# Fauna Call Recogniser Project

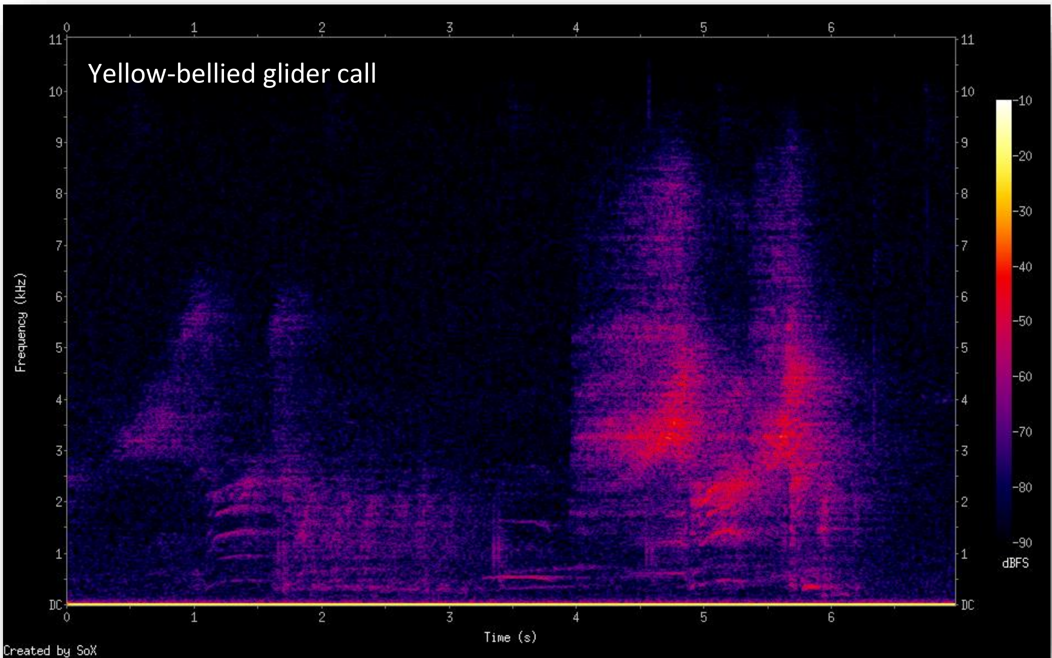
*Final report to NRC*

**16 June 2021**


*K. Thompson, L. Gonsalves, B. Law* — *NSW DPI*

*A. Truskinger, M. Towsey, P. Roe* — *QUT*

*N. Priyadarshani* — *AviaNZ*

## 1. Project outline

The aim of this project was to develop species recognisers to analyse acoustic data collected in forest monitoring programs and other research programs within NSW. The project aimed to compare two approaches to recogniser development. Recognisers for the 13 target species were successfully built in open-source software: AviaNZ (Marsland, 2020) and Queensland University of Technology's AnalysisPrograms.exe (henceforth abbreviated as AP) (Towsey, 2020) and tested in both AviaNZ and Egret (Truskinger, 2020). Preliminary tests during the project were also made using Kaleidoscope Software (Wildlife Acoustics Inc. 2017), however its performance was found to be poor relative to the two alternative approaches being tested. All recognisers were tested for recall and precision on three sets of data, event-level, segment-level and 'real-world' one-hour recordings. Each recogniser had varying levels of success. Additional high-recall wavelet filters were also built in AviaNZ for a subset of the species to use for convolutional neural network (CNN) training. CNN is expected to substantially improve precision results (fewer false positives). Testing for these high-recall wavelet filters has been undertaken at event-level and segment-level and CNN training has been applied for yellow-bellied gliders and is in progress for powerful owls. Metadata information has also been provided for all recognisers (both AviaNZ and AP) to accompany the recogniser when distributed to provide information on performance and sounds that will result in potential confusion.

## 2. Target species

The target species for this project are predominantly listed as vulnerable in NSW but also includes species that can easily be confused by an automatic recogniser (Table 1). For example, powerful owls, barking owls and boobooks have a two-note call in approximately the same frequency range and the species can often be confused by a recogniser. Our approach began with nocturnal species which are considered easier, more recognisable and have less background noise in recordings. The sound files used were recorded in a variety of forests across a range of regions in NSW including north-east and south-east regions. Inclusion of a variety of regions is important as they likely contain different background environmental sounds and include potential for geographic variation in calls of the target species. Majority of calls were extracted from DPI's existing call library, although calls for some species required additional sourcing by the project team (including FCNSW) via field recordings: greater sooty owl, masked owl, rufous scrub-bird and glossy black cockatoos, the latter sourced from Lauren Hook in DPIE.

**Table 1**. List of target species, their call frequency range and conservation status in NSW.

| No. | Species | Call freq. range (Hz) | Conservation status in NSW |
|---|---|---|---|
| 1 | Powerful owl *(Ninox strenua)* | 350-550 | Vulnerable |
| 2 | Barking owl *(Ninox connivens)* | 240-1300 | Vulnerable |
| 3 | Southern boobook *(Ninox boobook)* | 500-960 | Least concern |
| 4 | Yellow-bellied glider *(Petaurus australis)* | 200-10,000 | Vulnerable |
| 5 | Grey-headed flying-fox *(Pteropus poliocephalus)* | 1900-23,000 | Vulnerable |
| 6 | Greater sooty owl *(Tyto tenebricosa)* | 1200-10,500 | Vulnerable |
| 7 | Masked owl *(Tyto novaehollandiae)* | 1500-2500 | Vulnerable |
| 8 | Sugar glider *(Petaurus breviceps)* | 430-2600 | Least concern |
| 9 | Squirrel glider *(Petaurus norfolcensis)* | 300-2100 | Vulnerable |
| 10 | Glossy black cockatoo *(Calyptorhynchus lathami)* | 2000-8000 | Vulnerable |
| 11 | Brown treecreeper *(Climacteris picumnus)* | 2400-9300 | Vulnerable |
| 12 | Rufous scrub-bird *(Attrichornis rufescens)* | 1900-7600 | Vulnerable |
| 13 | Bell miner *(Manorina melanophyrs)* | 2000-10,000 | Least concern |

## 3. Development of fauna call recognisers

Files used for training of recognisers were sourced from different regions and sites and included a variety of loud and soft calls and variation in background noise (Table 2). AviaNZ requires calls to be annotated prior to building a recogniser and this was completed manually (Figure 1). The same annotated training files were also used to build the recognisers for AP. However, as the recognisers built for AP are handcrafted algorithms, not every training file may be used in the training process.

**Table 2.** Number of 30 second files and calls used for training data for the target species for both AviaNZ and AP recognisers. The table also identifies the species for which an additional CNN wavelet filter could feasibly be developed based on having sufficient calls.

| No. | Species | No. files | No. calls | Additional CNN filter |
|---|---|---|---|---|
| 1 | Powerful owl | 48 | 123 | Yes |
| 2 | Barking owl | 29 | 132 | Yes |
| 3 | Southern boobook | 12 | 95 | Yes |
| 4 | Yellow-bellied glider | 50 | 79 | Yes |
| 5 | Grey-headed flying-fox | 50 | 338 | No |
| 6 | Greater sooty owl | 27 | 79 | No |
| 7 | Masked owl | 50 | 58 | No |
| 8 | Sugar glider | 12 | 267 | Yes |
| 9 | Squirrel glider | 20 | 228 | Yes |
| 10 | Glossy black cockatoo | 18 | 64 | No |
| 11 | Brown treecreeper | 13 | 183 | No |
| 12 | Rufous scrub-bird | 28 | 76 | No |
| 13 | Bell miner | 13 | 162 | No |

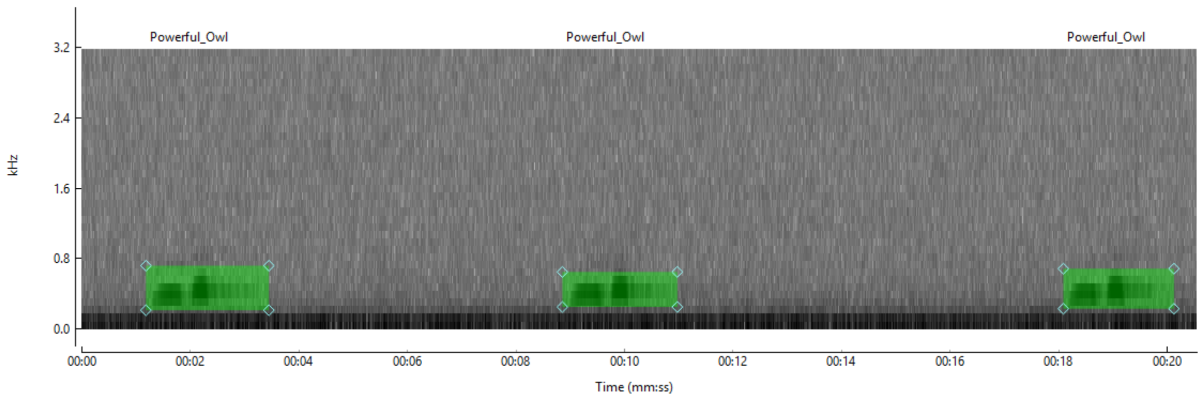



**Figure 1.** Manually annotated powerful owl calls for recogniser training in AviaNZ.

*AviaNZ*

Annotated training files underwent a clustering process in AviaNZ where annotations were grouped into five clusters of similar calls. Clusters were then combined prior to scanning by AviaNZ to identify and suggest settings for key parameters used by the recogniser, including minimum call length, maximum call length, average syllable length, maximum gap between syllables, lower frequency limit and upper frequency limit. The next stage included a cross-validation process in which AviaNZ generated a ROC curve illustrating the true positive rate (%TPR) vs false positive rate (%FPR) for the training data. The point chosen on the ROC curve was dependent on the species for which the recogniser was being built. For species that emit few calls in a short time sequence (powerful owl, yellow-bellied glider, grey-headed flying-fox, greater sooty owl, masked owl and squirrel glider) a point was chosen with slightly higher recall to maximise true positives. Whereas for species that call repeatedly (barking owl, boobook, sugar glider, glossy-black cockatoo, brown treecreeper, rufous scrub-bird and bell miner) a point was chosen with lower recall as the recogniser has a higher chance of picking up a call within a call sequence while aiming for higher precision. This is particularly important for diurnal birds calling within a complex soundscape which can produce higher numbers of false positives.

## Recogniser training using CNN

We aim to eventually improve wavelet filters by adding a deep learning step called a convolutional neural network (CNN). CNN uses thousands of calls from the target species and false positives to train the recogniser with the aim to significantly decrease the overall number of false positives the recogniser detects. We have discovered that the number of false positives is a major issue to overcome in the application of recognisers and CNN appears to be a good way forward to overcome this limitation.

The addition of a CNN high-recall wavelet filter for a species was decided upon based on low precision results at an event and segment-level and also the availability of a large amount of calls to be able to undertake the CNN training (Table 2). The CNN high-recall wavelet filter uses the same number of files and calls as other wavelet filters, but a different true positive and false positive rate is decided upon to give high recall, with the aim of CNN training being to improve precision.

For CNN training, thousands of calls from target species and false positive (referred to as ‘noise’) files were required. This requires considerable work prior to running a CNN, while the CNN training itself takes hours to days to complete. To date, a CNN recogniser for the yellow-bellied glider has been successfully completed and a CNN recogniser for the powerful owl is underway. For the yellow-bellied glider CNN recogniser, all available calls from the species (862 calls) (Table 3) were collated for training and were annotated manually. For the powerful owl CNN recogniser, a dataset was chosen that had previously been used for wavelet filter testing and all available powerful owl calls from that dataset (2702 calls) were collated and annotated manually.

For the noise files, false positives that were identified when training the original wavelet filter and CNN wavelet filter were collated from the DPI database. These files were checked to confirm that they didn’t contain calls of the target species. Throughout CNN training, AviaNZ will randomly select a subset of target and noise files to use. More noise files than files containing target species were used to increase the chances of them being chosen as the overall goal was to increase precision, rather than recall. In addition, a greater proportion of the noise files contained particularly difficult false positives to increase the chances of that false positive being selected for training. The list of false positives used for CNN training (Table 3) is not exhaustive and the false positives encountered in various data sets differ. We also aim to undertake ongoing refinement by adding false positives picked up when running the CNN recogniser and adding them to the noise files for the next iteration of a CNN recogniser i.e. a Version 2.

**Table 3.** Number of target files and calls used for the successfully built yellow-bellied glider CNN recogniser and the powerful owl CNN recogniser in which training is underway.

| Species | No. target files | No. calls | No. noise files | Types of noise files |
|---|---|---|---|---|
| Yellow-bellied glider | 629 | 862 | 8837 | Aeroplane, anthropogenic, brushtail possum, dog, fantail cuckoo, flying-foxes, frogs, insects, koala, koel, kookaburra, owlet nightjar, rain, sugar glider, white-throated nightjar, wind. |
| Powerful owl | 1001 | 2702 | 5252 | Aeroplane, anthropogenic noise, boobook, barking owl, dog, frogs, insects, koala, kookaburra, sugar glider, tawny frogmouth, wind. |

*QUT Ecoacoustics Analysis Programs (AP)*

To build the recogniser, exemplar calls of the target species from the training dataset were studied and heuristics manually derived that are indicative of its call (e.g. harmonics, two-second syllables, fundamental frequency, two syllables present in every call). The heuristics were then used as a starting point to build the recogniser in AP. When the recogniser is built, a configuration file is generated and named. A set of algorithms are then manually composed in the configuration file. Each algorithm will then focus on components of the call at a finer scale. For example, these components can include lines, whistles, clicks and harmonics. The components are then parameterised manually and configuration files are saved.

To train the recogniser, annotations from training files were imported to AP with a script and these annotations are used as specifications for the recogniser's performance. As the recogniser continues to be developed, scripts are used to automatically test larger portions of the training dataset. When a specification fails to produce the expected results, the recogniser is manually adjusted. This manual training process seeks to produce a 100% TP rate (100% recall) for the training dataset through repeated iterations of recogniser configuration. To avoid overfitting data, development is stopped when performance gains no longer warrant additional time investment.

## 4. Recogniser testing process

Definition of terms:

- True positive (TP): Recogniser and human agree there was a call.
- True negative (TN): Recogniser and human agree there was no call.
- False positive (FP): Recogniser says there was a call but the human did not.
- False negative (FN): Recogniser did not detect a call that the human found.

Three separate approaches to testing the AviaNZ and AP recognisers were followed (Table 4).

Event-level*:* 100 files, 30 seconds in length manually annotated for each species plus approximately 10 x 30sec files from numerous false positive groups with no target species calling. False positives used in testing were anthropogenic noise (highway, vehicle noise and sirens), biophonic noise (barking owl, boobook, brushtail possum, diurnal birds, dogs barking, fan-tailed cuckoo, frogs, grey-headed flying fox, insects, male koala bellows, kookaburra, masked owl, owlet nightjar, powerful owl, squirrel glider, sugar glider, yellow-bellied glider) and geophonic noise (rain and wind). False positive files varied for each species depending if the calls were confused with the target species by the recogniser, as identified during the training process. This was used to test the performance of each recogniser at the level of every event (hit). AviaNZ compared the auto generated annotations from the recogniser with the manual annotations provided, producing a results summary in the text file format based on the total number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). As such, performance was assessed against every annotation. Egret follows the same process.

<u>Segment-level:</u> The same 100 x 30sec files and 10 x 30sec false positives from each group as event-level but files were not annotated for AviaNZ testing. This was used to test the performance of each recogniser at segment-level, detecting an event (hit) anywhere in a 30sec file. For AviaNZ testing, the results were manually validated.

<u>Real-world:</u> Three one-hour recordings that contained the target species somewhere within the hour. This was used to test performance on 'real-world' recordings which contained a variety of sounds. Files were split into 30sec segments so the segment-level testing method could be applied. As file splitting does not always occur evenly, some species have more or less than 360 files, but the three hours total used for testing was consistent. Bell miner was the exception with only two hours available for real-world testing.

For wavelet recognisers built in AviaNZ, event-level and segment-level testing was undertaken in AviaNZ and segment-level results were validated manually. For recognisers built in AP, all testing was undertaken in Egret (Truskinger, 2020), software designed to test recognisers at an event-level and segment-level. AviaNZ real-world testing was also undertaken in Egret. No manual validation is required for Egret testing as all files are manually annotated beforehand. CNN wavelet filters were not tested on real-world files.

**Table 4.** Number of 30sec files used for testing at each level. The same number of files was used for both AviaNZ and AP recognisers.

| No. | Species | No. testing files | No. testing files | No. testing files |
|---|---|---|---|---|
| | | Event-level | Segment-level | Real-world |
| 1 | Powerful owl | 229 | 229 | 360 |
| 2 | Barking owl | 265 | 265 | 360 |
| 3 | Southern boobook | 249 | 249 | 360 |
| 4 | Yellow-bellied glider | 236 | 236 | 360 |
| 5 | Grey-headed flying fox | 206 | 206 | 359 |
| 6 | Greater sooty owl | 152 | 152 | 360 |
| 7 | Masked owl | 176 | 176 | 360 |
| 8 | Sugar glider | 250 | 250 | 360 |
| 9 | Squirrel glider | 239 | 239 | 360 |
| 10 | Glossy black cockatoo | 190 | 190 | 360 |
| 11 | Brown treecreeper | 190 | 190 | 362 |
| 12 | Rufous scrub-bird | 170 | 170 | 360 |
| 13 | Bell miner | 180 | 180 | 240 |

## 5. Recogniser test results

Recognisers for all 13 species were successfully built in AviaNZ and AP and tested in both AviaNZ and Egret. Recall and precision were calculated for all three levels of testing (Tables 4, 5 and 6).

- Recall = TP/(TP+FN). Number correctly labelled as positive / actual number of positive examples plus false negatives. Higher values indicate better performance (more hits of target species).
- Precision = TP/(TP+FP). Number correctly labelled as positive / number labelled as positive plus false positives. Higher values indicate better performance (fewer false positives).

*AviaNZ*

At event-level, precision was good overall with the lowest precision at 49.47% for rufous scrub-bird (Table 5). Lower precision would be expected for diurnal birds as there is more background noise and other species calling at that time, leading to more potential false positives. Recall mostly fell between 40% - 64% with the highest recall 71.37% for masked owl. Recall was purposively reduced for bell miner (16.45%), boobook (20.66%) and glossy black cockatoo (26.23%) to allow for a higher precision rate. Although at event-level, recall was quite low for these species in particular, at segment-level, the recall was much higher i.e. the recogniser was identifying at least one true positive within a 30sec file. Boobook recall was still the lowest for segment-level testing, but as this species calls continuously, it was deemed a sufficient trade-off for reasonable precision. All other species recall results fell between 70% - 100.00%. However, precision at segment-level did drop for all species except sooty owl and masked owl. This makes sense as deliberate false-positive files were included in segment-level testing. Species that had especially low-frequency calls such as barking owl, sugar glider and squirrel glider picked up most false positives (Table 6). The diurnal birds as well had lower precision sitting between 43% - 52%.

**Table 5.** Test evaluation results for DPI's AviaNZ recognisers (wavelet only), tested in AviaNZ (event and segment-level) and Egret (real-world). Event-level: recogniser performance at level of every event (hit). Segment-level and real-world: recogniser performance at detecting an event anywhere within a 30sec file. Bold and underlined real-world results were deemed good results.

| No. | Species | Event-level | | Segment-level | | Real-world | |
|---|---|---|---|---|---|---|---|
| | | Recall | Precision | Recall | Precision | Recall | Precision |
| 1 | Powerful owl | 41.33% | 94.99% | 71.00% | 56.35% | **<u>40.00%</u>** | **<u>73.47%</u>** |
| 2 | Barking owl | 45.57% | 65.12% | 82.00% | 30.26% | 79.41% | 11.95% |
| 3 | Southern boobook | 20.66% | 97.83% | 56.00% | 54.37% | 13.56% | 66.67% |
| 4 | Yellow-bellied glider | 49.02% | 93.31% | 76.00% | 53.52% | **<u>42.86%</u>** | **<u>81.82%</u>** |
| 5 | Grey-headed flying fox | 61.06% | 70.91% | 100.00% | 58.82% | **<u>97.62%</u>** | **<u>40.86%</u>** |
| 6 | Greater sooty owl | 63.03% | 52.47% | 89.29% | 55.56% | 40.00% | 5.56% |
| 7 | Masked owl | 71.37% | 58.70% | 96.00% | 65.31% | **<u>90.70%</u>** | **<u>36.79%</u>** |
| 8 | Sugar glider | 47.11% | 88.52% | 91.00% | 39.06% | 82.61% | 7.89% |
| 9 | Squirrel glider | 59.77% | 72.06% | 92.00% | 38.98% | 90.32% | 8.62% |
| 10 | Glossy black cockatoo | 26.23% | 59.48% | 73.00% | 44.51% | 93.10% | 10.34% |
| 11 | Brown treecreeper | 42.65% | 95.97% | 89.00% | 48.90% | **<u>96.85%</u>** | **<u>39.55%</u>** |
| 12 | Rufous scrub-bird | 53.17% | 49.47% | 98.00% | 43.75% | 100.00% | 8.94% |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 13 | Bell miner | 16.45% | 75.00% | 94.00% | 51.37% | **93.48%** | **94.85%** |

Real-world results were mixed and precision differed between recognisers. Recall dropped but precision improved from segment-level for powerful owl, boobook, yellow-bellied glider which were three species that had high precision overall (73.47%, 66.67% and 81.82%, respectively). Precision for a number of species decreased significantly from segment-level testing with precision results for sooty owl, sugar glider, squirrel glider and rufous scrub-bird all falling below 10%. Barking owl and glossy black cockatoo also dropped from 30.26% to 11.95% and 44.51% to 10.34% respectively. Extensive insect noise was present in the sooty owl real-world test and this led to a high false positive rate for this species and probably others as well. Bell miner performed particularly well in real-world testing with 93.48% recall and 94.85% precision.

False positives that were picked up in segment-level testing were collated and percentages calculated for each recogniser (Table 6). Geophonic noise (rain and wind) was not picked up by most recognisers, but anthropogenic noise was picked up by all recognisers. As noted previously, recognisers for low frequency species, in particular the barking owl, sugar glider and squirrel glider recognisers, have trouble with false positives in most files that were presented to them. Yellow-bellied gliders also proved to be a main false positive for most nocturnal species, though is not considered much of an issue as they don't call repeatedly and are listed as vulnerable in NSW. Kookaburras have proved another frequent false positive across all recognisers. They were included for nocturnal species recognisers as they are usually one of the first birds to start calling before dawn. This could be an issue with other nocturnal recognisers, particularly if there are several individuals or they are calling particularly loudly.

**Table 6.** Detailed false positive results for each AviaNZ recogniser (wavelet only). The table shows percentages of false positive test files that were falsely picked up by each recogniser. For example, the powerful owl recogniser found a false positive in 40% of available anthropogenic files. Sample size for all false positives is 10 x 30sec files except frog sample size is 15 x 30sec files (to include high and low-calling frogs). Dashes mean the false positive was not included in testing.

| | **Powerful owl** | **Barking owl** | **Boobook** | **Yellow-bellied glider** | **Grey-headed flying fox** | **Sooty owl** | **Masked owl** | **Sugar glider** | **Squirrel glider** | **Glossy-black cockatoo** | **Brown treecreeper** | **Rufous scrub-bird** | **Bell miner** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Anthropogenic | 40 | 70 | 10 | 10 | 10 | 20 | 10 | 70 | 80 | 10 | 70 | 20 | 10 |
| Barking owl | 40 | 0 | 10 | 0 | - | - | - | - | - | - | - | - | - |
| Boobook | 40 | 100 | - | 20 | 20 | - | - | 100 | 100 | - | - | - | - |
| Brushtail possum | - | 100 | 90 | 80 | 80 | - | - | 100 | 100 | - | - | - | - |
| Diurnal birds | - | - | - | - | - | - | - | - | - | 100 | 100 | 80 | 70 |
| Dogs | 70 | 100 | 50 | 10 | - | - | - | 100 | 100 | 0 | 30 | - | - |
| Fan-tailed cuckoo | - | - | - | 90 | 100 | 30 | 100 | - | - | 100 | 80 | - | 60 |
| Frogs | 20 | 73.33 | 6.67 | 6.67 | 26.67 | 20 | 20 | 66.67 | 53.33 | - | - | - | - |
| Grey-headed flying-fox | - | 100 | - | 90 | - | 90 | 100 | 100 | 100 | 90 | 100 | 100 | 90 |
| Insects | 40 | 90 | 10 | 40 | 70 | 60 | 50 | 80 | 100 | 40 | 90 | 40 | 50 |
| Koala | 70 | 100 | 20 | 40 | 40 | 40 | - | 100 | 100 | - | - | - | - |
| Kookaburra | 80 | 100 | 100 | 100 | 100 | - | - | 100 | 100 | 100 | 100 | 100 | 100 |
| Owlet nightjar | - | 90 | 30 | 50 | - | 30 | - | 90 | 80 | - | - | - | - |
| Powerful owl | - | 40 | 10 | - | - | - | - | 20 | - | - | - | - | - |
| Rain | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 80 | 100 | 0 | 50 | 0 | 0 |
| Sirens | 10 | 90 | 30 | - | - | - | - | - | - | - | - | - | - |
| Squirrel glider | - | - | - | - | - | - | - | 90 | - | - | - | - | - |
| Sugar glider | 10 | 100 | 0 | 20 | - | - | - | - | 100 | - | - | - | - |
| Wind | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 |
| Yellow-bellied glider | 90 | 100 | 60 | - | 100 | 90 | 100 | 100 | 100 | - | - | - | - |

Additional wavelet filters were developed for use by AviaNZ CNN and these were tested in AviaNZ with the aim for the filters to achieve high recall results, particularly at segment-level (Table 7). This was successful and the filters are now ready to be extended for CNN, which aims to reduce the false positive rate.

**Table 7.** Test evaluation results for DPI's high recall AviaNZ recognisers, tested in AviaNZ. These will be used for CNN training.

| No. | Species | Event-level | | Segment-level | |
|---|---|---|---|---|---|
| | | Recall | Precision | Recall | Precision |
| 1 | Powerful owl | 98.18% | 44.17% | 100.00% | 39.53% |
| 2 | Barking owl | 93.49% | 34.71% | 100.00% | 32.15% |
| 3 | Southern boobook | 60.37% | 83.11% | 92.00% | 35.38% |
| 4 | Yellow-bellied glider | 90.43% | 56.05% | 99.00% | 35.74% |
| 5 | Sugar glider | 88.12% | 88.90% | 100.00% | 37.74% |
| 6 | Squirrel glider | 91.65% | 48.62% | 100.00% | 33.56% |

CNN training for the yellow-bellied glider was successful and the CNN recogniser was tested at event-level, segment-level and on real-world files. Recall for the CNN was 24.41% and 20.00% lower at event-level and segment-level (Table 8), respectively, than the CNN wavelet filter (Table 7). But precision increased by 19.4% and 44.06% at event-level and segment-level, respectively. A further real-world test showed that precision for the yellow-bellied glider CNN recogniser was 100.00%, recording no false positives. Recall did drop from segment-level to real-world testing for the CNN recogniser, but was likely due to faint calls being present in the real-world data and these were difficult for all AviaNZ recognisers to detect. Overall, this is a positive result for the CNN training and shows that it can be used as a method to increase precision.

CNN training for the powerful owl was also undertaken and the CNN recogniser was tested at event-level, segment-level and on real-world files. This CNN recogniser was less successful than the yellow-bellied glider CNN as recall decreased at event-level and segment-level. However, the precision did increase from 39.53% to 53.85% at segment-level, but a higher increase in precision is what would be expected from CNN training. Real-world recall was good, but precision was poor. Further CNN training is required to improve this CNN recogniser further.

**Table 8.** Test evaluation results for DPI's AviaNZ CNN recogniser, tested in AviaNZ.

| Type | Species | Event-level | | Segment-level | | Real-world | |
|---|---|---|---|---|---|---|---|
| | | Recall | Precision | Recall | Precision | Recall | Precision |
| CNN | Yellow-bellied glider | 66.02% | 75.45% | 79.00% | 79.80% | 50.00% | 100.00% |
| CNN | Powerful owl | 61.41% | 41.27% | 77.00% | 53.85% | 85.06% | 32.03% |

*QUT Ecoacoustics Analysis Programs (AP)*

Overall, QUT's AP recognisers produced similar results to the AviaNZ recognisers (**Table *9***). At an event-level, the precision was slightly lower, but was slightly higher than the AviaNZ recognisers at the segment-level. Species that produced low segment-level precision were the same species that produced low precision in the AviaNZ recognisers - barking owl, boobook, sugar glider and squirrel glider. Recall at segment-level was high with recall ≥75% for all species except powerful owl (61%). Real-world results for most species differed from segment-level testing, particularly for precision. Overall, recall was good although it did drop for powerful owl, boobook, yellow-bellied glider and sooty owl and slightly for bell miner. Precision for powerful owl and bell miner was excellent, but – similar to the AviaNZ recognisers – was low for low frequency calling species barking owl, sugar glider and squirrel glider. Sooty owl – again, same as AviaNZ – was particularly low. A quick scan of the files showed lots of insect noise which may have been a factor. Except for bell miner, other diurnal bird recognisers had low precision which is expected for species that call in the dawn chorus.

**Table 9.** Test evaluation results for QUT's AP recognisers, tested in Egret. Event-level: recogniser performance at level of every event (hit). Segment-level and real-world: recogniser performance at detecting an event anywhere within a 30sec file. Bold and underlined real-world results were deemed good results.

| No. | Species | Event-level | | Segment-level | | Real-world | |
|---|---|---|---|---|---|---|---|
| | | Recall | Precision | Recall | Precision | Recall | Precision |
| 1 | Powerful owl | 47.44% | 78.31% | 61.00% | 81.33% | **<u>35.56%</u>** | **<u>100.00%</u>** |
| 2 | Barking owl | 76.05% | 17.37% | 97.00% | 44.09% | 94.12% | 18.93% |
| 3 | Southern boobook | 92.89% | 43.36% | 97.00% | 49.74% | **<u>79.66%</u>** | **<u>58.02%</u>** |
| 4 | Yellow-bellied glider | 37.14% | 21.76% | 76.00% | 52.78% | 19.05% | 40.00% |
| 5 | Grey-headed flying-fox | 61.36% | 45.03% | 98.99% | 59.39% | **<u>96.03%</u>** | **<u>42.16%</u>** |
| 6 | Greater sooty owl | 52.29% | 43.91% | 85.71% | 52.75% | 50.00% | 4.31% |
| 7 | Masked owl | 58.25% | 29.56% | 93.00% | 59.62% | 95.35% | 15.36% |
| 8 | Sugar glider | 80.57% | 68.66% | 99.00% | 46.70% | 100.00% | 19.01% |
| 9 | Squirrel glider | 74.05% | 34.10% | 98.02% | 46.70% | 83.87% | 10.08% |
| 10 | Glossy black cockatoo | 39.45% | 40.04% | 75.00% | 63.56% | 79.31% | 10.70% |
| 11 | Brown treecreeper | 65.14% | 37.40% | 100.00% | 54.95% | 100.00% | 35.67% |
| 12 | Rufous scrub-bird | 20.99% | 56.67% | 96.00% | 84.21% | 100.00% | 9.06% |
| 13 | Bell miner | 34.80% | 84.02% | 94.00% | 68.12% | **<u>85.51%</u>** | **<u>91.47%</u>** |

## 6. Discussion

Overall, 13 successfully built recognisers in both AviaNZ and AP (QUT) as well as an additional CNN recogniser and another underway is a major achievement in a field which is still considered relatively new. The process included sourcing and selecting training and testing data for each species, manually annotating all calls (25,918 calls in total across all species training and testing data and CNN), hours of testing wavelet filters and undergoing various iterations of CNN. AviaNZ software is new and DPI is one of its first users with features being updated throughout the project.

Performance of recognisers was highly variable depending on species and the method for testing performance. Some variation was also evident between the two approaches to recogniser development. Comparing approaches, AviaNZ does not pick up faint calls very well. In future, to help counteract this we could have separate 'call types' with louder and fainter calls, however, this would increase the run time for the recogniser. Quiet calls are common and most recognition technologies will struggle to detect such events. One approach is to ignore faint calls, but AP caters for quiet calls in two ways:

- With different 'profiles' within the same recogniser using different settings that cater for the altered spectral shapes of attenuated calls. The 'profiles' mechanism can also be used by AP to encode regional variation into a species recogniser.
- With arrays of decibel detection thresholds in each detection algorithm. The lowest or highest decibel detection threshold can be used depending on the need and these allow for differing faint or loud call detections.

Recogniser results for each species differed between the level of testing which makes it difficult to gauge how well they performed. Real-world testing is probably the most reliable however, three hours of data is a modest sample size. For this, we were constrained by the time taken to annotate all calls within the three hours. For example, just two hours of real-world testing files for bell miner consisted of 3691 annotations. The sooty owl recogniser, for example, performed particularly poorly in real-world tests (potentially due to insect noise) despite doing well at event-level and segment-level. Additionally, with unbalanced data sets, recall and precision cannot be directly compared between recognisers and results should only be compared by species. This is because the relative ratio of the positive and negative affect the interpretation of scores that measure the change in one class compared to the sum of both classes. For example, with many more negatives than positives, the chance of getting a false positive increases, since the false positive class has more of an effect on the result. There are metrics that can be used to compare imbalanced datasets fairly (ROC curves are a good way to visualise such results), however the simplest approach is to be cognizant of the number of positive and negative classes when comparing recall and precision across recognisers.

Typically pattern recognition tasks as used by QUT are trained on negative examples as well as positive target species. This is done so that potential false positives can be excluded before the recogniser is used on real data. While this project did include a significant number of negative test cases, they were only used in evaluation of test datasets. This means false-positives in the test dataset could not be considered for training and recognisers could not be changed (fairly at least) to exclude any false positive.

In AviaNZ, users are restricted in what changes can be made to the recogniser, with a few parameters available for adjustment (e.g., minimum and maximum frequency, average syllable length, etc.), this was created to make it easy and user friendly for wider bio acoustics community so that they can train their own recognisers for any target species of interest. However, AviaNZ provides capability to train CNNs, which can potentially improve precision. Initial CNN training of test species is underway and has so far produced mixed results. There has been difficulty in training powerful owl calls, but more success with yellow-bellied glider calls. Some key points we have learnt during the process:

- Deciding on a 'balanced' dataset of target species calls and false positives can be tricky and may take some time to adjust.
- Batch process a diverse dataset composed with selected recordings from multiple sources / areas. Identify key false positives in wavelet filter training so they can be used in CNN training instead of using a suite of calls that may only be encountered rarely
- Multiple CNNs can be created for specific datasets / sites / regions that may contain area/region specific false positives

Although only completed on two species to date, CNN training has proved to be a good addition to the yellow-bellied glider wavelet filter by reducing false positives, with an increase in recogniser precision at segment-level by 44%. As the CNN recogniser is used to scan existing and new datasets, additional false positives that are identified may be included in future iterations of the CNN. There is also potential for increasing the recall by including other yellow-bellied glider calls the recogniser may pick up from future datasets. More training is required for the powerful owl CNN.

*Recommendations*

Testing of both recogniser approaches revealed a high rate of false positives, which limits the routine implementation of these draft recognisers on audio recordings without extensive human validation. Indeed, some human level of validation is considered to be a requirement in the foreseeable future with any automated method. Future work should focus on improving recognisers with configuration and profile targeting and better grouping of multiple syllables into calls or CNN approaches where large datasets are available.

However, human validation with similar false positive rates has successfully been used for passive acoustic surveys and monitoring of koalas (Law et al. 2018) so the potential exists with current recognisers to analyse acoustic data with additional species. For large and extensive data-sets of recordings, species recognisers with a high precision will be most efficient to use. For smaller datasets, some precision can be sacrificed as there will be a smaller volume of calls to validate. Species that call repeatedly such as bell miners and boobooks can be adequately scanned by recognisers with a lower recall rate. Typically for rare or threatened species, recognisers with higher recall would be used, which comes at the cost of lower precision.

Considering the purpose of the recogniser before building is also useful. If the purpose is to identify most calls for a particular analysis (e.g. spatial count), then a high recall is required. Whereas if the aim is to undertake occupancy modelling, a lower recall may be acceptable as a trade-off for fewer false positives, thereby aiming for higher precision. Limiting false positives could also be achieved by avoiding sampling in particular areas (e.g. highways, creeks/dams where insect calling could be high) or to limit sampling to during the peak calling period of the target species e.g. limiting to a couple of hours a night or avoiding dawn and dusk chorus when diurnal birds may be problematic.

*Acknowledgements*

We thank Cassie Thompson of the Natural Resources Commission for overseeing this project and helpfully offering assistance wherever possible. We thank Rohan Bilney, Matt Dobson, Chris Slade, and Lauren Hook for providing additional calls of elusive target species.

## 7. References

Law B.S, Brassil, T., Gonsalves, L., Roe, P., Truskinger, A. & McConville, A. (2018) Passive acoustics and sound recognition provide new insights on status and resilience of an iconic endangered marsupial (koala *Phascolarctos cinereus*) to timber harvesting. *PLoS ONE* 13(10): e0205075. https://doi.org/10.1371/journal.pone.0205075

Marsland, S., Priyadarshani, N., Juodakis, J. & Listanti, V. (2020). *The AviaNZ Program*. Version 3.1. [Computer software]. http://www.avianz.net/

Towsey, M., Truskinger, A., Cottman-Fields, M. & Roe, P. (2020). *Ecoacoustics Audio Analysis Software.* Version 20.11.2.0. [Computer software]. http://doi.org/10.5281/zenodo.4274299

Truskinger, A. (2020). *Egret.* Version 1. [Computer software].

Wildlife Acoustics Inc. (2017). *Kaleidoscope.* Version 4.2.0. [Computer software]. https://www.wildlifeacoustics.com