Using feature vectors to detect frog calls in wireless sensor networks

Benjamin Croker
School of Information Technology and Electrical Engineering, University of Queensland, St. Lucia, Queensland 4072, Australia
benjamin.croker@gmail.com

Navinda Kottege[a)]
Autonomous Systems Laboratory, CSIRO ICT Centre, Pullenvale, Queensland 4069, Australia
navinda.kottege@csiro.au

Abstract: A method for detecting vocalization of giant barred frogs (*Mixophyes iteratus*) in noisy audio is proposed. Audio recordings from remote wireless sensor nodes were segmented into individual sounds and from each sound a small set of features was extracted. Feature vectors were compared to those of example calls using a Euclidean distance formula as a detection system. The system achieved a sensitivity of 0.85 with specificity of 0.92 when distinguishing *M. iteratus* calls from other species' calls and sensitivity of 0.88 with specificity 0.82 against background noise.

PACS numbers: 43.80.Jz, 43.28.Hr, 43.60.Bf [CM]
Date Received: February 28, 2012 **Date Accepted:** March 27, 2012

1. Introduction

Audio data recorded by wireless sensor networks (WSNs) can provide important information about species diversity and population. However, manual surveying of the data is laborious and often impractical, necessitating methods of automatic detection. Multiple instances of giant barred frog (*Mixophyes iteratus*) calls, an endangered species endemic to a small region on the east coast of Australia (Koch and Hero, 2007), were recorded by a remote WSN deployed along a stream in a reservoir catchment area over a period of six months. While there is a large body of literature on automatic species detection techniques, much of it concerns species-specific methods, that perform poorly with the typical noisy, low-quality audio recorded by WSNs.

This work presents a simple yet effective and computationally efficient method of accurately detecting *M. iteratus* calls in noisy recordings using both temporal and frequency features. Since no current work specifically addresses *M. iteratus* detection, our proposed technique is compared with spectrogram correlation used by Baumgartner *et al.* (2008) to identify instances of sei whale calls in noisy underwater recordings. The performance of the two methods are evaluated using audio recorded from the remote WSN and the results are compared.

Feature vectors have been previously used for classification of animal calls such as baleen whales (Baumgartner and Mussoline, 2011), but the work presented here shows a small set of features extracted directly from spectrograms is sufficient for detecting the frog calls. Simplicity was a desirable feature in designing the methodology to allow implementation on embedded devices as a next step.

2. Methodology

M. iteratus calls are short, low frequency "grunts" of about 0.1 to 0.2 s in duration with a peak frequency around 650 Hz. A sample spectrogram of a call is shown in Fig. 1(b).

a) Author to whom correspondence should be addressed.

The audio was recorded next to a running stream which dominated background noise in the 250 Hz to 1 kHz range—a similar frequency range to *M. iteratus* calls.

The audio was recorded at 16 kHz with 16-bit depth giving a full bandwidth of 8 kHz. Recording temperature ranged from 14.5 °C to 29.2 °C for full dataset. A bandpass filter was applied to the audio to isolate the frequency range of *M. iteratus* calls. The filter was designed and implemented as a digital infinite impulse response (IIR) filter, and was specified to have a passband with center frequency of 650 Hz and a bandwidth of 900 Hz. After filtering, spectrograms of the audio were calculated using a 256 sample window (16 ms) with a 50% overlap. Spectrograms in this paper are represented by S_{ij}, where $1 \le i \le N$ is the index of a frame and $1 \le j \le M$ is the index of a frequency band. N and M are the maximum frame and frequency indices, respectively.

2.1 Background noise rejection

Candidate sounds were found in each recording before being classified as *M. iteratus* calls by identifying recording segments with a higher energy level than the background noise. Summing each spectrogram frame over all frequencies gives short time average of the signal energy at a particular point in time (i.e., at a particular spectrogram frame). Therefore energy at frame i is $E_i = \Sigma_j S_{ij}$.

The recordings had mostly constant background noise levels punctuated by the occasional animal call or other sound. A naive method for identifying regions of audio with animal calls was to declare portions of audio with E_i greater than a threshold as sounds. This method failed when the energy of a call fluctuated above this level as was the case with *M. iteratus* calls. Lowering the threshold level only leads to this effect occurring with lower-energy calls.

The method used in this work specifies a maximum amount of time for which the energy level may dip below a threshold. This was modeled as a simple finite state machine [Fig. 1(a)]. A timeout period TO was specified, as was a threshold level h. The timeout was specified in seconds, but was converted into a number of spectrogram

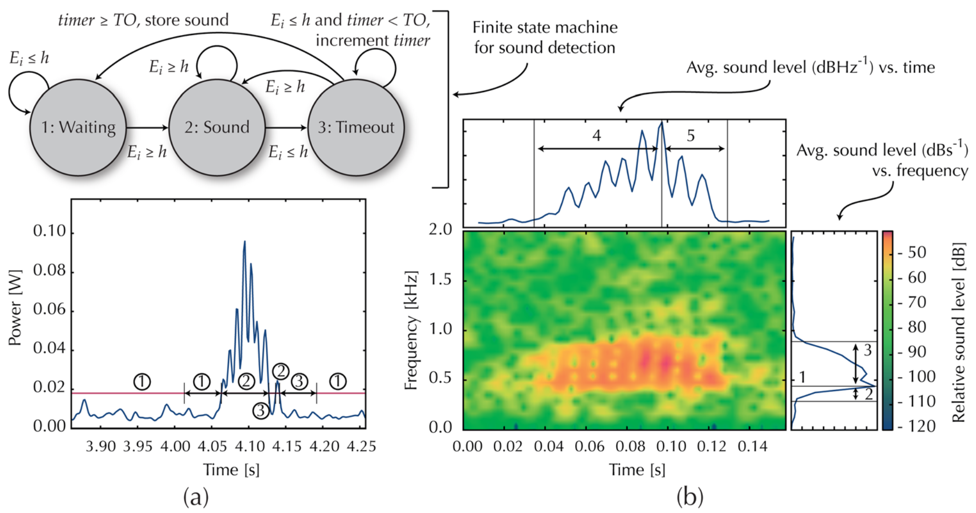

Fig. 1. (Color online) (a) Process of detecting a sound in the presence of background noise with the finite state machine (top) and sound levels annotated with a corresponding state (bottom): Initially the sound level is below the threshold (state 1) before rising above it (state 2). The level briefly falls below the threshold but not long enough to indicate the end of the sound (state 3 then state 2). Eventually the level remains below the threshold and the section of audio is marked as a sound. (b) Spectrogram of a sample *M. Iteratus* call from the WSN recording showing the five attributes used as a feature vector. Sound levels averaged along the time axis (right) are used for the spectral properties [(i)–(iii)] and sound levels averaged along the frequency axis (top) are used for the temporal properties [(iv),(v)].

frames. The following process for detecting sounds iterates over each spectrogram frame in the recording. The index i is initialized to 1 and incremented with each step.

State 1. If $E_i \leq h$, no action is taken.

State 2. If $E_i \geq h$, a sound is started and the start location $a = i$ is recorded.

State 3. When $E_i \leq h$, the sound may or may not have finished. The temporary end time $b = i$ is recorded. A timer is incremented every iteration. If $E_i \geq h$ at any point before the timeout is reached, the process returns to state 2, b is discarded, and the timer reset. If the timer is greater than TO, the sound is declared to have finished, with a and b marking the start and end points.

Using this method, a and b mark the exact start and end of the sound when compared to the threshold. Since animal calls typically ramp up and down in volume, a buffer of length TO was added to the start and end of each sound in an effort to capture the entire call. Therefore, the frames of the recording spectrogram from $a - TO$ to $b + TO$ were marked as a sound. In this work the timeout was set at 0.1 s and the threshold level set as 10 dB above the mean energy level of the recording. This allowed the threshold level to adaptively vary with the intensity of each recording. As a result, a higher threshold level was used for recordings with more intense background noise.

2.2 Feature vector detection

A feature vector was assigned to each individual sound identified using the method in Sec. 2.1. The features were chosen to include both temporal and spectral properties of a sound as well as some information about its shape. These features shown in Fig. 1(b) were (i) dominant frequency, (ii) frequency difference between the lowest and dominant frequency, (iii) frequency difference between the highest and dominant frequencies, (iv) time from the start of the sound to the peak volume, and (v) time from the peak volume to the end of the sound.

The start and end of a sound were defined as the points when the energy was 10 dB below the peak. The highest and lowest frequencies were defined similarly. Features relating to frequency [(i), (ii), and (iii)] were represented as a fraction of the maximum frequency—8 kHz in this case since the recordings were sampled at 16 kHz. This was done as a means for feature normalization. Time-related features [(iv) and (v)] were measured in seconds. Similar orders of magnitude were desired since feature vectors were compared in terms of Euclidean distance, and all features were to be weighted equally.

To classify a sound as a *M. iteratus* call, the sound's feature vector was calculated and compared to the feature vector of an example call. The Euclidean distance between the two was calculated and compared to a threshold level.

2.3 Evaluation

The system was evaluated by creating three sets of sounds. The first contained only *M. iteratus* calls and was constructed by segmenting 6 randomly selected recordings with *M. iteratus* calls into individual sounds, then manually removing all other sounds from the set.

The second set comprised other frog calls occupying the same frequency band as *M. iteratus* calls—50 Hz to 1 kHz. Since no such calls were found in the audio recorded by the WSN, sample recordings were manually constructed by mixing CD-quality recordings onto a recording of background noise in the area (Stewart, 1999). Care was taken to balance the frog call signal to noise levels to closely match the field recordings. The frog species used were *Litoria caerulea, Philoria loveridgei, Limnodynastes salmini, Philoria sphagnicola*, and *Philoria kundagungan.*

The third set contained background noise sounds, and was designed to simulate loud bursts of noise which may erroneously be detected as sounds by the mechanism described in Sec. 2.1. This was achieved by running the sound detection algorithm on a recording of background noise with a low threshold level.

Each set had 100 to 130 sounds, which was reduced to a random sample of 100 sounds for each set to provide uniform size. A sample vector was constructed by averaging the feature vectors from two *M. iteratus* calls considered to be "representative" of the species' calls. They were the two highest intensity calls picked from the first week of recordings. The Euclidean distance between each sound in each set and the sample vector was calculated with distances under a certain classification threshold (h_C) returning a positive classification. The different true and false positive rates achieved by varying h_C are summarized in Sec. 3. The performance was measured in terms of the ability to distinguish *M. iteratus* calls from other frog calls (sets 1 and 2), as well as from bursts of background noise (sets 1 and 3).

2.4 Spectrogram correlation

The results of feature vector detection were compared to spectrogram correlation—a method used by Mellinger and Clark (2000) to detect low frequency whale calls in the presence of background noise. The problem bears some similarity to detecting *M. iteratus* calls, and was used to evaluate the performance of the feature vector detection scheme proposed in this paper.

Spectrogram correlation involves cross correlating an example spectrogram (kernel) with a recording potentially containing calls of interest. High values in the resulting cross correlation indicate likely instances of the example call in the recording. It is desired for the kernel have zero sum over all frequencies for each frame, i.e., $\Sigma_j K_{i,j} = 0 \ \forall i$, where K is the kernel spectrogram. This ensures that the cross correlation of the kernel and white noise is 0. To achieve this, each kernel frame was constructed as a "Mexican hat" wavelet as described by Mellinger and Clark (2000) and Baumgartner *et al.* (2008). However, since the frequency distribution of the *M. iteratus* call is asymmetric [see Fig. 1(b)], different standard deviations were used for the positive and negative regions of the wavelet.

Two sample *M. iteratus* calls were taken, and an average spectrogram S calculated, then normalized. The kernel was calculated as

$$K_{ij} = C_i\left(1 - \frac{(j-f_i)^2}{\sigma^2}\right)\exp\left(\frac{(j-f_i)^2}{2\sigma^2}\right),$$

where $C_i = \max[S_{ij}]$, $1 \le j \le M$, $f_i = \arg\max_{1 \le j \le M}[S_{ij}]$, and σ is the distance (in indices) from the peak frequency to the frequency 7 dB down from the peak in either the positive (σ_p) or negative (σ_n) direction. Therefore, $\sigma = \sigma_p$ for $j \ge f_i$ and $\sigma = \sigma_n$ for $j < f_i$.

The kernel was cross correlated with individual sounds extracted from the recordings. Correlation scores exceeding a particular threshold indicated a match. The performance of this system was tested using the method described in Sec. 2 and results are discussed in the following section.

3. Results

The performance of the detection systems are summarized in Table 1, with receiver operator characteristic (ROC) curves shown in Fig. 2. ROC curves plot the false positive

Table 1. Overall performance of the feature vector detection (FV) and spectrogram correlation (SC) showing the max accuracy, sensitivity, and specificity.

Method	Set	Accuracy	Sensitivity	Specificity
FV	Frogs	0.89	0.85	0.92
	Noise	0.85	0.88	0.82
SC	Frogs	0.73	0.78	0.68
	Noise	0.94	0.94	0.94

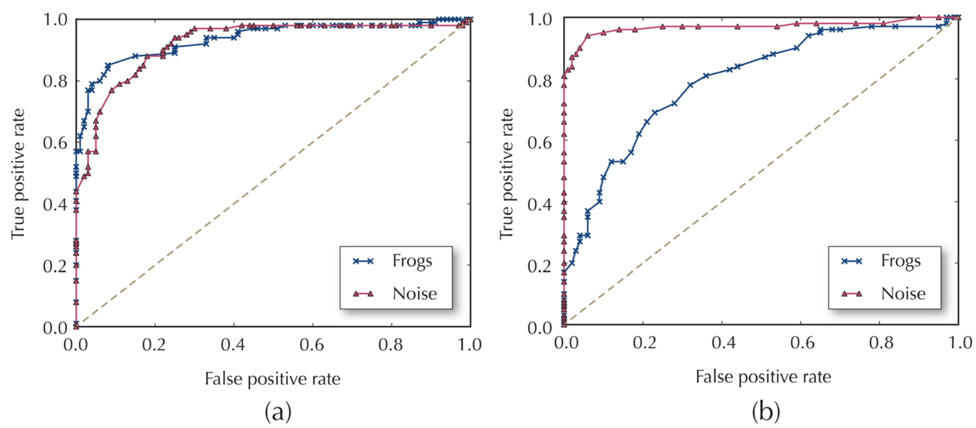

Fig. 2. (Color online) ROC curves of (a) detection with feature vectors, and (b) detection with spectrogram correlation showing the performance of each method distinguishing *M. iteratus* calls from other frog calls and from bursts of background noise while varying the classification threshold h_C. The dashed line represents the expected performance of a random guess.

rate $P_f/(P_f+N_t)$ against the true positive rate $P_t/(P_t+N_f)$ for different h_C values where P_t, P_f, N_t, and N_f are the number of true positives, false positives, true negatives, and false negatives, respectively. The accuracy of the system is given by $(P_t+N_t)/(P_t+P_f+N_t+N_f)$. Sensitivity of the system is the same as the true positive rate while specificity is 1 - false positive rate.

Compared to the established method of spectrogram correlation, feature vector detection performed better at detecting *M. iteratus* calls against other frog calls, but had lower performance against background noise. Depending on the application, performance against other species may be more critical, since most background noise is discarded during the sound segmentation process.

The improvement in sensitivity and specificity, shown by comparing the ROC curves in Fig. 2 indicates that feature vector detection is more effective than spectrogram correlation in applications where *M. iteratus* calls need to be distinguished from other frog calls.

In addition to comparable performance, feature vector detection requires less computation than spectrogram correlation. Feature vector detection requires sums over the frequency and time axes of the spectrogram, and iteration over the sums to find the features. This process results in $N(M-1)+M(N-1)$ addition and $N+M$ comparison operations for an individual sound with spectrogram size $N \times M$. Spectrogram correlation involves convolving a kernel with the sound and summing the result, requiring $(N-N_k+1)\ MN$ multiplications and $(N-N_k+1)\ MN$ additions, where the kernel spectrogram has dimensions $N_k \times M$.

4. Conclusions

A method for detecting *M. iteratus* calls in noisy audio was proposed based on comparing small sets of features extracted from individual sounds. Compared to spectrogram correlation, an established method for detecting low frequency animal calls in the presence of noise, the proposed method achieved higher performance in distinguishing calls of *M. iteratus* in the presence of other frog species calling in the same frequency band albeit with slightly less robustness against noise. Due to its simplicity and low computational cost, the proposed method is suitable for use in embedded devices such as WSN nodes.

Acknowledgments

The authors wish to thank SEQWater who funded the collection of audio data used in this work, and Jean-Marc Hero, Greg Lollback, and Jon Shuker from Griffith University for their assistance during this research.

References and links

Baumgartner, M. F., and Mussoline, S. E. (**2011**). "A generalized Baleen whale call detection and classification system," J. Acoust. Soc. Am. **129**, 2889–2902.

Baumgartner, M. F., Parijs, S. M. V., Wenzel, F. W., Tremblay, C. J., Esch, H. C., and Warde, A. M. (**2008**). "Low frequency vocalizations attributed to Sei whales (*Balaenoptera borealis*)," J. Acoust. Soc. Am. **124**, 1339–49.

Koch, A. J., and Hero, J.-M. (**2007**). "The relationship between environmental conditions and activity of the giant barred frog (*Mixophyes iteratus*) on the Coomera River, Southeast Queensland," Aust. J. Zool. **55**, 89–95.

Mellinger, D. K., and Clark, C. W. (**2000**). "Recognizing transient low-frequency whale sounds by spectrogram correlation," J. Acoust. Soc. Am. **107**, 3518–3529.

Stewart, D. (**1999**). "Australian frog calls: Subtropical east" (Nature Sound, Mullumbimby, NSW, Australia) [Audio CD].