Bioacoustics: The International Journal of Animal Sound and its Recording

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tbio20

A toolbox for animal call recognition

Michael Towsey [a] , Birgit Planitz [a] , Alfredo Nantes [a] , Jason Wimmer [a] & Paul Roe [a]

[a] School of Computer Science, Queensland University of Technology, Queensland, Australia

Version of record first published: 10 Feb 2012.

To cite this article: Michael Towsey , Birgit Planitz , Alfredo Nantes , Jason Wimmer & Paul Roe (2012): A toolbox for animal call recognition, Bioacoustics: The International Journal of Animal Sound and its Recording, 21:2, 107-125

To link to this article: http://dx.doi.org/10.1080/09524622.2011.648753

Bioacoustics
Vol. 21, No. 2, June 2012, 107–125

Taylor & Francis
Taylor & Francis Group

A toolbox for animal call recognition

Michael Towsey*, Birgit Planitz, Alfredo Nantes, Jason Wimmer and Paul Roe

School of Computer Science, Queensland University of Technology, Queensland, Australia

(*Received 27 September 2010; final version received 8 June 2011*)

Monitoring the natural environment is increasingly important as habit degradation and climate change reduce the world's biodiversity. We have developed software tools and applications to assist ecologists with the collection and analysis of acoustic data at large spatial and temporal scales. One of our key objectives is automated animal call recognition, and our approach has three novel attributes. First, we work with raw environmental audio, contaminated by noise and artefacts and containing calls that vary greatly in volume depending on the animal's proximity to the microphone. Second, initial experimentation suggested that no single recognizer could deal with the enormous variety of calls. Therefore, we developed a toolbox of generic recognizers to extract invariant features for each call type. Third, many species are cryptic and offer little data with which to train a recognizer. Many popular machine learning methods require large volumes of training and validation data and considerable time and expertise to prepare. Consequently we adopt bootstrap techniques that can be initiated with little data and refined subsequently. In this paper, we describe our recognition tools and present results for real ecological problems.

Keywords: environmental acoustic analysis; automated animal call recognition; sensor networks

Introduction

The increased availability, power and storage capacity of computing hardware has made it feasible to gather large volumes of audio data for ecological analysis. In addition, enhanced web services have made it possible to bring that audio data directly to the laboratory rather than have ecologists go to the field. However, it is impossible for ecologists to listen to even a small fraction of the audio data made so easily available (Agranat 2009): some degree of automated assistance is essential. In conjunction with ecologists, our laboratory has developed an online service (http://sensor.mquter.qut.edu.au) that offers a variety of tools for the examination and analysis of environmental recordings. We have described various aspects of our sensor network for collecting audio data in previous reports (Lau et al. 2008). In this report, we describe our approach to the problem of automated analysis and, in particular, to automated animal call recognition.

Perhaps due to the importance of birds as indicator species of environmental health, there is already a considerable body of work published on the detection of bird vocalizations (Anderson et al. 1996; McIlraith and Card 1997; Kwan et al. 2004; Chen and Maher 2006; Somervuo et al. 2006; Cai et al. 2007; Juang and Chen 2007; Brandes 2008; Acevedo et al. 2009). A common approach has been to adopt the well-developed tools of Automated Speech Recognition (ASR) but unfortunately it is not so easy to translate ASR to the analysis

*Corresponding author. Email: m.towsey@qut.edu.au

ISSN 0952-4622 print/ISSN 2165-0586 online

http://dx.doi.org/10.1080/09524622.2011.648753
http://www.tandfonline.com

of environmental recordings because there are far fewer constraints in the latter task. Two issues are noise and variability. ASR tasks are typically restricted to environments where noise is tightly constrained, e.g. over a telephone line. By contrast, environmental acoustics contain a wide variety of non-biological noises having a great range of intensities and a variety of animal sounds that have nothing to do with the task at hand. Furthermore, the sources can be located any distance from the microphone which greatly affects the acoustic properties of the recorded sound. Secondly, despite its difficulty, ASR applied to the English language requires the recognition of about 50 phonemes. By contrast, bird calls offer endless variety; variety of call structure between species, variety between populations of the one species and variety within and between individuals of the one population. Many species have multiple calls (in this paper we do not distinguish *calls* from *songs*) and many are mimics. To give some indication of the difficulty of bird call recognition, a state-of-the-art commercial system using an ASR approach that has been under development for more than a decade, achieves, on unseen test vocalisations of 54 species, an average accuracy (defined as the average of precision and recall) of 65–75% (Agranat 2009). This accuracy would not suffice for most ASR applications and is indicative of the difficulty of the unconstrained environmental acoustic problem. Some work has been done on the recognition of acoustic events in an urban setting (auditory scene analysis) but this task suffers from exactly the same difficulties (Cowling and Sitte 2003; Temko et al. 2006; Zhuang et al. 2008).

Our approach to animal call recognition has been dominated by a requirement to produce solutions that are practical for an ecologist who is not necessarily well-versed in the methods of machine-learning. In particular we draw attention to the following features of our work:

(1) Real world data: the results we describe are obtained with real-world recordings that have not been cleaned of artefacts. There is a world of difference between reporting classification results on carefully cleaned data with balanced training and test sets versus the raw recordings to which ecologists actually listen. Constructing a data set containing equal numbers of each call type that have been manually cut from recordings to exclude extraneous noise is not a realistic approximation to the real situation of very uneven class numbers and low call density in arbitrary background noise.
(2) Limited training data: despite the demonstrated accuracy of machine learning methods such as Neural Networks (NN) and Hidden Markov Models (HMM) on standard datasets, these methods do not necessarily adapt well to the real world of environmental recordings. Many bird species are cryptic and the large amount of data required to train a NN or HMM is not immediately available. (ASR does not suffer from this problem.) It is more practical to adopt methods that require just one or a few instances of a call type and bootstrap from there. This approach is particularly effective for species whose calls vary little within and between populations, e.g. the Lewin's Rail (Lau et al. 2008). Another consideration is the practicality of training of multi-class classifiers. Training a 50 bird-call classifier for a given locality may be possible but it is not practical if it must be repeated every time the ecologist wishes to incorporate a new call instance or call class.
(3) Multiple recognition methods: after an initial period where we adapted ASR methods to animal call detection, we came to the conclusion that a one-algorithm-fits-all approach could not deal with the enormous variety of environmental acoustic events. On the other hand, it is also not practical to construct individual recognizers tailor-made for all the calls and other acoustic events that one might

expect in an arbitrary recording of the environment. Consequently we adopted an intermediate approach – we developed a 'toolbox' of generic recognizers that identify commonly found features in calls of interest. The 'toolbox' approach allows us to mix and match feature extraction with classifiers to suit generic call types. The expectation is that if a set of features can be chosen that are appropriate to the call, then the classifier used for detection purposes can be made simpler than an HMM.

In this paper we describe a 'toolbox' of basic animal call recognition techniques developed in our laboratory (demonstrated at http://sensor.mquter.qut.edu.au/). This is not to say that more sophisticated tools have no value. Rather our tools can be viewed as filters to identify points of potential interest in long recordings that can then be interrogated with other techniques. We limit ourselves to terrestrial animals – in particular we exclude marine animals whose calls present a different set of problems (Rickwood and Taylor 2008). In addition we avoid birds that mimic – these can be difficult to recognize, even for humans.

In the Methods section, we describe some different call structures and the recognition algorithms appropriate to them. In the Results section, we describe the results of experiments with datasets obtained for selected animal calls. We conclude the paper with a discussion of our ongoing work.

Methods

Call structures

Many animal calls have a hierarchical structure. A complex bird call, for example, may be divided into phrases, the phrases into syllables and the syllables into one or more elements (Catchpole and Slater 1995). Each element may take the form of a whistle (single tone), chirp (slowly modulated tone), whip (rapidly modulated tone), click (appearing as vertical line in a spectrogram), vibrato, shriek, stacked harmonics (simultaneous multiple tones) or buzz (rapidly repeated click) (Catchpole and Slater 1995). The same syllable can be repeated multiple times. Figure 1 illustrates the structure of calls used in this work. Each image has been extracted from a spectrogram – the x-axis represents time, the y-axis frequency, and the grey scale represents acoustic intensity. Typically, bird calls are more variable than insect and frog calls and therefore present a bigger recognition problem. Nevertheless, each call type has some more or less invariant feature(s) on which recognition can focus. The same can even be said of certain diffuse, apparently structureless acoustic events, for example those caused by wind (Figure 1k) and canopy rain (Figure 1l).

Our toolbox contains a suite of classifiers each of which we have found appropriate for different call structures and syllable types (see the list in Table 1)). Call structure dictates feature selection, which in turn favours a particular classifier type. For example to detect a frequency-modulated whistle, we might use its spectrogram image for template matching. But this method will not work if the modulation varies significantly from bird to bird in which case other invariant features must be found.

Hardware

We employ both networked sensors and acoustic data loggers. Networked sensors provide near real-time sensing in locations with 3G connectivity. Due to bandwidth constraints and 3G transmission costs, they are configured to record at regular intervals (typically for 2 minutes at 30-minute intervals), upload data to a central repository and then deactivate until

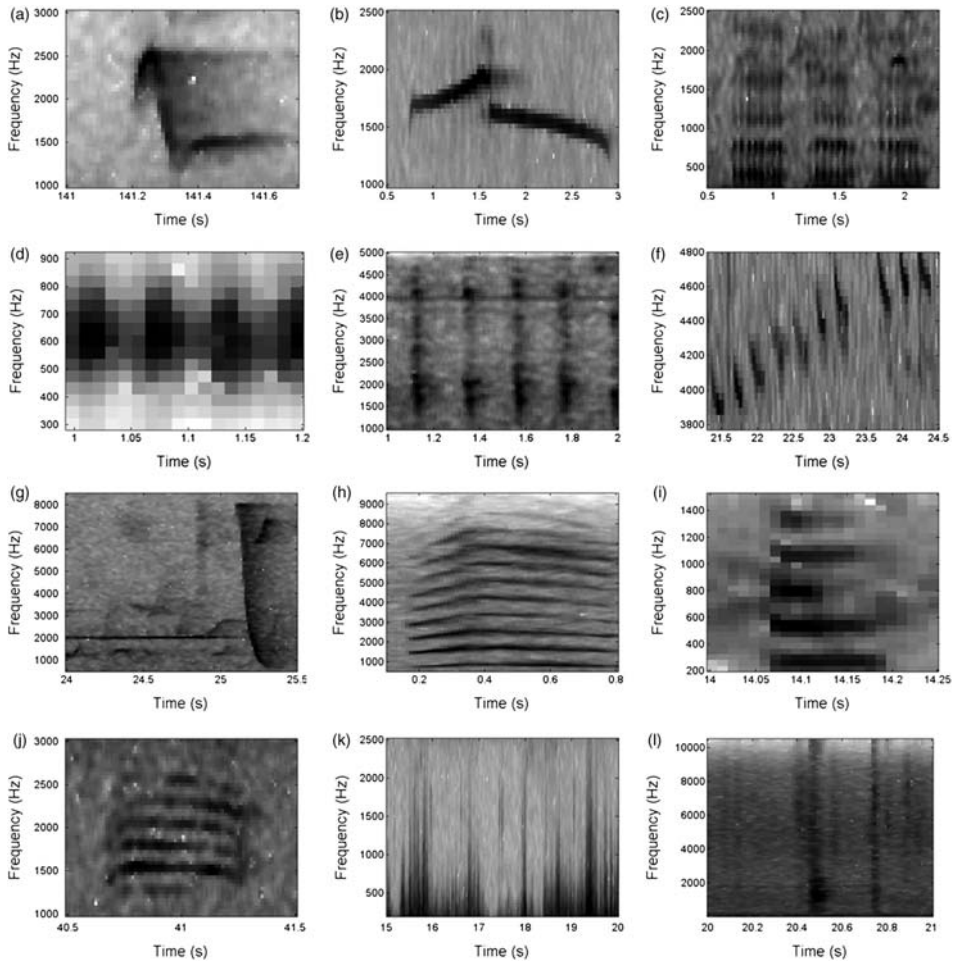

Figure 1. The spectral structure of calls studied in this work. Each image has been extracted from a spectrogram – the x-axis represents time in seconds, the y-axis frequency (Herz), and the grey scale represents acoustic intensity. (a) Currawong *Strepera graculina*; (b) Beach Stone-curlew *Esacus neglectus*; (c) male Koala *Phascolarctos cinereus*; (d) Cane Toad *Bufo marinus*; (e) Asian House Gecko *Hemidactylus frenatus*; (f) Ground Parrot *Pezoporus wallicus*; (g) Eastern Whipbird *Psophodes olivaceus*; (h) female Koala *Phascolarctos cinereus*; (i) human speech, vowel; (j) Torresian Crow *Corvus orru*; (k) wind gusts; (l) canopy rain.

the next scheduled recording. Our networked sensors consist of a 3G smartphone (HTC), an electret-style external microphone, pre-amplifier, DC-DC converter and an external power supply (solar panel or battery). The system is capable of operating continuously and autonomously for months at a time with minimal maintenance. Uninterrupted, continuous deployments of 18 months have been achieved to date (Wimmer et al. 2010).

Acoustic data loggers are designed to provide a short term or ad hoc high resolution acoustic sensing capability but can be configured to record continuously for extended periods (Wimmer et al. 2010). Acoustic data is stored internally and a device can provide up to 14 days continuous recording depending on the batteries provided. These sensors are highly portable and can also be powered using a solar supply for fixed deployments. As with the networked devices, data loggers are installed in an all-weather container.

Data sets

All recordings are sampled at 22,050 Hz (or subsequently down-sampled to this value) and a bit depth of 16. We prepared data sets for the calls of seven different animal species and for wind and rain events which frequently contaminate recordings in tropical Queensland.

Currawong *Strepera graculina*: the Currawong call (Figure 1a) consists of a frequency modulated whistle. The Currawong dataset consists of 29 4-minute recordings taken at 30-minute intervals from 00:00 to 07:00 and 17:00 to 00:00 on the 2 November 2009 at St Bees Island. This protocol was designed to encompass both sides of the morning and evening chorus. The dominant calling species in the recordings is the Beach Stone-curlew. However five of the 29 recordings also contain Currawong calls. Typically their calls are clustered, suggesting flocks of birds calling at one time.

Beach Stone-curlew *Esacus neglectus*: the curlew call consists of multiple syllables having simple to complex structure, with or without harmonics. The syllable types vary within the one call but the single modulated tone shown in Figure 1b is sufficiently well defined to warrant targeting. The curlew dataset is the same as the currawong data set. Nineteen of the 29 recordings contain curlew calls, although nine of them are of low intensity due to distance from the source.

Male Koala *Phascolarctos cinereus*: the male Koala has a complexly structured bellow consisting of a series of inhalations and exhalations lasting for 30 seconds or more. A small portion of its call (illustrating three oscillatory exhalations) is shown in Figure 1c). The exhalations have a snoring-like oscillatory component which offers a suitable target for recognition. Recordings were obtained from a wildlife sanctuary on St Bees Island off the coast of Queensland. Training data consists of 12 4-minute recordings, each containing a Koala call. Test data, totalling 460 minutes, are split into 115 4-minute recordings, 12 of which contain Koala calls. There are 18 calls, ranging from high to low intensity. This selection of recordings having low call density is representative of the real world situation.

Cane Toad *Bufo marinus*: the Cane Toad has a multi-syllable call consisting of a click rapidly repeated for 20 seconds or more (Figure 1d). Cane Toad data were collected in a suburban backyard (Brisbane, Queensland, January 2010) and in rural farmland (near Gympie, South East Queensland). Both locations are in the vicinity of permanent or semi-permanent water bodies with known Cane Toad populations. A total of 674 minutes of recording was split into 337 2-minute files. The dataset contains 83 cane toad calls in 53 files. The suburban recordings are ‘contaminated’ with a wide variety of extraneous sounds including traffic, air-conditioning, speech, dogs etc.

Asian House Gecko *Hemidactylus frenatus*: Asian House Geckos have multi-syllable calls consisting of some five to six clicks slowly repeated (Figure 1e). The periodicity of the clicks depends on temperature (Marcellini 1974). Calls were recorded in January and February 2010 just outside a suburban house (Brisbane, Queensland). 540 minutes of recording were split into 270 2-minute files, 77 of which contain a total of 84 calls.

Ground Parrot *Pezoporus wallicus*: the Ground Parrot call consists of about 10–13 syllables, each a brief descending chirp. Successive syllables increase in pitch (Figure 1f). Ground Parrot data were collected with data loggers placed in a nature reserve 100 km north of Brisbane. We acquired a total of 13 hours of recording, much of which was dominated by heavy rain but managed to recover 6 hours and 45 minutes of useable sound. This was divided into 1-minute sections and an ecologist tagged the calls. Of the 405 1-minute files, 32 contain calls. However nine of these are barely audible. It should be noted that 1–2 dB is the minimum perceptible audible difference between signal and

background noise (Lüscher 1951). Consequently Ground Parrot calls whose maximum intensity is less than 2 dB above background noise were ignored for testing purposes.

Eastern Whipbird *Psophodes olivaceus*: the whipbird has a two-syllable call consisting of a constant tone whistle (whose frequency may vary from bird to bird) followed by a whip (Figure 1g). The whip may be either ascending or descending. The dataset consists of 38 2-minute recordings, 14 of which contain whipbird calls.

Torresian Crow *Corvus orru*: the crow call appears as a set of stacked harmonics in a spectrogram (Figure 1j). The crow dataset consists of 20 4-minute recordings, 12 of which contain crow calls. Two other examples of calls incorporating stacked harmonics are the female Koala call (Figure 1h) and human speech vowels (Figure 1i) but we do not present results for these calls in this paper.

Wind events: these were obtained from all our recordings but primarily from those obtained on St Bees Island off the coast of Queensland. The training set consists of 142 'wind' and 142 'not-wind' events. The validation set consists of 383 'wind' events and 243 'not-wind' events. 'Not-wind' events include low frequency rumbling due to traffic and aircraft. The test data consist of 1235 1-minute audio files acquired with a variety of sensors at different locations. These recordings also include Koala bellows, ground parrot calls and many other kinds of acoustic events.

Rain events: canopy rain events were extracted from recordings taken at five different locations in Queensland, Australia. The events were labelled as 'rain', 'not-rain' or 'not-sure'. Events needed to be longer than 3 seconds for reliable labelling but even so, many events were labelled as 'not-sure'. The training and validation sets exclude 'not-sure' events thereby reducing the available data. Canopy rain events have a percussive content due to heavy rain drops hitting surfaces near the microphone. Consequently for the 'not-rain' instances we selected a variety of percussive sounds resulting from construction and human activity. The training set consists of 54 'rain' and 52 'not-rain' events. The validation set consists of 19 'rain' and 14 'not-rain' events. The test data consist of 247 1-minute files, 104 of which contain rain events.

Spectrograms

The call recognition algorithms described in this paper extract features from spectrograms prepared using the Short-Time Fourier Transform (STFT). The signal is framed using a window of 512 samples (23.2 ms) overlapping 50% which offers a reasonable compromise between time and frequency resolution. This choice of parameters is typical for ASR and is also used by others working with bird calls (Brandes et al. 2006; Brandes 2008), but we vary window overlap from 0 to 75% depending on the amount of fine structure in the call of interest. (Note: the terms *window* and *frame* are used interchangeably in this paper.) A Hamming window function is applied to each window prior to performing a Fast Fourier Transform (FFT), which yields amplitude values for 256 frequency bins, each spanning 43.07 Hz. The spectrum is smoothed with a moving average filter (window width = 3). The amplitude values (A) are converted to decibels using $\text{dB} = 20.\log10(A)$. Note that the decibel values at this stage are with respect to a hypothetical signal having unit amplitude in each frequency bin.

Segmentation

The speed of call identification can be much increased by skipping periods of silence. Our segmentation algorithm calculates the acoustic energy (dB) in the frequency band of

interest for every frame. We determine a baseline (modal) frame-energy and its standard deviation using the method described in Towsey and Planitz (2011). Segmentation involves retaining consecutive frames whose energy exceeds a user-threshold (defined in terms of standard deviations above the modal noise level). These acoustically 'active' frames are expected to contain calls of interest. In this work, segmentation was used only in conjunction with the oscillation detection method (see below).

Binary template matching

Some calls, in particular frequency modulated whistles (Figure 1a–b), can be recognized using a simple binary template. The user marquees the call of interest in a spectrogram and extracts a binary representation using an intensity threshold (typically around 4–6 dB above background noise). The template's *on*-cells define the call of interest and its *off*-cells contribute to an error function. The following metric works on the principle of matching shape and intensity profiles in images (Brunelli 2009) and is designed to pick out faint calls from background noise:

$$\text{Match score} = \Sigma_{\text{on}} intensity / c_{\text{on}} - \Sigma_{\text{off}} intensity / c_{\text{off}},$$

where $\Sigma_{\text{on}} intensity$ and $\Sigma_{\text{off}} intensity$ are the sums of the acoustic intensity in the *on* and *off* cells, respectively, and c_{on} and c_{off} are the count of *on* and *off* cells, respectively. In other words, this score calculates the difference in mean intensity between template *on*-cells and *off*-cells. The minimum perceptible difference is 1–2 dB and we find a suitable threshold score to adjust the recall/sensitivity trade-off is in the range 2.0 to 6.0 dB.

Our template extraction tool allows manual editing to clean up background noise and to idealize the shape. Although a binary template does not model variations in acoustic intensity, this lack of specificity can be an advantage if it generalizes over irrelevant call variability. The template is passed over all the frequency bins in a user specified band and a match score is calculated for each window position.

Oscillation detection

Many animal calls consist of a single repeating or oscillating element, for example the Lewin's Rail, the gecko and the male Koala bellow. Oscillation Detection is performed on the spectrogram using the Discrete Cosine Transform (DCT), which is a highly sensitive technique but must be used with caution (see Discussion). We apply it to the time series in each frequency bin (or row) of the spectrogram prior to noise removal, since noise removal also removes faint oscillations that are nevertheless detectable by DCT. For a more detailed description of this method see Towsey and Planitz (2011).

Two significant parameters for this technique are the time duration (number of coefficients in the DCT) and amplitude threshold required to register an oscillation 'hit'. In practice, it is not difficult to determine appropriate parameter values using training data as long as the call variability is within definable limits.

Call identification depends on recognizing concentrations of oscillation 'hits' within the call's frequency band. Typical parameter values for male Koalas, geckos and Cane Toads are shown in Table 2. The recall/sensitivity trade-off is controlled by adjusting the fraction of 'hit' bins (within the call's frequency band) required for a positive detection.

Event pattern recognition

Some multi-syllable animal calls can be modelled by the 2D distribution of their component syllables in the spectrogram rather than by the actual content of those syllables.

This approach has also been used by Kirschel et al. (2009). We apply this technique, which we term Event Pattern Recognition (EPR), to Ground Parrot calls, for which it is effective even when their calls are contaminated with other 'noise' events. Note that current ASR methods do not work well under such conditions.

This call detection technique requires two steps: first, the identification of acoustic events in a spectrogram and second, the recognition of groups of events having specific spatial relationships. Acoustic events are determined using the method of Towsey and Planitz (2011) and are defined by their start time, end time and minimum and maximum frequency. For the detection of Ground Parrot calls, events are ignored if they have an intensity of less than 3 dB above background and an area less than 100 spectrogram pixels.

A single training example is sufficient to construct a template which describes the call's component events (see Figure 2). In recognition mode, the template is passed over the acoustic events extracted from a test spectrogram (hereafter called 'test' events). To limit unnecessary computation, only 'test' events are considered whose centroid lies within the user defined frequency band (3.5–4.5 kHz for Ground Parrots). A match-score is calculated as the average overlap between the template events and the closest 'test' events whose centroids fall within the bounds of the template. The fractional overlap between a single template event and its closest 'test' neighbour is given by:

$$\text{overlap} = 1/2(x/T + x/E),$$

where $x =$ the overlapped area (in pixel units), $T =$ the (pixel) area of the template event and $E =$ the (pixel) area of the 'test' event. The overlap fraction lies between 0.0 (no overlap) and 1.0 (exact coincidence). The average overlap of all the events in the template gives rise to a score between 0.0 (complete mismatch of all events is actually not

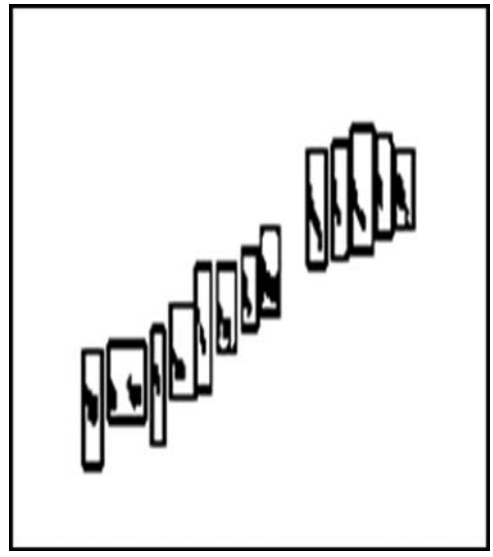

Figure 2. The Ground Parrot template. The call itself consists of 15 brief chirps ascending in frequency. The tenth chirp is missing in the above representation because its intensity did not exceed the user supplied intensity threshold. The template consists of 13 rectangular marquees placed using automated Acoustic Event Detection (Towsey and Planitz 2011). This automated approach has mistakenly placed one marquee around two consecutive chirps. Despite the 'mistakes' in this template due to automation, it nevertheless achieved a precision and recall of 87% in recordings containing many noise artefacts, especially heavy rain.

possible because at least the first template event must find some overlap) and 1.0 (complete coincidence of 'test' and template events).

The recall/sensitivity trade-off can be adjusted using an overlap threshold in the range [0, 1]. The optimum value for this threshold should be derived from an ROC curve and strictly speaking the data required to obtain the ROC curve has the status of training data. In our case we had so few calls in 6 hours of recording that the results reported below were obtained by optimizing the threshold on the available test data.

Syntactic pattern recognition

Syntactic Pattern Recognition (SPR) utilizes the ability to represent a temporal or sequential pattern as a sequence of symbols selected from a finite alphabet, each symbol representing a 'primitive' element of the composite pattern (Bunke and Sanfeliu 1990). This permits the representation of complex sequential patterns more accurately than can be achieved with 'flat' feature vectors of fixed dimensionality. The Eastern Whipbird has a simple two component call (a whistle followed by a whip; see Figure 1 g) that suggests two primitives, a horizontal line segment and a near-vertical line segment. Recognition depends on detecting a sequence of horizontal primitives (the whistle) followed by vertical primitives (the whip) where the whistle can be of varying durations and frequency and the whip can be either ascending or descending. In our implementation, the notion of a whipbird 'grammar' is implicit in the scoring algorithm.

Call recognition requires two steps: first, the identification of spectral tracks in the spectrogram and second, the recognition of a sequence of tracks that match the specified grammar. The method is described in detail in Towsey and Planitz (2011). The approach is to assign to each frame of the spectrogram two scores, a *whistle* score and a *whip* score. The whistle score for frame *N* is the fraction of frames over a *previous* (user-specified) time period traversed by a horizontal primitive. The whip score for frame *N* is the fraction of cells in the *subsequent* rectangle (enclosed by a user-specified time and bandwidth) traversed by a vertical primitive. The whipbird score for frame *N* is the average of its whistle and whip scores. Since both scores lie in the interval [0, 1] a threshold in [0, 1] can be used to adjust the recall/sensitivity trade-off for the combined score.

This method is only useful for birds whose calls appear as extended clean tracks in a spectrogram. Fortunately many bird calls have this characteristic. The method is not useful to detect parrot shrieks or diffuse events such as made by wind and rain.

Harmonic detection

Many animal calls display harmonics above a fundamental tone. Furthermore, the harmonic tracks often trace paths nearly parallel to the fundamental. Typical examples are the female Koala (Figure 1 h), human vowels (Figure 1i) and crows (Figure 1j). A serious difficulty with using harmonics as a feature for call recognition in environmental recordings is that higher harmonics drop out at a distance. Consequently the spectral signature of a bird depends on its proximity to the microphone. In ASR this problem does not arise because speakers remain at a fixed distance from the microphone. In environmental recordings, calls will be uttered at any distance. Nevertheless, Harmonic Detection can be useful if one knows that the calls will be uttered close by or if one wishes to determine the proximity of a known source by the number of high frequency harmonics.

The vowels in human speech appear as stacked harmonics in a spectrogram and the Discrete Cosine Transform has proved to be most useful in ASR to extract *cepstral*

coefficients as representations of vowel sounds (see Discussion). We did not find this approach suitable for our environmental recordings and instead we counted the number of harmonic tracks appearing within a user defined bandwidth. The score for a given frame is a measure of the average intensity of the harmonic peaks (dB above background) in a noise reduced spectrum (Towsey and Planitz 2011) discounted by a function of the difference between the observed number of harmonics and the expected number of harmonics:

$$\text{score} = (\Sigma_n a_n/N).w,$$

where a_n is the amplitude of the n^{th} spectral peak/harmonic, N is the number of observed spectral peaks in the user defined bandwidth, and w is a weighting factor that is a function of the difference between the observed number of peaks, N, and the expected number E:

$$\begin{aligned} w &= 1.0 && \text{if } (\text{abs}(N - E)) < 3 \\ &= 3/\text{abs}(N - E) && \text{otherwise.} \end{aligned}$$

The score array is smoothed with a moving average filter (window = 5) and a 'hit' is predicted where the score exceeds a user defined threshold for a user defined number of consecutive frames.

Hidden Markov models

As noted in the Introduction we did not find ASR techniques using HMMs to work well for our recognition tasks. Yet we started with this approach because it has been reported in the literature for bird call recognition. We include it in this report simply to compare it with other methods.

An HMM represents a bird call as a sequence of observations, each observation being a vector of spectral features derived from a single frame and its two neighbours. The observations are treated as emissions from a dynamical system whose state transitions are described by a Markov process. In a simple Markov process each observation would be assigned to a unique system state but animal calls are too variable for such a restrictive model. In an HMM each observation is modelled as a probabilistic function of the system state and each call type is modelled as a sequence of states. An HMM classifier returns the probability that the observed sequence would be emitted by a given call model.

We implemented the Hidden Markov Model Toolkit (HTK: http://htk.eng.cam.ac.uk/), a freely available software library for designing, training and testing HMMs (Young et al. 2006). Although tailored for speech recognition tasks, HTK has also been applied to biological sequences (Akhtar et al. 2007) and bird call recognition (Trifa et al. 2008; Kirschel et al. 2009). HTK automatically extracts the standard ASR feature vector consisting of *mel-frequency cepstral coefficients*. We use most of the default parameter settings, in particular 50% frame-overlap and 12 cepstral coefficients. The frequency band is constrained to match the call to be recognized. Rather surprisingly, we obtained best results when omitting signal energy and the dynamic delta and acceleration features.

Important HMM parameters are the number of model states, the number of emission categories, the number of training iterations and the estimation of an HMM model representing background noise. There are two possibilities for this last: (1) estimate a noise model from the silence periods in the training instances; or, (2) estimate a noise model from separate recordings appropriate to the operational environment. We tried both approaches and had more success with the latter.

Detecting wind gusts

Wind and rain are frequent acoustic 'contaminants' of environmental recordings. There are two reasons why automated recognition of these episodes might be useful. First, in most cases the user will want to minimize storage and computation by avoiding wind and rain episodes that mask useful information. Second, although wind and rain can be detected using meteorological instruments, hardware security is a problem in many locations – indirect evidence of wind and rain could help to interpret other features of a recording. While covering microphones with foam baffles can reduce the effect of wind, in practice we found that once wet they retain moisture.

We approached wind detection as a classification task, where the entities being classified are acoustic events extracted using the method of Towsey and Planitz (2011). Gusting wind events (e.g. Figure 1k) are found in the low frequency range. We explored a range of event features and adopted four; two describing the distribution of acoustic intensity and two describing acoustic entropy. For training and test data we extracted all acoustic events whose minimum and maximum frequencies were <500 Hz and <2 kHz respectively. Wind events had to be longer than 1 second for reliable tagging but the extracted features are duration independent and therefore the trained classifier is able to label events shorter than one second.

Detecting heavy rain

As with wind detection, we approached rain detection as a classification task using the method described in Towsey and Planitz (2011). The selected rain events consisted of heavy canopy rain. In particular, they excluded light rain and drizzle. During canopy rain, broadband percussive effects arise due to large rain drops striking surfaces near the microphone (see Figure 1l).

Experimental design

All recordings were obtained from various locations on the east coast of Queensland, Australia, using either networked sensors or data loggers. Most of the recordings were obtained in the context of three research projects, one on Koalas (FitzGibbon et al. 2009), another on Quolls (Belcher et al. 2008) and another at a university field station (Williamson et al. 2008).

Networked sensors returned recordings of 2 or 4 minutes depending on the protocol requested by the ecologist. Data loggers returned recordings of several hours duration which were split into lengths of 2–4 minutes.

Much consideration needs to be given to the reporting of recognizer accuracy. In the absence of generally accepted standardized datasets, recognition accuracy can be made arbitrarily high depending on the selection and prior cleaning of the recordings. We have endeavoured in these experiments to construct datasets that reflect the 'real-world' of sound which ecologists must process. Datasets were chosen according to their expected ecological significance rather than to obtain clean recordings. For example, recordings contaminated with wind and rain were not removed. Some of the recordings include the morning and evening chorus where there can be a cacophony of sound. Others contain traffic noises, air-conditioners (in city recordings), human speech, dogs barking and airplanes. Finally the tagging of calls was done by ecologists whose trained ears could detect faint distant calls that one would not expect to detect by automated means.

All our recognizers were designed to produce a binary output, that is, call identified or not identified. A critical issue in the context of our data is whether to express accuracy in terms of correctly identified *calls* or correctly labelled recording *segments*. We opted for the latter because the principle cost of an error is the time spent online by an ecologist loading a single file (a recording segment) to access the predicted call. Consequently we use the following standard definitions for recall and precision:

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}),$$
$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}),$$

but we define TP (true positives) as the number of short (2–4 minute) recordings correctly identified as containing one or more calls of interest; FN (false negatives) as the number of files containing positive calls none of which were identified; and FP (false positives) as the number of files incorrectly identified as containing one or more calls. Accuracy is defined as the per cent of total files (recording segments) correctly classified. Note that this definition can result in a value for accuracy that is greater or less than the average of recall and precision depending on the ratio of positive to negative files (those containing calls of interest or not).

Scoring on the basis of recording segments (files) has two effects on the reported performance values. Where a recognizer detects a TP yet makes an FP or FN error in the same file, we label that file correctly classified. On the other hand we observe many instances where multiple calls correctly recognized in one recording are offset by a single error in another recording. This situation arises because bird calls are frequently clustered. In short, the recognition rates that we present in the Results section must be regarded as indications of operational performance in a real problem as opposed to finely tuned estimates of accuracy.

Results

Currawong

We compared two recognition techniques, HMMs trained on standard ASR cepstral features and a simple binary template. As can be seen from the results in Table 1, neither method performed well. However there are two points to note here: first, much time was spent obtaining training and validation data for the HMM method whereas the binary template was quickly prepared using just one representative call, manually edited to remove artefacts. Second, the Currawong calls were numerous but clustered into just five files. The binary template recognized the great majority of the calls and therefore on a call basis its recall would have far exceeded that of the HMM. The low recall of the HMM approach was due to the difficulty of training a suitable noise model that covered the range of ambient noise situations. The low precision of the binary template was due to false positive identification of the more numerous curlew calls which sit in the same frequency band.

Beach Stone-curlew

A binary template achieved an accuracy of 76% (Table 1). The dominant errors were false negatives due to the large proportion of low intensity calls.

Male Koala

Using the method of oscillation recognition, recall of Koala bellows was 75% with the three false negative files containing distant bellows of very low intensity. To measure precision, we took into account that Koala bellows contain multiple oscillatory exhalations

Table 1. Toolbox for call recognition.

Generic call structure	Call	Features extracted from spectrograms	Classification algorithm	Duration of test recordings	Fraction of files with calls	Recall	Precision	Accuracy
Frequency-modulated whistle(s)	Currawong	Standard ASR features	HMM	2 h	5/29	40%	100%	90%
		Binary template	Template matching			100%	50%	83%
	Curlew	Binary template	Template matching		19/29	63%	100%	76%
Oscillatory components	Male Koala	Periodicity and amplitude of oscillations	Threshold	7 h 40 min	12/115	75%	75%	95%
	Cane Toad			11 h 14 min	55/337	93%	98%	99%
	Asian House Gecko			9 h	77/270	91%	90%	94%
Syllable events have fixed distribution in spectrogram	Ground Parrot	Duration and frequency bounds of acoustic events	Event Pattern Recognition	6 h 45 min	23/405	87%	87%	99%
Fixed syllable sequence (whistle and whip)	Whipbird	Orientation of spectral tracks	Syntactic Pattern Recognition	1 h 16 min	14/38	100%	67%	82%
Stacked harmonics	Crow	Frequency spacing of parallel spectral tracks.	Threshold	1 h 20 min	15/20	100%	71%	75%
Diffuse	Wind	Distribution of acoustic intensity & entropy	Diagonal linear hyperplane	20 h 35 min	792/1235	99%	96%	96%
	Rain	Acoustic properties of raindrops	Hyper-plane	4 h 7 min	104/247	75%	75%	79%

Table 2. Oscillation detection: parameter values and recognition results.

	Dataset		
	Male Koala	Cane Toad	Asian House Gecko
Parameters			
Frequency band (kHz)	0.1–1.0	0.5–1.0	1.5–3.0
Frame overlap	75%	75%	0%
DCT duration (seconds)	0.3	0.5	1.0
Oscillation bounds (Hz)	20–50	10–20	3–7*
Min DCT amplitude	0.6	0.6	0.5
Call duration (seconds)	0.5–2.5	0.5–20.0	1.0–6.0
Percentage of frequency bins with oscillations	20%	40%	30%
OD recognition results – without segmentation			
Total number of files	115	337	270
True positives files	9	49	70
False positives files	3	1	8
False negatives files	3	4	7
Recall	75%	93%	91%
Precision	75%	98%	90%
Accuracy	95%	99%	94%
OD recognition results – with segmentation			
Recall	75%	83%	91%
Precision	75%	98%	93%
Accuracy	95%	97%	96%
Processing time (percentage of the time without segmentation)	34%	56%	13%

*Asian House Gecko oscillation bounds were derived based on behavioural studies described in Marcellini (1974).

which will be detected in clusters. Consequently files containing only a single hit are likely to be false positives. Ignoring recordings with fewer than two hits resulted in a precision of 75%. False positives were mostly due to a bird (the Orange-footed Scrub Fowl) with a deep, chattering call, producing oscillations in a frequency band overlapping that of the male Koala. Although precision and recall give some indication of accuracy of the detection, these metrics ignore the large volume of data that was scanned to get the result. Based on total files correctly classified, we obtain an accuracy of 95%. For the ecologist this is a large saving in time and this recognizer is used by a Koala ecologist on a regular basis.

Cane Toad

The oscillation recogniser achieved a recall and precision of 93% and 98% respectively, and an accuracy of 99% over 337 files. These high accuracy rates are partly due to the fact that Cane Toads have consistent call characteristics which make it possible to fine tune the recognizer parameters. The false positives fall into two categories: Kookaburra calls that oscillate in the same frequency range as Cane Toads and very short hits due to background noise. The false negatives were all low intensity calls. Additionally, one call was masked by wind, an unwelcome contaminant that affects the accuracy of animal call recognition in all real world tasks.

Asian House Gecko

The oscillation recognizer achieved recall and precision rates on test Asian House Gecko recordings of 91% and 90% respectively with an accuracy of 94% over the 270 files.

The high accuracies confirm that geckos, like Cane Toads, have very consistent call characteristics. False negatives errors were due to missing calls of low intensity.

Segmentation prior to oscillation recognition

We repeated the recognition experiments for male Koalas, Cane Toads and geckos using segmentation to remove periods of silence. The results (Table 2) confirm that computation time can be greatly reduced whilst retaining high accuracy rates. In the case of the gecko, accuracy actually increased due to the filtering out of false positives ('noise'). However in the case of the Cane Toad data, segmentation removed some weak calls thus lowering recall. However, the main purpose of the segmentation filter is to reduce processing time – by 87% in the case of the gecko data (Table 2, bottom right cell).

Ground Parrot

We used Event Pattern Recognition to detect Ground Parrot calls. Note that ground parrots have other vocalizations that differ from our call of interest. The EPR algorithm calculates the per cent overlap between an event template (see Figure 2) and underlying acoustic events. The precision and recall rates cited below were calculated using a threshold overlap of 27%. This threshold was determined using the test data as there was not enough data for separate validation and test sets.

Precision and recall for the 23 files with audible parrot calls were both 87%. Two of the three false negatives detected were faint calls (compared to the true positives), and the other was missed because of edge effects (i.e. only half a call was present at the end of a recording). Three false positives resulted from incorrect detections due to rain which presented in the spectrogram as a random distribution of acoustic events. Total accuracy was 99%. We believe Event Pattern Recognition to be a particularly effective technique for multi-syllable calls having a characteristic distribution of events in the spectrogram.

Eastern Whipbird

Syntactic Pattern Recognition is a good approach to recognizing the Australian Eastern Whipbird because its call consists of two clearly defined spectral tracks. Recall and precision were 100% and 67% respectively. The accuracy of 82% shown in Table 1 is a misleading underestimate because whipbird calls are clustered and some of the recordings contained many true positive recognitions that counted only as one TP on a file basis.

Torresian Crow

The difficulty in recognizing crow calls is that the higher harmonics tend to drop out at a distance. Nevertheless, the crow is clearly a candidate for our Harmonic Detection algorithm. Recall and precision were 100% and 71% respectively with an accuracy of 75%. We have used the same algorithm (with different parameter values) for recognition of female Koala calls and the vowels in human speech but we do not present results due to lack of suitable data in our environmental recordings.

Wind events

A classifier was trained using MATLAB's classify.m class which manoeuvres a decision surface to optimally separate two classes. MATLAB's class offers a choice of three classifiers and best results on the validation set were obtained with a diagonal-hyperplane

classifier. The FP-FN trade-off for the wind detector can be adjusted using a constant which shifts the decision plane towards one class or the other – we used the default value of zero. Accuracy on test data was 96% on a one-minute recording basis (Table 1). False positive detections were due to male Koala bellows that have low frequency content.

Rain events

Canopy rain events were classified as 'rain', 'not-rain' using MATLAB's classify.m function. A Linear classifier provided best results on the validation set. When tested on 247 1-minute files, 104 of which contained rain events, the classifier achieved recall, precision and accuracy rates of 75%, 75% and 79%, respectively (Table 1). It is worth noting that rain detection is more difficult than wind detection, both for humans and for the machine classifier. The poorer performance of the rain classifier was probably due, in part, to inaccurate tagging of data.

Discussion and conclusions

In this work we have described a 'toolbox' of call recognition techniques to detect animal calls in environmental recordings. Our objective was to report performance results under experimental conditions that reflect the needs of ecologists having to process many hours of recordings. The work reported here arose out of an early realization that a one-recognizer-fits-all approach would not cope with the unconstrained variety of acoustic events that appear in environmental recordings. With the one-recognizer-fits-all approach, the same feature set must be extracted for all classes of call and consequently a powerful classifier is required because it is difficult to separate the classes in feature space. By contrast, given a set of features appropriate to the call of interest, a simple threshold or linear classifier is often good enough. This was most obviously demonstrated in the wind and rain classification tasks where MATLAB returned a linear classifier in preference to a quadratic.

A particular conclusion of our work is that the highly refined techniques of ASR are not suitable in the context of environmental acoustics. We have already mentioned two reasons for this. First, HMMs require much training data which is often not available for animal call recognition tasks. Kirschel et al. (2009) attribute the poorer performance of HMMs on one of their data sets to the small amount of available training data. In our own work we could not, for example, have used HMMs for the Ground Parrot task due to paucity of recordings. Even if training data is available, preparation and cleaning are time consuming.

Second, the ceptral feature set used in ASR has been tailored to suit a very constrained audio environment, typically a telephone channel where background noise can be controlled. Performance of current ASR technology is known to degrade rapidly where there are simultaneous speakers. ASR requires the additional training of an HMM noise model but in an arbitrary acoustic environment, background noise (due, for example, to rustling leaves, traffic and drizzle) is unpredictable. We believe that the poor performance of the HMM in our study was due to an inadequate noise model rather than to poor selection of training data. Again, Kirschel et al. (2009) note the importance of background noise profile in their implementation of HMMs.

Quite apart from the practical problems mentioned above there is also a sound theoretical reason for believing that ceptral features are ill-suited to the task of animal call recognition. The vowels in human speech are composed of many formants which appear as stacked harmonics in a spectrogram. Cepstral coefficients (obtained by the Discrete Cosine Transform) are able to capture vowel and speaker dependent transitions in rising

and falling formants. However, in our recordings we have not seen any bird or animal call that approaches the richness of the human voice in its formants. Female Koala and crow calls contain formants (stacked harmonics) but nothing comparable to human vowels. On the contrary, many bird calls incorporate a pure whistle (one frequency tone) which presents like an impulse to the Discrete Cosine Transform and thus returns spurious multiple frequency content giving no clue as to the true frequency of the whistle. It is for this reason that we do not use the Discrete Cosine Transform to detect bird calls with stacked harmonics. In our experience it returns too many false positive 'hits'.

Our recognizers have a number of features that reflect real world usage:

(1) Except for the HMM, they can be trained with just one or very few training instances. They can be refined when new instances become available. Even if sufficient data is available, large training and tests sets are time consuming to curate.
(2) The recognizers (except those for wind and canopy rain) are constructed as threshold classifiers trained on positive instances only. The difficulty with training an *N*-class classifier is how to select examples of the *null* class when the audio content is unconstrained.
(3) An additional difficulty with training an *N*-class classifier is that the discovery of a new instance or class typically requires the retraining of the entire classifier. In the case of *N* binary classifiers, only one of them needs retraining.
(4) For the most part, our classifiers have tuning parameters whose function is intuitively clear for the non-technical ecologist. For example, calls have a minimum and a maximum duration and amplitude thresholds are represented in decibels. The obvious exception is the cepstrum-HMM approach whose many parameters require a basic understanding of the theory.

There are two obvious extensions to our work. The first is to construct classifiers for additional call types. The second is more challenging. At the present time our recognisers return a score either in dB units or normalized in [0, 1]. We are not able to make a choice between two simultaneous 'hits' if the score is in different units. We would like to normalize all scores in order to disambiguate cases where a single acoustic event returns positive to more than one classifier.

Acknowledgement

The Microsoft QUT eResearch Centre is funded by the Queensland State Government under a Smart State Innovation Fund (National and International Research Alliances Program), Microsoft Research and QUT.

References

Acevedo MA, Corrada-Bravo CJ, Corrada-Bravo H, Villanueva-Rivera LJ, Aide TM. 2009. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. Ecological Informatics 4:206–214.

Agranat I. 2009. Automatically identifying animal species from their vocalizations. Paper presented at: Fifth International Conference on Bio-Acoustics; Holywell Park, Loughborough, UK.

Akhtar M, Ambikairajah E, Epps J. 2007. GMM-based classification of genomic sequences. Paper presented at: The International Conference on Digital Signal Processing; Cardiff, Wales, UK.

Anderson S, Dave A, Margoliash D. 1996. Template-based automatic recognition of birdsong syllables from continuous recordings. Journal of the Acoustical Society of America 100:1209–1219.

Belcher C, Jones M, Burnett S. 2008. Spotted-tailed quoll, *Dasyurus maculatus*. Chatswood, Sydney: New Holland Publishers Australia.

Brandes ST. 2008. Automated sound recording and analysis techniques for bird surveys and conservation. Bird Conservation International 18(S1):S163–S173.
Brandes T, Naskrecki P, Figueroa H. 2006. Using image processing to detect and classify narrow-band cricket and frog calls. The Journal of the Acoustical Society of America 120:2950–2957.
Brunelli R. 2009. Template matching techniques in computer vision: theory and practice. West Sussex: John Wiley & Sons.
Bunke H, Sanfeliu A, editors. 1990. Syntactic and structural pattern recognition – theory and applications. Singapore: World Scientific Publishing Co.
Cai J, Ee D, Pham B, Roe P, Zhang J. 2007. Sensor network for the monitoring of ecosystem: Bird species recognition. Paper presented at: Third International Conference on Intelligent Sensors, Sensor Networks and Information Processing; Melbourne, Australia.
Catchpole C, Slater P. 1995. Bird song: biological themes and variations. 2nd ed. Cambridge: Press Syndicate University of Cambridge.
Chen Z, Maher R. 2006. Semi-automatic classification of bird vocalizations using spectral peak tracks. The Journal of the Acoustical Society of America 120:2974–2984.
Cowling M, Sitte R. 2003. Comparison of techniques for environmental sound recognition. Pattern Recognition Letters 24:2895–2907.
FitzGibbon S, Ellis W, Carrick F. 2009. Mines, farms, koalas and GPS-loggers: assessing the ecological value of riparian vegetation in central Queensland. Paper presented at: The 10th International Congress of Ecology; Brisbane, Australia.
Juang C, Chen T. 2007. Birdsong recognition using prediction-based recurrent neural fuzzy networks. Neurocomputing 71:121–130.
Kirschel ANG, Earl DA, Yao Y, Escobar IA, Vilches E, Vallejo EE, Taylor CE. 2009. Using songs to identify individual Mexican Antthrush *Formicarius moniliger*: Comparison of four classification methods. Bioacoustics 19:1–20.
Kwan C, Mei G, Zhao X, Ren Z, Xu R, Stanford V, Rochet C, Aube J, Ho KC. 2004. Bird classification algorithms: theory and experimental results. Paper presented at: IEEE International Conference on Acoustics, Speech, and Signal Processing; Montreal, Canada.
Lau A, Mason R, Pham B, Richards M, Roe P, Zhang J. 2008. Monitoring the environment through acoustics using smartphone-based sensors and 3G networking. In Langendoen K, editor. Proceedings of the Second International Workshop on Wireless Sensor Network Deployments (WiDeploy08); 4th IEEE International Conference on Distributed Computing in Sensor Systems, DCOSS 2008, Santorini Island, Greece, Proceedings. p. 52–57.
Lüscher E. 1951. The difference limen of intensity variations of pure tones and its diagnostic significance. The Journal of Laryngology & Otology 65:486–510.
Marcellini DL. 1974. Acoustic behavior of the gekkonid lizard, *Hemidactylus frenatus*. Herpetologica 30:44–52.
McIlraith AL, Card HC. 1997. Birdsong recognition using backpropagation and multivariate statistics. IEEE Transactions on Signal Processing 45:2740–2748.
Rickwood P, Taylor A. 2008. Methods for automatically analyzing humpback song units. Journal of the Acoustical Society of America 123:1763–1772.
Somervuo P, Harma A, Fagerlund S. 2006. Parametric representations of bird sounds for automatic species recognition. IEEE Transactions on Audio, Speech, and Language Processing 14:2252–2263.
Temko A, Malkin R, Zieger C, Macho D, Nadeu C, Omologo M. 2006. Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems. Cough 65:5–11.
Towsey M, Planitz B. 2011. Technical report: Acoustic analysis of the natural environment. Brisbane: Queensland University of Technology, Available from: http://eprints.qut.edu.au/41131/ (last accessed 9 January 2012).
Trifa V, Kirschel A, Taylor C, Vallejo E. 2008. Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. The Journal of the Acoustical Society of America 123:2424–2431.
Williamson I, Fuller S, Marston C. 2008. A vertebrate survey of the Samford Ecological Research Facility. Brisbane: School of Natural Resource Sciences, Queensland University of Technology (QUT).
Wimmer J, Towsey M, Planitz B, Roe P. 2010. Scaling acoustic data analysis through collaboration and automation. Paper presented at: IEEE eScience 2010 Conference; Brisbane, Australia.

Young S, Evermann G, Gales M, Hain T, Kershaw D, Xunying L, Moore G, Odell J, Ollason D, Povey D, et al. 2006. The HTK book (for HTK Version 3.4). Cambridge: Cambridge University Engineering Dept.

Zhuang X, Xi Z, Huang TS, Hasegawa-Johnson M. 2008. Feature analysis and selection for acoustic event detection. Paper presented at: IEEE International Conference on Acoustics, Speech, and Signal Processing; Montreal, Canada.