EUSIPCO 2013 1569744753

UNSUPERVISED DISCOVERY OF ACOUSTIC PATTERNS IN BIRD VOCALISATIONS EMPLOYING DTW AND CLUSTERING

Peter Jančovič[1] **, Münevver Köküer*[2,1], *Masoud Zakeri*[1] *and Martin Russell*[1]

[1] School of Electronic, Electrical & Computer Engineering, University of Birmingham, UK
[2] Faculty of Technology, Engineering & Environment, Birmingham City University, UK
{p.jancovic,mxz848,m.j.russell}@bham.ac.uk, munevver.kokuer@bcu.ac.uk

ABSTRACT

This paper presents a method for an unsupervised discovery of acoustic patterns in bird vocalisations recorded in real world natural environments. The proposed method employs sinusoidal detection to provide frequency tracks which are used as features to characterise bird tonal vocalisations. A variant of dynamic time warping, capable of searching for multiple partial matchings, is used to segment the data based on these frequency track sequences. Agglomerative hierarchical clustering approach is then employed to cluster recurring segments. Evaluations are performed on audio recordings provided by the Borror Laboratory of Bioacoustics. The obtained results indicate that structurally distinct stereotyped acoustic units can be determined.

Index Terms— unsupervised, clustering, segmentation, dynamic time warping, bird, vocalisation, sinusoid, tonal

1. INTRODUCTION

Bird vocalisations can be considered to be composed of subunits of different levels, such as elements (also referred to as notes), syllables, phrases and songs. Elements can be taken as the smallest structurally distinct stereotyped acoustic units produced by birds, and these can be thought of similarly as phonemes in the context of speech processing. While large amount of phoneme (or higher) level of annotated data exists for speech, there are no wide range publically available annotated data for bird vocalisations. Such annotated bird acoustic data and the inventory of units of bird vocalisations are important both for bioacousticians, for instance, to study differences between individuals and populations or behaviour contexts, and for development of more advanced automated systems for processing of bird vocalisations.

Unsupervised processing of time series data and searching for recurring patterns relates to current research in various fields, from computational biology to audio summarisation. A recent review of time series matching approaches was presented in [1]. We focus here on works in speech and audio processing. An unsupervised derivation of variable-length acoustic units from speech signal employing hidden Markov models was investigated in [2]. The authors in [3] employed dynamic time warping (DTW) and neural networks for an unsupervised categorisation of isolated vocalisations of dolphins and whales. The work in [4] employed a segmental variant of DTW for unsupervised processing of speech data to automatically extract words and linguistic phrases from recordings of academic lectures. In [5], the segmental DTW and K-means clustering was employed for unsupervised learning of acoustic events, with evaluations presented for spoken digits and non-speech sounds in meeting rooms. In [6], a similarity matrix approach was used to summarise music data.

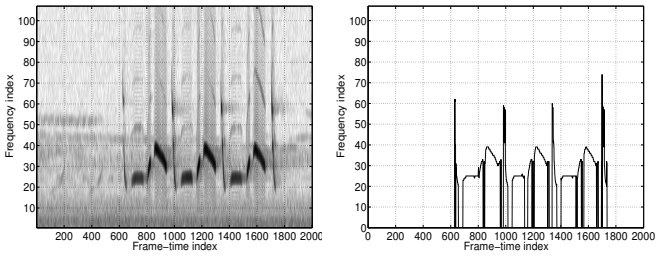
Automatic processing of bird vocalisations is a relatively recent research field [7, 8, 9]. The data used in many studies up to date consists of recordings of relatively isolated bird vocalisations without noise. Some studies used continuous recordings and split the signal into smaller segments either by human intervention of spectrograms [9] or automatically using an energy-based threshold decision in time or time-frequency domain [7, 10, 11, 12]. Such energy-based segmentation may be difficult to obtain accurately in recordings of bird vocalisations in their natural habitat due to being usually contaminated by various background noise or vocalisations of other birds or animals.

In this paper, we propose an approach for unsupervised discovery of acoustic elements in bird vocalisations. As we are dealing specifically with bird tonal vocalisations, we employed an algorithm, which we introduced in [13, 14], to decompose the entire acoustic scene into sinusoidal components. This is then used for detection and estimation of frequency tracks that are used in this paper as temporal sequences for further processing stages. Note that the further stages of the processing are not dependent on the type of features and thus the presented work could also be applied to birds producing non-tonal vocalisations. We developed a variant of DTW which can search for multiple partial matchings within given sequences. The resulted segments are then, based on their DTW measured similarity, clustered using a hierarchical clustering approach. Experimental evaluations show that the proposed method can provide a set of structurally distinct stereotyped bird vocalisation patterns.

2. ESTIMATION OF FREQUENCY TRACKS

As we consider birds producing tonal vocalisations in this paper, we can describe an audio signal in terms of sinusoidal components. Based on the detected sinusoidal components, we then characterise the signal in terms of the frequency of the most prominent sinusoidal component detected at each frame-time. This section gives a brief summary of the method we employed for detection of sinusoidal components, which we introduced in [13] and further improved in [14] and employed for processing of speech [15] and bird signals in [8]. Note that the presented method could be directly employed for dealing with simultaneous bird vocalisations, this however is not the aim of this paper. As such we simply consider that the bird of interest produces the loudest sinusoidal component.

An example of a spectrogram of an audio field recording from the Borror data [16] and the estimated frequency track is depicted in Figure 1. It can be seen that the frequency track corresponds well to the tonal vocalisation of the bird.



(a) (b)

Fig. 1. An example of a spectrogram (a) of audio field recording and the corresponding estimated frequency track (b).

2.1. Method outline

We consider that the signal may consist of an unknown number of sinusoidal components. The method tackles the detection problem as a pattern recognition problem. Each spectral peak is considered as a potential sinusoidal component. A set of features, extracted from the short-time spectrum, is obtained for each spectral peak. The decision whether the peak is detected as a sinusoid or not is based on calculating the probability of the extracted set of features on a model corresponding to sinusoids and to noise.

2.2. Spectral magnitude and phase features

Let us denote by $S_l(k)$ the short-time spectrum of the l^{th} frame of the signal. Denote by k_p the frequency index of a spectral peak found in the short-time magnitude spectrum. For each peak, a multivariate feature vector $\mathbf{y}$, capturing the spectral magnitude shape and phase continuity information around the peak, is extracted. The magnitude shape features are obtained by using a normalised spectral magnitude values over the range of frequency bins from $k_p - M$ to $k_p + M$, i.e., $\mathbf{y}$=($|\tilde{S}_l(k_p - M|, \ldots, |\tilde{S}_l(k_p - 1)|, |\tilde{S}_l(k_p + 1)|, \ldots |\tilde{S}_l(k_p + M|$), where $|\tilde{S}_l(k)|$ is the magnitude spectrum $|S_l(k)|$ normalised by the magnitude value at the peak $|S_l(k_p)|$ and M denotes the number of bins around the peak to be used. The phase continuity features are obtained by using the spectral phase difference values over the range of frequency bins from $k_p - M$ to $k_p + M$, i.e., $\mathbf{y}$=($\Delta\phi_l(k_p - M), \ldots, \Delta\phi_l(k_p + M)$). The phase difference between the current and previous signal frame is defined as $\Delta\phi_l(k) = \phi_l(k) - \phi_{l-1}(k) - 2\pi k_p L/N$, where $\phi_l(k)$ and $\phi_{l-1}(k)$ denote the phase of the frequency point k at frame-time l and $l-1$, respectively, and L is the frame-shift in samples.

2.3. Probabilistic modelling

Various classification approaches could be employed for making the decision about the peaks. In this paper, we employed Gaussian mixture models (GMMs) which have been extensively and successfully used in speech and audio pattern processing. We are currently also investigating the use of discriminative approaches, such as support vector machines.

The GMM models the distribution of the multivariate feature vector $\mathbf{y}$, representing the spectral magnitude shape and phase continuity. A large collection of features $\mathbf{y}$ corresponding to spectral peaks of noise and of sinusoidal signals at various SNRs are used as the training data to estimate the parameters of the GMM of noise, denoted by λ_n, and of sinusoidal signals, denoted by λ_s.

A given unknown audio signal is processed as described in the previous section to extract the features for each spectral peak. The decision whether a spectral peak at a given signal frame corresponds to a sinusoidal signal or not is based on the maximum likelihood criterion, i.e., the peak is detected as a sinusoid if $p(\mathbf{y}|\lambda_s) > p(\mathbf{y}|\lambda_n)$.

3. UNSUPERVISED SEGMENTATION

The application of the sinusoidal detection method described in Section 2 results in a form of an initial segmentation of the signal; for instance, signal-frames in which no sinusoid is detected indicate no presence of tonal vocalisations. However, these initial segments may contain several repetitions of vocalisation elements and/or other tonal sounds, which could be anywhere within the detected segments. As such, the use of the conventional DTW that searches for similarity of whole sequences is not suitable. Instead, we need to search for partial and multiple matchings within a given pair of segments.

3.1. Dynamic Time Warping

Conventional DTW algorithm can find the optimal global alignment, or warping, path between two whole sequences, while utilising some distance measure and constraints. The

accumulated distance between the two sequences along that path can be used as a basis for comparison of the sequences.

Consider two sequences representing a time series of vectors, $X = (\mathbf{x}_1, \ldots, \mathbf{x}_{N_X})$ and $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_{N_Y})$, where N_X and N_Y is the length of each sequence, respectively. A warping path, $W = (w_1, \ldots, w_K)$, defines a mapping between the sequence X and Y. The k^{th} element of W is defined as $w_k = (i_k, j_k)$, where i_k and j_k are frame-time indices for X and Y sequence respectively. The globally optimal warping path W is such that minimises the cumulative distance

$$D_W(X, Y) = \sum_{k=1}^{K} d(\mathbf{x}_{i_k}, \mathbf{y}_{j_k}) \quad (1)$$

where $d(\mathbf{x}_i, \mathbf{y}_j)$ represents a distance measure between the vectors $\mathbf{x}_i$ and $\mathbf{y}_j$, with the Euclidean distance being often used and also employed in this paper.

A variety of constraints may be imposed on the warping path W in order to avoid a warping of the time-axis which is considered as undesirable in a given specific task. A commonly used relations between two consecutive points on the warping path, which we also employed here, specifies that $w(k-1)$ is one of the following $(i_k, j_k - 1)$, $(i_k - 1, j_k - 1)$ or $(i_k - 1, j_k)$, i.e., only a single frame-time move in horizontal, diagonal or vertical direction is allowed, respectively. We also employed a constraint on the possible relation among several consecutive moves on the warping path, specifically, we do not allow more than three consecutive moves in horizontal or vertical direction. The constraints on w_1 and w_K determine the possible starting and ending point.

Since the distance $D_W(X, Y)$ as expressed in Eq. 1 is accumulating over the warping path, the use of this distance value directly would cause that the length of the sequence affects the value. Thus, we normalise this distance by the length of the warping path.

3.2. Partial and multiple matching using a modified DTW

This section presents modifications to the conventional DTW algorithm that we employed in order to obtain multiple partial alignment paths for two given sequences.

In order to search for partial paths, we consider that the starting and ending points can be anywhere within the $N_X \times N_Y$ matrix. This is implemented by calculating several DTW searches in parallel, each considering a different starting point on one of the sequence, let's say X, and allowing the start anywhere on the other sequence Y. For clarity, let us consider only one such DTW search corresponding to a starting point i_r on the sequence X. As the DTW calculation progresses, the cumulative distance values are obtained for subsequent points in the matrix. The values of the normalised cumulative distance can be examined at the frame-time $i_r + L_{min}$ on the sequence X and any frame-time j on the sequence Y. L_{min} is the minimum length of a sequence we consider for matching and this is employed to avoid short accidental match. If there is no j such that the normalised cumulative distance is below a given threshold D_{thr}, i.e., $D(i_r + L_{min}, j) \geq D_{thr}$ for all $j = 1, \ldots, N_Y$, then, we can consider that no minimum-length partial match is found for the DTW search starting at the i_r on the sequence X and as such we can stop proceeding with this DTW calculation further. On the other side, if the minimum-length match is found, this DTW search will continue until the normalised cumulative distance becomes greater than D_{thr}.

Based on the above procedure, we obtain a set of partial paths matchings within the two sequences X and Y. For each of the found partial warp paths, we have an associated starting and ending points and the normalised cumulative distance. If there is more than one path falling within the rectangular area defined by the starting and ending points of the given partial warp path, we consider only the path with maximum length.

An example of the result obtained by the above procedure on a pair of real-world bird recordings is given in Figure 2. The lines indicate all the partial matchings found. For simplicity, the lines are drawn by connecting the starting and ending points of the found match. It can be seen that the procedure found 13 partial matches between the given two sequences that allign well to each other.

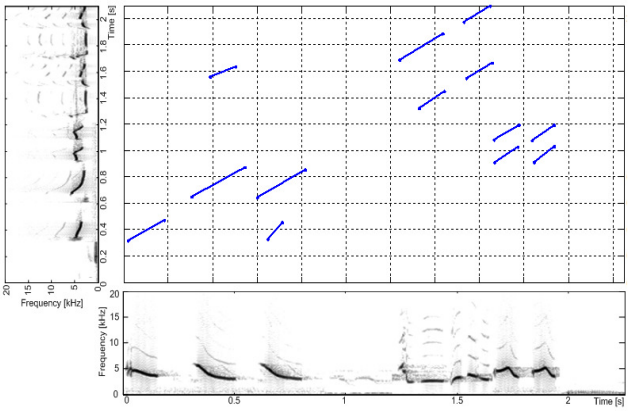


Fig. 2. An example of the output of multiple partial matchings found when comparing two bird recordings.

4. CLUSTERING OF SEGMENTS

The output of the partial DTW is a large collection of segments and their associated distances one to another. Here, we use agglomerative hierarchical clustering method to identify and group together all structurally similar segments that were produced by a particular bird.

Agglomerative method is a bottom-up hierarchical method, where clusters at one level are merged as clusters at the next level. Initially each segment is assumed as a distinct cluster. Based on the distances between the segments (as obtained by DTW), the two closest ones are merged into a larger clus-

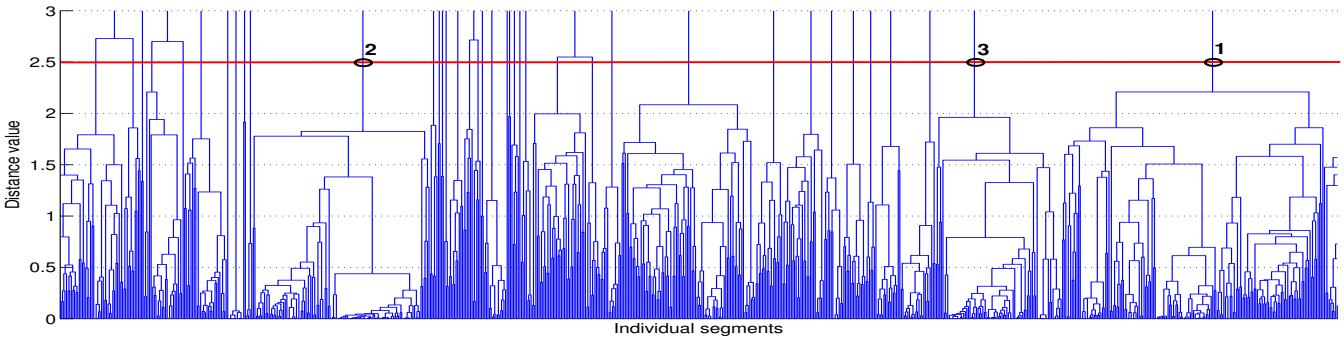


Fig. 3. A part of the obtained hierarchical clustering tree of the segments found by the partial DTW segmentation.

ter. The pairwise similarity between the new cluster and the remaining clusters is then calculated, by taking the average distance of these two clusters to the remaining clusters, and distance matrix is updated. This procedure of merging into larger clusters and correspondingly updating the distance matrix is repeated recursively until the termination criterion is reached. The termination may be based on the distance exceeding a pre-defined value.

5. EXPERIMENTAL EVALUATIONS

This section presents experimental evaluation of the entire proposed system for learning bird vocalisation patterns. We performed experiments using audio recordings from [16]. These recordings were collected over several decades, mostly in the western United States. There are several files for each bird specie, and each file is typically between one to ten minutes long. The recordings are encoded as mono 16-bit wav files, with sampling rate of 48 kHz. They are field recordings in real world natural habitats of birds, and as such, there is also present a various level of background environmental noise, vocalisations of other birds/animals and human speech. There is no labeling information indicating the times of bird singing accompanied with the acoustic data.

We divided the signal into frames of 256 samples with a shift of 64 samples between adjacent frames. The frame length corresponds to approximately 5.3 ms. Similarly short signal frames were found suitable for processing of bird acoustic signals also in our previous research [8]. Hamming analysis window is used and the DFT size is set to 512 points, i.e., the signal is appended by 256 zeros in order to provide a finer DFT sampled spectrum. In partial DTW search, the value of D_{thr} and L_{min} was set to 2 and 15, respectively. The results presented below are for the bird specie 'Carolina Wren'. As no annotation of the data is available, evaluations are performed by visually inspecting if the segments assigned to the same cluster are similar to each other.

The partial DTW search provided 1500 variable-length segments. The clustering of these segments resulted in a hierarchical tree, part of which is depicted in Figure 3. The hierarchical clustering can provide various levels of categorisation of these segments, based on a threshold value used. This threshold is directly related to the distance value from the DTW search, with zero indicating a perfect match. We observed the value around 2.5 to be suitable (depicted as a red line in Figure 3). The use of this threshold resulted in grouping of the 1500 segments into 142 clusters. The occupancy of clusters is depicted in Figure 4. It can be seen that out of these 142 clusters, there is a large number of clusters with a very low occupancy. We have observed that these low-occupancy clusters corresponded to a variety of acoustic events, such as tonal noise, speech or infrequent vocalisations of other birds/animals, which are due to the recordings coming from real world natural environments. Out of these 142 clusters, for instance, there are 99 clusters with an occupancy of 6 acoustic segments or less. These 99 clusters attracted altogether only 241 acoustic segments. The remaining 43 higher-occupancy clusters attracted 1259 segments out of 1500. These clusters are considered to correspond to the bird vocalisations named in the recording.

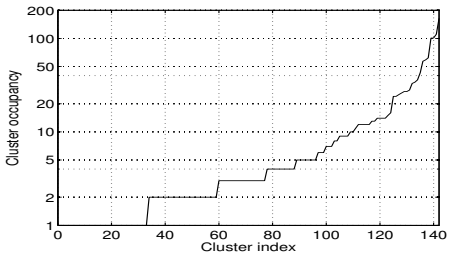


Fig. 4. Occupancy of clusters.

A part of the obtained clustering result is presented in Figure 5. In the figure, each row corresponds to an individual cluster found and each column shows an example of the frequency track of a DTW found partial segment associated with that cluster. The first three rows correspond to clusters with the highest occupancy, each containing over 100 segments. These are marked in the hierarchical tree in Figure 3 by number from 1 to 3. As can be seen from Figure 5, frequency

tracks within each cluster show great similarity to each other, while across clusters show clearly distinctive patterns.

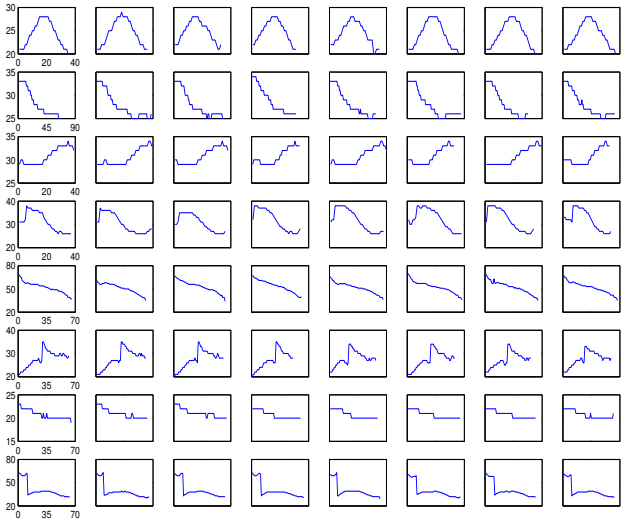


Fig. 5. A part of the outcome of the unsupervised clustering depicting several examples of frequency tracks (where the x- and y-axis corresponds to the frame-time and frequency index, respectively) of partial segments associated with eight different clusters (corresponding to each row).

6. CONCLUSION

In this paper, we presented an approach for unsupervised discovery of bird vocalisation patterns. The proposed approach employed frequency tracks as features to characterise bird tonal vocalisations. These frequency tracks were estimated by employing a method for detection of sinusoidal components, without requiring any information about noise estimate. We developed a modified dynamic time warping algorithm that allowed to search for multiple and partial machings between the given sequences. The obtained distances between the sequences, as outcome of the DTW search, were then used in a hierarchical clustering. It was demonstrated that the obtained clusters showed good coherence and provided a set of structurally distinct bird vocalisation patterns. The presented work can also be applied to birds producing non-tonal vocalisations, or audio signals in general, by using a different set of features, instead of the frequency tracks.

Acknowledgement

Data provided by Borror Laboratory of Bioacoustics, The Ohio State University, Columbus, OH, all rights reserved.

7. REFERENCES

[1] T. Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 1, pp. 164–181, 2011.

[2] S. Deligne and F. Bimbot, "Inference of variable-length acoustic units for continuous speech recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, Apr. 1997, vol. 3, pp. 1731–1734.

[3] Volker B. Deecke and Vincent M. Janik, "Automated categorization of bioacoustic signals: Avoiding perceptual pitfalls," *The Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 645–653, 2006.

[4] A.S. Park and J.R. Glass, "Unsupervised pattern discovery in speech," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 16, no. 1, pp. 186–197, Jan. 2008.

[5] J. Schmalenstroeer, M. Bartek, and R. Haeb-Ubbach, "Unsupervised learning of acoustic events using dynamic time warping and hierarchical k-means++ clustering," in *Interspeech, Florence, Italy*, Aug. 2011, pp. 27–31.

[6] M. Müller and F. Kurth, "Enhancing similarity matrices for music audio analysis," *IEEE Int. Conf. on Acoustics, Speech, and Signal Proc., Toulouse, France*, vol. V, pp. 9–12, May 2006.

[7] P. Somervuo, A. Härmä, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 14, no. 6, pp. 2252–2263, Nov. 2006.

[8] P. Jančovič and M. Köküer, "Automatic detection and recognition of tonal bird sounds in noisy environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, pp. Article ID 982936, 2011.

[9] F. Briggs, B. Lakshminarayanan, L. Neal, X.Z. Fern, R. Raich, S. J.K. Hadley, A.S. Hadley, and M.G. Betts, "Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach," *The Journal of the Acoustical Society of America*, vol. 131, no. 6, pp. 4640–4650, 2012.

[10] S. Fagerlund, "Bird species recognition using support vector machines," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. Article ID 38637, Jan. 2007.

[11] A. Selin, J. Turunen, and J.T. Tanttu, "Wavelets in recognition of bird sounds," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. Article ID 51806, Jan. 2007.

[12] T.S. Brandes, "Feature vector selection and use with hidden Markov Models to identify frequency-modulated bioacoustic signals amidst noise," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 16, no. 6, pp. 1173–1180, Aug. 2008.

[13] P. Jančovič and M. Köküer, "Estimation of voicing-character of speech spectra based on spectral shape," *IEEE Signal Processing Letters*, vol. 14, no. 1, pp. 66–69, Jan. 2007.

[14] P. Jančovič and M. Köküer, "Detection of sinusoidal signals in noise by probabilistic modelling of the spectral magnitude shape and phase continuity," *IEEE Int. Conf. on Acoustics, Speech, and Signal Proc., Prague, Czech Republic*, pp. 517–520, May 2011.

[15] P. Jančovič and M. Köküer, "Incorporating the voicing information into hmm-based automatic speech recognition in noisy environments," *Speech Communication*, vol. 51, no. 14, pp. 438–451, May 2009.

[16] "Borror Laboratory of Bioacoustics," *The Ohio State University, Columbus, OH, all rights reserved.*, www.blb.biosci.ohio-state.edu.