Detection of anuran calling activity in long field recordings for bio-acoustic monitoring

Authors

***Abstract*—This paper presents a system to analyze long field recordings with low signal-to-noise ratio (SNR) for bio-acoustic monitoring. A method based on spectral peak track, Shannon entropy, harmonic structure and oscillation structure is proposed to automatically detect anuran (frog) calling activity. Gaussian mixture model (GMM) is introduced for modelling those features. Four anuran species widespread in Queensland, Australian, are selected to evaluate the proposed system. A visualization method based on extracted indices is employed for detection of anuran calling activity which achieves high accuracy.**

Keywords—Anuran calling activity detection; canetoad detection; frog detection; spectral peak track; Gaussian mixture model

I. INTRODUCTION

Global biodiversity is decreasing due to habitat loss, natural resource depletion, invasive species, climate change and so on. Therefore monitoring biodiversity is becoming increasingly important. Due to the development of sensor technique, sensors have been widely deployed in nature for monitoring biodiversity, which produces large volumes of acoustic data. However, traditional manual methods (based on ecologists spending extensive time in the field) are both costly and time consuming [1]. Hence, it is essential to develop new automated and semi-automated methods [2].

Several papers have already described automated methods for detection and classification of animal calls. Almost all prior work adopts the following structure: (1) pre-processing, (2) syllable segmentation, (3) feature extraction, (4) classification. In particular, features are all extracted from syllables, so the accuracy of syllable segmentation directly affects detection accuracy. Dayou et al. introduced a k nearest neighbor (k-NN) classifier to classify frog calls [1]. Spectral centroid and two entropy features were extracted from syllables for classification of frog calls. Chen et al. [3] combined spectral centroid, signal bandwidth and threshold crossing rate to do frog classification. The k-NN and support vector machines classifiers were introduced for classifying frog calls. Chen et al. [4] proposed a method based on syllable duration and multi-stage average spectrum for frog call recognition. Tyagi et al. [5] proposed the spectral ensemble average voice print to do bird call recognition, the recognition accuracy was improved by dynamic time warping. Lee et al. [6] introduced a recognition method based on spectrogram analysis to detect each syllable and calculate the Mel-frequency cepstrum coefficients (MFCCs). All averaged MFCCs of each frame were defined as

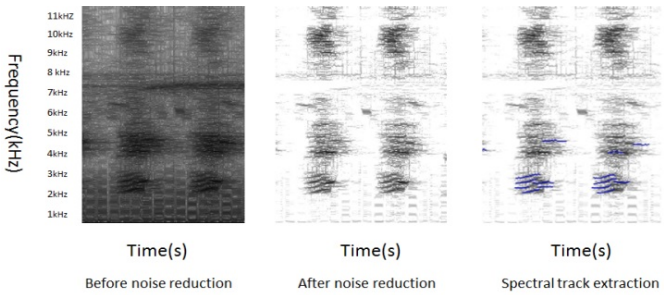


Fig. 1. The spectrogram result of spectral peak track extraction.

features. The linear discriminant analysis was used for classifying 30 kinds of frog calls and 19kinds of cricket calls. All the work achieves a high accuracy rate in recognition and classification, but it assumes that there is no overlap of syllables in the time domain.

Unfortunately, it is difficult to follow these methods for recordings with low SNR and contain much overlapping vocal activity, because they are hard to be split into syllables.

This study adopts an approach which does not depend on syllable segmentation because our recordings have low SNR and much overlapping vocal activity. We demonstrate the method with frog recordings because frogs are inclined to sing in choruses which make syllable segmentation difficult if it is possible. Furthermore, frogs are considered as a vital group in studies of environmental healthy and biodiversity. Our approach splits long field recordings into fixed one-minute segments and then extracts features directly from the one-minute segments. GMM is employed for constructing frog calls for frog calling activity detection. Meanwhile, the extracted features are incorporated into a visual representation of the recordings. Thus the ecologists can make a single image of an over-night recordings to quickly identify individual frog species which are active.

This paper is organized as follows: Section II describes the methods of recording acquisition. In section III, we describe spectrogram analysis. Section IV describes signal processing, feature extraction, and frog calling activity modelling. Section V describes the results on two overnight 12-hour recordings. Conclusions are offered in section VI.

II. ACQUISITION OF ANURAN CALL RECORDINGS

Digital recordings were obtained around ponds in the ground of James cook university campus, Townsville, using a

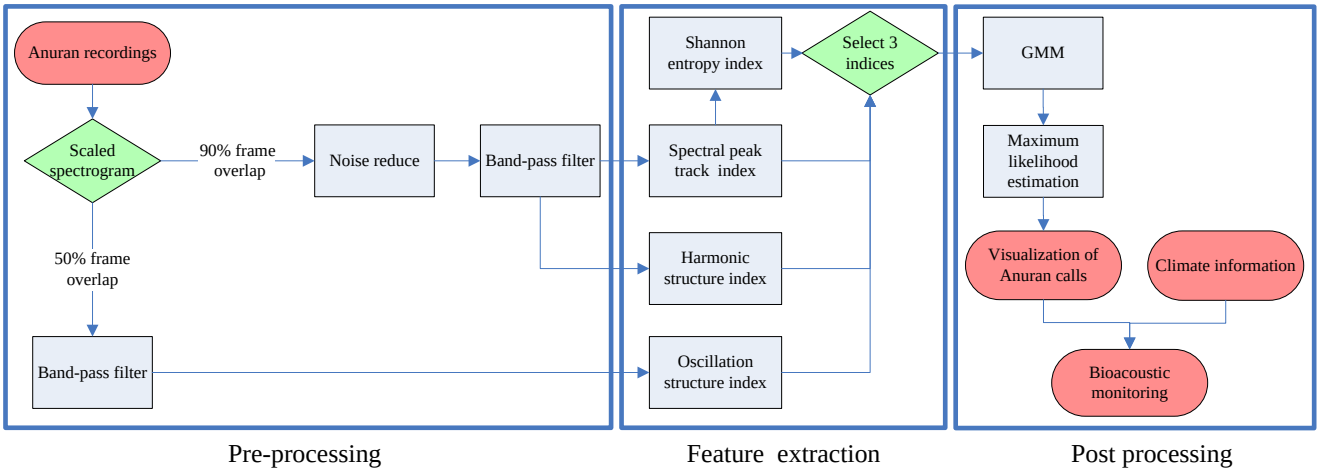


Fig. 2. Flowchart of anuran call recognition

battery-powered acoustic sensor (stored in a weather proof metal box) with an external microphone. The recordings were stored on 16GB SD cards in 64 kbps MP3 mono format. The recordings started around sunset, finished around sunrise every day and have 12 hour duration. In this study, two 12-hour recordings are used. Canetoad and Litoria gracillenta are recorded on 2nd March 2013; another recording is recorded on 16th January 2013 which contains Litoria nasuta and Litoria caerulea.

III. Spectrogram Analysis

The structure of most frog calls in spectrogram (90% frame overlap) is a curve (named spectral peak track in this study) or multiple curves (harmonic structure) which is wide in time domain but short in frequency domain (see Fig.1). The structure of some frogs in spectrogram (50% frame overlap) is a curve which is wide in frequency domain but short in time domain. In our research, four indices are selected: (1) spectral peak track index (SPI), (2) harmonic structure index (HSI), (3) Shannon entropy index (SEI), (4) oscillation structure index (OSI). SPI, SEI and HSI are extracted from spectrogram with higher frequency resolution (90% frame overlap). For OSI, time domain resolution is more important (50% frame overlap). For detecting the anuran calling activity, the spectrogram information of different anuran species is first analyzed for choosing suitable indices. The result is shown in Table I. According to the result, the combination of SPI, OSI and SEI is suitable for detecting Canetoad and Litoria caerulea. SPI, HSI and SEI are combined for detecting Litoria nasuta; For Litoria gracillenta, SPI and SEI are used.

TABLE I. Characteristic of Anuran Calls Based on Spectrogram

Anuran species	Frog index			
	Main frequency band (Hz)	Duration of one syllable(ms)	Syllable oscillation rate(cycle/s)	Harmonic structure
Canetoad	400-900	~40	12-25	N/A
Litoria nasuta	1000-3500	~60	N/A	Yes
Litoria gracillenta	1800-3200	1500-2000	N/A	N/A
Litoria caerulea	300-2000	~150	3-8	N/A

IV. Frog Calling Activity Detection System

The anuran calling activity detection system consists of pre-processing, feature extraction, detection and bio-acoustic monitoring. The procedure is depicted in Fig.2 and explained analytically in the following sections.

A. Signal processing

Each 12 hour recording is divided into 720 one-minute segments. For each one-minute recording, Short-time Fourier transform (STFT) is used to obtain the spectrogram. Features are extracted from the one-minute segment. Due to the low SNR of our recordings, noise reduction is essential for extracting indices. We implement the algorithm described in [7] to do noise reduction. This method first produces a spectrum of background noise values, one value for each frequency bin. These values are then subtracted from the 'raw' decibel values with negative values truncated to zero. The result is a spectrogram of positive decibel values with zero decibels representing the base level of the background noise (Fig.1).

B. Feature extraction

In this study, four indices are extracted from one-minute recording for anuran calling activity detection. They are SPI, HSI, SEI and OSI. Among them, SPI, HSI and OSI are first explored for anuran calling activity detection.

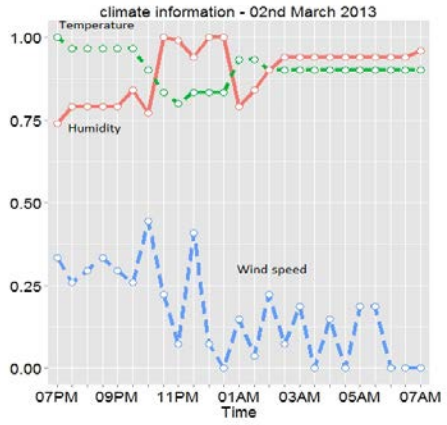
- *Spectral peak track index*

Due to its ability in dealing with noise and competing background noise, spectral peak track is extracted from spectrogram as the index here. We modified the method for extracting spectral peak track based on [8]. The details are described as follows:

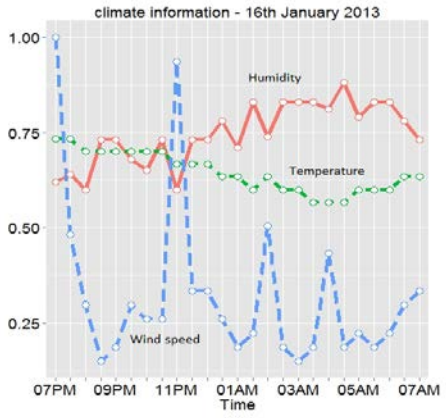
(1) Apply band-pass filter to spectrogram for locating particular frequency band.

$$S[i,j] = \emptyset * B[i,j] \quad (1)$$

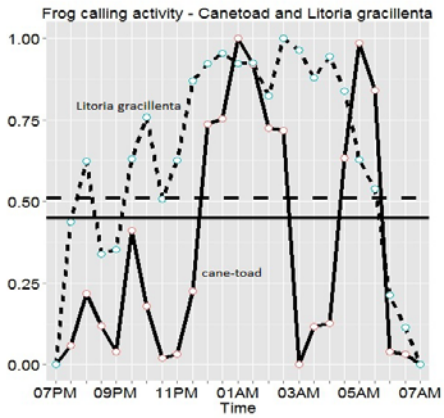
where $S[i,j]$ is filtered spectrogram, $\emptyset$ is band-pass filter, $B[i,j]$ is the original spectrogram.



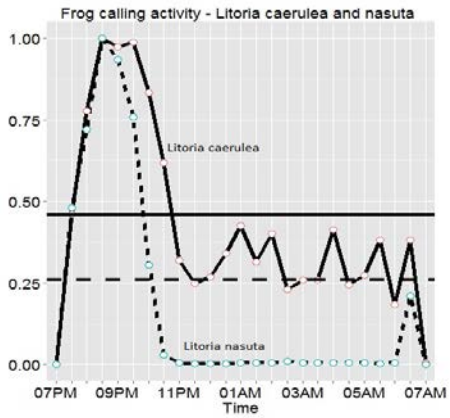
(a)



(b)



(c)



(d)

Fig.3. Anuran calling activity and climate information. The horizontal line is the frog calling activity threshold Both the audio and weather data are selected from 7pm to 7am.

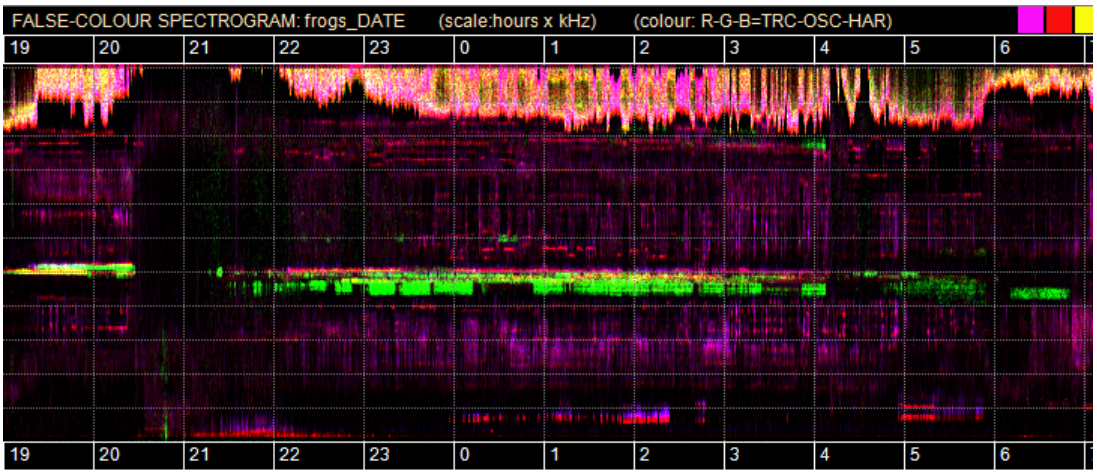


Fig.4. False-colour spectrogram based on spectral peak track, oscillation structure and harmonic structure --- 2nd March 2013. The x-axis is time scale, the audio data is from 7pm to 7am; the y-axis is the frequency, each row represents one frequency.

(2) Extract local peaks from each frame that satisfy the following conditions

$$\begin{cases} S_t[j] - S_t[j-1] > \Delta - 3 \\ S_t[j] - S_t[j+1] > \Delta - 3 \\ S_t[j] - S_t[j-3] > \Delta \\ S_t[j] - S_t[j+3] > \Delta \end{cases} \quad (2)$$

where Δ is the interval threshold of harmonic structure, j is the frequency bin index, t is the time frame index.

(3) Calculate the intervals in time and frequency domain directions of first two peaks. Once the intervals satisfy the corresponding thresholds, new track will be born.

(4) After a new track is born, linear regression will be implemented for predicting the next position of this track. Then calculate the intervals between predicted position and peaks. Then, the intervals and corresponding thresholds are compared to decide whether to grow the track or not.

(5) Once the track stops growing, the duration and slope of the track will be calculated to decide whether to keep the track or not.

- *Harmonic structure index*

Harmonic structure index (HSI) is a feature extracted from successive frames. One efficient algorithm derived from [9] is proposed for detecting harmonic structure. The DCT-based method is used to isolated different frequency components of each frame. The DCT of each frame $S_t[j]$ is defined as

$$y(k) = w(k)\sum_{j=1}^{N} S_t[j]\, cos\frac{\pi(2j-1)(k-1)}{2N}, k = 1, \dots, N \quad (3)$$

where

$$w(k) = \begin{cases} \frac{1}{\sqrt{N}}, & k = 1 \\ \sqrt{\frac{2}{N}}, & 2 \le k \le N \end{cases}$$

Then, noise reduction is realized by one threshold value (T).

$$T = 3\sigma_y \quad (4)$$

where σ_y is the estimated standard deviation of $y(k)$,

$$\sigma_y = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(y(k) - \frac{1}{N}\sum_{k=1}^{N} y(k)\right)^2} \quad (5)$$

Next, we derive sequence $y_f(k)$ from $y(k)$ as follows:

$$y_f(k) = \begin{cases} y(k), |y(k)| \ge T \\ 0, \ |y(k)| < T \end{cases} \quad (6)$$

It means that $y_f(k)$ keeps the most significant component of $y(k)$.

After removing the white noise, different frequency components need to be segmented. The segments are determined by the following conditions:

$$\begin{cases} y_f(k_{s,i}) \ne 0 \ and \ y_f(k_{s,i} - r) = 0, for\ r = 1 \\ y_f(k_{e,i}) \ne 0 \ and \ y_f(k_{e,i} - r) = 0, for\ r = 1,2,3,4 \\ k_{s,i} \le k_{e,i} \end{cases} \quad (7)$$

After each segment is split, we do iDCT and return to time domain, $y_i(k) \xrightarrow{iDCT} S_t[j]$. Then, we apply the zero-crossing rate (ZCR) to detect harmonic structure. The ZCR positions $Z_t(l)$ of $S_t[j]$ for $l = 1,2, \dots, L$ should be found first. Then, the increments of $Z_t(l)$ are obtained as

$$T_t(l) = 2\big(Z_t(l) - Z_t(l-1)\big), l = 1,2,3, \dots, L \quad (8)$$

and calculate the mean period $\overline{T_t}$ and the standard deviation σ_{T_l} of $T_t(l)$ as

$$\overline{T_t} = \frac{1}{L-1}\sum_{l=2}^{L} T_t(l)$$

$$\sigma_{T_l} = \sqrt{\frac{1}{L-2}\sum_{l=2}^{L}(T_t(l) - \overline{T_t})^2} \quad (9)$$

Based on the derived mean and standard deviation, a harmonic structure can be considered to be regular when

$$\frac{\sigma_{T_l}}{\overline{T_t}} < \frac{1}{3} \quad (10)$$

- *Shannon entropy index*

Shannon entropy is introduced in this study for describing the intensity of anuran calling activity. The procedure for Shannon entropy extraction is described as follows:

The histogram H is first calculated based on the decibel value of the peaks in each frequency band.

$$H(n) = \sum\{S_L(n), S_H(n)\} \quad (11)$$

where $S_L(n)$ and $S_H(n)$ determine the decibel range.

Then we normalize the histogram to [0,1] and calculate the normalization factor (NF)

$$H_f(n) = \frac{H(n) - min(H(n))}{max(H(n)) - min(H(n))} \quad (12)$$

$$NF = \frac{In(L)}{In(2)} \quad (13)$$

where L is the length of the histogram.

At last, we calculate the Shannon entropy index

$$SEI = \frac{-\sum_i^N log_2(NF_i)}{log_2(N)} \quad (14)$$

- *Oscillation structure index*

Oscillation structure widely exists among anuran calls. The windowed correlation technique is applied here for the detection [10].

Let $E[i,j]$ be the windowed power spectrum of signal S. Then the energy $N_{l,h}$ for a frequency range from l to h at time t is defined as

$$N_{l,h}[E](j) = \sum_l^h E(i,j)(E(i,j) - E(i,j+1))^2 \quad (15)$$

The squared difference is chose for emphasising large change in energy over small ones.

From the energy change, we derive one criterion $B[E](t)$ for the presence of the frog calls at time t which is used for smoothing the noise energy.

$$B[E](t) = N_{l,h}[E](j) - \theta * N_{l_b,h_b}[E](j) \quad (16)$$

After extracting $B[E]$, we calculate the windowed autocorrelation $A(\tau, t)$ of $B[E]$.

where τ is the autocorrelation lag, and the time t gives the centre of the window.

At last, one feature sequence is derived by

$$\tilde{A}(\tau, t) = \frac{1}{h}\left(\sum_{\tau=a}^{a+h-1} A(\tau, t)\right) - \frac{1}{k}\left(\sum_{\tau=b}^{b+k-1} A(\tau, t)\right) \quad (17)$$

Then the typical call oscillation structure can be measured by the strength of the autocorrelation at lags $a, \dots, a+h-1$.

C. Modelling of anuran calling activity

Let $F_k^c[SPI, HSI, SEI, OSI]$ be the vector representing the segment belonging to the same class c where $c = 1, \dots, C$ and C is the total number of classes. For one particular frog species, if the particular index is not extracted, then the feature value will be set to zero. We model F_k^c with a probability density

TABLE II. DETECTION RESULT OF ANURAN CALLING ACTIVITY FOR DIFFERENT ANURAN SPECIES

In this table, the ground truth means the number of minutes that contains corresponding anuran species. Three measurements are employed to verify the result. TP: True positive, FP: false positive, FN: false negative.

Anuran species	**Ground truth**	**TP**	**FP**	**FN**	**Precision**	**Recall**
Canetoad	163	157	89	6	63.82%	96.31%
Litoria nasuta	138	128	34	10	79.01%	92.75%
Litoria gracillenta	345	264	87	59	75.21%	76.52%
Litoria caerulea	91	88	18	3	83.01%	96.70%

function with a weighted sum of Gaussian component densities. The Gaussian mixture mode of F_k^c can be given by

$$p(F_k^c|\lambda) = \sum_{i=1}^{M} w_i \, g(F_k^c|\mu_i, \sum_i) \quad (18)$$

where $w_i, i = 1, \dots, M$ are the mixture weights, the component Gaussian densities are $g(F_k^c|\mu\_i, \sum_i), i = 1, \dots M$. Each component density is a Gaussian function of the form with mean vector μ_i and covariance matrix $\sum_i$.

$$g(F|\mu_i, \sum_i) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} \exp\{-\frac{1}{2}(F - \mu_i)' \sum_i^{-1}(F - \mu_i)\} \quad (19)$$

The mixture weights satisfy the constraint: $\sum_{i=1}^{M} w_i = 1$. Note that each covariance matrix represents the anuran calling activity of one specific species.

During recognition, let T be the feature vector of the unknown segment. The maximum likelihood estimation is employed for making the decision.

$$D(c) = \operatorname{argmax} P(c)P(T|c), c = 1, \dots, C \quad (20)$$

where $c = 1, \dots, C$ and C is the total number of classes. In our study, there are only two possibilities of the detection result: (1) with frog calling activity, and (2) without frog calling activity. Therefore, $P(c) = \frac{1}{2}$. The frog calling activity is finally determined by finding the segments with maximum likelihood value.

V. EXPERIMENT RESULT

In this experiment, long audio recordings are split into one-minute segment to identify frogs' calling activities. Precision and recall are calculated for evaluating the detection system. The ground truth is generated by an anuran expert who labels two 12-hour recordings containing anuran calls. Four indices are extracted from each one-minute segment. Then, GMM is used for detecting the activity combining maximum likelihood estimation. The detection result is displayed in Table II with a high precision and recall. The changes of climate information [11] and anuran calling activity through 12 hours are displayed in Fig.3. Many parameter values are normalized for better display: The anuran calling activity value is normalized to $[0, 1]$ according to the duration of the spectral peak tracks; the temperature is normalized from $[0{,}35\,℃]$ to $[0{,}1]$; the wind speed is normalized from $[0{,}50 km/h]$ to $[0{,}1]$; the humidity is normalized to relative $[0{,}1]$. The anuran calling activity is detected based on the combination of SPI, HSI, SEI and OSI indices. Fig.3 shows the correlation between anuran calling activity and climate information. Fig.3 (c) and (d) demonstrate that different anuran species tend to call together.

Based on extracted indices, a visualization technique named False-colour spectrogram is applied for visualizing the detection result. For different anuran species, three suitable normalized indices are mapped to three primary colours: red, green and blue (RGB) respectively [7]. The x-axis extends from 7:00 pm to 7:00 am. Compared with the standard spectrogram, the x-axis scale is compressed more than 1000 times. However, the frequency scale is unchanged. From the visualization result (Fig.4), we can find that there is thunderstorm from 9:00 pm to 11:00 pm which is highly correlated with the climate information in Fig.3 (a). The distributions of Canetoad and Litoria gracillenta calling activities are also well displayed. The visualization result shows that this system can be used to assist ecologists locate anuran calling activity quickly and accurately.

VI. CONCLUSION

This study develops a new technique to detect anuran calling activities in long field recordings with low SNR. Different from previous methods, syllable segmentation is not employed in our method. The anuran calling activity is detected based on the summary indices of one-minute recordings. Four indices, spectral peak track index, harmonic structure index, Shannon entropy index and oscillation structure index are employed. The anuran calling activity is then modelled by GMM for detection. The False-colour image technique is used to visualize the detection result which shows a high accuracy rate. In the future, we will apply this technique to more anuran species.

REFERENCES

[1] Han, N.C., S.V. Muniandy, and J. Dayou, Acoustic classification of Australian anurans based on hybrid spectral-entropy approach. Applied Acoustics, 2011. 72(9): p. 639-645.

[2] Wimmer, J., et al., Sampling environmental acoustic recordings to determine bird species richness. Ecological Applications, 2013. 23(6): p. 1419-1428.

[3] Huang, C.-J., et al., Frog classification using machine learning techniques. Expert Systems with Applications, 2009. 36(2): p. 3737-3743.

[4] Chen, W.-P., et al., Automatic recognition of frog calls using a multi-stage average spectrum. Computers & Mathematics with Applications, 2012. 64(5): p. 1270-1281.

[5] Tyagi, H., et al. Automatic identification of bird calls using spectral ensemble average voice prints. in Proceedings of the 13th European signal processing conference. 2006.

[6] Lee, C.-H., et al., Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis. Pattern Recognition Letters, 2006. 27(2): p. 93-101.

[7] Towsey, M., et al., Visualization of Long-duration Acoustic Recordings of the Environment. Procedia Computer Science, 2014. 29: p. 703-712.

[8] Roch, M.A., et al., Automated extraction of odontocete whistle contours. The Journal of the Acoustical Society of America, 2011. 130(4): p. 2212-2223.

[9] Li, X., et al., The DCT-based oscillation detection method for a single time series. Journal of Process Control, 2010. 20(5): p. 609-617.

[10] Bardeli, R., et al., Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. Pattern Recognition Letters, 2010. 31(12): p. 1524-1534.

[11] http://www.wunderground.com,retrieved on 11th Dec..