Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference

Timed Probabilistic Automaton: A Bridge between Raven and Song Scope for Automatic Species Recognition

Shufei Duan, Jinglan Zhang, Paul Roe, Jason Wimmer, Xueyan Dong, Anthony Truskinger, Michael Towsey
Microsoft QUT *e*Research Center, Queensland University of Technology
shufei.duan@student.qut.edu.au

Abstract

Raven and Song Scope are two, state-of-the-art automated sound analysis tools, based on machine learning techniques for detection of species vocalisations. Individually, these systems have been the subject of a number of reviews; however, to date there have been no comparisons made of their relative performance. This paper compares the tools based on six aspects: theory, software interface, ease of use, detection targets, detection accuracy, and potential applications. Examining these tools, we identified that they fail to detect both syllables and call structures, since Raven only aims to detect syllables while Song Scope targets call structures. Therefore, a Timed Probabilistic Automata (TPA) system is proposed which separates syllables and clusters them into complex structures.

Deleting this Example of a Section Heading will Remove the Copyright Line

Introduction

Animal call recognition plays a significant role in environmental monitoring where it can be used as an indicator of species diversity, abundance and overall environmental health (Towsey et al., 2012). Manual analysis is effective for single species identification but can not deal with datasets over large spatiotemporal scales. Automated species recognition tools greatly facilitate animal call recognition especially over large datasets by reducing processing time, and increasing the efficiency.

Two state-of-the-art applications, Raven (Bioacoustics Research Program, 2011) and Song Scope (Wildlife Acoustics, 2011) have been developed to assist ecologists in dealing with large amounts of data. However, though they have been widely used for years, very little research has compared their performance on real-world data set (Crothers, Gering, & Cummings, 2011; Depraetere et al., 2012). The aim of this paper is to explore the performance of these tools and potential application areas using real-world datasets.

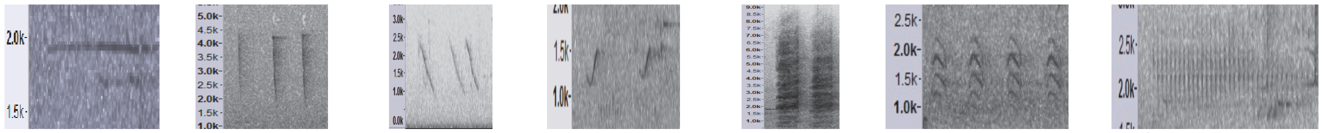
Examining the systems we found that Raven and Song Scope are built for recognizing either syllables or call structures, not for both. To build a bridge between them, we present Timed Probabilistic Automata (TPA) to join syllable and call structure detection together.

Call Structures

Many animal calls have hierarchical structures. A typical bird song is divided into phrases, syllables, and elements (Somervuo, Harma, & Fagerlund, 2006). Generally, syllables mean timestamps in an audio stream (Zhuang et al., 2010) while call structures consist of single or multiple syllables.

Since animal call structures are comprised of some common patterns, there are many attempts to define these typical components (McCallum, 2010; Scott Brandes, 2008). Different from component definitions in phonetics, Duan defined broad acoustic components according to their appearance in the spectrogram (Duan et al., 2011). These components can be classified into two types: primitive and composite. The primitive components include whistle (a horizontal line), click (a vertical line), slur (from the whip to a slow chirp), warble (moduled in one direction and then back again), and blocks (energy concentrated rectangular or triangular areas). The composite components include stacked harmonics (vertical stacks of lines or warbles spaced equally) and oscillations (horizontally repeated acoustic components). Specifically, Figure 1shows the appearance of components and the typical species which are comprised of these components.

The primitive components as well as stacked harmonics are similar to the common definition of syllables. They are inseparable in time and can be used to construct call



EasternWhipbird, Little Bronza-Cuckoo, Gray Fantail. (a)Whistle

Eastern Yellow Robin, Eastern Whipbird, Mistletoebird. (b) Click

Little Friarbird, White-browed Scrubwren. (c) Slurs

Indian Peafowl, Eastern Koel, Australian Owlet-Nightjar. (d) Warble

Rainbow lorikeet, Masked Owl. (e) Block

Australian Owlet-Nightjar, Corellla sp, Crow. (f) Stacked Harmonic

Lewin's HoneyEater, Rufous Whistler, White-browed Scrubwren, Little Friarbird. (g) Oscillation

Figure 1. Acoustic Components' Appearance and Representative Species

structures. Oscillation is a special component which is a quite common call structure among animal calls. It consists of primitive components typically clicks or stacked harmonics. Duan (2011) categorized it as a component because detecting this pattern is also fundamental in animal call recognition. The names of these components follow McCallum's definition.

Software

Raven

Raven, produced by the Cornell Lab of Ornithology, is a software program for the acquisition, visualization, measurement, and analysis of sounds (Charif et al., 2010). Raven can render audio files as waveforms and spectrograms, and allows users to apply a set of analysis tools. Historically, it has been designed for birdsong analysis, and provides tools to perform band-pass filters and manual or semi-automatic syllable segmentation (Stowell & Plumbley, 2011). Raven has an intuitive user interface and is relatively easy to learn. It also has very powerful play and cut modules so users can focus on the specific fraction of a sound clip that they need to analyze. In terms of target detection, Raven has two detectors: a band limited energy detector and an amplitude detector.

- Band Limited Energy Detector:
  Estimates the background noise of a signal and uses this to find sections of the signal that exceed a user-specified signal to noise ratio threshold in a specific frequency band, during a specific time.
- The Amplitude Detector:
  Detects regions of a signal where the magnitude of the waveform's envelope exceeds a threshold value.

These detectors are relatively simple to configure, and clear instructions are provided in the manual. The band limited energy detector is based on the spectrogram while the amplitude detector works on the waveform.

Raven can also perform batch processing which allows users to run the detector over a large datasets. This is a considerable advantage when performing analyses on large volumes of data.

Primarily, Raven aims to detect syllables. Multiple detectors can be run over one spectrogram (waveform), which allows one to build separate detectors for different syllables. However, it is limited for detecting call structures that contain multiple syllables. Even if different syllables can be detected, they are not joined together to form a call structure.

Song Scope

Song Scope, produced by Wildlife Acoustics, Inc., is a sophisticated digital signal processing application designed to quickly and easily scan audio recordings made in the field, and automatically locate vocalizations made by specific bird species and other wildlife (Song Scope 4.0 User's Manual, 2011).

Compared with Raven, Song Scope does not have general purpose recording or play back controls. Furthermore, it does not allow users to replay particular sections without annotating these sections and saving them as new files. Song Scope also centers on audio files viewed as waveforms and spectrograms. The user interface is simple and spectrograms are rendered in colour. One potential drawback of the colour spectrogram is the potential to influence the user's perception and interpretation of the data compared with a gray-scale alternative (Rogowitz, Treinish, & Bryson, 1996).

Song Scope is aimed at detecting call structures, which is different approach to Raven. The Song Scope classification algorithms are based on Hidden Markov Models using spectral feature vectors similar to Mel Frequency Cepstral Coefficients as these methods have been proven to work effectively in robust speech recognition applications (Agranat, 2009).

We observed that Song Scope segments the syllables first and then clusters the related segments to form call structures. However, this approach is very sensitive to the purity of syllables. If syllables are polluted by non-target species or background noise, the model is very sensitive, thereby affecting the recognition accuracy.

Using Song Scope effectively requires some background knowledge of signal processing to understand and setup the parameters. Song Scope also supports batch processing to deal with large datasets.

Regarding the annotation work, both Raven and Song Scope cannot accept existing call tags. This is inconvenient to share work among different research groups. In our case, we have already collected a library of tags which were labeled by bird watchers. The quick and convenient way for us to use software is to directly import these tags into the software so that we do not need to label them twice.

Experiments

To evaluate the call detection performance of Raven and Song Scope, we configured each system and tested them independently. .

Dataset

The testing dataset was collected from the Samford Valley (20 kilometres north-west of Brisbane, Queensland, Australia) during the dawn chorus from 4am to 9am, 14th, Oct, 2010. This is a dataset tagged by a team of bird watching enthusiasts. During the dawn period 46 species were detected and annotated by bird observers using a custom-built online acoustic analysis tool (Wimmer et al 2012). Among these species, five representative samples were selected to characterize different types of call structures as mentioned in section 2 (see Figure 1), Lewin's Honeyeater (*Meliphaga lewinii*) for oscillations, Eastern Whipbird (*Psophodes olivaceus*) for whistles and clicks, Eastern Koel (*Eudynamys orientalis*) for warbles, Torresian Crow (*Corvus orru*) for stacked harmonics, and Rainbow Lorikeet (*Trichoglossus haematodus*) for blocks. There are in total 131 minutes among five hour recordings which contain HoneyEater labels, 167 minutes contain EasternWhipbird labels, 237 minutes contain Eastern Koel labels, 67 minutes contain Torresian Crow labels, and 93 minutes contain Rainbow Lorikeet labels. The training dataset were selected from the same site but from a different day (15th Oct, 2010). Each species has 25 samples for training. This dataset is accessible on request.

Figure 2 shows the signal to noise ratio (SNR) distribution for the five hours of dawn chorus. The x-axis represents the time range in 10 minute interval from 240th minute (4am) to 540th minute (9am). The y-axis represents the SNR (in dB). The average SNR is 13 dB, while the maximum is 33 dB and the minimum is 3.7 dB. As we can see, there are three peaks located at periods: (290, 310), (390,420), and (450, 470). The minimum value for peak time is about 23 dB.

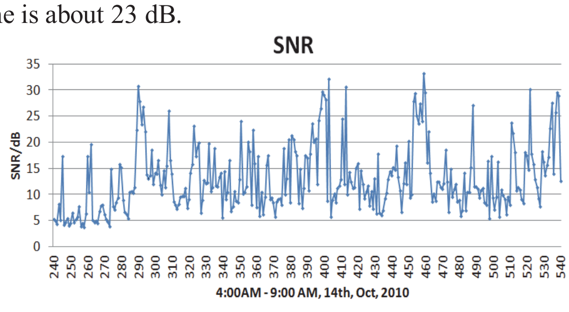


Figure 2. Noise Distribution over Dawn Chorus (4am-9am), 14th, Oct, 2010.

There are many species calling at the same time which can potentially cause inaccurate detection results. Due to the high number of species vocalising simultaneously during the dawn period, detecting targets during the dawn chorus is challenging for both automated tools and humans.

Software Version and Parameter Setting

The software versions tested were: Raven Pro 1.4 and Song Scope 4.1.1, respectively. Both systems require extensive configuration for each individual species recognizer. Specific configuration details for each system are detailed in the respective user manuals. Twenty five training samples for each species were selected from the same site but over different days. Table I lists the parameters for each species and each tool.

Table I. Parameters for Different Species and Tools.

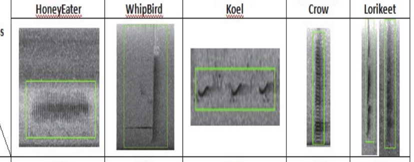
Tools	Parameters	Lewin's HoneyEater	Eastern WhipBird	Eastern Koel	Torresian Crow	Rainbow Lorikeet
Raven	Min Frequency (Hz)	1000	290	765	443	1327
	Maximum Frequency (Hz)	3100	4135	1800	3216	9125
	Minimum Duration (s)	2.57	1.0	0.25	0.19	0.20
	Maximum Duration (s)	10.82	2.5	1	0.33	0.46
	Minimum Separation (s)	0.38	0.17	0.15	0.06	0.06
	Minimum Occupancy (%)	15	20	20	20	20
	SNR threshold (dB)	5	5	5	5	5
	Block size (s)	33	10	3.5	1.0	1.5
	Hop size (s)	11	5	2.5	0.7	1.0
	Percentile	10	10	20	30	20
Song Scope	FFT size	256	1024	1024	512	512
	FFT overlap	½	½	½	½	½
	Frequency Minimum	11	20	60	12	38
	Frequency Range	24	140	37	104	167
	Amplitude Gain (dB)	0	0	0	0	0
	Background Filter (s)	1	4	1	1	1
	Max Syllable (ms)	23	2000	736	448	360
	Max Syllable Gap (ms)	12	488	10	32	46
	Max Song (ms)	2601	2067	800	448	453
	Dynamic Range (dB)	15	20	20	18	20
	Maximum Complexity	48	32	48	32	48
	Maximum Resolution	6	6	6	6	20
	Total training result	76.80±9.57%	83.32±2.27%	82.64±4.11%	81.47±3.09%	81.84±4.99%

Results

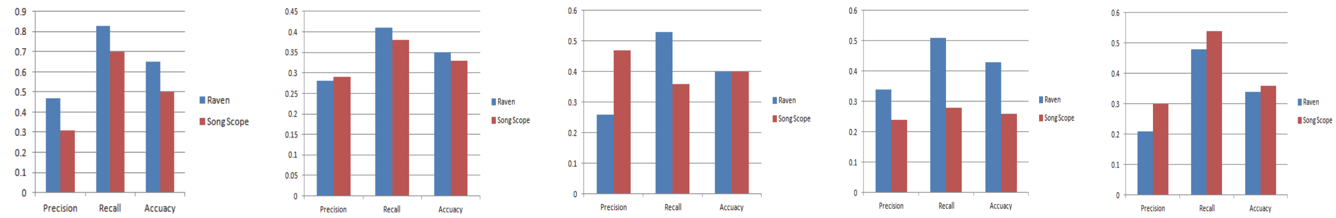
Table II presents the accuracy of each tool for the five species tested. Figure 3 shows the comparison of precision, recall, and accuracy for each species. Precision and recall are defined in (Olson & Delen, 2008).

Table II demonstrates that on average, the performance of Raven and Song Scope was similar for the species tested. Raven had a higher detection rate for the Lewin's Honeyeater (oscillations) and Torresian Crow (stacked harmonics). Song Scope performed slightly better for block-type calls (Rainbow Lorikeet). Detection rates for Eastern Koel (warbles) and Eastern Whipbird (lines) were roughly equivalent for both systems.

Table II. Accuracy of Tools for Sample Species Detection.

Species / Tools	Lewin's HoneyEater	Eastern WhipBird	Eastern Koel	Torresian Crow	Rainbow Lorikeet
					
Raven	0.65	0.35	0.40	0.43	0.34
Song Scope	0.50	0.33	0.40	0.26	0.36

Overall, the average accuracy of Raven is approximately 0.43 while 0.37 for Song Scope. This is because Raven detects syllables while Song Scope works on call structures with multiple syllables. Raven focuses on small sections of energy while Song Scope models the structure among



(A) Lewin's Honeyeater (B) Eastern Whipbird (C) Eastern Koel (D) Torresian Crow (E) Rainbow Lorikeet

Figure 3. Precision Recall and Accuracy of each Species.

syllables. 'Noisy' training samples affect the test results of Raven, but the effect is much greater on Song Scope as Hidden Markov Model treats noisy signals as syllables and models them as a call structure.

The average precision of Raven is approximately 0.25 (Figure 3), which means the false positive rate is high. This is reasonable due to the feature of energy. Acoustic events exceeding the threshold will be detected as Raven disregards the internal structure. Recall is high, which is approximately 0.50, which indicates that it detected approximately half the calls. If users are aiming to detect the activity level of targets, Raven may be more suitable.

The precision of Song Scope is around 0.32 which is relatively higher than Raven and reflects a low false positive rate. The recall is much lower than Raven, which means the ability of Song Scope to detect calls is less than Raven. However, once a call is detected, the signal is more likely to be a true positive. If users want to detect the presence of targets, Song Scope may be suitable.

Timed Probabilistic Automata (TPA)

Raven targets syllables, however, it cannot join these syllables together to form call structures. Song Scope detects call structures by clustering syllables, but fails to accurately separate syllables. These tools were developed in an attempt to detect a wide range of syllables and call structures. Overall, given the generic nature of the tools, we consider that the average performances satisfactory. The critical gap here is the lack of an approach to join the two aspects together in terms of a better recognition result.

Timed Probabilistic Automata (TPA) are developed to address this problem. It not only allows users to run the syllable detectors, but also give users the ability to build call structures by themselves.

Theory and Process

TPA was adapted from the theories of Syntactic Pattern Recognition and the Markov Model. Though the syntactic complexity of birdsongs cannot be directly compared with human speech due to a lack of semantics and lexicon (Berwick et al., 2011), the call structures of many avian species can be modeled by low-order Markov chains. This implies that the full power of human speech recognition is probably not needed. For many applications very simple recognizers may be suitable.

These observations inform the design of TPA, which breaks with current practice in several ways. First the algorithm locates specific shapes in the spectrogram. From the shapes identified in the spectrogram a sequence of *acoustic components* is derived, each of which is characterised by a tuple: (*shape, start time, duration, minimal frequency, maximal frequency*). Components map to symbols in the alphabet of a call-specific language described by a probabilistic timed finite automaton. A recogniser for the call-specific language, tuned with parameter values obtained from components obtained from a training set of positive examples, is used for classification of previously unlabelled input.

The core part of TPA is acoustic component detectors. These detectors are developed specifically for five types of call components: lines, warbles, blocks, oscillations, and stacked harmonics. Acoustic component detectors work as filters in the spectrogram. They are parametric and relatively easy to configure according to the specific targets. The processes of TPA for automatic animal call recognition are:

(1) According to the target's call structure shown in the spectrogram, select the proper acoustic component detector.
(2) Execute the component detector. The result of the detector is a list of components found in the spectrogram. These components are characterized by a tuple: (shape, start time, duration, minimal frequency, maximal frequency).
(3) Component filtering. Choose the training samples and train them to filter out components that do not belong to the target species.
(4) Using a timed automaton to model and control time duration of the whole target call.
(5) Apply probabilistic automata to represent the target species call structure in a sentence way.
(6) Similarity matching. Match the testing representation with the training one. If the probability falls in the training probability distribution, a target call is recognized.

Eastern Whipbird

To illustrate our approach we examine an application of TPA to the call of the Eastern Whipbird. The call structure of an Eastern Whipbird contains a whistle and a click (see Table II). The state transition diagram for TPA detection of the Eastern Whipbird is shown in Figure 5. P(*w*), P(*g*), P(c) denote the probability of whistle, gap, and click, respectively.

$$P(Whipbird) = \mathrm{P}(w) \times \mathrm{P}(g) \times P(c) \qquad (1)$$

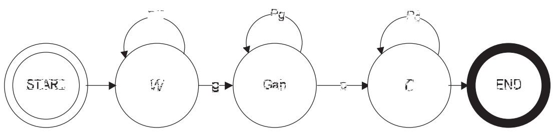


Figure 5. TPA for Eastern Whipbird.

First we call the whistle and click detectors for the component toolbox to detect whistles and clicks. Model whistle, click and gap in-between using frequency and time information which have already collected in the tuple. The TPA is applied as follows:

(1) Whistle filtering. Calculate the probability of all testing whistles. Compare this probability with the training probability. If the testing probability value is between the maximum and minimum value of training probability, a confirmed Whipbird whistle hit. Remove all irrelevant whistles.

(2) Click filtering. Calculate the probability of all testing clicks. Compare this probability with the training probability. If the testing probability value is between the maximum and minimum value of training probability, a confirmed Whipbird click hit. Remove all irrelevant clicks.

(3) Gap filtering. Calculate the probability of all gaps between Whipbird whistles and clicks. Compare this probability with training gap probability. If the testing probability value is between the maximum and minimum value of training probability, a confirmed whipbird gap hits. According to this confirmed gap value, keep pairs of whistle and click which have the confirmed gap. Remove all irrelevant whistles and clicks.

(4) Marquee the left pairs of whistles and clicks as Eastern Whipbird call.

Figure 6 illustrates experimental results of Eastern Whipbird recognition. Blue dots are signals left after noise removal. Green lines represent whistle and clicks. The red marquee covers Whipbird call.

Comparison with Raven and Song Scope

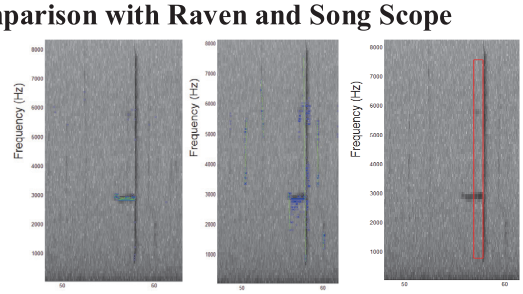


Figure 6. Eastern Whipbird Recognition by TPA.

To test the performance of TPA, we have compared it with Raven and Song Scope. Experiments were executed under the same conditions using the same training and testing dataset as in section 4. The training process of TPA was conducted on both acoustic component detectors and automata. The parameters were manually configured. Table III lists the statistics of these three tools. To better illustrate points, we graph the comparison results and add the error bars with standard deviation in Figure 7.

Table III. Statistics of Tools for Eastern Whipbird.

	Raven	Song Scope	TPA
Precision	0.28	0.29	0.45
Recall	0.41	0.38	0.59
Accuracy	0.35	0.33	0.52

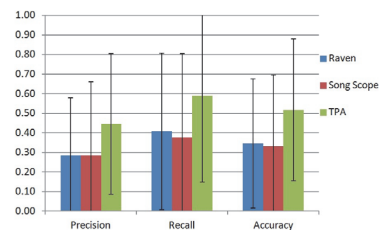


Figure 7. Comparison among Raven, Song Scope and TPA.

Clearly, TPA outperforms Raven and Song Scope under the same conditions during dawn chorus (4am to 9am) for whip bird detection. Precision, recall, and accuracy all have dramatic increase. The error bars show the distribution of precision, recall and accuracy. As we can see, the distribution is consistent among three indices. However, the standard deviation is high. This is because the testing data is from dawn chorus when many species call at once. In total of 114 minutes where there were whip bird calls, the signal is either too weak or too noisy; this noise causes tools fail to detect. Therefore, the precision and recall are all zero. Zero precisions have a strong negative impact on the mean and increase the standard deviation. Even the rest of minutes have better recognition results, the percentage over the total minutes is small. Table IV shows the number and percentage of zero precisions after detection of tools. From this table, we are convinced that even under noisy situation, the recognition ability of TPA is still better than Raven and Song Scope. However, we admit that detecting targets during dawn chorus is a difficult research problem: the accuracy of TPA is still only 0.52.

Table IV. The number of zero precisions

	Raven	Song Scope	TPA
Precision (0)	44	58	32
Percentage Over total minutes (114)	39%	51%	28%

Discussion and Conclusion

Raven and Song Scope are well-developed tools for bioacoustic monitoring. We compared their performance using real data, collected in typical, challenging acoustic environments.

In theory, Raven utilizes two different detectors to locate the syllables in the spectrogram while Song Scope can detect the call structures using feature vectors and Hidden Markov Models.

Compared with Song Scope, Raven has a more intuitive user interface and more powerful control modules. Because Song Scope requires expertise about signal processing to configure parameters, it is more difficult to use than Raven. In terms of the recognition ability for five types of call components, Raven had relatively better performance than Song Scope with accuracy of 0.43 and 0.37, respectively. The precision of Song Scope is higher but the recall is lower. This indicates that Raven can be applied to detect the activity of animals while Song Scope can be used to detect the presence of a target.

Raven detects only syllables and Song Scope only detects call structures; TPA is different. It is designed not only for building acoustic component (syllable) detectors separately, but also for using Syntactic Pattern Recognition and Markov chains to cluster the components in order to form call structures. This clustering can provide the basis for a user operated tool that will allow users to run the component filters and build call structures according to their specific targets. Compared with Song Scope and Raven, the precision, recall and accuracy are all increased with TPA. Even in the noisy environment (dawn chorus), TPA picks up an extra 10% of signals than the other two tools for whipbird detection.

This paper is part of an ongoing research project for automatic species recognition. The TPA approach is still under testing and construction. Further research for multiple sites and multiple days, with more species with complex call structures are required.

It is a difficult task to recognize targets during dawn chorus in the automated species call recognition research area based on existing machine learning techniques. TPA has achieved only approximately 50% accuracy. It has not yet reached a level of reliability that allows ecologists to use the methods without careful verification of results. Much work is required for the real applications in future.

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

References

Agranat, I. (2009). Automatically Identifying Animal Species from their Vocalizations. Paper presented at the Fifth International Conference on Bio-Acoustics.

Berwick, R. C., Okanoya, K., Beckers, G. J. L., & Bolhuis, J. J. (2011). Songs to syntax: the linguistics of birdsong. *Trends in Cognitive Sciences, 15*(3), 113-121.

Bioacoustics Research Program, 2011. Raven Pro: Interactive Sound Analysis Software (Version 1.4) [Computer software]. Ithaca, NY: The Cornell Lab of Ornithology. Available from http://www.birds.cornell.edu/raven.

Charif, RA, LM Strickman, AM Waack., 2010. Raven Pro 1.4 User's Manual. The Cornell Lab of Ornithology, NY.

Crothers, L., Gering, E., & Cummings, M. (2011). Aposematic Signal Variation Predicts Male–Male Interactions in A Polymorphic Poison Frog. *Evolution*, 65(2), 599-605.

Depraetere, M., Pavoine, S., Jiguet, F., Gasc, A., Duvail, S., & Sueur, J. (2012). Monitoring animal diversity using acoustic indices: Implementation in a temperate woodland. Ecological Indicators, 13(1), 46-54.

Duan, S., Towsey, M., Zhang, J., Truskinger, A., Wimmer, J., & Roe, P. (2011, 6-9 Dec. 2011). Acoustic component detection for automatic species recognition in environmental monitoring. Paper presented at the Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2011 Seventh International Conference on.

McCallum, A. (2010). Birding by ear, visually. Part 1: Birding acoustics. *Birding*, 42, 50-63.

Olson, D. L., & Delen, D. (2008). Advanced Data Mining Techniques (1 ed.): Springer.

Rogowitz, B. E., Treinish, L. A., & Bryson, S. (1996). How not to lie with visualization. *Comput. Phys., 10*(3), 268-273.

Scott Brandes, T. (2008). Automated sound recording and analysis techniques for bird surveys and conservation. *Bird Conservation International*, 18(S1), S163-S173.

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(6), 2252-2263.

Song Scope 4.0 User's Manual, 2011. Wildlife Acoustics, Inc, USA.

Stowell, D., & Plumbley, M. (2011). Birdsong and C4DM: A survey of UK birdsong and machine recognition for music researchers: Centre for Digital Music, Queen Mary, University of London.

Towsey, M., Planitz, B., Nantes, A., Wimmer, J., & Roe, P. (2012). A toolbox for animal call recognition. *Bioacoustics*, 1-19.

Wildlife Acoustics, 2011. Song Scope: Bioacoustics Software (Version 4.1.1) [Computer Software]. USA: Wildlife Acoustics, Inc. Available from http://www.wildlifeacoustics.com/products/analysis-software.

Zhuang, X., Zhou, X., Hasegawa-Johnson, M. A., & Huang, T. S. (2010). Real-world acoustic event detection. *Pattern Recognition Letters, 31*(12), 1543-1551.