A new optimization method of the geometric distance in an automatic recognition system for bied vocalisations

Un nouveau procédé d'optimisation de la distance géométrique dans un système de reconnaissance automatique de chants d'oiseaux

Michihiro Jinnai
Kagawa National College of Technology
355 Chokushi-cho
761-8058 Takamatsu
Japan
E-mail : jinnai@t.kagawa-nct.ac.jp

Neil Boucher
SoundID
PO Box 649
Maleny, 4552
Queensland
Australia
E-mail : neil.boucher@soundid.net

Minoru Fukumi
University of Tokushima
2-1 Minami-josanjima
770-8506 Tokushima
Japan
E-mail : fukumi@is.tokushima-u.ac.jp

Hollis Taylor
Wissenschaftskolleg zu Berlin
Institute for Advanced Study
Wallotstrasse 19
14193 Berlin
Germany
E-mail : wiko@wiko-berlin.de

Abtract
We describe a fully automated, PC based wildlife monitoring and survey system that is used for diverse species studies. The system as described is fully functional and operational at this time (download it at www.soundid.net). It uses a wide-area recorder that can record over areas of up to several square kilometres. The recorder can run, unattended, for more than a month. The recordings can either be analysed in real time to produce a particular response (e.g. send an SMS if a rare parrot is detected), or can be analysed post-recording on a PC. Any number of different species can be analysed simultaneously. In survey mode, calls can be counted and recognised with a summary of species and calling rate produced. The system has been successfully tested with the dawn chorus (which itself can be used for censuses) and against human surveys with impressive results. The software is equally effective for animal and non-animal sound sources. It can analyse calls at a rate of more than 100,000 per second.
(This is equivalent to comparing 100,000 spectrograms per second.)

Résumé
Nous décrivons ici un système de surveillance de la faune et d'enquête, entièrement automatisé sur PC, qui est utilisé pour étudier diverses espèces d'oiseaux. Il utilise un capteur qui permet d'enregistrer sur des zones de plusieurs kilomètres carrés. L'appareil peut fonctionner, sans surveillance, pendant plus d'un mois. Les enregistrements peuvent être analysées en temps réel pour produire une réponse particulière (par exemple envoyer un SMS si un perroquet rare est détecté), ou peuvent être analysés plus tard sur un PC. N'importe quel nombre d'espèces différentes peuvent être analysés simultanément. En mode enquête, les appels peuvent être pris en compte et reconnus avec un résumé des espèces et le taux d'appels produits. Le système a été testé avec succès lors du chant de l'aube (qui lui-même peut être utilisé pour les recensements) et malgré une surveillance humaine avec des résultats impressionnants. Le logiciel est également efficace pour les sources d'origine non-animale. Il peut analyser plus de 100 000 appels par seconde. (Ceci est équivalent à comparer 100 000 spectrogrammes par seconde.). Ce système est entièrement fonctionnel et opérationnel (le télécharger à www.soundid.net).

Beginning in 2002 we set out to design a recording and PC analysis package that could allow 24/7 acoustic surveillance of areas that were suspected to be prime sites for the irregular appearance of a rare Australian parrot, the Coxen's Fig Parrot (Cyclopsitta diophthalma coxeni). From the outset we decided to build the system to be a general purpose sound detection and recognition system, with the intention that it could be used for other acoustics purposes.

An initial survey of the technology available at the time indicated that, although many had attempted to build such a system before, there was little evidence that anything existed that could meet our requirements, either for the recording devices or the PC analysis. In particular, our recording device needed to operate for months at a time in a sub-tropical rainforest and the PC recognition system needed to be capable of distinguishing our target parrot species from 12 other co-habiting parrot species, some of which were known to be challenging to separate by ear even to an experienced human listener.

Past failures by others convinced us quite early that we needed to be innovative and explore new technologies up to and including new mathematics if necessary.

The Recognition Concept

The basic concept of the system was to transform a sound into a spectral image and then use pattern matching techniques. The matching was to be mathematical and not using AI or similar techniques.

Ultimately it was decided to use a library of reference calls, stored as mathematical images of their spectrums and then to find the best matching reference to each sound. Initially the Euclidian distance was used but, while it proved adequate for perfect or near perfect matches, it was not good at handling similarity. At this point we adopted a method pioneered by one of us (Jinnai) which he called the Geometric Distance (GD).

Running the two methods side by side, both returned the same result (a distance of zero) for perfect matches, but the GD proved much more effective and robust for the identification from similar (as opposed to exact) matches to the reference.. The GD is computationally more expensive than the Euclidean Distance but the advantages were soon obvious.

Calls can by analysed at a rate as high as 100,000 per second per processor but more typically the rate is 2,000 -3,000 per second, when a large number of different reference templates are used simultaneously. The rate depends on the call settings and the number of reference calls in a reference template, with the rate increasing for larger templates. A template is a file that is a mathematical image representing a collection of WAV file examples of the call. Essentially a template contains the information to build the images in Figures 1 and 2, for a collection of WAV reference files.

The Geometric Distance Concept

The most common measure of similarity is the Euclidean Distance and as the name implies it uses the linear distance between two patterns as a measure of the difference. The Geometric Distance is measured as the angle between two vectors that are the result of transforms on the original data. For our purposes the GD is measured in degrees with 90 degrees being the distance between two totally dissimilar images. Differences of 3 to 3.5 degrees in CD quality sounds are found between different sounds (as subjectively judged by a human listener). In real world soundscapes sounds that are similar are typically within a GD of 5-6 or less of each other.

Dimensionality

Most of the early work that we did used a 2-dimensional image (see the example below of two different Australian bird species in Figure 1: a Kookaburra call in Figure 1A compared to a Pale-headed Rosella call in Figure 1B). It is easy to see that the images are rather different and it is these transformed images of the call that we compare.

To derive these images, the Linear Predictive Coding (LPC) transform is used to calculate the frequency vs amplitude spectrum of a frame typically of 2001 data points. This was found to work well for most sounds, but it inherently loses temporal information which is sometimes important. The GD concept is N-dimensional (See Jinnai et al [7]) and so it is possible to employ it on the more conventional 3-dimensional spectrogram (see the 3-D image below in Figure 3), again calculated using the LPC. The 2-dimensional GD process was found to be a lot faster than the 3-D and so we use both, choosing the 3-D only where it is needed. It is also easier to visualise how the pattern matching can be done using the 2-dimensional LPC than with higher dimensional ones.

The Spectrum Transform

Initially the Fast Fourier Transform (FFT) was used as the default transform and while some success was had with it, we became concerned that the artifacts of the transform were making the matching process less exact. By running a lot of tests against the LPC using the same data, we concluded that the LPC transform, though significantly slower computationally than the FFT, was a more appropriate transform for our purposes.

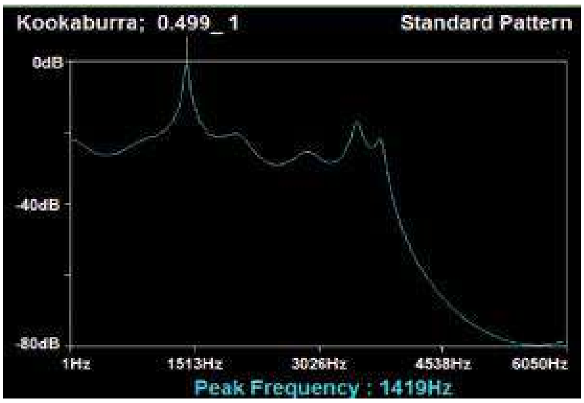


Fig. 1A : The 2-D image of a Kookaburra call
Une image en 2-D d'un chant de Kookaburra

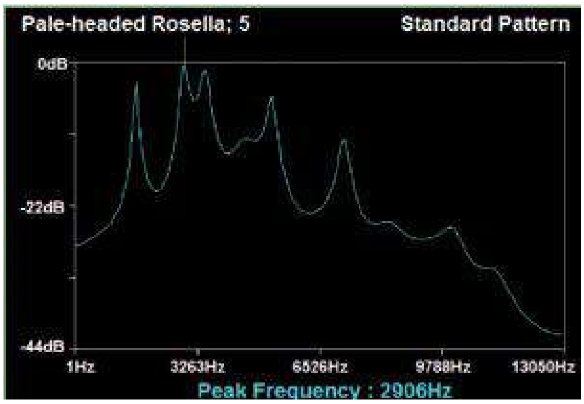


Fig. 1B : The 2-D image of a Rosella call
Une image en 2-D d'un chant de Rosella

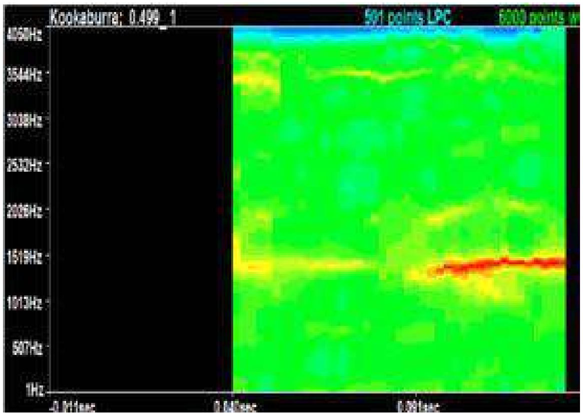


Fig. 2 : The 3-D image of the same Kookaburra call as in Figure 1A
Une image en 3-D du même chant de Kookaburra que pour la figure 1A

The LPC, first mooted in 1966 by S. Saito and F. Itakura of NTT Japan, is widely used as a telecommunications speech compression transform. Its use as a spectral transform gives results that are largely consistent with the FFT but with fewer artifacts. It can also be used to resolve small signal fragments without the same loss of spectral resolution that is characteristic of the FFT. It is of course subject to the limitations of the uncertainty principle and, as implemented by us, does produce some artifacts.

Early Testing

Early testing (see Boucher et al [6]) was based on a CD of parrot recordings consisting of 20 different sets of parrot calls with about 600 instances of calls in total. It took some years but the time came when with the initial CD the system was 100% accurate with 0% false positives (a few that were initially found turned out to be due to the wrong parrot making a guest appearance in some of the tracks). This was a good result but the real world is not CD quality and a lot more work was needed to get similarly good results in the field.

Real World Challenges

The real world was far more challenging. The first thing to be learnt is that birds (and we mainly worked with birds at first) have a woeful sense of pitch. Young birds clearly practice by both overshooting and undershooting the target frequencies by gliding up and down through the frequency range. We had to allow for this.

To make matters worse it seems birds do not have a sense of scales (in the musical sense) and so in different geographical regions they are likely to settle on a different "normal" frequency for a particular call.

In the search for rare species, even a poor recording is valuable and so it is necessary to work in a high S/N environment. Initially we aimed at 10 dB S/N but found that number to be too conservative. Sometimes today we are working at -20 dB S/N and getting helpful results.

To be really useful the system needs to be able to work with multiple species at the same time. This was something that again was easy enough using CD recordings but proved much more difficult when the calls were mixed with significant noise.

Next we found that most species had dozens of call types and some even more. Therefore we would need multiple examples of each type of call to get good matches. Typically we recommend that the reference files contain at least 10 examples of each call type.

The original system was designed to handle up to 32,768 different reference calls, but this was later extended to over 1 billion. Which brings us to the next problem - where are all these reference calls to come from? It is not so easy to get good clean reference calls. Here it is worth noting that although the system can work in poor S/N levels, it is rather important to have good references as otherwise the system will try to match on the call + the noise and in very poor S/N environments can end up matching noise with noise.

Reference File Sources

The majority of our users have had at least one attempt to collect their own reference calls. The results are usually disastrous. In order to find out why, we met with some professional sound recordists to get their views on this.

It seems that patience is the main ingredient. Professional recordists will spend days or sometimes even weeks to get that special recording. Additionally they use top quality equipment and have years of accumulated experience.

For the amateur the best advice is to get as close to the target as possible and record at the lowest volume level (to keep out extraneous sounds).

However the real answer is to get the professionals onside and have them provide the reference files. We have found that in Australia, at least, this is readily achieved.

The Recorder

Initially we assumed that the recorder would be something that was available off the shelf. And while it is true that many good quality recorders are commercially available few are designed to operate for long periods without attention and in any kind of weather.

Two problems emerged as most salient: the need for weatherproofing the housing and the need to power the unit for long periods (months at a time). It was soon realized that these two factors were interdependent as lower power consumption meant less heat, smaller batteries and so a smaller housing.

Recorder Characteristics

Most recorders are designed to work only at close range. That is, the user either actually holds the recorder or places it nearby. For wildlife recording, in order to get sufficient information, the recorder needs to operate over a wide area. To allow for this we designed a high-gain AGC amplifier, that had an effective range from 0.5 metres to several kilometers (under the right conditions).

The recorder has a large 27 AH battery that can power the Olympus LS-11 for about 2 months while the timer can be activated with up to 8 settings a day. Power can also be provided from and external socket that can double as a solar panel charging point. The battery is designed to be removed in seconds for easy field replacement, as is the LS-11. The LS-11 is controlled entirely from its external ports and there are no modifications to it.

On board is a high-gain AGC amplifier, a three frequency filter (60, 120 and 200 Hz) and a switchable gain control. The microphones clip into a pair of XLR sockets. The microphones are standard low noise electrets, but ultra-low noise ones are available if the yare needed.

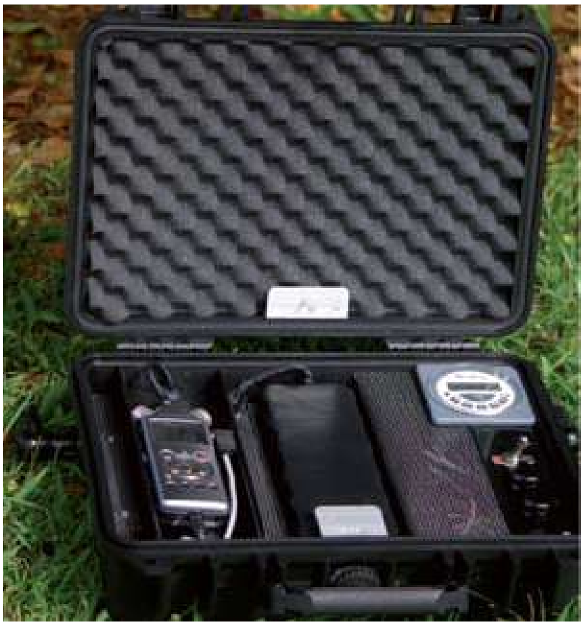


Fig. 3: The recorder based on the Olympus LS-11, with large external battery, timer and AGC card all housed in a waterproof case.
L'enregistreur est composé d'un Olumpus LS-11 avec une grosse batterie externe, un chronomètre et une carte AGC placés dans une valise étanche

Testing with the Dawn Chorus

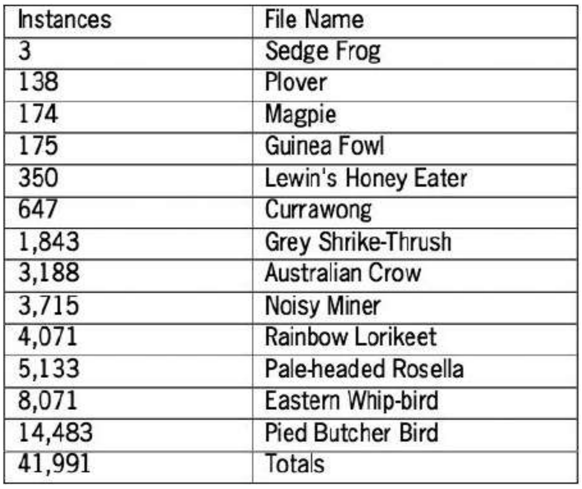
By November 2011 we began testing the system with the dawn chorus. One of us (Boucher) lives in a semi-rural location in Australia that was suitable for the testing. The recordings were made with a tripod mounted LS-11 which was set to record for a few hours on successive mornings.

The following species are known to be participants of the chorus.

- Australian Crow (*Corvus* spp.)
- Pied Currawong (*Strepera graculina*)
- Eastern Whipbird (*Psophodes olvaceus*)
- Grey Shrike-thrush (*Colluricincla harmonica*)
- Guineafowl (*Numida* spp.)
- Kookaburra (*Dacelo* spp.)
- Lewin's Honeyeater (*Meliphaga lewinii*)
- Magpie (*Gymnorhina tibicen*)
- Noisy Miner (*Manorina melanocephala*)
- Pale-headed Rosella (*Platycercus adscitus*)
- Pied Butcherbird (*Cracticus nigrogularis*)
- Spur-winged Plover (*Vanellus* spp.)
- Rainbow Lorikeet (*Trichoglossus haematodus*)
- Eastern Sedgefrog (*Litoria fallax*)

There were 14 reference templates consisting of a total of 566 individual examples of calls (from various sources) loaded into the software. Some of the templates were rather scant (e.g. the Eastern Sedgefrog was represented by only 3 examples of the call), but most had 25 or more.

In a random sample of the dawn chorus field recordings totaling just over an hour (1 hr 8 minutes) the following results were returned



Instances	File Name
3	Sedge Frog
138	Plover
174	Magpie
175	Guinea Fowl
350	Lewin's Honey Eater
647	Currawong
1,843	Grey Shrike-Thrush
3,188	Australian Crow
3,715	Noisy Miner
4,071	Rainbow Lorikeet
5,133	Pale-headed Rosella
8,071	Eastern Whip-bird
14,483	Pied Butcher Bird
41,991	Totals

Table 1 : Recognised calls from 1 hour 8 minutes of recording
Chants reconnaissables au bout d'1 heure 8 minutes d'enregistrement

So we have a total of 41,991 recognised calls or 37,000 per hour. The Kookaburra (which can be heard in the original recordings) alone was not in the references and so was not found.

The very large number of calls in such a short time partly reflects the fact that the recognisor looks for call segments of about 0.05 seconds (it varies by call type) in duration. Typically the recognisor will find 3 call segments per call.

For other dawn chorus recordings the recognisor typically returned 10,000 to 40,000 identifications per hour with better than 95% accuracy (as determined by a human listener). That accuracy rate can be improved with more and better references.

Soon after doing this test a CD arrived that contained 45 minutes of professionally recorded frog calls. A quick check revealed that the Sedgefrog was included. From the calls a reference file of 218 calls was made and run. These were then compared to the original 3 and no match was found between any of the two groups. A listening test confirmed that the calls did not match. A quick study revealed that "Sedgefrog" is a generic name for a number of different species of frog and even though both recordings came from the same region they were not of the same frog.

Just to see what would happen, the new reference file was run against a 3 hour dawn chorus recording (that included the one above) and a total of 27 matches were found. They were all false positives which is a false positive rate of 0.00014%. The three instances with the original file (Table 1 above) were confirmed to be a correct matching.

Performance

The software is suitable for processing terabytes of data and can be set to run all of the files on a HDD. It essentially searches the nominated HDD for all .WAV files and processes them sequentially. The approximate speed of processing on a 3.0 GHz processor is 50 times faster than real-time per reference file. This typically translates to about 10 times faster than real-time for about 5 reference files.

A template could be a collection of files for one species, but more generally it can be any collection of reference files (perhaps of multiple species) for which there is a common setting. The setting is the window through which we view each target call. Thus if we look at Figure 1A and 1B (which are displays of the viewing window) we notice that they span a different frequency band and they have different noise floors. The frequency and noise floor are two of the settings so these two species would have unique settings. Other settings include target GD, weighting vector, and frame size (the number of points for matching on). Changing even one of these makes the reference file unique.

The exact processing speed depends on the number of signals on the recording (the software looks for significant energy levels and interprets these as signals to be processed) and it also depends on the settings used.

The speed of processing can be compromised by noise such as wind and rain, both of which generate significant energy levels. A good microphone windshield is thus strongly recommended.

The accuracy depends primarily on the quality of the reference recordings, but is also dependent on the settings. Correctly set up the accuracy should easily exceed 95% (a figure which corresponds to human accuracy when asked to recognise a group of words out of context).

Accuracy also increases as the number of reference calls in the template is increased. For example if there are 100 calls of a certain species, for each signal on the recording every one of the 100 recordings is compared in turn. Each one will be assigned a GD and the smallest GD is the one that is declared the match.

This goes even further when there are multiple species. Assuming each species has its own template then as each of the templates are run the lowest GD becomes the "best match". Some species sound similar and as an example the Australian Currawong sometimes imitates the Grey Shrike Thrush (both species considered above). So on a first pass the software may assign a match of a call to the Currawong (when it is running the corresponding template) . However, later it might run the Grey Shrike Thrush (which we assume here is the calling bird). Now to all but the most expert human ear these are the same. But the software will not be fooled and will assign a lower GD to the Grey Shrike Thrush which at the exact time of the call will now over-ride its first "guess" of a Currawong with the Grey Shrike Thrush. When the run is completed the correct assignment will have been made.

So the software accuracy improves not only with more examples of the target species, but also with more examples of other species that might be calling in the area.

Trade-offs

As the software developed it became clear that it was both possible and desirable to trade CPU time for greater accuracy. Increasing the number of templates (each with their own settings) certainly improved the accuracy but increased CPU usage almost in direct proportion to the number of templates.

The 2-D analysis for most calls closely approximates the 3-D and since it runs faster it is the preferred mode. The 3-D mode is suited best to those situations where the temporal signature is important (for example in estimating the number of frogs calling in a chorus).

Noise performance of the system is good and it is possible to trade off accuracy for the ability to work in a noisy environment. The system will perform well at S/N levels of 10 dB, but can still give useful results in conditions as noisy as -20 dB S/N if required. In this instance it is found that by matching just the peak energy part of the call (and hence by setting the parameters to focus on the peak energy section of the call) it is possible to get good matching. However, by doing this, a lot of information about the rest of the call is not used and the uniqueness of the call is totally searched for in a small portion of the energy peak, so that false positives will increase.

After a lot of testing we found that the PC clock speed was the most important indicator of the total run-time. In recently years clock speeds seem to have saturated and 3.8 GHz seems to be about as fast as a PC will run without over-clocking. Modern PCs tend to be adding processors rather than ramping up clock speed and this is a minor dilemma for the software. While most 64 bit code has powerful parallel implementations the older 32 bit code usually does not. We therefore decided to limit the 32 bit code to one processor (although on a multiprocessor machine multiple instances of the code can be run at any time and this is recommended for processing very large file collections).

Field Applications

It is well established that employing bioacoustic methods for animal surveys in ecological studies offer a number of advantages over using visual-type surveying methods. These advantages include: a reduced need to disturb or handle animals, an ability to survey visually cryptic species, achieving effective monitoring during inclement weather, and cost savings through reducing the amount of times a site needs to be visited by highly skilled specialists.

Despite its clear utility and widespread use (see for example work by Rountree et al[1],Payne et al[2], Riede[3]), the application of bioacoustic monitoring can by somewhat constrained by "back end" data analysis requirements. That is, studies using bioacoustics can generate large amounts of acoustic data, often thousands of hours of recordings [1]. Processing large amounts of data that is in an acoustic form can be laborious and time consuming. As such, the time, resource and cost savings gained through implementing bioacoustics monitoring over other forms of survey method may be completely eroded.

Automated analysis of bioacoustic data is promoted as the answer to circumventing manual data processing and analysis obstacles. In practice however, the development of automated sound recognition is challenging, primarily because the vocalisations of many species can be complex. For example, some bird species can sing in duets, while others can intentionally mask their calls, perform vocal mimicry, have regional dialects, have large song repertoires and can perform improvisational songs [4]. Furthermore, despite widespread reporting of successful automated sound recognition in the literature (including for birds species [5]), the utility, practicality and accessibility of these systems to field ecologists seemed limited.

Field Implementation

As discussed, the recorder and automated sound recognition system was initially designed as an acoustic surveillance tool for rare parrots, and in this regard is an advancement in acoustic survey techniques for the conservation of rare and threatened fauna species. To illustrate, the system can be deployed at key sites on a long term basis (e.g. over the fruiting season of certain food source trees within the known range of the target species) to make high quality recordings over a distance of hundreds of metres. Recordings can then either be accurately analysed by the software in real-time to produce a particular response, for example send an SMS notification over a mobile phone network if a rare parrot is detected, or analysed post-recording on a PC to extract information on timing and frequency of site visitation and potentially species abundance.

Clearly, however, the system has additional utility across the fields of natural resource management. For example, in most Australian jurisdictions a Development Application requires the completion of an Environmental Impact Assessment (EIA). In these cases, and particularly for large scale developments such as mines, a comprehensive survey is required of local fauna to determine the impact the development will have on wildlife populations. Because the automated sound recognition system can be used to recognise the sound/vocalisation of any species or group of animals it can be used in conjunction with conventional techniques to enable a more accurate census of wildlife to be undertaken in the EIA process with minimal additional resourcing.

Thus there is a reduction in the possibility that an important species or aspect of the faunal community has been overlooked.

The system also has clear application in biological and ecological sciences. To illustrate, the system is currently being used to measure changes in amphibian communities at different points within a catchment and between catchments in order to establish clear links between frog species diversity and catchment health. Similarly, colleagues have been working on applying the recorders and automated sound recognition system to biological studies in Borneo, and others targeting a better understanding of Dolphin vocalisations. More recently the developers of the software have been approached by University researchers wanting to apply the method to studying mosquitoes.

Conclusion

A valuable new tool for bioacoustics has been described, and it is currently in use for wide-area acoustics surveys, rare parrot studies, bat and frog surveys and it has found uses in industrial and medical environments. The system as described is based on 32 bit code and for many applications will require some trade-off between accuracy and processing time. The 64 bit code under development (also the subject of a paper at this conference) will address this shortcoming and will also add many more features.

Références bibliographiques

[1] Rountree, R.A., Gilmore, R.G., Goudey, C.A., Hawkins, A.D., Luczkovich, J.J., Mann, D.A. Listening to Fish: Applications of Passive Acoustics to Fisheries Science. Fisheries. 31: 433-446. (2006)

[2] Payne, K.B., Thompson, M., Kramer, L. Elephant calling patterns as indicators of group size and composition: the basis for an acoustic monitoring system. African Journal of Ecology. 41: 99-107 (2003)

[3] Riede, K. Acoustic monitoring of Orthoptera and its potential for conservation. Journal of Insect Conservation. 2: 217-223.(1998)

[4] Brandes, T.S. Automated sound recordings and analysis techniques for bird surveys and conservation. Bird Conservation International. 18: 163-173. (2008)

[5] Chesmore, D. Automated bioacoustic identification of species. Annals of the Brazilian Academy of the Sciences. 76: 435-440. (2004)

[6] Boucher, N J. Jinnai, M. Gynther I. Design Considerations in a Sound Recognition System for Wildlife Indentification. Australian Institute of Physics Brisbane (2006)

[7] Jinnai, M., Boucher N J., Robertson J., Kleindorfer S., Design Considerations in an Automatic Classification System for Bird Vocalisations using the two dimensional Geometric Distance and Cluster Analysis, International Congress of Acoustics, Sydney. (2010)