Ecological Informatics 39 (2017) 99–108

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Automated bird acoustic event detection and robust species classification

Zhao Zhao[a,b,*], Sai-hua Zhang[a], Zhi-yong Xu[a], Kristen Bellisario[b], Nian-hua Dai[c], Hichem Omrani[b,d], Bryan C. Pijanowski[b]

[a] *School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China*
[b] *Department of Forestry and Natural Resources, Purdue University, West Lafayette IN47907, USA*
[c] *Institute of Biological Resources, Jiangxi Academy of Science, Nanchang 330096, China*
[d] *Urban Development and Mobility Department, LISER, Luxembourg*

ARTICLE INFO

Keywords:
Bioacoustics monitoring
Automated acoustic event detection
Robust bird species classification
Gaussian mixture model
Autoregressive model

ABSTRACT

Non-invasive bioacoustic monitoring is becoming increasingly popular for biodiversity conservation. Two automated methods for acoustic classification of bird species currently used are frame-based methods, a model that uses Hidden Markov Models (HMMs), and event-based methods, a model consisting of descriptive measurements or restricted to tonal or harmonic vocalizations. In this work, we propose a new method for automated field recording analysis with improved automated segmentation and robust bird species classification. We used a Gaussian Mixture Model (GMM)-based frame selection with an event-energy-based sifting procedure that selected representative acoustic events. We employed a Mel, band-pass filter bank on each event's spectrogram. The output in each subband was parameterized by an autoregressive (AR) model, which resulted in a feature consisting of all model coefficients. Finally, a support vector machine (SVM) algorithm was used for classification. The significance of the proposed method lies in the parameterized features depicting the species-specific spectral pattern. This experiment used a control audio dataset and real-world audio dataset comprised of field recordings of eleven bird species from the Xeno-canto Archive, consisting of 2762 bird acoustic events with 339 detected "unknown" events (corresponding to noise or unknown species vocalizations). Compared with other recent approaches, our proposed method provides comparable identification performance with respect to the eleven species of interest. Meanwhile, superior robustness in real-world scenarios is achieved, which is expressed as the considerable improvement from 0.632 to 0.928 for the F-score metric regarding the "unknown" events. The advantage makes the proposed method more suitable for automated field recording analysis.

1. Introduction

Biodiversity monitoring can provide essential information for conservation action used to mitigate or manage the threats of climate change and high rates of species' loss. Since birds have been widely used as biological indicators for ecological research, the observation and monitoring of birds are increasingly important for biodiversity conservation (Aide et al., 2013; Dawson and Efford, 2010; Potamitis, 2014). Traditional human-observer-based survey methods for collecting data on birds involve a costly effort and have very limited spatial and temporal coverage (Brandes et al., 2006; Swiston and Mennill, 2009). A promising alternative is acoustic monitoring that possesses many advantages including increased temporal and spatial resolution, applicability in remote and difficult-to-access sites, reduced observer bias, and potentially lower cost (Blumstein et al., 2011; Brandes, 2008a; Ganchev et al., 2015; Krause and Farina, 2016; Ventura et al., 2015).

The deployment of acoustic sensor nodes that work continuously as soundscape recording units (Sedláček et al., 2015) is restricted practically only by data storage capacity and/or battery life. Therefore, the volume of collected data is significantly large. Manual analysis of acoustic recordings can produce accurate results, however the time and effort required to process recordings can make manual analysis prohibitive (Swiston and Mennill, 2009; Wimmer et al., 2013). Recently, a number of automated approaches have been proposed to analyze vast amounts of field recordings. According to their objectives, the applications of these approaches roughly fall into two categories: species richness survey (e.g., Eichinski et al., 2015; Pieretti et al., 2015; Sedláček et al., 2015; Wimmer et al., 2013) and species-specific survey (e.g., Aide et al., 2013; Brandes, 2008b; Chen and Maher, 2006; Frommolt and Tauchert, 2014; Kaewtip et al., 2013; Keen et al., 2014; Potamitis et al., 2014; Towsey et al., 2012; Trifa et al., 2008; Wei and Alwan, 2012). The species richness category is also related to a

[*] Corresponding author at: School of Electronic and Optical Engineering, Nanjing University of Science and Technology, 200 Xiaolingwei Road, Xuanwu District, Nanjing 210094, China.
E-mail address: zhaozhao@njust.edu.cn (Z. Zhao).

http://dx.doi.org/10.1016/j.ecoinf.2017.04.003
Received 24 November 2016; Received in revised form 7 April 2017; Accepted 18 April 2017
Available online 19 April 2017

new research area – soundscape ecology (Pijanowski et al., 2011a, 2011b). Both categories require efficient analysis methods including bird vocalization detection and classification to deal with volumes of data. As for bird vocalizations, calls usually refer to isolated, short monosyllabic sounds, while songs are composed of several syllables which consist of elements or notes (Marler, 2004). The classification of birdsongs can be conducted either on an entire song strophe for species with low to medium song complexity, or on smaller entities, i.e. syllables, which can build up different song strophes in species with higher song complexity (Ruse et al., 2016). Here, a strophe usually contains a few syllables and subsequent strophes are separated by pauses of about the same duration (Gill, 2007; Thompson et al., 1994). In this paper, an acoustic event refers to either a call or a syllable.

Intensive studies have been conducted in the field of bioacoustics classification by employing different measurements and methods. To date, based on the ways to classify avian vocalizations, those numerous methods fall into two general categories: template and feature-based. Template-based methods utilize spectrogram-based template matching techniques (e.g., Ehnes and Foote, 2014; Frommolt and Tauchert, 2014; Kaewtip et al., 2013; Meliza et al., 2013; Swiston and Mennill, 2009; Towsey et al., 2012) while feature-based methods calculate a set of spectro-temporal measurements to characterize bird vocalizations. These feature measurements are then fed into a selected automatic classifier with options ranging from simple clustering techniques such as nearest neighbor (e.g., Fagerlund and Harma, 2005) or Euclidian distance between measurements (e.g., Schrama et al., 2008), to more complex algorithms including Gaussian mixture model (GMM) (e.g., Lee et al., 2008), support vector machine (SVM) (e.g., Andreassen et al., 2014; Fagerlund, 2007), decision trees (e.g., Acevedo et al., 2009), Hidden Markov Models (HMMs) (e.g., Aide et al., 2013; Brandes, 2008b; Potamitis et al., 2014; Trifa et al., 2008; Ventura et al., 2015), and random forest (e.g., Neal et al., 2011; Ross and Allen, 2014). Feature-based methods, rather than template-based methods, are more appropriate for dealing with challenging signals such as field recordings containing environmental noise (Keen et al., 2014).

Spectro-temporal measurements employed in feature-based methods can be calculated in each frame or event, which results in frame-level features and event-level features, respectively. Recently, various frame-level features have been employed including peak frequency and short-time frequency bandwidth, as well as their changes between adjacent frames (Brandes, 2008b), Mel-frequency cepstral coefficients (MFCCs) and linear predictive coding coefficients (LPCCs) (Trifa et al., 2008), the combination of LPCCs and a lattice model (Wei and Blumstein, 2011), and a 51-dimensional vector, namely PLP_E_D_A_Z (Potamitis et al., 2014). More recently, a robust frame selection method was proposed which made use of morphological filtering applied to the spectrogram in order to exclude portions of audio with dominant environmental noise (Oliveira et al., 2015; Ventura et al., 2015). Nevertheless, the temporal evolution of frame-level features among consecutive frames is commonly modeled by HMMs. The HMMs implementation in these studies rely on the Hidden Markov Model Toolkit (HTK) (Gales and Young, 2008; Young et al., 2006) which is not a stand-alone recognizer, and its performance depends greatly on the knowledge and experience of the user in pipelining such sophisticated tools (Potamitis et al., 2014).

On the other hand, event-level features have been adopted in many methods, which allow for circumventing the complicated modeling of frame-to-frame variation. Event-level features focus on a whole acoustic event, rather than a single frame within it, and contain a variety of measurements to characterize the time-frequency properties of the event. Some time-frequency features tested include different combinations of descriptive measurements such as central frequency, highest frequency, lowest frequency, initial frequency, loudest frequency, average or maximum bandwidth, duration, type of blur filter used, average frequency slope, maximum power, frequency of maximum power in eight portions of the segment, component shape, and specific narrow-band energy with accumulation in time (e.g., Acevedo et al., 2009; Bardeli et al., 2010; Brandes et al., 2006; Duan et al., 2012; Pedro and Simonetti, 2013; Schrama et al., 2008). Besides these descriptive measurements, many other event-level features have also been studied including amplitude and frequency trajectory (Harma, 2003), harmonic structure (Harma and Somervuo, 2004), spectral peak tracks (e.g., Chen and Maher, 2006; Jančovič and Köküer, 2011, 2015), and the MPEG-7 angular radial transform (ART) descriptor (Lee et al., 2012). However, these methods are restricted to deal with tonal or harmonic vocalizations, or susceptible to environmental noise. Recently, another approach was investigated using regions of interest (ROI) in a spectrogram and the multi-instance multi-label (MIML) framework for machine learning (e.g., Briggs et al., 2012; Potamitis, 2014). The experimental results of classifying 40 bird species field recordings in Mato Grosso, Brazil, proved the performance of ROI-based method unsatisfactory (Ventura et al., 2015).

Many of these experimental methods and evaluations for multiple species classification were usually conducted using datasets that only involved the species of interest—that is, each instance in the dataset belongs to one of the species of interest. However, an important aspect of classifying real-field recordings is that the classifier will encounter some acoustic events, namely “unknown” events, not well suited to any existing classes. In this work, we propose a new automated field recording analysis method robust to the “unknown” events. We designed a reject option scheme in classification motivated by Keen et al., 2014. The major contributions are listed as follows: 1) devised a complete automated analysis procedure, 2) incorporated an event-energy-based sifting procedure after the conventional GMM-based frame selection, and 3) utilized a novel event-level parameterized feature consisting of the coefficients from AR modeling of temporal evolution within each subband to depict the species-specific spectral pattern.

In the rest of this paper, Section 2 describes the field recording database and illustrates the proposed method. Section 3 briefly outlines the reference approach and describes the common experimental protocol and performance metrics. The experimental evaluation results are provided in Section 4, which demonstrate the robust performance of our method for field recordings. Further discussion is presented in Section 5. Finally, Section 6 concludes this work.

2. Materials and methods

2.1. Field recordings database

The field audio recordings used in this work were downloaded from the Xeno-canto Archive (http://www.xeno-canto.org/), a website for sharing recordings of sounds of wild birds from all across the world. A subset of 11 common and widespread North American bird species were selected. It is worth mentioning that these are real-world recordings and each recording potentially contains vocalizations of several animal species and competing noise originating from wind, rain, or anthropogenic interference.

There were five basic sound unit shapes categorized by Brandes (2008a), ranging from tonal or harmonic vocalizations to inharmonic or noise-like bird sounds contained in the recordings. To be more specific, the spectrogram types of acoustic events of the 11 species included constant frequency (CF), frequency modulated whistles (FM), broadband pulses (BP), broadband with varying frequency components (BVF), and strong harmonics (SH). According to the principles of reproducible research, we provided the detailed description of the dataset used in this study in Table 1, which enables other researchers to perform and assess comparative experiments. For simplicity, these species from No. 1 to No. 11 are denoted as B-J, S-S, M-W, C-YT, C-S, A-Y-W, G-B-H, A-C, C-WW, H-F and I-BT in the sequel, respectively.