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
1574-9541/

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.

Table 1
Details of species and corresponding field recordings adopted in this work.

No.	Bird species	Call/ Song	Sound unit shape[a]	Number of events	XC-number of recordings
1	Blue Jay *Cyanocitta cristata* (B-J)	Call	SH	251	155439,109601,296398,282296,165317,164549,165323,165321,116374,116373,122442,149337, 174959,312995,311958,282295
2	Song Sparrow *Melospiza melodia* (S-S)	Call	SH	259	298454,159908,201405,201421,159703, 301851,290096,289225,293818,125719,287645,287881
3	Marsh Wren *Cistothorus palustris* (M-W)	Call	BP	249	141876,159791,210943,21380,169208,173719,199955,157097,159258,191069
4	Common Yellowthroat *Geothlypis trichas* (C-YT)	Call	BP	256	20980,51772,86759,278352,277441, 260184,62257,112583,298965,30788,30787,293121,83414,293315,283273
5	Chipping Sparrow *Spizella passerine* (C-S)	Call	FM	253	229853,69969,132415,148066,288985,12580, 139574, 294349,294362,169162,265803,97367,149900,296964,199341, 131296,131373,285564,289019,294360,31365,294363,269236,254450
6	American Yellow Warbler *Setophaga aestiva* (A-Y-W)	Call	FM	247	206064,313884,304776,31191,13616,59160,190462,229678,302232,229680,233725,253392,274184,278071,287997
7	Great Blue Heron *Ardea Herodias* (G-B-H)	Call	BVF	247	210695,143575,155021,160005,162099,125836,173601,196154,192288,147334,190802,130681,210126,37259
8	American Crow *Corvus brachyrhynchos* (A-C)	Call	BVF	253	29506,90534,69128,71555,181497, 309641,197259,206082,195540,76435,188842,12786,92058,92057
9	Cedar Waxwing *Bombycilla cedrorum* (C-WW)	Call	CF	246	313684,305407,303683,298968,31159, 36559,65794,121797,34859,121800,296871,306661,107623,121805,90143,135078
10	House Finch *Haemorhous mexicanus* (H-F)	Song	SH	249	306935,306936,314417
11	Indigo Bunting *Passerina cyanea* (I-BT)	Song	FM	252	144653,167461,142681,124889,124497

[a] Five sound unit shapes are constant frequency (CF), frequency modulated whistles (FM), broadband pulses (BP), broadband with varying frequency components (BVF), and strong harmonics (SH), respectively.

Fig. 1. Overall block diagram of the proposed method.

2.2. Method

In general, field recording analysis involving automated detection and classification of bird vocalizations follows a typical analysis workflow of four steps: preprocessing, automated segmentation, feature extraction, and classification. In line with this workflow, the individual processing steps of our method are described in the following subsections. The overall block diagram of our method is depicted in Fig. 1.

2.2.1. Preprocessing

Audio recordings selected from the database are resampled to a uniform sampling frequency of 32 kHz, which facilitates the following signal representation and processing. Spectrograms are calculated using the resampled recording files with frame length of 10 ms and a shift of 160 samples between the adjacent frames. Hamming window and 512-point FFT are used to implement the short-time Fourier transform (STFT). Finally, the corresponding spectrogram $S(k, l)$ is fed into the next step with k the index of the Fourier coefficients and l the index of temporal frames.

2.2.2. Automated segmentation

One of the major challenges in the automated bird species recognition is the audio segmentation which aims to select portions of audio that are considered promising and eliminate silent segments or others that contain predominantly background noise. Various energy-based acoustic activity detection methods have been studied for audio segmentation, among which the band-limited-energy detectors are commonly used based on the comparison between the short-term signal energy and a threshold. The threshold is adjusted to select only those frames with high energy for further processing and can usually produce an acceptable true positive rate, but often results in high false positive rates (Oliveira et al., 2015; Ross and Allen, 2014). A more sophisticated approach for acoustic activity detection is based on modeling the distribution of log-energies of frames with a GMM of two mixtures. In this model, one mixture component is fitted to the distribution of the low-energy frames and the other is fitted to the distribution of the high-energy frames. Then, the cross-point of the two components is often selected as the decision threshold, which provides the data-adaptive selection in a recording and makes it convenient for practical use (Alam et al., 2014; Sahidullah and Saha, 2012). Finally, a cluster of consecutive selected frames are grouped into a single event, i.e. a segment, and those segments with length less than 20 ms are discarded in this work.

Note that according to our preliminary study, the initial candidate acoustic events, namely AE_i, $i = 1, 2\ldots K$ with K the number of events, obtained from the conventional GMM-based detection usually own significantly different event-energies. Since only those events with high signal strengths can ensure ornithologists identification with certainty when using traditional audiovisual survey methods (Oliveira et al., 2015), an event-energy-based sifting procedure was conducted to eliminate the events with faintish energies facilitating the interpretation of the observed acoustic activity patterns (Fig. 2). After the event-energy-based sifting, a set of the remaining acoustic events is fed into a parameterized feature extraction step.

After manual inspection of automated segmentation results of the recordings listed in Table 1, an average number of 251 acoustic events, were available for each of the 11 species from a total sample size of 2762. In addition, the number of detected "unknown" events (corresponding to noise or unknown species vocalizations) was 339.

An illustrative example of segmentation using a portion of a recording is given in Fig. 3 where the panel (a) shows the original spectrogram after resampling to 32 kHz and the panels (b)–(c) present the selected frames for the spectrogram using the conventional GMM-based method and the proposed segmentation containing sifting procedure, respectively. The pane (d) zooms in on the box labeled in the panel (c). It can be observed that the vocalizations of dominant species as well as strong sounds of other species are preserved for both methods, however those acoustic events corresponding to environmental noise or soft sounds originating from other species are eliminated in panel (c). Note that this process will benefit the bird species recognition system by reducing the computational demands as well as potentially improving the classification accuracy.

Input: Initial candidate acoustic events set
$D = \{AE_1, AE_2, \cdots, AE_K\}$ with K the number of events

Procedure:

1: Calculate each event-energy with $e(l)$ the l-th frame energy

$$EAE_k = 10\log_{10}\left(\sum_{l \in AE_k} e(l)\right), k = 1, \cdots, K;$$

The maximum is denoted as $ME = \max_k EAE_k$

2: Let $RD = \varnothing$

3: **for** k=1,2,…,K **do**

  if $ME - EAE_k \le 20dB$ **then**

    $RD = RD \bigcup \{AE_k\}$

  end if

end for

Output: Events set RD

Fig. 2. The event-energy-based sifting procedure.

2.2.3. Parameterized feature extraction

Each acoustic event in RD is represented by a feature vector described in this subsection. A Mel-scaled filter bank containing a set of 32 equal-height triangular band-pass filters is first applied to the spectrogram of the event. f_{low} = 1 kHz and f_{high} = 16 kHz are respectively the low and high boundary frequencies for the entire filter bank.

Given an event in RD, the spectrogram is re-calculated with a shift of 64 samples between the adjacent frames to obtain a higher temporal resolution. Meanwhile, the frame length is still 10 ms. Afterwards, the vector corresponding to the short-time power along the frequency in the l-th frame is denoted as

$$\mathbf{sp}(l) = [|S(0,l)|^2, |S(1,l)|^2, \cdots, |S(N/2-1,l)|^2]^T \quad (1)$$

where the superscript T denotes transpose and N = 512 is the number of frequency bins. Then, the power matrix of the event is

$$\mathbf{SP} = [\mathbf{sp}(1), \mathbf{sp}(2), \cdots, \mathbf{sp}(L)] \quad (2)$$

where L is the total number of frames in the event. The frequency response of each band-pass filter is denoted as

$$\mathbf{h}(i) = [h_i(0), h_i(1), \cdots h_i(N/2-1)]^T,\ i = 1, 2, \ldots, 32 \quad (3)$$

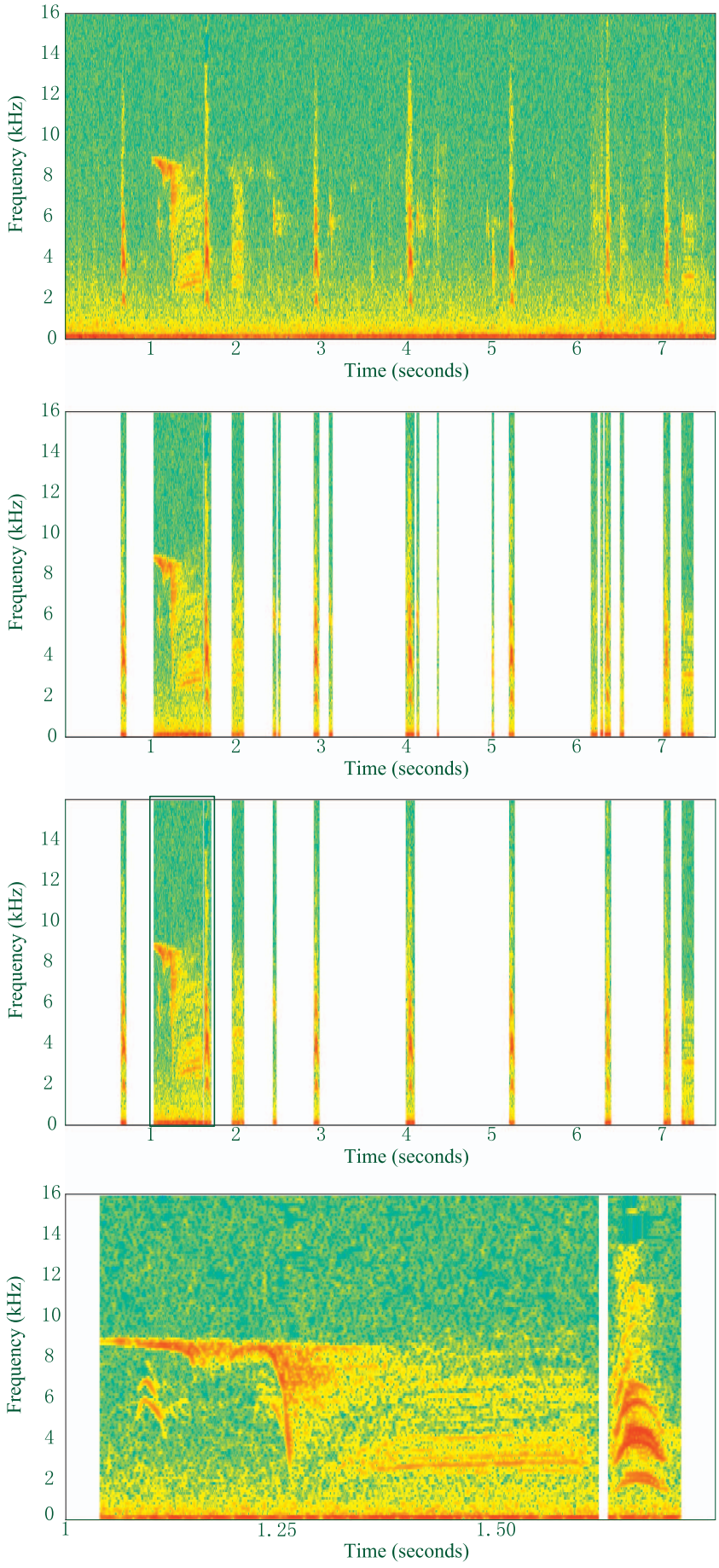
Note that the non-zero coefficients of each $\mathbf{h}(i)$ only locate within the frequency range of (f_{low}, f_{high}). Following Eq. (3), the frequency response matrix of the Mel-scaled filter bank is

$$\mathbf{H} = [\mathbf{h}(1), \mathbf{h}(2), \cdots, \mathbf{h}(32)]^T \quad (4)$$

Then, the filtering output of the event is given by

$$\mathbf{Y} = \mathbf{H} \cdot \mathbf{SP} \quad (5)$$

where $\mathbf{Y} = [\mathbf{y}(1), \mathbf{y}(2), \cdots, \mathbf{y}(\mathrm{L})]$ and $\mathbf{y}(l), l = 1, 2, \ldots, L$ contains the output of each band-pass filter with respect to the l-th frame. For the i-th band-pass filter, the temporal evolution of the output is represented by the time-series



(caption on next page)

Fig. 3. An illustrative example for the segmentation using a portion of a recording: (a) the original spectrogram after resampling to 32 kHz; (b) conventional GMM-based method; (c) the proposed segmentation containing sifting procedure; (d) zoom-in of the box labeled in (c).

$$\mathbf{u}_i^T = \underset{i}{row}(\mathbf{Y}) - mean\left(\underset{i}{row}(\mathbf{Y})\right) = [u_i(1), u_i(2), \cdots, u_i(L)],\ i = 1, 2, \ldots, 32 \tag{6}$$

where $\underset{i}{row}(\cdot)$ denotes the i-th row of the matrix and $mean(\cdot)$ is the mean value.

From a signal processing point of view, the spectrum of $\mathbf{u}_i$ can be considered as the envelope-spectrum-like information within the specific frequency range. In other words, for the spectrogram of the event, the temporal evolution of those powers within different Mel-scaled bandwidth can be characterized by appropriate time-series models. Considering the envelope-spectrum-like property, an AR model is appropriate for each $\mathbf{u}_i$. The model order is determined by the Akaike information criterion (AIC) (Gardner, 1988). According to most recent study in cognitive science (Bregman et al., 2016; Shannon, 2016) spectral pattern accounts for the recognition of the bird sounds. It is worth noting that by AR modeling, the species-specific feature of spectral pattern with respect to certain acoustic event is conveyed by the model coefficients. Furthermore, this parameterization process can deal with a variety of bird acoustic events with either different durations or different sound unit shapes. In brief, a M_i-th order AR model for $\mathbf{u}_i$ is the difference equation

$$u_i(l) + a_1^{(i)} u_i(l-1) + \cdots + a_{M_i}^{(i)} u_i(l - M_i) = z_i(l) \tag{7}$$

where z_i (l) is a zero-mean white noise excitation and the coefficients $\{a_1^{(i)}, \cdots, a_{M_i}^{(i)}\}$ constitute the parameterized feature corresponding to $\mathbf{u}_i$.

As one of the parametric methods for spectral analysis, AR modeling theory has been widely used for decades. Meanwhile, it is computationally inexpensive thanks to the particularly efficient recursive algorithm called the Levinson-Durbin algorithm (Gardner, 1988). Finally, the spectral pattern of the event is characterized by the following feature vector

$$\mathbf{v} = [\mathbf{v}_1^T, \mathbf{v}_2^T, \cdots \mathbf{v}_{32}^T]^T \tag{8}$$

$$\mathbf{v}_i = [a_1^{(i)}, a_2^{(i)}, \cdots, a_M^{(i)}]^T,\ i = 1, 2, \ldots, 32 \tag{9}$$

where $M = 10$ according to our preliminary study where an AR model with maximum order 10 is good enough for fitting the time-series $\mathbf{u}_i$ with respect to the field recordings dataset used in this work. Note that for those models with order M_i less than M, the last $(M - M_i)$ coefficients in Eq. (9) are set to zeros.

2.2.4. SVM-based classification

The classification stage is fed with feature vectors obtained in Section 2.2.3. As we consider a multi-class classification task, species-specific datasets are required and split into training and testing sets. Once all models have been trained, the system is ready for classifying recordings of interest. The features extracted from the instances in the testing dataset are compared to reference species-prototypes in order to find possible matches. The label of the best matching reference becomes the classification result of the corresponding instance. In this work, we employed multi-class SVM based on the "one-versus-one" strategy (Hsu and Lin, 2002; Knerr et al., 1990) for species classification.

As stated previously, a classifier has to deal with some instances not well suited to any existing classes when working with real-world dataset. Conventionally, a threshold $_\theta$ based on the posterior probability is introduced as the reject option. However, since those "unknown" events contain vocalizations of other animal species and competing noise originating from wind, rain, or anthropogenic interference, an empirically determined threshold may influence the performance of classifiers to a great extent. It is also difficult to use the error-reject tradeoff curves due to the lack of a priori knowledge of a corresponding distribution (Hansen et al., 2000). Note that Keen et al. (2014) recommended the exploration of an additional "unknown" class when working with field recordings, although it was not implemented in that study. Motivated by the recommendation, we designed a reject option scheme in which all the detected "unknown" acoustic events we can obtain from the field recordings were attributed to a new class, namely unknown. In both training and classification, all the classes corresponding to those species of interest and the unknown were treated equally. Although the coverage of the unknown class is currently limited by the recordings used in this work, new collected "unknown" acoustic events in future research can further enrich the dataset as well as improve the performance of classifiers.

3. Experimental setup

We used the LIBSVM (Chang and Lin, 2011) implementation of the multi-class SVM, which applies the "one-versus-one" strategy. Specifically, the C-Support Vector Classification (C-SVC) using the Radial Basis Function (RBF) kernel within the LIBSVM package version 3.18 (https://www.csie.ntu.edu.tw/~cjlin/libsvm/) was employed to perform species classification. The RBF layer has two parameters, gamma and cost, which were set to 0.0625 and 8 in this work, respectively. The comparative performance evaluation of the proposed method with the reference method is illustrated in two steps. First, a control experiment was conducted in which those acoustic events in *RD* belonging to the unknown class were *not* included in the dataset. To our best knowledge, this is the common case in literature for multiple species classification—that is, each instance in the dataset belongs to one of the classes of interest. To further investigate the performance of the two methods in a more real-world scenario, a second experiment was conducted in which 339 acoustic events in *RD* belonging to the unknown class were also included. In both experiments, 10 trials were carried out. In each trial the dataset was split randomly into 60% training set and 40% testing set to obtain statistically relevant results. While splitting, the properties of the dataset was kept as same as 60:40—that is, each class percentage was maintained for both training and testing sets which is called stratified sampling (Zhou, 2012).

3.1. Reference method

In many related studies, the MFCCs-plus-HMMs approach is commonly used to classify bird sounds (e.g., Oliveira et al., 2015; Potamitis et al., 2014; Trifa et al., 2008; Ventura et al., 2015). In the feature extraction stage, a set of equal-height triangular band-pass filters was first applied to each frame, followed by a discrete Cosine transformation of log-energies of the filter bank to compute the MFCCs. The full feature vector corresponding to each frame usually also included the first and second time-derivatives (delta and delta-delta coefficients). For the recognition of an acoustic event, the feature vectors of all frames within the event are fed into a species-specific HMM-based classifier. In this work, a full feature vector of a length of 48 parameters, and 12 three-state species-specific HMMs were built in line with the aforementioned studies. Furthermore, the number of mixtures in each state should be small in order not over-optimize on the training set (Potamitis et al., 2014). Thus, we adopted 4 mixtures for our experiment. HTK version 3.4.1 was used for the HMMs implementation.

3.2. Experimental protocol and performance metrics

A common experimental protocol was used to compare the proposed method with the reference approach outlined in Section 3.1. Both methods were evaluated by means of three performance metrics

including precision (P), recall (R) and F-score which are denoted as:

$$P = \frac{TP}{TP + FP} \tag{10}$$

$$R = \frac{TP}{TP + FN} \tag{11}$$

$$F\text{-}score = \frac{(1 + \beta^2)PR}{\beta^2 P + R} \tag{12}$$

where TP is the number of detected true positive events for each class. FP and FN are the numbers of false positive and false negative events for each class, respectively. For the F-score which is the weighted harmonic mean of these two metrics in Eqs. (10) and (11), we used $\beta = 1$ implying that recall and precision play the same important role in this work. After 10 trials, averaged metrics for each method were calculated.

4. Results

In the following subsection, we analyze the experimental evaluation results for our two treatment groups, a control and real-world one. We performed 10 trials on the control group and calculated the precision, recall and F-score metrics for each species (Table 2).

There is a difference between the performance of the proposed method and the MFCCs-plus-HMMs approach for some classes. For some species, such as S-S, C-YT and A-Y-W, using the F-score as a measure, the MFCCs-plus-HMMs approach outperforms the proposed method. However, for the species C-WW and I-BT, the F-score shows that the proposed method is the better one. For the class B-J, the two methods perform almost the same. Note that hypothesis test is usually employed for statistically reliable comparison between algorithms. Specifically, considering the multi-class classification case in this work, we used the Bowker's test (Bowker, 1948; Krampe and Kuhnt, 2007) which is a generalization of the McNemar's test.

After each trial, the proposed method and the MFCCs-plus-HMMs approach judged the instances of the testing set into 11 different species and the resulting data were given in a two-dimensional 11×11-contingency table. Based on a χ^2-approximation of the distribution of the test statistic, the Bowker's test decision is to answer the question whether or not there is significant difference between the two methods with a significance level α which was set to 0.05 in this work. A value of $p < \alpha$ is considered statistically significant. The computations for the Bowker's test were conducted using the commercial SPSS statistical software version 19.0 (IBM Corp., Armonk, NY, USA). Within each test, we always found the null hypothesis was accepted, suggesting that our method is statistically comparable with the reference approach. To illustrate the procedure in detail, a contingency table after one of the trials is given in Table 3 and the corresponding estimate value $p = 0.121$ which is larger than 0.05.

In the second experiment, 10 trials with the same splitting percentage as that in the control experiment were also carried out and the averaged metrics are shown in Table 4. Consider the comparison between Tables 4 and 2, as for those species of interest, there is almost no performance change for the proposed method except a decrease for S-S, however distinct degradation of the MFCCs-plus-HMMs approach can be observed, especially for C-S, C-WW and I-BT. More importantly, one should note that there is a significant difference between the two methods with respect to the unknown class. To be more specific, the proposed method achieved 0.928 for the F-score metric which is considerably larger than 0.632 obtained by the MFCCs-plus-HMMs approach. Meanwhile, as for the precision and recall scores, our method also significantly outperformed the reference approach with 93.9% versus 73.6% and 91.7% versus 55.5%, respectively.

From a signal processing perspective, those instances belonging to the unknown class can be considered as outliers for algorithms. The performance of algorithms with regard to outliers represents the level of "robustness" (Zoubir et al., 2012). As a result, the proposed method is more robust in real-world scenarios. A possible reason for this result is that the parameterized feature depicting the species-specific spectral pattern, rather than the MFCCs, is capable of decreasing the similarity between unknown class and those species of interest. Note that this advantage makes the proposed method more suitable for automated analysis on field recordings.

Table 2
Averaged performance metrics for each species in the control experiment. They show that there is a difference between the performance of the proposed method and the reference approach for each species. Each instance of the dataset in this experiment belonged to one of the 11 species of interest.

Classes	Precision (%)		Recall (%)		F-score	
	This work	MFCCs-plus-HMMs	This work	MFCCs-plus-HMMs	This work	MFCCs-plus-HMMs
B-J	97.5	97.7	98.7	98.7	0.981	0.982
S-S	93.5	99.8	94.6	98.9	0.940	0.994
M-W	93.8	95.9	92.7	95.3	0.932	0.955
C-YT	89.0	96.9	90.7	95.5	0.898	0.962
C-S	94.7	95.8	94.4	96.4	0.945	0.961
A-Y-W	91.6	94.6	92.7	98.5	0.921	0.965
G-B-H	95.8	99.0	93.9	97.7	0.948	0.983
A-C	98.7	99.6	98.0	99.6	0.983	0.996
C-WW	99.8	97.4	97.2	97.1	0.985	0.972
H-F	94.0	95.7	95.1	95.9	0.945	0.958
I-BT	94.6	93.0	94.0	91.7	0.943	0.923

5. Discussion

Our work differs from previous related work where AR modeling was also used (Potamitis et al., 2014). In that study, the cube root was first applied to the outputs of each Mel-scaled band-pass filter with respect to the l-th frame, i.e. $y(l), l = 1, 2, \ldots, L$ in Eq. (5). Then, an AR model was employed to approximate the power spectrum represented by the cube root sequence in each frame. Finally, in a frame-by-frame manner, the coefficients of the resulting model were transformed to cepstral coefficients. By contrast, in this work, AR models are used to characterize the temporal evolution of those powers within different Mel-scaled bandwidth as illustrated in Section 2.2.3. With the AR modeling of the time-series representing the temporal evolution information, the species-specific feature of spectral pattern with respect to certain acoustic event is conveyed by the model coefficients.

It is worth remarking that the number of triangular band-pass filters in the Mel-scaled filter bank may bring an impact on the performance of our method due to variations in characterization of the spectral pattern. Two other options, i.e. 24 and 40, for the number of band-pass filters were employed to investigate the influence and the corresponding experimental results of the F-score metric are depicted in Fig. 4. Note that 10 trials were carried out as well. It is clear that for the options of 24 and 32, corresponding metrics are all not less than 0.9 while there are six items inferior to 0.9 for the option of 40 filters. This implies excessive number of filters is not a better choice for appropriately characterizing the species-specific spectral pattern.

Although the experimental results show that our new method is promising, there are still some limitations. For multiple concurrent bird acoustic events overlapping in time or even both time and frequency, it is not recommended to employ the proposed method directly. Considering that microphone arrays have been increasingly used in acoustic monitoring in terrestrial environments (Blumstein et al., 2011), multiple acoustic sources separation can be conducted as the first step which utilizes the spatial cues. A case in point for bird sounds separation is the Voxnet using the approximate maximum likelihood (AML) algorithm (Cai et al., 2013). However, considering the fact that the vocalizations of a large number of birds occupy broad bands (e.g., 1 kHz–16 kHz), the spatial aliasing effect has to be taken into account

Table 3
A contingency table after one of the trials. In this trial, the number of the instances in the testing set was 1112 which approximates 40% of the total number of acoustic events, i.e. 2762.

		MFCCs-plus-HMMs approach										
		B-J	S-S	M-W	C-YT	C-S	A-Y-W	G-B-H	A-C	C-WW	H-F	I-BT
This work	B-J	97	0	0	0	0	0	1	0	0	0	0
	S-S	3	105	0	1	0	0	4	0	0	0	2
	M-W	1	0	84	6	0	1	1	1	0	0	2
	C-YT	1	0	3	93	4	9	1	0	0	0	0
	C-S	0	0	4	1	92	4	1	0	3	0	0
	A-Y-W	0	0	2	7	2	84	1	0	0	1	0
	G-B-H	0	1	0	0	0	0	89	0	1	0	0
	A-C	0	0	0	0	0	0	0	102	0	0	0
	C-WW	0	0	0	0	2	0	0	0	91	0	2
	H-F	0	1	0	0	0	0	1	0	1	88	8
	I-BT	0	0	0	0	2	3	0	0	0	8	90

Table 4
Averaged performance metrics for each species in a more real-world scenario. Through the comparative evaluation of *F*-score metric with Table 2, there is almost no performance change for the proposed method except a decrease for S-S, however distinct degradation of the MFCCs-plus-HMMs approach can be observed. Advantageous performance for unknown class is particularly evident.

Classes	Precision (%)		Recall (%)		*F*-score	
	This work	MFCCs-plus-HMMs	This work	MFCCs-plus-HMMs	This work	MFCCs-plus-HMMs
B-J	97.8	92.8	97.9	96.9	0.979	0.948
S-S	88.5	94.7	92.1	97.7	0.903	0.962
M-W	94.5	91.9	90.2	93.5	0.922	0.926
C-YT	89.7	92.0	90.8	94.2	0.902	0.930
C-S	93.7	86.1	93.8	91.7	0.937	0.888
A-Y-W	90.7	92.2	93.4	96.8	0.919	0.944
G-B-H	94.7	93.1	91.1	95.9	0.928	0.944
A-C	96.6	97.5	97.7	98.0	0.971	0.977
C-WW	98.9	90.2	95.4	93.4	0.971	0.917
H-F	92.4	94.7	94.7	94.3	0.935	0.945
I-BT	93.1	87.3	94.1	89.7	0.936	0.884
Unknown	93.9	73.6	91.7	55.5	0.928	0.632

and some spatial-aliasing-suppression approaches have been proposed (e.g., Xu and Zhao, 2016). After acoustic sources separation, the proposed method can be employed for the recordings of each source.

Furthermore, although the SVM algorithm is applied in this work, we make no claim of optimality for the SVM as a classifier. The motivating factors including the high computational efficiency, the simplicity of the algorithm, the availability of the LIBSVM package, and its applications in previous work (e.g., Andreassen et al., 2014; Fagerlund, 2007), are main reasons for us to choose it. The classification performance of SVM was convincing, however other machine learning methods might have more predictive power. Additionally, one may combine several classifiers to improve the prediction accuracy. This combination of classifier has proven its value in many disciplines but not yet in sound pattern recognition.

6. Conclusions

Since it is usually not feasible for human experts to hear and/or visually inspect huge amounts of audio data, automated processing of the sound files is a prerequisite for analyzing information in a timely manner. In this work, a complete analysis workflow was devised and we proposed a new method for automated field recordings analysis with improved automated segmentation and robust bird species classification in real-world scenarios. After the conventional GMM-based frame selection, an event-energy-based sifting procedure was employed to select more representative acoustic events which can benefit the bird species recognition system by reducing the computational demands as well as potentially improving the classification accuracy. A novel event-level feature, capable of dealing with a variety of bird acoustic events with either different durations or different sound unit shapes, was extracted to characterize the species-specific spectral pattern.

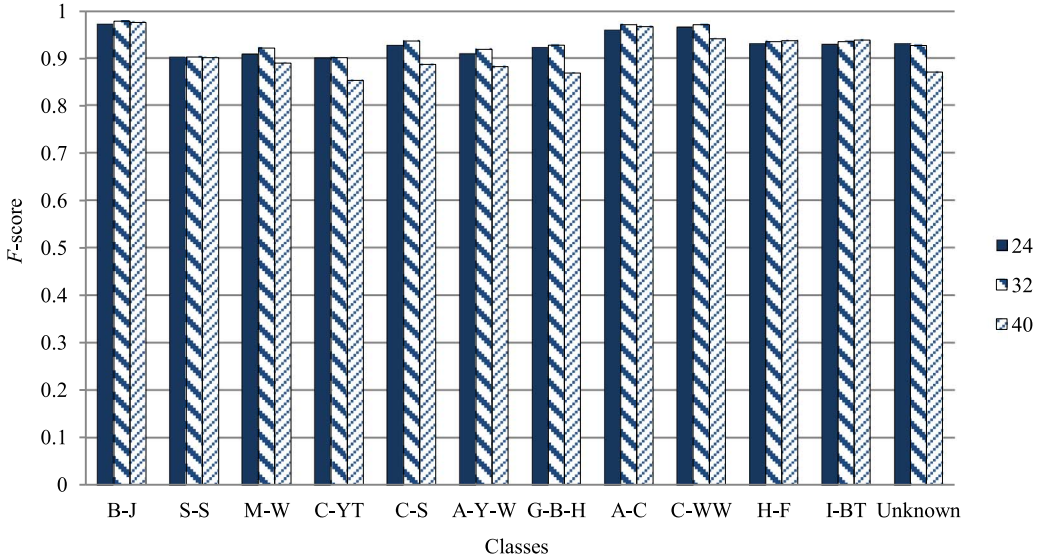


Fig. 4. *F*-score metrics for different numbers of band-pass filters in the same real-world scenario as Table 4. 10 trials were carried out as well.

Experimental results demonstrated that the proposed method provided comparable identification performance with the widely used MFCCs-plus-HMMs approach and superior robustness for those acoustic events not well suited to any existing species of interest. This advantage makes the proposed method more suitable for the applications of acoustic monitoring in terrestrial environments to promote the widespread use of automated field recording analysis technologies.

Acknowledgements

This work was supported by the National Natural Science Foundations of China [grant numbers 61401203, 61171167]; the Natural Science Foundation of Jiangsu Province [grant number BK20130776]; and the State Scholarship Fund of China [grant number 201606840023]. The field audio data used here were downloaded from Xeno-canto (www.xeno-canto.org).

References

Acevedo, M.A., Corrada-Bravo, C.J., Corrada-Bravo, H., Villanueva-Rivera, L.J., Aide, T.M., 2009. Automated classification of bird and amphibian calls using machine learning: a comparison of methods. Ecol. Inform. 4, 206–214.
Aide, T.M., Corrada-Bravo, C., Cerqueira, M.C., Milan, C., Vega, G., Alvarez, R., 2013. Real-time bioacoustics monitoring and automated species identification. PeerJ 1, e103.
Alam, M.J., Kenny, P., Ouellet, P., Stafylakis, T., Dumouchel, P., 2014. Supervised/Unsupervised voice activity detectors for text-dependent speaker recognition on the RSR2015 Corpus. The Speaker Lang. Recogn. Workshop 123–130.
Andreassen, T., Surlykke, A., Hallam, J., 2014. Semi-automatic long-term acoustic surveying: a case study with bats. Ecol. Inform. 21, 13–24.
Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K.H., Frommolt, K.H., 2010. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. Pattern Recogn. Lett. 31, 1524–1534.
Blumstein, D.T., Mennill, D.J., Clemins, P., Girod, L., Yao, K., Patricelli, G., Deppe, J.L., Krakauer, A.H., Clark, C., Cortopassi, K.A., Hanser, S.F., McCowan, B., Ali, A.M., Kirschel, A.N.G., 2011. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. J. Appl. Ecol. 48, 758–767.
Bowker, A.H., 1948. A test for symmetry in contingency tables. J. Am. Stat. Assoc. 43, 572–574.
Brandes, T.S., 2008a. Automated sound recording and analysis techniques for bird surveys and conservation. Bird Conserv. Int. 18, S163–S173.
Brandes, T.S., 2008b. Feature vector selection and use with hidden markov models to identify frequency-modulated bioacoustic signals amidst noise. IEEE Trans. Audio Speech Lang. Process. 16, 1173–1180.
Brandes, T.S., Naskrecki, P., Figueroa, H.K., 2006. Using image processing to detect and classify narrow-band cricket and frog calls. J. Acoust. Soc. Am. 120, 2950–2957.
Bregman, M.R., Patel, A.D., Gentner, T.Q., 2016. Songbirds use spectral shape, not pitch, for sound pattern recognition. PNAS Early Edit. 1–6.
Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X.Z., Raich, R., Hadley, S.J., Hadley, A.S., Betts, M.G., 2012. Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. J. Acoust. Soc. Am. 131, 4640–4650.
Cai, S., Collier, T., Girod, L., Lee, J.Y., Hudson, R.E., Yao, K., Taylor, C.E., Bao, M., Wang, Z., 2013. New wireless acoustic array node for localization, beamforming and source separation for bio-complexity bird data collection and study. IEEE China Summit Int. Conf. Sign. Inform. Process 210–214.
Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 27:1–27:27.
Chen, Z., Maher, R.C., 2006. Semi-automatic classification of bird vocalizations using spectral peak tracks. J. Acoust. Soc. Am. 120, 2974–2984.
Dawson, D.K., Efford, M.G., 2010. Bird population density estimated from acoustic signals. J. Appl. Ecol. 46, 1201–1209.
Duan, S., Zhang, J., Paul, R., Michael, T., 2012. Timed and probabilistic automata for automatic animal call recognition. IEEE Int. Conf. Patt. Recogn. 2910–2913.
Ehnes, M., Foote, J.R., 2014. Comparison of autonomous and manual recording methods for discrimination of individually distinctive ovenbird songs. Bioacoustics 24, 111–121.
Eichinski, P., Sitbon, L., Roe, P., 2015. Clustering acoustic events in environmental recordings for species richness surveys. Procedia Comp. Sci. 51, 640–649.
Fagerlund, S., 2007. Bird species recognition using support vector machines. EURASIP J. Adv. Sign. Process. 7, 1–8.
Fagerlund, S., Harma, A., 2005. Parametrization of inharmonic bird sounds for automatic recognition. In: 13th EUSIPCO, pp. 1–4.
Frommolt, K.H., Tauchert, K.H., 2014. Applying bioacoustic methods for long-term monitoring of a nocturnal wetland bird. Ecol. Inform. 21, 4–12.
Gales, M., Young, S., 2008. The application of hidden Markov models in speech recognition. Found. Trends Sign. Process 1, 195–304.
Ganchev, T.D., Jahn, O., Marques, M.I., Figueiredo, J.M., Schuchmann, K.-L., 2015. Automated acoustic detection of *Vanellus chilensis lampronotus*. Expert Syst. Appl. 42, 6098–6111.
Gardner, W.A., 1988. Statistical Spectral Analysis: A non-probabilistic Theory, Chapter 9. Prentice-Hall, Englewood Cliffs, New Jersey.
Gill, F.B., 2007. Ornithology, third ed. W.H. Freeman, New York Chapter 8.
Hansen, L.K., Liisberg, C., Salamon, P., 2000. The error-reject tradeoff. Open. Syst. Inf. Dyn. 4, 159–184.
Harma, A., 2003. Automatic identification of bird species based on sinusoidal modeling of syllables. IEEE Int. Conf. Acoust. Speech Sign. Process. 545–548.
Harma, A., Somervuo, P., 2004. Classification of the harmonic structure in bird vocalization. IEEE Int. Conf. Acoust. Speech Sign. Process. 701–704.
Hsu, C.-W., Lin, C.-J., 2002. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. 13, 415–425.
Jančovič, P., Köküer, M., 2011. Automatic detection and recognition of tonal bird sounds in noisy environments. EURASIP J. Adv. Sign. Process. 1–10.
Jančovič, P., Köküer, M., 2015. Acoustic recognition of multiple bird species based on penalized maximum likelihood. IEEE Signal Process Lett. 22, 1585–1589.
Kaewtip, K., Tan, L.N., Alwan, A., Taylor, C.E., 2013. A robust automatic bird phrase classifier using dynamic time-warping with prominent region identification. IEEE Int. Conf. Acoust. Speech Sign. Process. 768–772.
Keen, S., Ross, J.C., Griffiths, E.T., Lanzone, M., Farnsworth, A., 2014. A comparison of similarity-based approaches in the classification of flight calls of four species of North American wood-warblers (*Parulidae*). Eco. Inform. 21, 25–33.
Knerr, S., Personnaz, L., Dreyfus, G., 1990. Single-layer learning revisited: a stepwise procedure for building and training a neural network. In: Neurocomputing: Algorithms, Architectures and Applications. Springer-Verlag, Berlin Heidelberg, pp. 41–50.
Krampe, A., Kuhnt, S., 2007. Bowker's test for symmetry and modifications within the algebraic framework. Comput. Statist. Data Anal. 51, 4124–4142.
Krause, B., Farina, A., 2016. Using ecoacoustic methods to survey the impacts of climate change on biodiversity. Biol. Conserv. 195, 245–254.
Lee, C.-H., Han, C.-C., Chuang, C.-C., 2008. Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients. IEEE Trans. Audio Speech Lang. Process. 16, 1541–1550.
Lee, C.-H., Hsu, S.-B., Shi, J.-L., Chou, C.-H., 2012. Continuous birdsong recognition using Gaussian mixture modeling of image shape features. IEEE Trans. Multimedia 15, 454–464.
Marler, P., 2004. Chapter 5–bird calls: a cornucopia for communication. In: Natures Music: Science of Birdsong. Elsevier, New York, pp. 132–177.
Meliza, C.D., Keen, S.C., Rubenstein, D.R., 2013. Pitch- and spectral-based dynamic time warping methods for comparing field recordings of harmonic avian vocalizations. J. Acoust. Soc. Am. 134, 1407–1415.
Neal, L., Briggs, F., Raich, R., Fern, X.Z., 2011. Time-frequency segmentation of bird song in noisy acoustic environments. IEEE Int. Conf. Acoust. Speech Sign. Process. 2012–2015.
Oliveira, A.G.D., Ventura, T.M., Ganchev, T.D., Figueiredo, J.M.D., Jahn, O., Marques, M.I., Schuchmann, K.L., 2015. Bird acoustic activity detection based on morphological filtering of the spectrogram. Appl. Acoust. 98, 34–42.
Pedro, A.R., Simonetti, J.A., 2013. Acoustic identification of four species of bats (Order Chiroptera) in central Chile. Bioacoustics 22, 165–172.
Pieretti, N., Duarte, M.H.L., Sousa-Lima, R.S., Rodrigues, M., Young, R.J., Farina, A., 2015. Determining temporal sampling schemes for passive acoustic studies in different tropical ecosystems. Trop. Conserv. Sci. 88, 215–234.
Pijanowski, B.C., Farina, A., Gage, S.H., Dumyahn, S.L., Krause, B.L., 2011a. What is soundscape ecology? An introduction and overview of an emerging new science. Landsc. Ecol. 26, 1213–1232.
Pijanowski, B.C., Villanuevarivera, L.J., Dumyahn, S.L., Farina, A., Krause, B.L., Napoletano, B.M., Gage, S.H., Pieretti, N., 2011b. Soundscape ecology: the science of sound in the landscape. Bioscience 61, 203–216.
Potamitis, I., 2014. Automatic classification of a Taxon-Rich community recorded in the wild. PLoS One 9, e96936.
Potamitis, I., Ntalampiras, S., Jahn, O., Riede, K., 2014. Automatic bird sound detection in long real-field recordings: applications and tools. Appl. Acoust. 80, 1–9.
Ross, J.C., Allen, P.E., 2014. Random forest for improved analysis efficiency in passive acoustic monitoring. Ecol. Inform. 21, 34–39.
Ruse, M.G., Hasselquist, D., Hansson, B., Tarka, M., Sandsten, M., 2016. Automated analysis of song structure in complex birdsongs. Anim. Behav. 112, 39–51.
Sahidullah, M., Saha, G., 2012. Comparison of speech activity detection techniques for speaker recognition. J. Immunother. 33, 609–617.
Schrama, T., Poot, M., Robb, M., Slabbekoorn, H., 2008. Automated monitoring of avian flight calls during nocturnal migration. Int. Exp. Meeting IT-based detect. Bioacoustic Patt. 131–134.
Sedláček, O., Vokurková, J., Ferenc, M., Djomo, E.N., Albrecht, T., Hořák, D., 2015. A comparison of point counts with a new acoustic sampling method: a case study of a bird community from the montane forests of mount cameroon. Ostrich 86, 213–220.
Shannon, R.V., 2016. Is birdsong more like speech or music? Trends Cogn. Sci. 20, 245–247.
Swiston, K.A., Mennill, D.J., 2009. Comparison of manual and automated methods for identifying target sounds in audio recordings of pileated, pale-billed, and putative ivory-billed woodpeckers. J. Field Ornithol. 80, 42–50.
Thompson, N.S., LeDoux, K., Moody, K., 1994. A system for describing bird song units. Bioacoustics 5, 267–279.
Towsey, M., Planitz, B., Nantes, A., Wimmer, J., Roe, P., 2012. A toolbox for animal call recognition. Bioacoustics 21, 1–19.
Trifa, V.M., Kirschel, A.N., Taylor, C.E., Vallejo, E.E., 2008. Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. J. Acoust. Soc. Am. 123, 2424–2431.
Ventura, T.M., Oliveira, A.G.D., Ganchev, T.D., Figueiredo, J.M., Jahn, O., Marques, M.I.,

Schuchmann, K.-L., 2015. Audio parameterization with robust frame selection for improved bird identification. Expert Syst. Appl. 42, 8463–8471.
Wei, C., Alwan, A., 2012. FBEM: a filter bank EM algorithm for the joint optimization of features and acoustic model parameters in bird call classification. IEEE Int. Conf. Acoust. Speech Sign. Process 1993–1996.
Wei, C., Blumstein, D.T., 2011. Noise robust bird song detection using syllable pattern-based hidden Markov models. IEEE Int. Conf. Acoust. Speech Sign. Proc. 345–348.
Wimmer, J., Towsey, M., Roe, P., Williamson, I., 2013. Sampling environmental acoustic recordings to determine bird species richness. Ecol. Appl. 23, 1419–1428.
Xu, Z.-Y., Zhao, Z., 2016. Spatial aliasing suppression for pooled angular spectrum using a widely spaced microphone array. IEEE Int. Conf. Sign. Process 443–447.
Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X.-Y., Moore, G., Odell, J.L., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2006. The HTK Book (for HTK Version 3.4). Cambridge University Press, Cambridge, UK.
Zhou, Z.-H., 2012. Ensemble Methods, Foundations and Algorithms. Taylor & Francis, Boca Raton, FL.
Zoubir, A.M., Koivunen, V., Chakhchoukh, Y., Muma, M., 2012. Robust estimation in signal processing: a tutorial-style treatment of fundamental concepts. IEEE Signal Process. Mag. 29, 61–80.