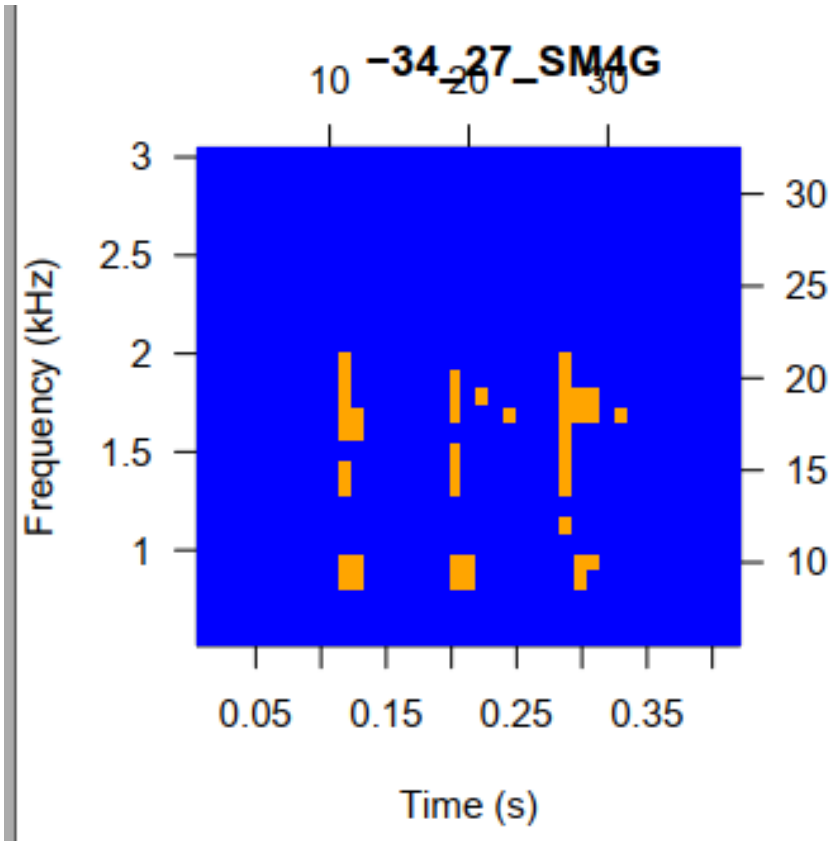
To build a training dataset of calls for recogniser development, we first manually selected and extracted a minimum of 100 reference calls for each species using Adobe Audition CC and Raven Pro 1.5 software. To maximise representativeness, we (a) selected calls from as many surveys sites as possible, to capture geo-

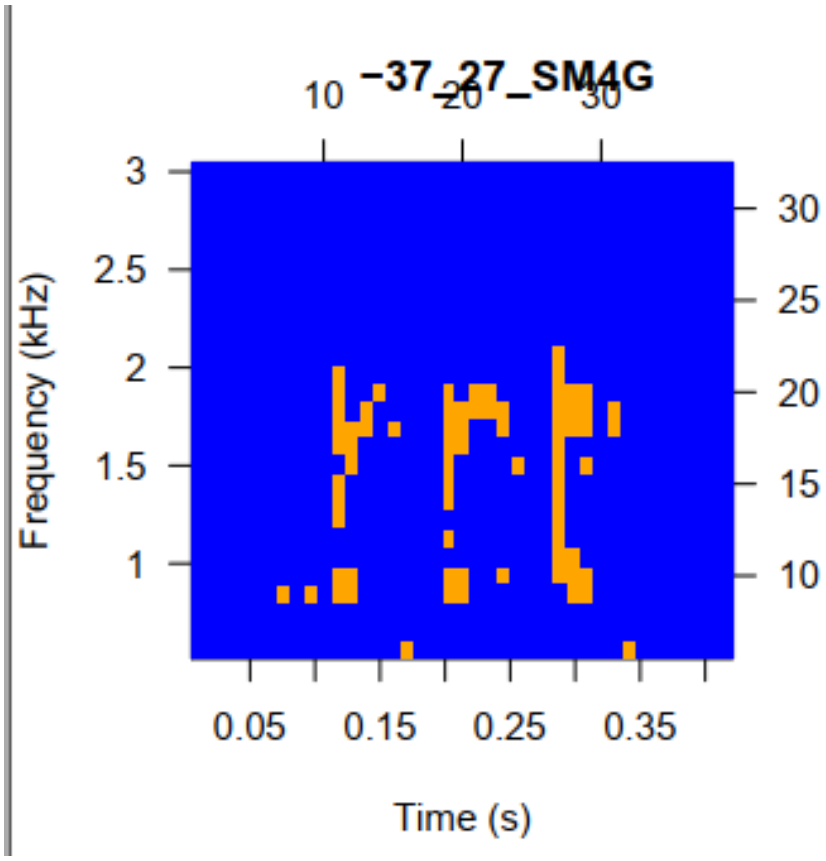
graphical variation, and (b) selected calls of varying quality and amplitude, to capture soundscape variation. These are important steps to improve the similarity between call templates and 'real-world' sound data. Building a recogniser solely from calls that are loud and clear would perform poorly if the species' calls are rarely loud and clear in field recordings. Given the complexity of frog choruses, variations in ambient noise and differences in amplitude among calls (e.g. from variations in the distance of the frog from the sound recorder), capturing diversity in call templates is a critical component of recogniser construction.

Recogniser construction

To construct the recognisers, we used the technique 'binary template matching' – a technique that first converts a spectrogram into a binary template and then matches 'on' and 'off' points of the template to the file the recogniser is applied to (Katz et al., 2016b; Towsey et al., 2012). This was done using the monitoR package in R, which provides flexibility in its construction parametres (Katz et al., 2016b). To build the initial recognisers, templates were constructed from a minimum of 10 reference calls (from the pool of 100 candidate calls) that were both clear and representative of the variation in calls and environmental conditions. Binary templates comprise 'on' and 'off' regions (call and non-call), which are based on a user-defined amplitude cut-off (Figure 1). Each template's amplitude cut-off was determined through manual inspection of templates using the makeBinTemplate function in monitoR (Katz et al., 2016b). Amplitude cut-offs were set arbitrarily and progressively altered and reviewed. A cut-off that clearly showed call structure and was not masked by background noise, was deemed appropriate, with some background noise deemed acceptable.







a)

Figure 1 Example of amplitude cutoffs of a) -37dB, b) -34dB, c) -31 dB for a single *Crinia signifera* template. Orange indicates "on" (call) regions and blue indicates "off" (non-call) regions.

The lower and upper bounds of the template's frequency limits were manually chosen to capture as much of the call as possible, while minimising potential overlap with common noise sources (e.g. crickets). Most templates were constructed with a window size of 512 samples. Templates for *C. signifera* used a window size of 256 samples to improve the resolution of the species' highly pulsatile call. Both the frequency limit and the window size affect the number of on and off points in the template and, therefore, the template's speed. For*L. tasmaniensis,* we trialled templates with both window sizes, but chose to use the 512 sample templates as they showed the call's structure more clearly.

Recogniser evaluation and score cut-off

For each template we tested a range of score cut-offs, which is a user-defined similarity threshold at which a template will return a detection. This threshold alters the proportion of true positive and false positive detections and is, therefore, an important part of optimising call recognisers (Katz et al., 2016a). For each species, we tested the recognisers at a low score cut-off of 3; thus, any call instance that scored 3 or higher was returned as a detection by monitoR. Optimal score cut-off for each template was determined by constructing receiver operating characteristic (ROC) curves, a diagnostic tool that optimises that trade-off between false positive and true positive rates (Figure 2). We calculated true and false positive rates at score cut-off increments of 0.2 and determined the optimum as the score cut-off where true positives were greatest relative to false positives (i.e. the peaks in Figure 2). We then retained these score cut-offs for the recogniser evaluation.

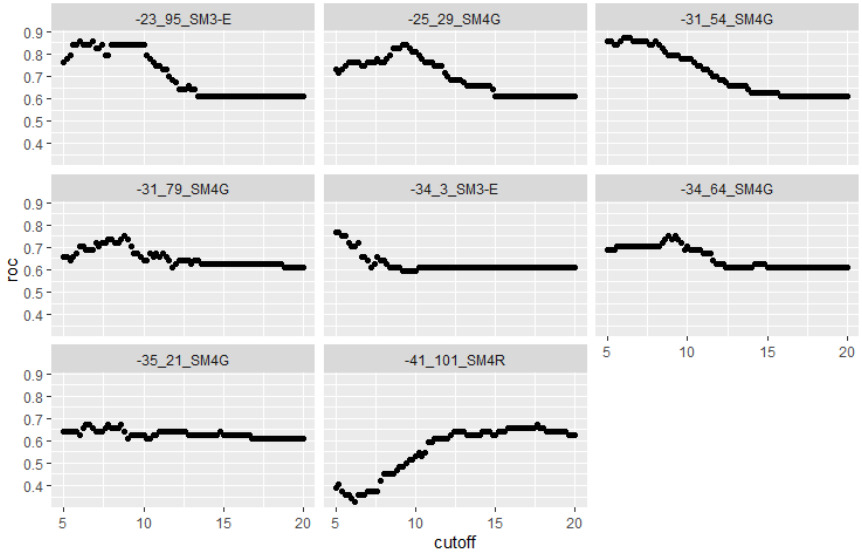


Figure 2 ROC plots for different templates to detect *Limnodynastes fletcherii* . Templates 1, 2, and 3 were chosen with the score cut-off corresponding to the highest ROC.

Recogniser evaluation and detection of false positives

Recogniser performance was evaluated on a manually verified and balanced subsample of 1-minute sound recordings that were categorised for each species' presence or absence. The sample size of evaluation files varied among species. Any detection returned in a recording where the species was present was taken to be a true positive detection. All other detections were deemed false positives. For each template, we quantified the number of call detections in sound files where the species was present (true positive count; Count TP) and absent (false positive count; Count FP) and the number of sound files in which the presence/absence (PA) of species was correctly detected (true presence: PA TP), incorrectly detected (false presence; PA FP), missed (false absence; PA FN) or correctly undetected (true absence; PA TN). Using these values, for each template we calculated precision, recall and ROC value, which are given by the formulas:

Precision = True Positives / (True Positives + False Positives)

Recall = True Positives / (True Positives + False Negatives)

ROC = (True Positives + True Negatives) / Number of evaluation files

Sample detections of false positives were manually verified. To determine the source of false positive, files that the recognisers deemed as having the frog species present were cross-examined against the manual detection list provided by the Forestry Corporation. This presented a list of unverified false positive detections – where the recognisers identified a call, but the original manual detection list indicated the frog species was not present. A subsample of these false positive detections was manually verified through audio and visual cues on Adobe Audition CC. The false positive detection files were reviewed until the source categories of these false positive detections were considered consistent.

Results

Recogniser performance evaluation

Recognisers performed well for most species (Table 3). Performance was high (ROC > 0.8) for templates of *L. dumerilii* ,*L. fletcheri,* and *L. peronii.* Performance was also high for *N. sudelli* and *L. raniformis* but the sample sizes of their evaluation files were relatively small (Table 2). Performance was moderately high

(ROC > 0.7) for most templates of *C. signifera* and *C. parinsignifera.* Conversely, performance was poor for *L. tasmaniensis* (ROC < 0.6), for which most templates showed low precision and moderate recall.

The *L. fletcherii* recogniser comprised two templates from two sites. All templates performed well. The first template had very high precision with only two false positive detections in one sound file. However, it had the greatest number of false negatives (i.e. poorest recall). The highest performing template had a ROC value of 0.897 and was moderately sensitive, but yielded fewer false positives. The third template was excluded. The *L. dumerilii* recogniser comprised four templates from two sites. Three of the templates had very high ROC values (0.87-0.81). The fourth template displayed the greatest number of false positive detections and the poorest ROC value (0.824) was excluded from the recogniser. The *L. peronii* recogniser comprised three templates from a single site. Two templates, had the same ROC value of 0.847 and all templates had relatively high precision.

The *C. signifera* recogniser comprised three templates from two sites, two from site S and one from site H (Appendix 1). All three templates performed moderately well, with ROC values between 0.767 and 0.793, modest survey precision and good recall. The *C. parinsignifera* recognisers were constructed from three templates, stemming from two sites. The template from the first site performed poorly, with a ROC value of 0.671. This template detected over 6700 calls in 58 sound files where the species was absent (low precision). The other templates were more precise and performed moderately well, with ROC values around 0.7-0.73.

The *L. tasmaniensis* recognisers performed poorly, with ROC values of 0.562, 0.596 and 0.582. One template performed moderately well for survey recall (0.822) but had poor precision and a very high number of detections in sound files where the species was absent. The other templates performed worse.

Two other species had limited validation data. The *L. raniformis*recogniser comprised three templates from site R. All performed highly, with two templates having no false positives (precision of 1.00), and one template having no false negatives (recall of 1.00). All templates had a ROC of 0.917, however these performance metrics were calculated from only 24 evaluation sound files and should be interpreted cautiously. The *N. sudelli* recogniser comprised three templates from site M. All templates had a high ROC value of 0.984. All templates had false positive detections in only one file, although the number of detections varied. However, performance was evaluated on only 62 sound files.

Table 2 Performance evaluation of final recognisers. Counts are the count of call detections in sound files where the species was absent (Count FP) or present (Count TP). Presence-absence (PA) values are the number of sound files where the species was missed (PA FN), correctly detected (PA TP), incorrectly detected (PA FP) or correctly undetected (PA TN). Performance metrics are precision (TP/TP+FP) for calls (Call Precision) and PA in files (Survey Precision), recall (TP/TP+FN) for PA in files and ROC. * Template excluded from final analyses.

Species	**Template name**	**Count FP**	**Count TP**	**Call Precision**	**PA FP**	**PA T**
Crinia signifera	-43_129_SM4S	1048	7963	0.884	28	116
Crinia signifera	-46_130_SM4S*	4432	18488	0.807	41	127
Crinia signifera	-47_183_SM4H	2211	18609	0.894	40	120
Limnodynastes tasmaniensis	512 5khz_-27_95_SM3-E+0	10509	4245	0.288	37	46
Limnodynastes tasmaniensis	512 5khz_-32_45_SM4G*	18671	9747	0.343	46	60
Limnodynastes tasmaniensis	512 5khz_-32_90_SM3-E+0	1807	1134	0.386	25	37
Limnodynastes fletcherii	-23_95_SM3-E	52	9438	0.995	7	36
Limnodynastes fletcherii	-25_29_SM4G	2	1041	0.998	1	27
Limnodynastes fletcherii	-31_54_SM4G	381	6531	0.945	6	37
Limnodynastes fletcherii	-34_3_SM3-E*	903	1866	0.674	14	38
Limnodynastes dumerilii	-18_104_SM3-E*	47	3253	0.986	11	35
Limnodynastes dumerilii	-18_106_SM3-E	7	1964	0.996	4	32
Limnodynastes dumerilii	-28_97_20181022 (SM4G)	3	4043	0.999	2	33
Limnodynastes dumerilii	-35_30_SM4G	19	3452	0.995	5	36

Litoria peronii	-30_256_60_SM3-E	31	1874	0.984	9	45
Litoria peronii	-31_512_65_SM3-E	9	806	0.989	5	46
Litoria peronii	-38_256_75_SM3-E	2	678	0.997	1	42
Crinia parinsignifera	-30_37_SM4S	1273	6201	0.830	33	100
Crinia parinsignifera	-33_38_SM4S	1520	4563	0.750	26	91
Crinia parinsignifera	-41_47_SM3-E0+1*	6707	10575	0.612	58	109
Neobatrachus sudelli	-23_512_94_SM4M	105	3032	0.967	1	31
Neobatrachus sudelli	-39_512_100_SM4M	36	2645	0.987	1	31
Neobatrachus sudelli	-52_512_1_SM4M	3	2097	0.999	1	31
Litoria raniformis	-36_512_24_SM4R	0	28	1.000	0	10
Litoria raniformis	-42_512_85_SM4R	8	962	0.992	2	12
Litoria raniformis	-44_512_99_SM4R	0	100	1.000	0	10

Sources of false positive detections

The major sources of false positive detections varied between frog species (Table 3). Insects caused the highest proportions of false positive detections for *C.signifera, L. fletcherii,* and *C. parinsignifera* . Calls by other frogs also caused many false positives. Other frogs returned the highest false positives for*L.tasmaniensis, N. sudelli, L. raniformis* , and second highest for*C. signifera* and *L. fletcherii* . False positive detections resulting from birds were high for *L. dumerilli, L. peronii* , and*L.raniformis* . Additionally, weather caused most false positive detections for *L. tasmaniensis* .

These false positive detection sources that generated some of the highest error – insects, other frogs, and weather – often caused high detections due to their continuous nature - as opposed to just episodic events. These categories of error routinely lasted the entire five-minute recording, hence resulting in very high false positive detections in a single file. Conversely, false positive detections from sources such as anthropogenic, aquatic, birds, nature, or other animals remained relatively low in quantity, as the intervals at which these sounds repeated were infrequent.

Table 3: False positive detections per category per frog species given as a percentage of total false positive detections of that species.

Sources of false positive detections (%)	*C. signifera*	*L. tasmaniensis*	*L. fletcherii*	*L. dumerilii*	*L. peronii*	*C. par*
Anthropogenic	0.1	9.9	3.3	33.3	0	0
Water sounds	0.2	0.1	0	4.8	7.0	0
Birds	10.5	0	9.1	52.4	72.1	0
Insects	46.5	0	48.1	0	0	66.1
Other geophony	0	0.1	0	0	2.3	33.9
Other animals	0	0	0.4	0	0	0
Other frogs	41.7	50.6	35.3	0	4.6	0
Weather	0.8	39.3	3.7	9.5	14.0	0

Discussion

Single call annotation, whether manual or via recognisers, is a viable alternative to acoustic indices for monitoring ecological restoration (Linke and Deretic, 2020). While recognisers are commonly treated as one analysis class, there is a gradient in both effort and performance of auto-detectors. This ranges from largely automated recognisers – typically built-in software packages such as 'Kaleidoscope' (Wildlife Acoustics, 2017) – to completely custom-built software (Towsey et al., 2012). In all cases, various parameters alter recogniser performance; these may be left as defaults in software or manipulated by the end-user. Differences in

recogniser construction alters performance and this can manifest as poor agreement among recognisers built using different software (Lemen et al., 2015). Relying on recognisers without properly understanding how they operate can be problematic (Russo and Voigt, 2016). In this study, we took a semi-custom approach; we used a pre-programmed matching algorithm (Towsey et al., 2012; Ulloa et al., 2016) within the R package monitoR (Katz et al., 2016b), but actively investigated three important parameters that are often overlooked – or, at least, are rarely reported on - in recogniser construction. These parameters were call template selection and representativeness, template construction (including amplitude cut-off) and the threshold of similarity at which a detection is returned (score cut-off). We argue that there is a need to establish thorough construction and evaluation mechanisms for building recognisers, and for these to be properly reported in the literature.

First, choices pertaining to call template selection are crucial (Katz et al., 2016a; Teixeira et al., 2022). Studies typically report the source of call templates (e.g. whether calls were collected from wild or captive animals), but usually fail to explain the decisions underlying the selection of the exact calls used. For example, were calls free of background noise – and how did this affect recogniser performance? Animal calls exist not in isolation, but within an overall soundscape. As such, representing calls within the context of the soundscapes that we seek to monitor may be important. While our call recognisers perform well overall (Table 2), they are also prone to species-specific errors. For example, *L. tasmaniensis* recognisers produce false positives for rain events, whereas erroneous detections of *C. parinsignifera* mainly found birds and insects (Table 3).

In this study, we attempted to represent common background noises, such other species' calls and non-biological sounds (e.g. running water). Although we selected calls that were relatively clear in their structure, we maintained a 'buffer' (or a margin) around each selection, in both the time and frequency domains. Since any manual selection of candidate calls will incur a level of human bias, we chose to extract between 100 and 200 templates per species, from which a minimum of 10 were tested and only two or three were chosen for the final recogniser. Although for some rare or cryptic species, call templates can be difficult to acquire, we argue that, as much as possible, recognisers should be built following the testing of many candidate templates.

Another important consideration is the representativeness of species' call types and behaviours (Priyadarshani et al., 2018). For species that exhibit large vocal repertoires, decisions must be made about the call types to feature in recognisers. This should be driven by a program's objectives or research questions; for example, monitoring breeding may require only one or two breeding-associated call types to feature in the recogniser (Teixeira et al., 2019). Further, geographic variation in call structure (e.g. regional dialects) may also impact recogniser performance, and should be investigated when recognisers are intended for use at spatial scales over which call types may vary (Kahl et al., 2021; Lauha et al., 2022; Priyadarshani et al., 2018). If recognisers are used among discrete or isolated populations, call templates may need to represent each area. In this study, we attempted to represent inter-site variability by selecting candidate call templates from every site where the species was recorded. For several species, the final recognisers comprised templates from more than one site.

Once call templates are chosen, decisions must be made about their construction for use in a recogniser. In binary point matching, call templates are created from a grid of on and off points (i.e. call and non-call points), which are manipulated by the amplitude cut-off set by the user (Katz et al., 2016b). In monitoR, the impact of altering amplitude cut-off can be easily visualised (Figure 1). In this study, we manipulated amplitude cut-off to show both the call structure and some background noise. Since the recogniser 'matches' both the on and off points, finding a suitable balance between these is important. Although visualising and selecting amplitude cut-off is a manual and somewhat arbitrary process, we considered that the large sample size of candidate templates tested in this study would minimise any bias from this process. However, for studies that test a smaller number of candidate templates, we recommend that each template is tested at several different amplitude cut-offs.

Finally, an appropriate score cut-off, which sets the threshold of similarity at which a detection is returned

(Figure 2), must be set for each call template. Score cut-off alters the template's sensitivity and therefore, greatly affects performance. A higher score cut-off will reduce false positive detections, but may increase false negatives (Katz et al., 2016a). Conversely, increasing sensitivity by lowering score cut-off will reduce false negatives, but it may reduce precision by returning more false positives. Here, we tested every call template at score cut-off increments of 0.2 from a low of 3, and measured performance by ROC value. For most species examined, high ROC values indicated that call templates were able to sufficiently trade-off false positives and false negatives while maximising true positives. This rigorous approach to score cut-off testing allowed us to set highly specific cut-offs in the final recognisers. However, for species that are rarer or more cryptic, returning sufficient true positives may require a lower score cut-off with a poorer ROC value. Where detecting most, if not all, calls is important, other performance metrics like recall should be given due consideration. Ultimately, decisions about score cut-off should be driven by a study's objectives, but we argue that general metrics like ROC values are a good starting point in most cases.

We argue that ecoacoustic researchers and practitioners need to stop treating recognisers like a black box and actively develop, improve and test processes that help evaluation. From the literature, it is currently unclear how reliable recognisers are. Many studies report poor performance but this may be more a function of inappropriate construction, rather than recognition methods per se. Especially recogniser testing is often ignored and recogniser performance is reported by number of detections in a larger dataset. Even when performance is reported, it is often unclear what the source of low recogniser accuracy is. We demonstrated that this could have multiple causes, from badly selected templates to a lack of template calibration, for example amplification or detection cut-offs. We recommend that recognisers are not treated as a static product. They can be refined and adapted as more monitoring data become available. Using this study as an example, we are currently working on a refinement for the recogniser for *L. tasmaniensis* that is based on better template recordings. A complete recommended workflow could start with a recogniser built for a particular species in a particular region, then enhanced by data from other environments, followed by a performance evaluation and refinement as necessary.

Conclusion and future directions

In this study, we have demonstrated the possibility of building high quality single call recognisers for monitoring ecosystem restoration. While testing the recognisers on data collected before and after watering would determine their precision in responding to actual restoration, we expect the response to be sharp as precision and recall scored high for most species. If sharp responses are found, we encourage testing transferability of recognisers to other locations as site specificity of recognisers has rarely been investigated. Amphibians are also not the only soniferous taxonomic group that responds to environmental watering. Birds were responsible for the bulk of the manually annotated response curves described by Linke & Deretic (2020). We strongly encourage additional studies that build call recognisers for water-dependent birds or alternatively trial the performance of a pre-built recogniser such as BirdNet (Kahl et al., 2021). This has the potential to lay the foundation for a real-time monitoring system that can detect the ecological outcomes of environmental water allocations in the Murray-Darling Basin – a milestone in the management of a stressed, but recovering, river.

Acknowledgements and funding

The authors received funding from Forestry Corporation of NSW to develop the recognisers.

Author Contributions

SL and DT secured the funding, SL, DT and KT conceptualised, managed the data, conducted formal analysis and validated the results. SL and DT developed the core recogniser optimisation algorithm. KT led writing the final client report this paper was based on, SL led the first draft of the paper. SL and DT contributed to revisions of the paper.

Conflict of interest

None

Data availability

monitoR recognisers can be made available on request, primary sound data is the property of the Forestry Corporation of NSW and needs to be requested directly.

Permission to reproduce materials from other sources

None

References

Aristotle (1910) Historia animalium. translated by Thompson, DW.

Bohnenstiehl BDR, Lyon LRP, Caretti CON, Ricci RSW and Eggleston EDB (2018) Investigating the utility of ecoacoustic metrics in marine soundscapes. Journal of Ecoacoustics 2(2). Doi: 10.22261/JEA.R1156L.

Brandes TS (2008) Automated sound recording and analysis techniques for bird surveys and conservation. Bird Conservation International 18(SupplementS1), S163-S173. Doi: doi:10.1017/S0959270908000415.

Brodie S, Allen-Ankins S, Towsey M, Roe P and Schwarzkopf L (2020a) Automated species identification of frog choruses in environmental recordings using acoustic indices. Ecological Indicators 119, 106852. Doi: https://doi.org/10.1016/j.ecolind.2020.106852.

Brodie S, Yasumiba K, Towsey M, Roe P and Schwarzkopf L (2020b) Acoustic monitoring reveals year-round calling by invasive toads in tropical Australia. Bioacoustics, 1-17. Doi: 10.1080/09524622.2019.1705183.

Campos IB, Landers TJ, Lee KD, Lee WG, Friesen MR, Gaskett AC and Ranjard L (2019) Assemblage of Focal Species Recognizers—AFSR: A technique for decreasing false indications of presence from acoustic automatic identification in a multiple species context. Plos One 14(12), e0212727. Doi: 10.1371/journal.pone.0212727.

Cragg JL, Burger AE and Piatt JF (2015) Testing the effectiveness of automated acoustic sensors for monitoring vocal activity of marbled murrelets *Brachyramphus marmoratus* . Marine Ornithology 43(2), 151-160.

Crump PS and Houlahan J (2017) Designing better frog call recognition models. Ecology and Evolution. Doi: 10.1002/ece3.2730.

Dema T, Towsey M, Sherub S, Sonam J, Kinley K, Truskinger A, Brereton M and Roe P (2020) Acoustic detection and acoustic habitat characterisation of the critically endangered white-bellied heron (Ardea insignis) in Bhutan. Freshwater Biology 65(1), 153-164. Doi: 10.1111/fwb.13217.

Dutilleux G and Curé C (2020) Automated acoustic monitoring of endangered common spadefoot toad populations reveals patterns of vocal activity. Freshwater Biology 65(1), 20-36. Doi: 10.1111/fwb.13111.

Grafton RQ, Pittock J, Williams J, Jiang Q, Possingham H and Quiggin J (2014) Water Planning and Hydro-Climatic Change in the Murray-Darling Basin, Australia. Ambio 43(8), 1082-1092. Doi: 10.1007/s13280-014-0495-x.

Greenhalgh JA, Stone HJR, Fisher T and Sayer CD (2021) Ecoacoustics as a novel tool for assessing pond restoration success: Results of a pilot study. Aquatic Conservation-Marine and Freshwater Ecosystems 31(8), 2017-2028. Doi: 10.1002/aqc.3605.

Kahl S, Wood CM, Eibl M and Klinck H (2021) BirdNET: A deep learning solution for avian diversity monitoring. Ecological Informatics 61, 101236. Doi: https://doi.org/10.1016/j.ecoinf.2021.101236.

Katz J, Hafner SD and Donovan T (2016a) Assessment of Error Rates in Acoustic Monitoring with the R package monitoR. Bioacoustics 25(2), 177-196. Doi: 10.1080/09524622.2015.1133320.

Katz J, Hafner SD and Donovan T (2016b) Tools for automated acoustic monitoring within the R package monitoR. Bioacoustics 25(2), 197-210.

King AJ, Gawne B, Beesley L, Koehn JD, Nielsen DL and Price A (2015) Improving Ecological Response Monitoring of Environmental Flows. Environmental Management 55(5), 991-1005. Doi: 10.1007/s00267-015-0456-6.

Knight EC, Hannah KC, Foley GJ, Scott CD, Brigham RM and 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.

Lauha P, Somervuo P, Lehikoinen P, Geres L, Richter T, Seibold S and Ovaskainen O (2022) Domain-specific neural networks improve automated bird sound recognition already with small amount of local data. Methods in Ecology and Evolution 13(12), 2799-2810. Doi: https://doi.org/10.1111/2041-210X.14003.

Lemen C, Freeman PW, White JA and Andersen BR (2015) The Problem of Low Agreement among Automated Identification Programs for Acoustical Surveys of Bats. Western North American Naturalist 75(2), 218-225. Doi: 10.3398/064.075.0210.

Linke S and Deretic J-A (2020) Ecoacoustics can detect ecosystem responses to environmental water allocations. Freshwater Biology 65(1), 133-141. Doi: 10.1111/fwb.13249.

Linke S, Gifford T and Desjonquères C (2020) Six steps towards operationalising freshwater ecoacoustic monitoring. Freshwater Biology 65(1), 1-6. Doi: 10.1111/fwb.13426.

Mitchell SL, Bicknell JE, Edwards DP, Deere NJ, Bernard H, Davies ZG and Struebig MJ (2020) Spatial replication and habitat context matters for assessments of tropical biodiversity using acoustic indices. Ecological Indicators 119, 106717. Doi: https://doi.org/10.1016/j.ecolind.2020.106717.

Ng M-L, Butler N and Woods N (2018) Soundscapes as a surrogate measure of vegetation condition for biodiversity values: A pilot study. Ecological Indicators 93, 1070-1080.

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

Ruff ZJ, Lesmeister DB, Duchac LS, Padmaraju BK and Sullivan CM (2020) Automated identification of avian vocalizations with deep convolutional neural networks. Remote Sensing in Ecology and Conservation 6(1), 79-92. Doi: 10.1002/rse2.125.

Russo D and Voigt CC (2016) The use of automated identification of bat echolocation calls in acoustic monitoring: A cautionary note for a sound analysis. Ecological Indicators 66, 598-602. Doi: http://dx.doi.org/10.1016/j.ecolind.2016.02.036.

Salamon J, Bello JP, Farnsworth A, Robbins M, Keen S, Klinck H and Kelling S (2016) Towards the automatic classification of avian flight calls for bioacoustic monitoring. PLoS ONE 11(11). Doi: 10.1371/journal.pone.0166866.

Souchon Y, Sabaton C, Deibel R, Reiser D, Kershner J, Gard M, Katopodis C, Leonard P, Poff NL and Miller WJ (2008) Detecting biological responses to flow management: missed opportunities; future directions. River Research and Applications 24(5), 506-518.

Sueur J and Farina A (2015) Ecoacoustics: the Ecological Investigation and Interpretation of Environmental Sound. Biosemiotics 8(3), 493-502. Doi: 10.1007/s12304-015-9248-x.

Sueur J, Farina A, Gasc A, Pieretti N and Pavoine S (2014) Acoustic Indices for Biodiversity Assessment and Landscape Investigation. Acta Acustica united with Acustica 100(4), 772-781. Doi: 10.3813/aaa.918757.

Teixeira D, Linke S, Hill R, Maron M and van Rensburg BJ (2022) Fledge or fail: Nest monitoring of endangered black-cockatoos using bioacoustics and open-source call recognition. Ecological Informatics 69, 101656. Doi: https://doi.org/10.1016/j.ecoinf.2022.101656.

Teixeira D, Maron M and van Rensburg BJ (2019) Bioacoustic monitoring of animal vocal behavior for conservation. Conservation Science and Practice 1(8), e72. Doi: 10.1111/csp2.72.

Towsey M, Planitz B, Nantes A, Wimmer J and Roe P (2012) A toolbox for animal call recognition. Bioacoustics 21(2), 107-125.

Ulloa JS, Gasc A, Gaucher P, Aubin T, Réjou-Méchain M and Sueur J (2016) Screening large audio datasets to determine the time and space distribution of Screaming Piha birds in a tropical forest. Ecological Informatics 31, 91-99. Doi: http://dx.doi.org/10.1016/j.ecoinf.2015.11.012.

Wildlife Acoustics (2017) Kaleidoscope (version 4.0. 4)[Software]. Wildlife Acoustics Inc. Massachusetts, United States Retrieved from https://www. wildlifeacoustics. com/download/kaleidoscope-software.

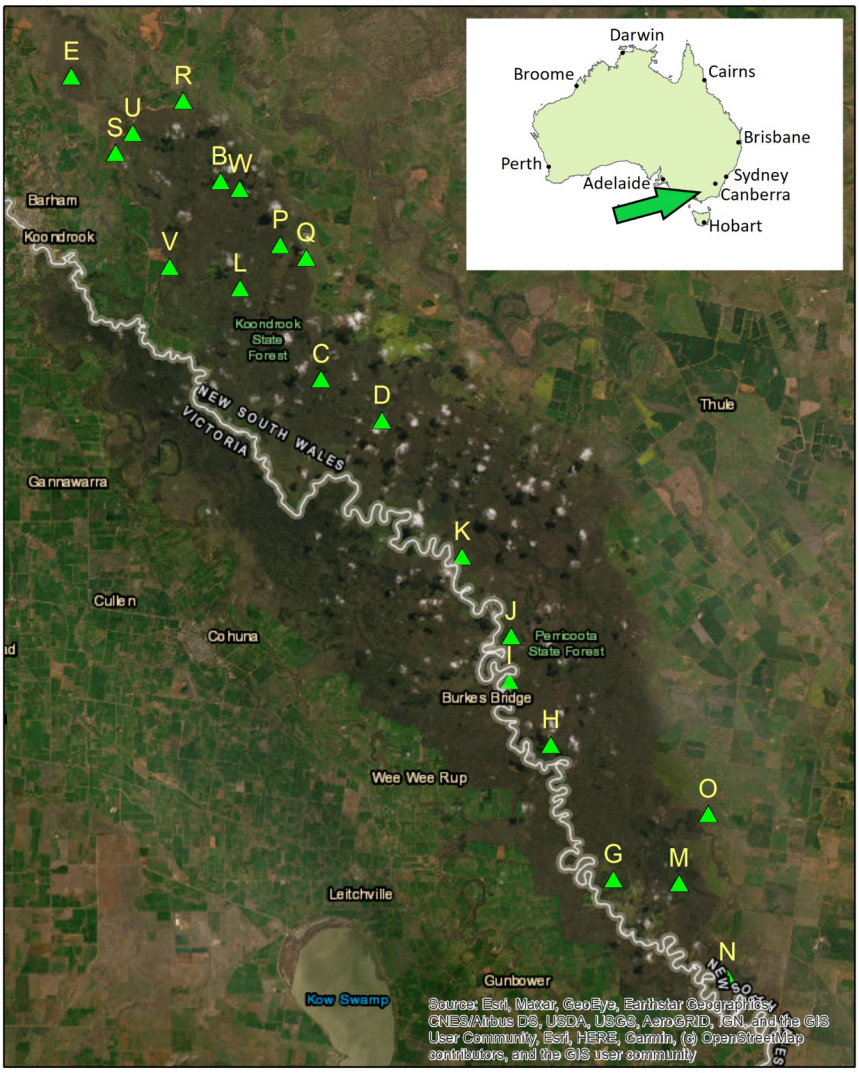
Wright WJ, Irvine KM, Almberg ES and Litt AR (2020) Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity. Methods in Ecology and Evolution 11(1), 71-81. Doi: 10.1111/2041-210x.13315.

Zhong M, LeBien J, Campos-Cerqueira M, Dodhia R, Lavista Ferres J, Velev JP and Aide TM (2020) Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling. Applied Acoustics 166, 107375. Doi: https://doi.org/10.1016/j.apacoust.2020.107375.

Znidersic E, Towsey M, Roy WK, Darling SE, Truskinger A, Roe P and Watson DM (2020) Using visualization and machine learning methods to monitor low detectability species—The least bittern as a case study. Ecological Informatics 55, 101014. Doi: https://doi.org/10.1016/j.ecoinf.2019.101014.

Znidersic E and Watson DM (2022) Acoustic restoration: Using soundscapes to benchmark and fast-track recovery of ecological communities. Ecology Letters 25(7), 1597-1603. Doi: https://doi.org/10.1111/ele.14015.

Zou KH, O'Malley AJ and Mauri L (2007) Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. Circulation 115(5), 654-657. Doi: doi:10.1161/CIRCULATIONAHA.105.594929.



Appendix 1. Songmeter sites in Koondrook-Pericoota Forest