Machine learning algorithms to classify bird calls from acoustic data

Michael Charleston, Scott Whitemore, Greg Jordan and Sue Baker
School of Natural Sciences, University of Tasmania

Our work on this project has significantly improved solutions to the problem of identifying the presence of bird species in forest regions where large-scale manned surveys are not practicable. Such regions can be inexpensively monitored by placing audio recorders to collect passive (non-targeted) environmental recordings for analysis later. We have created a rapidly trainable, high-quality classifier using Machine Learning that can determine with significant confidence which birds are calling (vocalising) in an environmental audio recording.

Summary

This ambitious project aimed to create a bird call recogniser, a computer program that, given passive environmental audio recordings obtained with recording units positioned in forest areas, could aid the identification of which bird species were vocalising during that recording.

The research for this project was largely undertaken by Mr Scott Whitemore as his PhD under the supervision of Professors Michael Charleston and Greg Jordan and Dr Sue Baker at the University of Tasmania. The achievements of this project can be roughly divided into four major outcomes:

1. In collaboration with IT students at UTAS, we created a web-based platform ("BirdSong") that allows user to annotate audio recordings in both time and frequency with the species of bird(s) vocalising. This was critical to the creation of our training data (below).
   url: `https://birdsong.ecoacoustics.science/`.

2. Created two world-class labelled acoustic datasets produced from surveys performed in North-west Tasmania and Victorian Central Highlands, provided by STT and VF.

3. Developed and tested highly accurate machine learning recognisers, optimised for these two data sets. To our knowledge few, if any, recognisers can do as well.

4. Developed a deployment option (via the Docker platform) that allows end-users to apply the recognizer to their own data.
   url: `https://github.com/scottwhitemore/UTAS_Project110300_BirdNet`

Overview of the Project

The importance of avian species within the forest context is widely appreciated: in addition to their own biodiversity values, birds are frequently used as measures of the health and biodiversity of many terrestrial ecosystems, and forests typically house a wide variety of bird species. Birds are a popular subject for monitoring: they contribute to important ecosystem functions, their populations respond rapidly and measurably to environmental changes and, being highly vocal, are relatively easy to detect. However, the detection of bird species in the field is a highly specialised task, generally requiring expert ornithologists to be present at regular surveys. This practice is extremely expensive and cannot scale up to exploit the thousands of hours of environmental audio recordings that are now available. In order to derive information about avian community composition and behaviour we saw a need to develop an automated system to estimate which species are present across sampled spatial and temporal locations.

Environmental recordings are made with passive acoustic recording units: these are easily deployed for extended time periods of up to several weeks, including night-time recording for nocturnal birds, and require

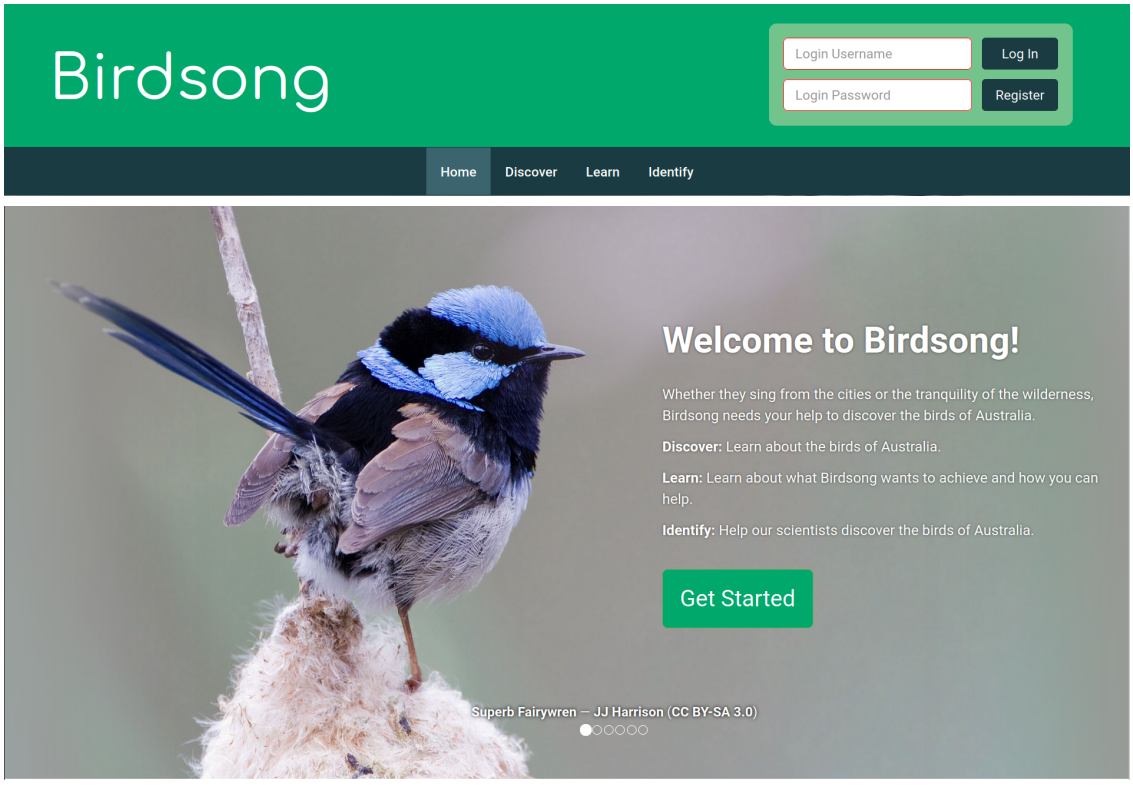


Figure 1: Birdsong banner and example landing page image

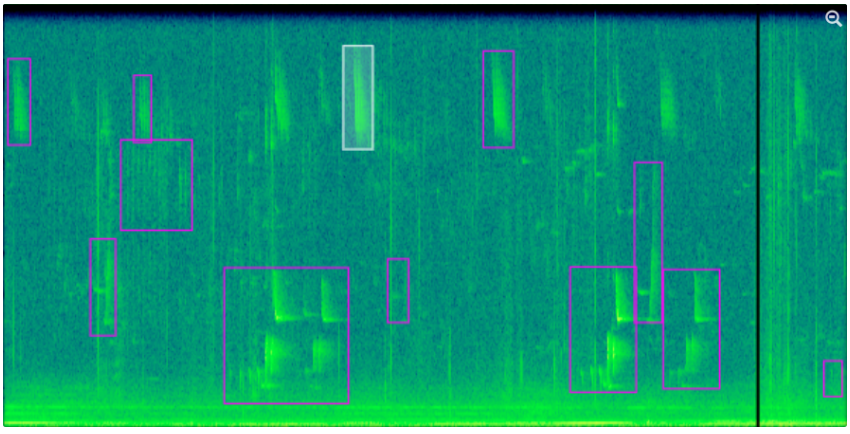


Figure 2: Tagged spectrogram such as used for producing our training data

little maintenance other than retrieval of solid state (SD) cards on which recordings are made, and changing standard batteries. However, current software for acoustic bird recognition is unable to recognise the majority of forest species, greatly limiting the functionality of this approach. While there have been significant advances made in directed recordings – where there is a specific bird species 'targeted' by a recording exercise and other sounds minimised – our task involved the significantly more difficult task of using non-targeted recordings in which many bird and other sounds occur in the same recording. This project has used state-of-the-art machine learning to develop species recognition algorithms which achieve our goal.

Stage 1

Problem

We approached bird call recognition as a *supervised learning problem* (a statistical problem of classifying input, using training data): as such we considered the fundamental components of the project to be: *training data* (that is, labelled examples of bird vocalisations), a *learning machine* (called a *classifier*) and some meaningful *assessment criteria*. Our problem could then be broken down to the following:

1. Acquire training data
2. Build a learning machine
3. Determine an appropriate scoring function

Thereby we sought to combine all of the above to get a performance-evaluated bird call recogniser. This recogniser would use audio signals in common `WAV` format as input, convert them to frequency spectra (akin to images where the x-axis is time, the y-axis is frequency, and the pixel color indicates intensity or volume), and return probabilities of each bird species being present vocalising across two-second segments of the recording.

Solution

1. Sustainable Timber Tasmania (STT; then Forestry Tasmania) supplied audio files from a survey project in NW Tasmania. Andrew Hingston was employed using external funds to produce gold standard annotations, resulting in approximately 10000 instances of calls from 36 species (27 having sufficient representation for learning and evaluation) across 950 minutes of recording.
2. Our team built the "Birdsong" website to enable tagging of spectrograms with bird species (Figure 1).
3. Based on the relevant literature and results from earlier machine learning competitions, a "template matching with bagged decision trees"-type recogniser design was selected and implemented in the Python programming language.
4. The industry standard at this stage to gauge performance was *precision* (accompanied by *recall*; both of these terms are defined below in the Appendix), F-score (that is, the harmonic mean of precision and recall) or Area Under the Curve (AUC). At this stage we found the argument for the use of *informedness* (unbiased *recall*) to be the most persuasive, and adopted that as our measure.

Presentation

Results from initial experiments, including some breakdown by species and a look at common cases of confusion, were presented at the **FPA 2018 Research Update**.

Stage 2

Problem

1. We also required training data from Victorian forests (VF).

2. The previous recogniser design, while functional and having certain advantages, had a low ceiling for possible improvement and was computationally inefficient. Hence we turned to "Deep Learning," referring to large Neural Networks (NNs), approaches to recognition. NNs were becoming more popular, have a higher performance ceiling and encode complex semantics much more efficiently. The downside was that deep learning is much trickier to get working, and comes with a host of challenges that have to be addressed, including:

   (a) Design of neural network architecture: finding the right architecture is something of a "black art".

   (b) Design of training protocol: there are many design decisions and parameters governing how well the classifier "learns" and how generalisable it is from training to testing.

   (c) Sensitivity to data imbalance: e.g., with rare *vs* common instances of bird species, the ability to classify the rare birds can be swamped by classifying the common ones, regarding the rare birds as "noise".

3. High *informedness* values were not correlated with high *markedness* on a per-species basis. In this case the markedness (and likewise precision, see Appendix) was low, resulting from the recogniser being too liberal in predicting the (positive) presence of species. This is both undesirable behaviour and likely indicative of the recogniser being poorly calibrated (predicted probabilities do not match the actual probability that a species is present). Without addressing calibration explicitly, we needed a measure that penalises this informedness-markedness mismatch.

Solution

1. VicForests supplied audio data from Victorian Central Highlands. Kerryn Herman was employed by VicForests to produce annotations, resulting in approximately 1400 instances across 40 species (19 above threshold were suitable for training) present in 70 mins of recordings.

2. (a) We adopted the ResNet50 architecture, a popular NN design, because it is stable and effective across multiple domains. Applying to bird call recognition required custom input and output layers.

   (b) We used pre-trained weights from a different problem domain to reduce the required duration of a fixed training schedule.

   (c) To ensure performance is maintained from training to testing we employed basic data augmentation, including time and pitch shifting.

   (d) For a combination of reasons we opted to tackle data imbalance with *resampling* – whereby input data is sampled with replacement in such a way as to reduce imbalance in what the machine 'sees'. Standard resampling algorithms exhibit undesirable behaviour when labels are correlated (in our situation, where species co-occur more or less than expected by chance if they were independent), so we created a novel algorithm that uses a linear algebra solver that avoids such behaviour.

3. A natural way to penalise the informedness-markedness mismatch was to take their (geometric) mean, which results in the conservative Matthews Correlation Coefficient (MCC). As a measure of classification performance, MCC is well regarded by statisticians and the machine learning community, and yet is relatively uncommon in the scientific literature.

Presentation

Results from initial experiments, including comparisons of resampling algorithms and various model parameters, were presented at the **Ecological Society of Australia 2019 Conference**.

Stage 3

Problem

1. We sought even better performance, since we knew the Deep Learning approach could achieve this.

2. Whilst we stand by our previous use of MCC, we wanted a more sophisticated approach that would include consideration of model calibration, in order to be able to assess whether the predicted probabilities were really reliable.

3. Deployment: we had always wanted the classifiers based on both NW Tasmanian and VicForest data to be available to our sponsors, and we sought to provide an accessible tool for this purpose.

Solution

1. We employed advanced neural network architectures (called "EfficientNets"), training policies and abstract data augmentation (creating synthetic samples) to improve performance without significantly increasing training time or model size.

2. We conclude our investigation of performance measures with the following remarks:

    - Calibration refers to the reliability of a model's predictions; i.e., a prediction of 90% probability of species A being present should result in species A being present 90% of the time. Calibration is both inherently valuable for the end-user and also satisfies requirements for certain desirable behaviours of other performance measures. The state-of-the-art technique to judge whether or not a model is calibrated is to conduct a *hypothesis test*. Unfortunately at this stage assessing the "calibrated-ness" of a model is confounded by the data imbalance problem, and can only be applied to more common species.
    - Target (importance) weighting. All of the measures we have mentioned, and all of the others we haven't mentioned, assume that all outcomes are equivalent, e.g., identifying that a forest raven is calling is just as valuable as identifying that a superb fairy-wren is not calling. For the end users of our recognisers, this equivalency is almost certainly not valid. This issue remains a challenge with no obvious solution: while *applying* relative weights to the different bird species' classification is practicable, the determination of what those weights should *be* is not.

3. We opted on deployment via a Docker Image – see next Section.

Presentation Scott prepared a video demonstrating the recogniser, which was played at the 2020 FPA Research Update, and a working (pre-Docker) deployment recogniser was demonstrated live to stakeholders.

Completed Product

We have created world-class bird call recognisers for Northwest Tasmanian and Victorian Central Highland acoustic surveys. Our recognisers can identify many of the bird species found in these regions with high accuracy. Further performance details are below.

Results: North-West Tasmania

The NW Tasmania data includes labels for 38 different classes that have been indicated as being present in the recordings. The hierarchical class "Thornbills and Scrubwrens" had to be split and the resulting ambiguity necessitated the collapse of "Tasmanian Thornbill" and "Brown Thornbill" into a single "Thornbill" class. The classes "Other", "Anuran", "Brown Falcon", "Yellow-tailed black cockatoo", "Grey butcherbird" and "White-breasted sea eagle" were removed as they did not have sufficient representation for performance evaluation. A further 3 classes ("Superb fairy-wren", "Black-faced cuckoo-shrike" and "Dusky robin") were removed due to extreme performance variation across splits. The remaining 27 classes were deemed to have sufficient representation for learning and evaluation.

Table 1 shows the precision and recall for the NW Tasmania data. Precision and Recall are described in more detail later, but *precision* can be thought of as the proportion of correct positive predictions out of all the positive predictions, and *recall* can be thought of as the proportion of correct positive predictions out of all the positive occurrences in the data.

Table 1: Precision and Recall of the recogniser for NW Tasmanian birds. Precision and Recall values of 0.9 or more are in **bold**. Count is the number of unique annotations for each bird.

Species	Count	Precision	Recall
Thornbill	2010	0.716	0.698
Striated pardelote	1160	**0.936**	0.638
Grey shrike-thrush	1142	0.788	0.754
Grey fantail	1111	0.829	0.766
Golden whistler	930	0.813	0.771
Crescent honeyeater	808	0.833	0.641
Forest raven	554	0.875	0.851
Tasmanian scrubwren	259	0.776	0.352
Green rosella	253	0.881	0.673
Pink robin	227	0.814	0.729
Fan-tailed cuckoo	175	0.843	0.754
Silvereye	171	0.789	0.882
Black-headed honeyeater	139	0.743	0.743
Black currawong	138	0.537	0.725
Yellow-throated honeywater	135	0.618	0.477
Common blackbird	122	0.606	0.571
Flame robin	83	0.85	0.567
Sulphur-crested cockatoo	72	0.852	**1**
Strong-billed honeyeater	68	**0.923**	0.461
Australian shelduck	63	**0.955**	0.7
Eastern spinebill	56	0.76	0.704
Olive whistler	56	0.846	0.55
Laughing kookaburra	46	**0.944**	0.548
Scrubtit	22	**1**	0.286
Shining-bronze cuckoo	18	**1**	0.625
Insect	79	0.48	0.387
Human	29	**1**	0.789

Results: VicForests

The VicForests data includes labels for 40 different classes that have been indicated as being present in the recordings. Considering that we have an order of magnitude fewer labels for this dataset, the VicForests data exhibits much greater species diversity. Unfortunately this means that 21 of the 40 classes do not have sufficient representation for learning and evaluation. Where with the NW Tasmania data we excluded three species for exhibiting extreme performance variance, this behaviour is seen for **all** VicForests species. As such, although we are able to do some learning and evaluation, we would like more confidence that the results that follow are meaningful for the application of the recognisers to data beyond those on which they was trained, e.g., from the same region at a different location or time period. This is a limitation that could potentially be addressed with more expert annotations.

Table 2 provides measures of identification accuracy for the Victorian data.

Deployment

We are currently able to deploy our recognisers in two forms: python script and docker image.

Table 2: Precision and Recall of the recogniser for Victorian birds.

Species	Count	Precision	Recall
Grey fantail	241	**0.941**	0.525
Golden whistler	223	0.533	0.145
Spotted pardalote	130	0.333	0.095
Brown thornbill	90	**1**	0.267
Grey currawong	62	**1**	0.692
Fan-tailed cuckoo	61	**1**	0.476
Eastern yellow robin	47	0.8	0.8
Superb lyrebird	47	**1**	0.5
Crimson rosella	43	0.286	0.25
Pied currawong	42	**1**	0.375
Grey butcherbird	37	0.5	**1**
Brown-headed honeyeater	33	**1**	0.333
Crescent honeyeater	32	0	0
Grey shrike-thrush	27	**1**	0
Whitebrowed scrubwren	26	0	0
Silvereye	21	**1**	0.2
Striated pardelote	21	0.375	0.375
White-throated treecreeper	21	0.857	0.4
Human	91	**0.909**	0.323

Python

The python script can be run from a terminal without any familiarity with the language, but it does require that additional software be installed to meet dependencies (python3 and multiple libraries) and is not guaranteed to work on any particular machine or operating system.

Docker

The docker image only requires that the user have docker installed, and almost surely works consistently across most machines and operating systems.

Details on both methods, and instructions on how to acquire them, are available at `https://github.com/scottwhitemore/UTAS_Project110300_BirdNet`

It is able to operate on a input audio sample in WAV format (which is the same as that used for the passive acoustic recorders) and produce predictions of which bird species are present (at 2 sec resolution) in convenient CSV and JSON formats.

Related Outputs

Birdsong

In collaboration with students and academics at the University of Tasmania and with support of our partners, we created the "**Birdsong**" website[1].

[1] `https://birdsong.ecoacoustics.science`

The Birdsong site accesses hundreds of hours of recordings (many thousands of 30 second samples) from Tasmania and VicForests, via a graphical interface that shows a frequency spectrogram (see Figure 2) for each sample and plays the sample, simultaneously showing where in the sample the sound is playing. The simple interface allows users to highlight rectangles on the spectrogram and mark them as being from any of the bird species in our list (or just "bird").

World-class tagged data

The accuracy of our classifier could not have been accomplished without the world-class training data we have acquired through the Birdsong website and the tagging efforts of expert ornithologists. This data set now comprises 56252 sequences of recordings, to a total of 562266 audio samples; from these, 18924 tags have been applied by 232 users. The bulk of tagging was performed by Andrew Hingston and Kerryn Herman, who are not only human experts at identifying birds, but were also physically present, surveying the locations while many of the recordings were being made. This enabled tags to be applied to the spectrograms (as in Figure 1)

Opportunities

Deployment

Our current means of deployment (as described above) is the fastest way for us to provide immediate access of our recognisers to stakeholders. Below we suggest three more advanced forms of deployment that address particular end-use scenarios, but that would require significant (software) development outside the scope and capability of the current project.

- Cloud-served Web Application – online deployment would enable users to access our recognisers from anywhere that has internet availability without having to install any software locally – and thus would be hardware agnostic. This would also allow us to maintain an up-to-date stable of recognisers as development of new and better recognisers continues. Such a service could become a central location for all recognition needs: data management, procurement of labelling experts, training of new recognisers, analytics and various modelling applications.

- Mobile Application – Often people interested in bird call recognition will not be in locations with reliable internet access, and likely their only device on hand will be a mobile phone. Our recognisers are efficient in terms of computation and memory usage, making them well suited to running on mobile devices. Likely of more interest to non-professionals, mobile deployment may have potential as a component of outreach programs.

- Edge ML – Edge ML refers to deployment of machine learning software on small scale hardware at the point of data collection, i.e., attached to the sensors. In terms of bird call recognition this means having our recognisers installed on low-power devices with audio sensors that will replace the Acoustic Recording Units (ARU) currently used for monitoring. Recognition would be performed in real time and on location; time-stamped species presence recorded instead of audio signal. This would drastically reduce the amount of memory required (factor of 144 under our current construction, assuming 22kHz sample rate and steady species occurrence rate) and the associated cost (varies with cost of storage and labour). This is a vital step as we eventually move from the current practice of stand-alone ARUs with regular maintenance to networks of sensors requiring negligible storage and maintenance; wireless transmission of audio signals is too energy intensive to be used in low power devices.

Future Expansion and Improvement

As has always been the case, the performance of these recognisers is highly dependent on the relevance, quality and quantity of training data available. While we do stress that more data will be required as the recognition task advances from engineering to application, recent advances suggest a means of reducing this data dependency.

Our current approach to recognition as a supervised learning (SL) problem could be shifted slightly to the adjacent problem of semi-supervised learning (SSL): training with a combination of labelled and unlabelled data. Being a much more difficult problem, advances in SSL have trailed SL somewhat and have garnered

much less attention. This has changed recently, with several big papers reporting state-of-the-art classification performance on benchmarks traditionally untouched by SSL. The methods used still require a significant amount of labelled data, but the ability to leverage unlabelled data to outperform the existing SL methods is promising. It is almost certain that in the future some form of SSL will be critical to expanding the use of automatic bird call recognisers, as it might ameliorate the critical dependency on *quantity of* training data.

Acknowledgements

Funding: Australian Research Council Linkage grant LP140100075; **Partner organisations**: Sustainable Timber Tasmania, VicForests, Tasmanian Land Conservancy, NRM South, Forest Practices Authority; University of Tasmania.

Tim Wardlaw guided project conception. *Marie Yee* at STT and *Lyz Pryde* at VicForests facilitated deployment of acoustic recorders.

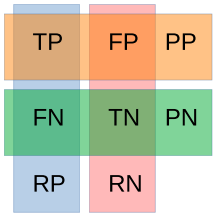
UTAS ICT students and staff, including *Nick Forbes-Smith*, *Saurabh Garg*, *James Montgomery* and undergraduate students for building the Birdsong website.

Andrew Hingston, *Kerryn Herman* and numerous other citizen scientists for tagging the bird calls.

Appendix

We include a few important details of terminology here.

- **Precision and Recall**
  We imagine a population of data in which an item, e.g., a bird species calling, is present or absent in each datum (e.g., a recording): we refer to the number of these items as Real Positive (RP) and Real Negative (RN) respectively. Our classifier will make a prediction for each datum as to whether that bird is present or absent, and ideally the classifier will predict a 'positive' whenever there is one (the bird is predicted to be present, and it was), and a 'negative' wherever there is one (the bird is not present, and the classifier agrees). The total number of predicted positives (PP) should, ideally, match the number of real positives (RP); similarly the total number of predicted negatives (PN) should match the number of real negatives (RN).

  Note that PP = TP + FP, PN = FN +PN, RP=TP+FN, and RN=FP+TN.

  Recall is the proportion of all the predicted positives that are correct; that is, Recall = TP/RP.

  Precision is the proportion of real positives that the classifier correctly identifies: Precision = TP/PP.

- **Informedness and Markedness**

  These are alternatives to recall and precision that are adjusted for bias when PP $\neq$ RP $\neq \frac{N}{2}$. Informedness is a measure of how informed the classifier is about both positives and negatives and is given by TP/RP - FP/RN (recall - inverse recall). Markedness is a measure of the trustworthiness of both positive and negative predictions, and is given by TP/PP - FN/PN (precision - inverse precision).

- **Calibration**:
  A model is *calibrated* when the proportion of times a presence or absence is called by the classifier matches the true presence/absence prooportions in the testing data. For example if currawongs are present in 30% of the testing data, we would want the classifier to predict currawongs being present 30% of the time. We use a hypothesis test by resampling from the data to give a p-value. The p-value of this hypothesis test tells us not how good we are at discriminating (between presence/absence of a bird) we are, but rather, the value produced by the recogniser (sometimes referred to as a confidence score) is an accurate estimate of our uncertainty. It is well known that a calibrated model does perform better on new data than an uncalibrated one: it is more reliable.