This full text paper was peer-reviewed at the direction of IEEE Instrumentation and Measurement Society prior to the acceptance and publication.

Acoustic Detection of Bees in the Field Using CASA with Focal Templates

David Heise
Dept. of CS, Technology & Mathematics
Lincoln University
Jefferson City, Missouri, USA
heised@lincolnu.edu

Nicole Miller-Struttmann
Biological Sciences Department
Webster University
Webster Groves, Missouri, USA
nmillstrutt@gmail.com

Candace Galen, Johannes Schul
Division of Biological Sciences
University of Missouri
Columbia, Missouri, USA
{galenc, schulj}@missouri.edu

***Abstract*—This paper describes a method of detecting buzzes of bees in field audio. Detecting the buzzing of bees from environmental recordings is an instance of sound scene analysis. In this work, we build upon prior work in computational auditory scene analysis (CASA), employing spectral clustering techniques to mitigate the weakness of the target signal, coupled with a newly-introduced concept of "focal templates". This system yields promising results on a previously acquired collection of environmental recordings, yielding results consistent with human performance, and, in some cases, improving upon human performance. Our success in this task suggests that the combination of focal templates and spectral clustering may prove valuable in other sound scene analysis tasks, especially when the target may be well-defined but may suffer from low signal-to-noise ratio (SNR). Survey recordings with manual (visual and acoustic) annotations were processed, and the algorithm yielded very favorable results. The potential for deploying this approach into a low-cost pollinator monitoring system is discussed.**

Keywords—computational auditory scene analysis; CASA; focal templates; buzz detection

I. Introduction

Pollinator monitoring is a field of growing interest in ecology, agriculture, and conservation. Pollinators, and bees in particular, have an enormous impact on our environment through the pollination services they deliver. Recent reports of honeybee hive collapse have brought the issue into the public eye. The health of bee pollinators (including honeybees, bumble bees, and other native species) is paramount to a secure food supply and stable economy. Given this, a method to monitor bees non-invasively and economically is highly desirable. In this paper, we discuss a system of *acoustic* monitoring for pollinators that meets these parameters.

Detecting the buzzing of bees from environmental recordings is an instance of sound scene analysis. As with many sound scene analysis tasks, one must separate the target acoustic components from everything else (i.e., the noise), and this must often occur in very low signal-to-noise ratio (SNR) circumstances, including instances of occlusions in the time-frequency domain. In this work, we employ spectral clustering techniques to mitigate the weakness of the target signal, coupled with a newly-introduced concept of "focal templates".

II. Background

A. State-of-the-art Monitoring Methods

The state-of-the-art in bee monitoring involves a combination of trapping, netting, and visual observation. This is a labor-intensive process, and it also removes bees from the environment (destructive sampling). While this is currently the only way to confidently identify exactly what species of bee are occupying an area, these methods are not practical for low-cost, widespread deployment.

B. Prior Work in Acoustic Monitoring

Some researchers have previously engaged in various forms of acoustic monitoring. Most, however, either rely on manual segmentation of an audio signal prior to further processing, or rely heavily on laboratory settings to normalize the sound inputs. Burkart et al. conducted a study on the flight and pollination buzzing of neotropical bees, providing some guidance related to the expected frequencies of each [1]. Gradisek et al. have demonstrated the ability to classify bee sounds (with constraints) using a labeled database and machine learning [2]. No work to date, however, has attempted to automate detection of buzzing in the wild.

C. Computational Auditory Scene Analysis

To detect buzzes within field recordings, we have developed a Computational Auditory Scene Analysis (CASA) approach to processing the signals. CASA is a developing field which attempts to implement the principles of auditory scene analysis (ASA) via computer algorithms that can "listen" in a similar way to humans. ASA is based upon the observed principles of how humans differentiate sound events and "streams" within an audio mixture [3]. It is clear that humans are very good at separating and identifying sounds within a complex audio mixture, especially by applying attention to a particular sound source, but the task of automating this is not straightforward. (The classic example demonstrating the human capacity for sound source separation is the "cocktail party problem", where humans are readily able to discern a particular conversation even in the midst of many interfering sounds in a noisy environment.) Many approaches to CASA have been attempted (e.g., [4]), but there does not yet exist a general solution to this challenging problem. Spectral

This work is supported by the National Science Foundation under awards IIA-1355406 and HRD-1410586. Niwot Ridge LTER (NSF DEB-1027341) and Mountain Area Land Trust (Pennsylvania Mountain) provided access to the research sites to collect data.

978-1-5090-3202-0/17/$31.00 ©2017 IEEE

clustering [5] has been applied to audio signals [6], and Martins has developed a framework for applying spectral clustering to musical signals using well-established perceptual cues [7]. In this work, we extend Martins' framework, adding focal templates as a means of applying attention to our sound source of interest (i.e., bee buzzing).

D. Digital Audio Representation and Focal Templates

Digital audio is recorded as a time-series of samples, with each sample representing the amplitude of a sound signal at a particular point in time; in our data, these samples were recorded at 44100 samples/second. This time-domain signal may be transformed into a time-frequency representation using a technique such as the discrete Fourier transform (DFT). Such a representation may be referred to as a spectrogram. The spectrogram represents the energy of the audio signal within time-frequency (T-F) bins, where the magnitude of each bin corresponds to the energy within a particular frequency band occurring during a narrow frame of time. By analyzing the pattern of energy, across frequencies and over time, one can detect complex patterns that correspond to events arising from sound sources in the audio mixture. Observing the patterns that correspond to bees buzzing leads us to our concept of *focal templates*.

III. Algorithm

The approach we take in processing the signal is illustrated in Fig. 1. The algorithm has been implemented in MATLAB and makes use of built-in functions (e.g., spectrogram) whenever possible.

The process currently works off-line; that is, it uses previously recorded audio. There is nothing to prevent the process from being implemented in real-time (save reading a 20-second buffer to use as a texture window). The input signal is resampled down to 4000 samples/second; this sampling frequency is chosen to preserve audio frequencies up to 2000 Hz while minimizing the amount of data (and later, the size of the affinity matrix) to process.

The signal is divided into 20-second *texture windows*, which is the macro temporal unit used to analyze the signal. The remainder of the process applies to each and every texture window.

A spectrogram is produced from the texture window, using 100ms windows with 90ms overlap (giving micro time resolution of 10ms). This results in a frequency resolution of 10 Hz. After the spectrogram is produced, the energy within four bands (0-500 Hz, 500-1000 Hz, 1000-1500 Hz, and 1500-2000 Hz) is assessed. If the energy within the first band (0-500 Hz) is greater than the energy within the other three bands combined, the signal (from the texture window) is passed through a high-pass filter before reproducing another spectrogram. We found that signals with very high low-band energy yielded poor detection results, and we observed that this problem could be mitigated via high-pass filtering at the texture window level. This process will repeat until the energy within the first band falls below the combined energy of the other three bands.

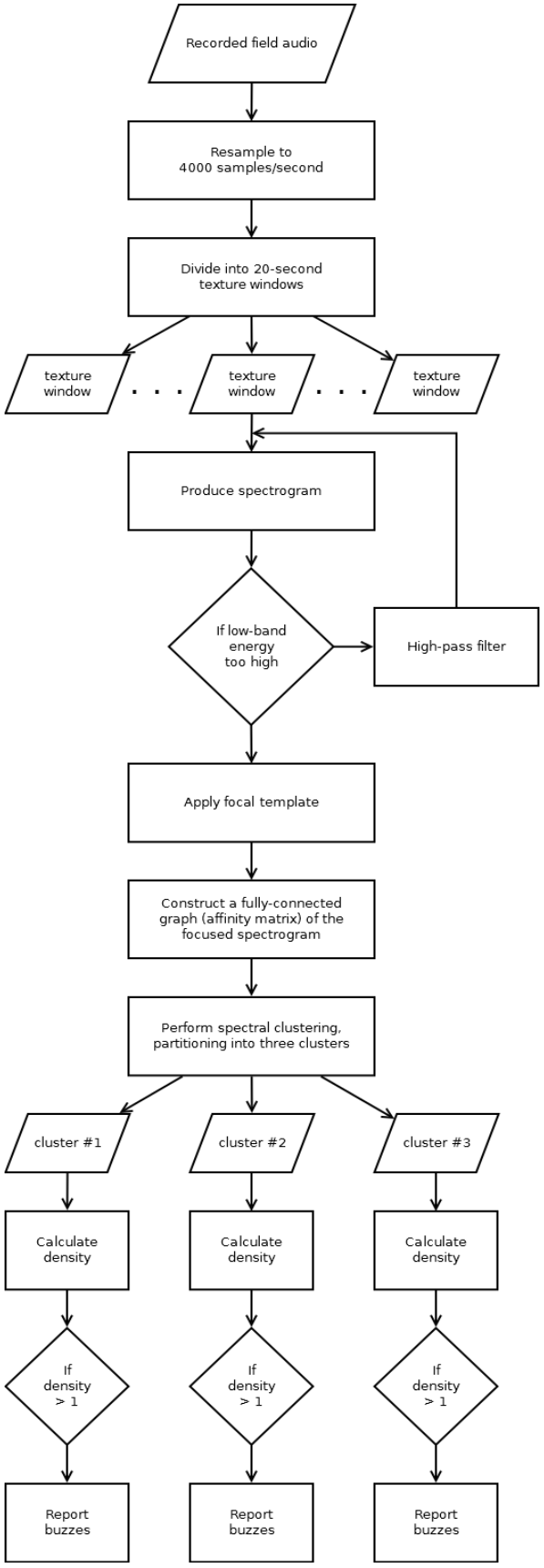

Fig. 1. Flowchart for buzz detection algorithm.

The *focal template* may be thought of as a type of dynamic T-F filter that capitalizes on certain properties of the target sound of interest. It is known that bee buzzing (along with many other natural sounds) is highly harmonic. Further, from observing spectrograms of buzzes, we note that while the (fundamental) frequency of a buzz can change over the course of the buzz, it generally remains relatively constant or changes gradually. Thus, our focal template (to focus attention on bees buzzing within the audio mixture) will look for harmonically related time-frequency elements within 10ms time slices of the spectrogram, will filter-out (eliminate from consideration) non-conforming T-F elements, and will look for periods of time when the fundamental frequency (across multiple 10ms time slices) remains relatively steady or changes gradually. Further, we expect that no buzzes will occur at frequencies below 120 Hz, so T-F elements corresponding to these frequencies are immediately eliminated.

Field audio contains many sources of noise, so the focal template must be robust to low SNR. To combat this, for each 10ms time slice within a texture window, the algorithm determines whether the four T-F bins having the highest energy are found at frequencies having a harmonic relationship with one another (that is, are the frequencies corresponding to the selected T-F bins integer multiples of the same fundamental frequency). Note that the algorithm does not require identification of consecutive harmonics; the algorithm is robust to missing or occluded harmonics. If the four T-F bins having highest energy are *not* all in harmonic relationship with one another, there is not a prominent (or prominent enough) buzz present in this time slice and no T-F bins will be transmitted to the next (spectral clustering) step. If, however, all four T-F bins *are* in harmonic relation, they are retained in the spectrogram for clustering in the next step.

A fully connected graph (affinity matrix; see [8]) is constructed from the preserved T-F elements of the focused spectrogram. This graph (or affinity matrix) is constructed by determining the similarity (or affinity) between every pair of T-F elements remaining after the focal template is applied. Various similarity measures have been proposed (see [7]), but here we calculate and use only the *time similarity* between two T-F bins as

$$time\_sim_{AB} = exp(-(time\_slice_A - time\_slice_B)^2) \quad (1)$$

where $time\_slice_X$ represents the particular 10ms unit of time within the texture window in which the T-F element X was found. Eq. 1 could also be expressed as

$$time\_sim_{AB} = exp(-(time\_dist_{AB})^2) \quad (2)$$

where *time_dist* of 1 corresponds to the distance between consecutive 10ms time slices of the spectrogram.

Spectral clustering is performed by calculating the eigenvectors of the affinity matrix, which corresponds to a *normalized cut* [8]. We partition the focused spectrogram into three clusters by assigning the T-F bins according to a k-means clustering of the smallest three eigenvectors. We have found that clustering into three clusters yields the best results. This is because there may be more than one bee buzzing at the same time, and these buzzes (if at different frequencies) may be allocated to separate clusters. Further, one or more clusters may contain a noise residual after a buzz has been clustered separately.

The *density* of each resulting cluster is then calculated by taking the maximum of

$$cl\_size * mean\_cl\_dist / (median\_cl\_dist^2) \quad (3)$$

or

$$median\_cl\_dist / mean\_cl\_dist \quad (4)$$

where *cl_size* is the cluster size, *mean_cl_dist* is the mean distance between all pairs of T-F elements within the cluster, and *median_cl_dist* is the median distance between all pairs of T-F elements within the cluster. (If the *median_cl_dist* equals 0, *cl_dist* is defined to be 1.) The distance between any two T-F elements within a cluster is defined as the Euclidean distance between the points in the spectrogram, considering the distance between contiguous frequency bins as *freq_dist* of 1 and the distance between contiguous time slices as *time_dist* of 1 (as in Eq. 2), yielding

$$dist_{AB} = sqrt(time\_dist_{AB}^2 + freq\_dist_{AB}^2) \quad (5).$$

If a cluster has density greater than or equal to 1, it is processed to report potential buzzes represented by the cluster. To assess whether a group of T-F elements in a cluster constitutes a buzz, the entire cluster is convolved with a 20-element vector and then summed across frequency into a one-dimensional "smashed cluster" vector (in the time-domain). This has the effect of smoothing over time to mitigate against low SNR (and to guard specifically against buzzes being occluded over one or more time slices). Peaks within the resulting "smashed cluster" are detected; if a peak reaches 1.0, it is considered a buzz. The time boundaries (start and stop) of the buzz are determined by determining the extent of the base of the detected peak (when the base reaches zero). These start/stop times are recorded for writing to an output file.

IV. Experimental Results

The algorithm has undergone extensive trials during development to ensure robustness to a variety of conditions. The algorithm has been used to process approximately 80 hours of data collected in alpine meadows of the Colorado Rocky Mountains during July 2015. These recordings were taken simultaneously with visual observations, and buzzes were manually annotated after-the-fact using the Audacity audio file editor. Thus, we have human-annotated ground truth available to validate the automatic buzz detection method.

At this time, we have tabulated results for 77 separate audio recordings; as may be expected in field audio, there is considerable variation from recording to recording with respect

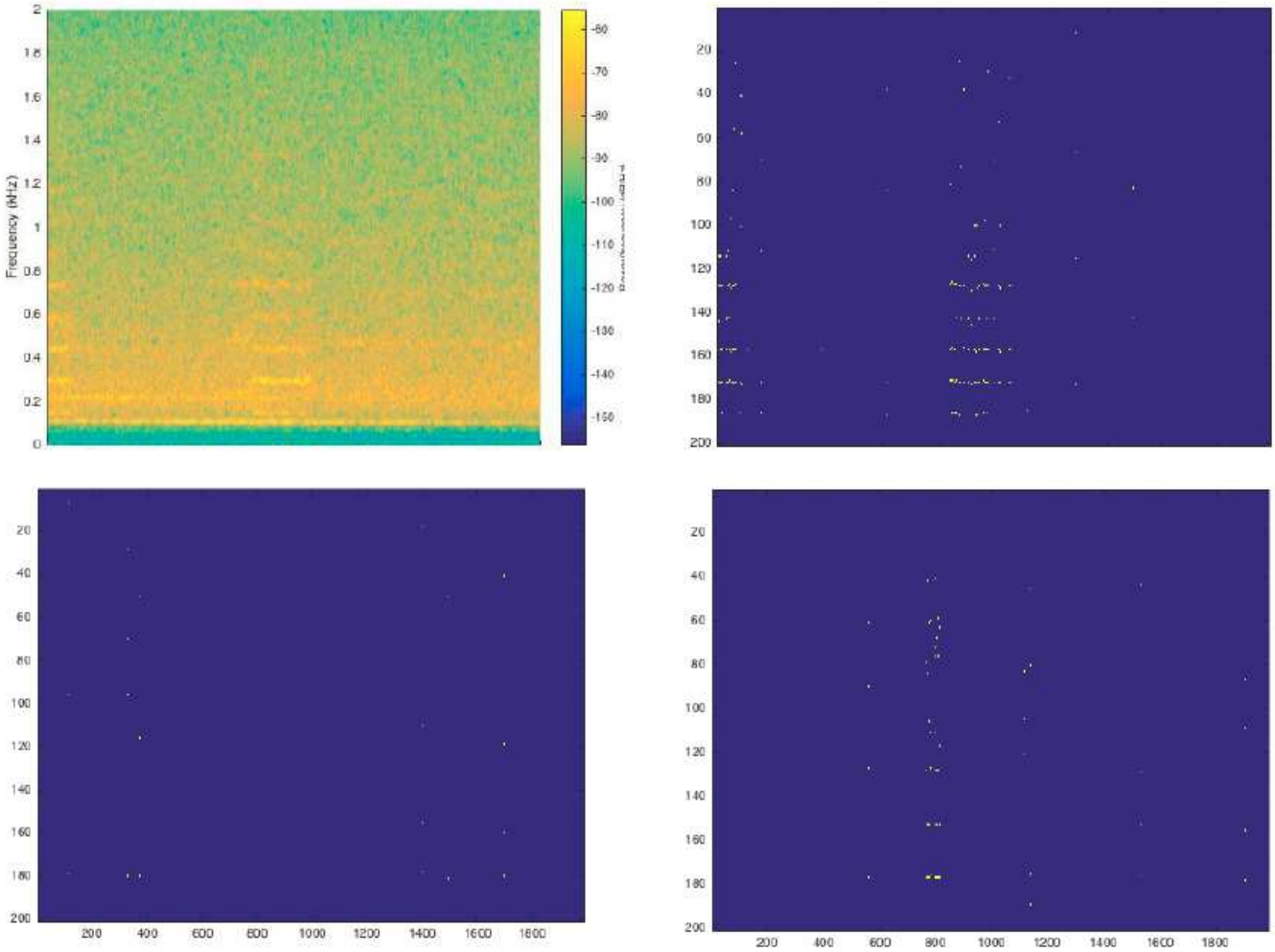

Fig. 2. "Spectrogram" output from the proposed method of spectral clustering with focal templates. The x-axis on each graph represents time (20-second texture windows shown); the y-axis represent frequency (ranging form 0 to 2000 Hz). Clockwise, from upper left: a) a spectrogram of noisy field audio containing buzzing from two different bees; one buzz is moderately weak and the other is very faint, b) an output cluster containing the time-frequency (T-F) elements of the stronger buzz; two distinct buzzes are detected by the algorithm in this cluster (at the beginning of the texture window and near the middle), c) an output cluster containing the T-F elements of the very faint buzz; a buzz is detected by the algorithm around the 8 second mark, and d) a third cluster containing only noise T-F elements.

to number of buzzes, prevalence of interfering noise (primarily airplanes, human voice), and overall fidelity. No attempt was made to "clean" the signals prior to processing. Even so, we find that the method is successful with 68.0% sensitivity and 61.4% selectivity (relative to the human-annotated ground truth). These statistics are encouraging for a new method based upon CASA, but closer examination reveals that even greater success may be claimed. First, across the 77 analyzed recordings, sensitivity ranged from a minimum of 19.5% up to a maximum of 93.8%. Selectivity ranged from a minimum of 5.4% up to a maximum of 95.3%. Recordings yielding low sensitivity tended to also yield low selectivity, suggesting that the overall quality of the recording (and subsequent SNR) was low. Additionally, some recordings contained very few buzzes (in one case, only 4 human-annotated buzzes), which tended to yield poorer results. Individual inspection (listening) of false negatives (missed buzzes compared to ground truth) suggested that many of the missed buzzes occurred when the buzzes were very, very faint. Finally, and perhaps most promising, individual inspection of false positives (detected buzzes that did not correspond to a manually annotated buzz in the ground truth) revealed a surprising number of buzzes that were missed in the manual annotation. That is to say, in some cases, the automated method performed *better* than human listening in an intentional annotation effort. Further work is necessary (and underway) to ascertain the extent to which buzzes may be missing in the "ground truth". Overall, we find these results to be very encouraging, and we are motivated to continue the analysis of this dataset while also looking to further validate the algorithm on other collected recordings.

Conclusions

We have presented a new method of detecting buzzing of bees from field audio. Initial results are encouraging, and detailed analysis suggests the method holds signficant promise, with some recordings yielding sensitivity and selectivity well above 90%. This approach to acoustic detection of bee

buzzing makes it possible to scale bee monitoring to applications heretofore impractical, such as routine monitoring of pollination services available to farms or orchards. This method also has positive implications for conservationists and ecologists who may wish to monitor bee populations non-destructively. We can envision developing a low-cost system that can conveniently capture the environmental audio, process the results, and provide time-stamped information to stakeholders about the bees that are present at any given time. Presently, this algorithm reports number of buzzes detected and total time of buzzing, but further development may lead to more detailed information that could lead to acoustically discerning functional traits of bees, or perhaps even species.

Acknowledgment

Elizabeth Hedrick (Lincoln University) was instrumental in collecting the data in July 2015 and manually annotating the data to enable validation of our method. Darrion Long, Zachary Knuth, and Derrick Parker (also of Lincoln University) contributed to preparation of data for automated processing.

References

[1] A. Gradisek, et al., "Predicting species identity of bumbles bees through analysis of flight buzzing sounds", Bioacoustics, pp. 1-14, 2016.

[2] A. Burkart, K. Lunau, and C. Schlindwein, "Comparative bioacoustical studies on flight and buzzing of neotropical bees", Journal of Pollination Ecology, vol. 6, no. 16, pp. 118-124, 2011.

[3] A. Bregman, Auditory Scene Analysis: the perceptual organization of sound, MIT Press, 1990.

[4] D. Wang and G. Brown, Computational Auditory Scene Analysis: Principles, Algorithms, and Applications, IEEE Press, 2006.

[5] U. von Luxburg, "A Tutorial on Spectral Clustering", Statistics and Computing, vol. 27, no. 4, pp. 395-416, 2007.

[6] M. Lagrange and G. Tzanetakis, "Sound Source Tracking and Formation Using Normalized Cuts", Proceedings of the 2007 International Conference on Acoustics, Speech, and Signal Processing, pp. 61-64, 2007.

[7] L. Martins, "A Computational Framework for Sound Segregation in Music Signals", PhD dissertation, University of Porto, 2008.

[8] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888-905, 2000.