

Monitoring Frog Communities: An Application of Machine Learning

Andrew Taylor

Computer Science and Engineering
University of New South Wales
andrewt@cse.unsw.edu.au

Graeme Watson

Zoology
University of Melbourne
Graeme.Watson@muwayf.unimelb.edu.au

Gordon Grigg and Hamish McCallum

Zoology, University of Queensland
{ggrigg, HMccallum}@zoology.uq.edu.au

Abstract

Automatic recognition of animal vocalisations would be a valuable tool for a variety of biological research and environmental monitoring applications. We report the development of a software system which can recognise the vocalisations of 22 species of frogs which occur in an area of northern Australia. This software system will be used in unattended operation to monitor the effect on frog populations of the introduced Cane Toad.

The system is based around classification of local peaks in the spectrogram of the audio signal using Quinlan's machine learning system, C4.5 (Quinlan 1993). Unreliable identifications of peaks are aggregated together using a hierarchical structure of segments based on the typical temporal vocalisation species' patterns. This produces robust system performance.

Problem Description

Since the unfortunate introduction of the Cane Toad (*Bufo marinus*) to Australia, its abundance and continuing spread through northern Australia have been the cause of considerable concern. It is a voracious predator taking a wide range of prey. Cane Toads also possess poison glands which can kill unwary animals which attempt to prey on them. Although there is great public alarm at the effect Cane Toads are having or will have on Australia's native fauna, there is actually no conclusive data available establishing a detrimental effect on the population of a native species.

The reason is that censusing populations of most of Australia's native fauna is a difficult and expensive undertaking. Biologists have been unable to collect sufficient suitable data to properly address the question of the Cane Toad's impact on native fauna. This is very unfortunate as it makes it difficult to determine the appropriateness, size and nature of efforts to control the Cane Toad.

The censusing of animals which make frequent distinctive vocalisations is more tractable but it is still

expensive and time consuming. Australia's species of native frogs are one of the groups most likely to be affected by Cane Toads and can be aurally censused. However there are a number of problems with such an approach.

The most desirable study location is an area in front of the advancing Cane Toads, allowing frog population censuses to be obtained before and after the Cane Toad's arrival. Unfortunately the main Cane Toad front is currently in a remote area of Australia's Northern Territory. This is an area whose weather is dominated by a wet season of monsoonal rains. Most of the frog species of the area are only active during this wet season so censuses must be conducted during this time.

Field work during the wet season is difficult as roads are often impassable and extreme heat and humidity is combined with intense irregular storms. The activity of most species is irregular, depending primarily on rainfall. Some species are active only during particular parts of the wet season. As a result short field trips to these areas are not a reliable method of censusing the frog populations. Continuous manual censusing through the wet season is not feasible so we have developed automatic methods which will be used to census frog populations at a number of sites through the wet season.

There are 22 frog species present in our study area. Their vocalisations range in length from less than 20 milliseconds to over a second. An example of the spectrogram of the vocalisation of a single frog can be seen in Figure 1.

Some species repeat their vocalisations incessantly, other species usually make only occasional isolated vocalisations. Many of the species tend to call in choruses with hundreds of individuals from a number of species present. There is also considerable background noise from insects, some species of which have vocalisations somewhat similar to some frog species.

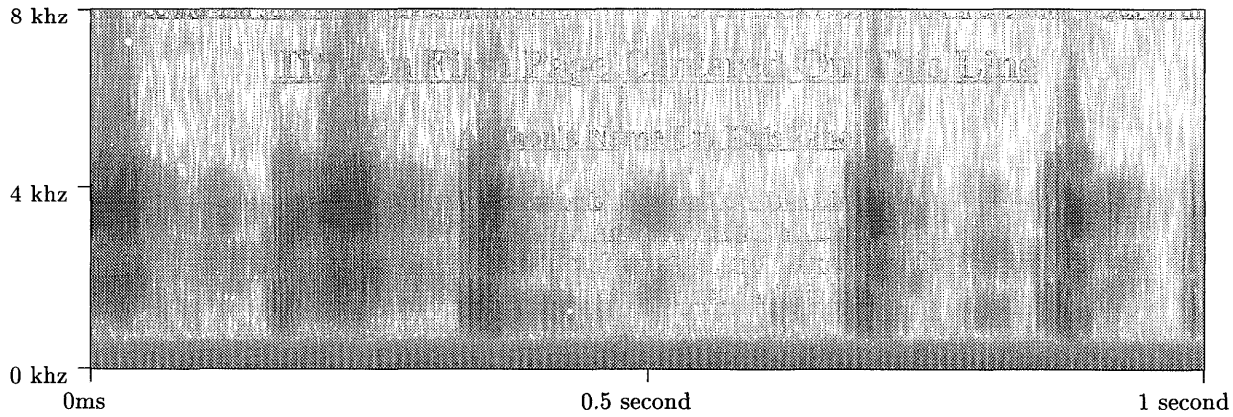


Figure 1: Spectrogram of a *Litoria nasuta* individual

Rain is another source of background noise which can not be ignored as some frog species call primarily during rain. An example of the spectrogram of a frog chorus recorded at our study site can be seen in Figure 2. The vocalisations of at least 11 individuals of 6 species of frog are apparent in this one second spectrogram. There is also noise from at least 3 species of insects present in the spectrogram. Figure 3 contains the spectrogram of a chorus with a similar species composition to Figure 2 but with a much larger number of individuals calling.

Most of the frog vocalisations function as an advertisement to other members of the same species and hence have evolved to be species-specific. Experiments on other frog species have shown a variety of properties can be used by frog species to recognise the vocalisations of their own species (Gerhardt 1988). These include call rate, call duration, amplitude-time envelope, waveform periodicity, pulse-repetition rate, frequency modulation, frequency and spectral patterns.

The vocalisations of some species have stereotypical properties which are apparently not used by members of that species for recognition (Gerhardt 1988). Little is known in these respects of the frog species in our study area. In any case the properties suitable for the wetware of a frog's hearing system and brains may not be the the best properties for our software and audio hardware.

Application Description

Automatic recognition of animal vocalisations would be valuable for a variety of biological research and environmental monitoring applications but it is an area which has seen little work and only preliminary results produced (Mills 1995), (Fristrup & Watkins 1995), (Taylor 1995).

One animal is an exception. There is a huge body of work devoted to distinguishing the complex vocalisations of *Homo sapiens*. This is usually termed speech recognition.

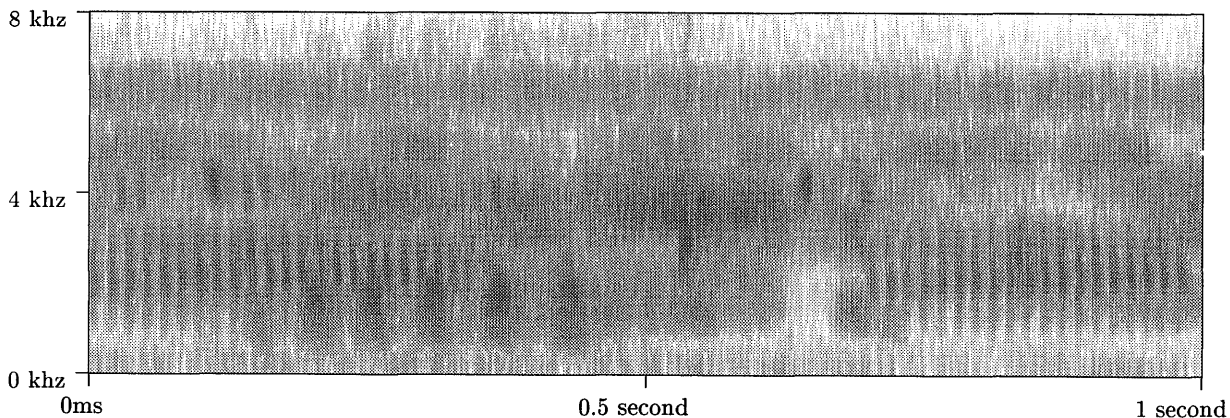


Figure 2: Spectrogram of a Frog Chorus

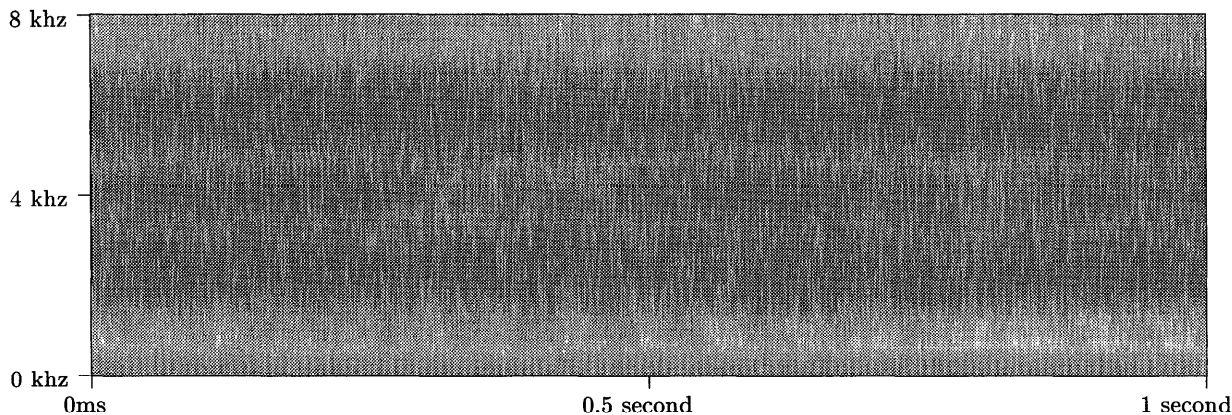


Figure 3: Spectrogram of a Frog Chorus

The frog vocalisations we wish to recognise are much simpler than those of humans. Their recognition would be an easy problem if it was conducted under similar conditions to that of most successfully deployed speech recognition systems: a single cooperative individual close to the microphone in a quiet environment (Deller, Proakis & Hansen 1993).

None of these conditions are met in our problem domain. Instead we must recognise simpler vocalisations but under much more difficult conditions.

Attributes

This has led us to adopt a different and simpler approach to that typically employed in speech recognition. Our system makes no attempt to segment or isolate individual vocalisations. It works entirely from the spectrogram of the incoming audio signal. A Fast Fourier Transform is used to produce a spectrogram of the signal with time-frequency pixels which are roughly 1 millisecond by 50 hertz.

Each time slice of the spectrogram is examined for pixels which contain more energy than any nearby (in frequency) pixels in the same time slice. There may be zero or more local peaks in a single time slice. If there are also local peaks at similar frequencies in several preceding and succeeding time slices then it is assumed the peak is part of a vocalisation and it is passed to the next stage of the system to be individually classified. Figure 4 contains a call with the local peaks marked.

Our system will examine each of the 40 local peaks in Figure 4 individually and classify it as belonging to a particular species. Information from the spectrogram surrounding the peak is used to construct attributes for classification. The information used includes the frequency of the peak, the relative frequency of nearby peaks in preceding and succeeding time slices and the relative height of pixels nearby in the same time slice and in preceding and succeeding time slices. There are a great number of way attributes might be constructed from this information.

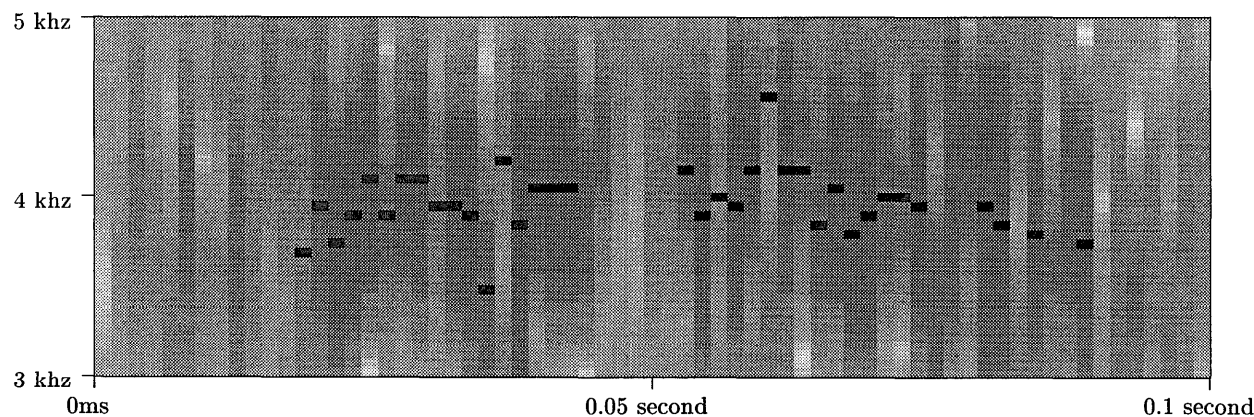


Figure 4: Local peaks of a *Litroia inermis* Call

During development we constructed a set of approximately 70 possible attributes and used a greedy search similar to what (John, Kohavi & Pfleger 1994) term *forward selection* to choose a subset of 15 of the attributes for the system to employ.

Training

Quinlan's machine learning system, C4.5 (Quinlan 1993), is used to construct the classifier. C4.5 is supervised learning system which, given a set of classified cases and a number of attributes for each case as training data, produces a decision tree to classify further cases. The training data for C4.5 was extracted from vocalisations of each of the 22 species in our study areas. These vocalisations were high quality recordings of single individuals. These had been gathered in previous biological research and were not from our study area. A number of vocalisations were selected manually from each recording for use in training. This ensured only vocalisations from the required species were present in each piece of training data. These vocalisations totalled 5 to 20 seconds of sound for each species.

We also introduced as training data sounds from cricket species which occur in our study area. These have similar qualities to frog vocalisations and their explicit classification improved system performance.

The decision tree produced by C4.5 has approximately 5000 nodes. Here is a small fragment of the tree.

```
vert2 <= 18:
|  freq-4 <= -8: Uperoleia lithomoda
|  freq-4 > -8:
|  |  verta+vertb-vert <= 7: Litoria bicolor
|  |  verta+vertb-vert > 7 :
|  |  |  timef+4 <= 0: Uperoleia lithomoda
|  |  |  timef+4 > 0: Litoria caerulea
vert2 > 18 :
|  horiz <= -50: Litoria tornieri
|  horiz > -50: Uperoleia inundata
```

We automatically translate the decision tree to data suitable for inclusion in a C program.

The identifications of individual peaks is, of course, unreliable as effectively only a tiny fragment of sound is being examined. The error rate in local peak classifications approaches 50%.

Voting

Our system aggregates these unreliable identifications to produce reliable recognition of vocalisations. The model used is simple. If within a certain time period a threshold number of local peaks are identified as be-

longing to a given species then a vocalisation of that species is assumed to be present.

The obvious time period to choose is the typical length of the species' vocalisations. We could not obtain suitable system performance by attempting to recognise vocalisations within the typical period of some species' vocalisation, mainly because some species have very short vocalisations. We remedied this by adopting a hierarchical structure of time segments based on the typical temporal patterns of the species vocalisations.

For example, a species might have a vocalisation typically lasting 300 milliseconds containing a number of 30 millisecond "notes" and it might usually produce 4 or more vocalisations in 3 seconds. Our system models this with 3 levels of segments. The level 0 segments will be 30 milliseconds long. If a threshold number of local peaks occur in that time period then the species is regarded as present in that level 0 segment, in other words we assume we have recognised a single "note" belonging to the species.

The level 1 segment will be 300 milliseconds long. If a threshold number of level 0 segments are identified as containing the species within that time period then the species is regarded as present in the level 1 segment, in other words we assume we have recognised a single vocalisation of the species.

Similarly the level 2 segment will be 3 seconds long and a threshold number of level 1 segments will be required to regard the species as present in the level 2 segment and hence reliably identified.

In practice, it was only necessary to specify the three level hierarchy described above for a few species. For most species, a one or two level hierarchy was sufficient. These hierarchies are based on the typical calling patterns of an individual but, in practice, perform well even when multiple individuals are present. This temporal structure could, in principle, be acquired automatically by the system but, in practice, it was much easier to provide it manually.

It is much more desirable for our system to fail to recognise a vocalisation (a *false negative*) than to incorrectly indicate the vocalisation of a particular species is present (a *false positive*). It is crucial then to choose thresholds such that false positives are unlikely. Recent work in recognising individual humans from their vocalisations, usually termed speaker recognition, has examined similar problems at some length; for an overview see (Gish & Schmidt 1994). We were concerned the assumption involved in this work would not be sufficiently valid in our domain so we instead adopted an empirical process.

It is easy to provide large amounts of training data

which does not contain the vocalisations of a given species. This does require the time consuming manual extraction of vocalisations that positive training data does. The thresholds for a particular species are estimated by applying the local peak classifier to negative training data and examining how often peaks are (incorrectly) classified as belonging to the particular species.

Hardware Platform

Our monitoring stations are to be setup at the start of each wet season and dismantled four months later at the end of each wet season allowing the data to be collected. They use a rugged single board PC-compatible computer designed for industrial applications. The 25mhz Intel 486 CPU allows approximately 25% of incoming sound to be classified, i.e 15 seconds of sound takes approximately 1 minute to process. The speed is limited by the signal processing, not the call recognition. Power is supplied by a solar panel. Every 5 minute period, details of the frog species heard in that interval are logged to flash memory. The monitoring stations also log rainfall, temperature and humidity data. Data in flash memory should survive most modes of system failure.

Application Use

Our system will be deployed at 12 sites next wet season. We can report testing on field data. This wet season we collected 29 recordings of frog choruses in our study area. The recording ranged in length from 3 to over 30 minutes. An inexpensive dynamic microphone, similar to that employed in our stations, was used for these recordings. The recordings were made under as varied conditions as possible. The distance to the nearest frog varied from 2 to 70 metres. Significant amounts of noise are present in some recordings from insects, rain and human speech or other human activity.

At the time of each recording it was noted what species were present and these were placed into two categories: species which were conspicuous to the human ear and species which were not. Some species were placed in the second category because they made only a few isolated vocalisations during the time of the recording, others because their vocalisations were difficult to discern because they were distant or obscured by other species or both. Each recording contains 1 to 10 species. The number of species in each track placed in category 1 varies from 0 to 4.

In total the 29 recordings contain 9 different species which occur on at least one track in category 1. Another 7 species occur in the recordings but only as cat-

egory 2. It was felt successful recognition of the category 1 vocalisations was a minimum requirement for our system.

System Performance

For all 29 recordings our system recognised the category 1 species with two exceptions. One species, *Uperoleia lithomoda* was never recognised. It has extremely brief vocalisations which sound like stones being clicked. A single *Uperoleia lithomoda* vocalisation, centred at 3 khz, can be seen just over 0.5 seconds into the spectrogram in Figure 2. We are currently investigating remedying this.

During our fieldwork collecting the recordings, it became apparent that it was very difficult for humans to separate the vocalisations of 3 of the species in our area. The frogs themselves of these species are also difficult to separate in the field. This makes collection of accurately labelled training and testing data difficult. In particular, even if you positively identify one species as dominating a chorus it is very difficult to exclude the presence of individuals of the other two species.

We have lumped these three species together in our system. Fortunately, this is acceptable for our project's purposes because of ecological similarities between these species. The lumped species was successfully recognised whenever it occurred in category 1.

System performance with category 2 species was varied. In approximately one third of cases category 2 species were recognised as being present in recordings. We are actively examining improving this performance. In one instance, a frog species was recognised as present even though the nearest individual was more than 70 metres distant. This exceeds the performance of the first author.

There were several misidentifications (i.e. *false positives*) of one species and one misidentification of a second species. We have since remedied this by modifying the temporal segments used for recognising these species.

We believe the above results demonstrate that the software component of our system will be successful.

Application Development and Maintenance

Application development has taken approximately two man months spread over 12 months. Some software from previous work was employed. The system could be trained for a new set of frogs in less than 1 day but performance evaluation and tuning would probably require up to a week.

We hope eventually to encapsulate and distribute the training software so biologists can construct their own identification systems. We also hope to apply the techniques to other taxa, including bats and cetaceans.

Acknowledgements

This work was supported by an Australian federal government grant. We thank the local land owners and the Northern Territory government for permitting our work.

References

J.R. Deller, J.G. Proakis, and J.H.L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan, 1993.

K.M. Fristrup and W.A. Watkins. Marine animal sound classification. *Journal of the Acoustical Society of America*, 97(5):3369–3370, May 1995.

H.C. Gerhardt. Acoustic properties used in call recognition by frogs and toads. In Fritzsche et al., editor, *The Evolution of the Amphibian Auditory System*, pages 455–483. John Wiley, 1988.

H. Gish and M. Schmidt. Text-independent speaker identification. *IEEE Signal Processing Magazine*, pages 18–32, October 1994.

G.H. John, R. Kohavi, and K. Pflieger. Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine Learning*, pages 121–129. Morgan Kaufman, 1994.

H. Mills. Automatic detection and classification of nocturnal migrant bird calls. *Journal of the Acoustical Society of America*, 97(5):3370–3371, May 1995.

J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman, 1993.

A.J. Taylor. Bird flight call discrimination using machine learning. *Journal of the Acoustical Society of America*, 97(5):3370–3371, May 1995.