Applied Acoustics 80 (2014) 1–9

Contents lists available at ScienceDirect

Applied Acoustics

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

Automatic bird sound detection in long real-field recordings: Applications and tools

Ilyas Potamitis [a,*], Stavros Ntalampiras [b], Olaf Jahn [c], Klaus Riede [c]

[a] Technological Educational Institute of Crete, Department of Music Technology and Acoustics, Crete, Greece[1]
[b] Department of Electronics & Information, Polytechnic of Milano, Italy
[c] Zoological Research Museum Alexander Koenig, 53113 Bonn, Germany

ARTICLE INFO

Article history:
Received 8 January 2013
Received in revised form 13 August 2013
Accepted 8 January 2014
Available online 2 February 2014

Keywords:
Birdsong detection
Bird recognition
Computational ecology

ABSTRACT

The primary purpose for pursuing this research is to present a modular approach that enables reliable automatic bird species identification on the basis of their sound emissions in the field. A practical and complete computer-based framework is proposed to detect and time-stamp particular bird species in continuous real field recordings. Acoustic detection of avian sounds can be used for the automatized monitoring of multiple bird taxa and querying in long-term recordings for species of interest for researchers, conservation practitioners, and decision makers, such as environmental indicator taxa and threatened species. This work describes two novel procedures and offers an open modular framework that detects and time-stamps online calls and songs of target bird species and is fast enough to report results in reasonable time for non-processed field recordings of many thousands files and is generic enough to accommodate any species. The framework is evaluated on two large corpora of real field data, targeting the calls and songs of American Robin *Turdus migratorius*, a Northamerican oscine passerine (true songbird) and the Common Kingfisher *Alcedo atthis*, a non-passerine species with a wide distribution throughout Eurasia and North Africa. With the aim of promoting the widespread use of digital autonomous recording units (ARUs) and species recognition technologies the processing code and a large corpus of audio recordings is provided in order to enable other researchers to perform and assess comparative experiments.

1. Introduction

In our rapidly changing world the monitoring of animal communities is becoming increasingly important. Reliable estimates of the range, population size, and population trends are critical for assessing the conservation status of the species. Only if we know their real population status, species-specific conservation measures can be implemented and extinctions can be avoided. However, the high costs of classical observer-based survey techniques and their temporal limitation are often a major problem for the protection of wildlife. A potential solution is the automated acoustic monitoring of sound-emitting animals, as they can provide continuous real-time information on the presence/absence of target species and on the general status of the biodiversity of an area.

Consequently, in recent years biologist started to use autonomous recording units (ARUs) to survey different taxonomic groups of sound-producing animals, such as mammals [1], birds [2,3], amphibians [4], and insects [5]. Considering that these ARUs can be operated in 24/7 modus and that several recorders can be used simultaneously, huge amounts of audio data can be gathered in relatively short periods of time, meaning that it is usually not feasible for human experts to hear or visually inspect the complete sample of recordings. Thus (semi-)automatic processing of the sound files is a prerequisite for analyzing the information in a timely manner.

The operation of autonomous remote audio recording stations and the automatic analysis of their data can assist decision making in a wide spectrum of areas, such as:

(1) *Monitoring of range shifts of animal species due to climate change*. Greenhouse warming is projected to profoundly change the distribution pattern of plants and animals worldwide. For instance, the average distributional range of European birds might shift nearly 550 km north-east by the end of this century [6–9]. In the same period, about 75% of the avian species might suffer range declines and the overlap of the current and future distribution might be only 40%.

* Corresponding author. Address: Department of Music Technology & Acoustics, Technological Educational Institute of Crete, E. Daskalaki, Perivolia, Rethymno 74100, Crete, Greece. Tel.: +30 28310 21900.
E-mail address: potamitis@staff.teicrete.gr (I. Potamitis).
URL: http://scholar.google.com/citations?user=gWZ4dTUAAAAJ&hl=en I. Potamitis).
[1] http://www.teicrete.gr/mta/en/index.php?q=node/32.

0003-682X/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.apacoust.2014.01.001

Thus, monitoring of animal populations can help us to document, understand, and mitigate the impacts of climate change.

(2) *Biodiversity assessment and inventorying of an area*: Classical observer-based audiovisual surveys typically give good spatial coverage; however, in remote areas they are very difficult and costly to perform, are often incomplete, cover only short periods of time and can be obtrusive. By contrast, passive acoustic monitoring is cost-effective, unobtrusive and can be set up for systematic seasonal, altitudinal, and longitudinal long-term environmental monitoring, allowing the automatic inventorying and assessment of sound-producing animal species [10].

(3) *Estimation of species richness and species abundance*: Global environmental crisis is manifested in declining species richness and decreasing population sizes, particularly of habitat-specialized taxa. The number of species, their threat status and population trends can be used for drawing conclusions on the conservation status of habitats and landscapes. Therefore, long-term acoustic monitoring of indicator species could serve as a proxy of environmental health [11] and the status of ecosystems [12] (e.g., Living Planet Index [14]).

(4) *Assessing the status of threatened species*: Monitoring the presence/absence and abundance of rare and threatened species with large networks of autonomous monitoring stations would considerably improve our knowledge of their true population status and trends, and thus facilitate the implementation of site-specific and species-specific conservation measures.

(5) *Alarming of specific atypical sound events* related to potentially hazardous events (e.g., wildfires) and human activities (e.g. gun shooting, tree felling).

During the last decade the progress of bioacoustic technology is evident especially in the field of hardware development, particularly of programmable and affordable ARUs. Modern models are powered by solar energy, equipped with large storage capacity, weather–proof normal and ultrasound microphones, wireless sensors for data transmission, and microphone arrays for the estimation of population size. In parallel to hardware progress, large amount of bird recordings are becoming available (e.g., Maculay Library[2], Tierstimmenarchiv[3], and Xeno-canto[4]) and acoustic signal processing and pattern recognition of bioacoustics signals for the purpose of detecting and identifying bird species is currently a very active research area [12,13].

Pattern recognition of bird sounds has a long history and many signal transformations, feature extraction techniques [15–21] as well as pattern recognition approaches [22–30] have been applied to the problem of automatic bird detection and identification. Most studies typically make use of recordings either from databases or – fewer – make use of real field recordings (i.e. unedited recordings). However, in either case these recordings are manually selected, evaluated and annotated by experts and are of high quality. Moreover, in the majority of the reported literature the recognizers are usually trained and tested on manually pre-segmented data that either contains the target signal or a background species exclusively [22,25]. This exclusive situation allows the recognizer to take a forced choice decision on the majority of the observations in a file (by adding class log-likelihoods of all processed frames). This procedure, though helpful in investigating and comparing features and classifiers, is not possible to be applied in a stream of data with unknown time boundaries.

A recent trend makes the leap into facing the complexity of the real world [28–30]. This work also puts the emphasis on bird detection in unprocessed real-field recordings gathered by ARU's. We adopt feature extraction and classifier approaches that are among the standard state-of-the-art choices in speaker recognition and we elaborate on this framework, where needed, in order to accommodate the idiosyncrasies of bird species. Our contribution is to offer an open modular framework that detects and time-stamps online calls and songs of target bird species and is fast enough to report results in reasonable time for non-processed field recordings of many thousands files and is generic enough to accommodate any species. We present two novel techniques associated with our framework:

(a) Automatic tagging and time-stamping of training data in order to initialize the statistical models of the detector and, (b) a semi-automatic procedure that allows extracting suitable training data from real-field recordings to enhance the detectors' discrimination ability (i.e., audio segments of the target as well as background species). The emphasis of this work is on addressing the practical challenges of real-field recordings and on evaluating the system effectiveness in supporting decision making in conservation issues. To this end, we evaluate our approach using two large real-field corpora, including one that was recorded by our ARUs and is made available along with our detection results and expert annotations in order to serve as a benchmark for other approaches.

It is our shared belief that in order to have solid progress in this research field, researchers should depart from using private data and focus on unprocessed data as typically recorded in nature. As regards the software implementing our approach is included as an online supplement to serve as a freely available benchmark for other approaches. One should note that to the knowledge of the authors there are only three commercially available programs Song Scope, sold by Wildlife Acoustics[5]; XBAT, provided by Cornell's Lab of Ornithology; and SyrinxPC, provided by the University of Washington.

This paper is organized as follows: In Section 2 we discuss about bird vocalizations from the signal processing point of view. In Section 3 we present a principled way to construct the target and background training folders. In Section 4 we review the feature extraction process used in this study while in Section 5 we analyze the fine-tuning of the pattern recognition methods. In Section 6, we perform experiments with real-field data and we analyze the results from the current work. A discussion of the results in Section 7 summarizes the implications of the results. Finally we conclude this work in Section 8.

2. Signal processing of bird songs

Birds use acoustic vocalization as a very efficient way to communicate as the sound does not require visual contact between emitter and receiver individuals, can travel over long distances, and can carry the information content under low visibility conditions, such as in dense vegetation and during night time hours. Birds produce a variety of sounds to communicate that can be divided into three large categories: calls, songs, and mechanical sounds. The latter refer to species-specific audio signals that are not produced in the bird's syrinx but by mechanical movements of certain body parts or structurally adapted feathers, viz. the bill clapping of storks, drumming of woodpeckers, wing clapping of pigeons and doves, and the "tail humming" of certain taxa of snipes. In this paper we will focus only on sounds produced in the vocal

[2] http://macaulaylibrary.org/index.do (date last viewed 15.08.13).
[3] http://www.tierstimmenarchiv.de/ (date last viewed 15.08.13).
[4] http://www.xeno-canto.org/ (date last viewed 15.08.13).
[5] http://www.wildlifeacoustics.com/ (date last viewed 15.08.13).

organ of birds. Calls usually refer to simple frequency patterns of short monosyllabic sounds that may have many functions [31] (see Fig. 1 top), which are expressed by biologists with a descriptive terminology referring to the behavioural context in which the vocalizations are emitted, e.g., begging calls, warning and alarm calls, territorial calls, and contact calls [41]. While all birds emit calls, although with different variability and frequency, only some birds also produce songs. In difference to calls, songs are longer, acoustically more complex, and often have a modular structure (see Fig. 1, second row).

In general, the songs of the non-passerines are less complex than the songs of the passerines (perching birds). The basic elementary unit of songs are simple non-separable segments of the spectrum (also called elements, pulses, or notes). Different elements make larger units called syllables. Syllables, in turn make phrases. Syllables in the same phrase form a statistically repetitive sequence. The various hierarchies of phrases and its subunits constitute the song. The number of different syllables and phrases constitutes its syllabic-phrases repertoire while the recombination of phrases makes the song. Song complexity resembles the composition of human language, though at a simpler level, as the recombination of song elements and phrases demonstrate syntactical and grammatical structure of the human language which is also modular: different sentences can be constructed, only by changing the ordering of words, while words themselves are produced by recombination of phonemes.

Research on songbirds has demonstrated that human language and birdsongs have striking analogies in terms of both vocal articulation and neural functionality [32]. We briefly overview the underlying process of the vocal articulation mechanism in human and birds that is of interest to our work (see Fig. 2). The speech production apparatus is composed of the lungs (as a source of air pumping during speech), the larynx (the vibrating vocal cords, when triggered produce a quasi-periodic flow of air called the source sound characterized by its fundamental frequency and harmonic overtones), the pharynx (throat), the oral (mouth) and nasal cavities. The vocal tract, consisting of the nasal and oral cavities in association with the lips, tongue, jaw and teeth forms cavities that act as resonators. The amplitude of the harmonics of the larynx are modulated and amplified near the resonances [33]. The birds in turn, possess the lungs, the syrinx that has similar function to the human vocal cords (and not the larynx) while the trachea mouth and beak form cavities as in humans. Although this analysis is simplified (see Fagerlund [24] and Baptista and Kroodsma [34] for a detailed analysis on the articulation of birds) it grasps the central idea of modelling vocalizations as a quasi-periodic flow of air that is modulated by a linear time-varying tube that acts as a filter. This source filter model that represents the resonances of the vocal tract is often used as a basis system for speech production, avian vocalization, and wind instruments. The highest complexity of bird songs is found in the oscine passerines (suborder Passeri, songbirds), which can control both branches of the trachea

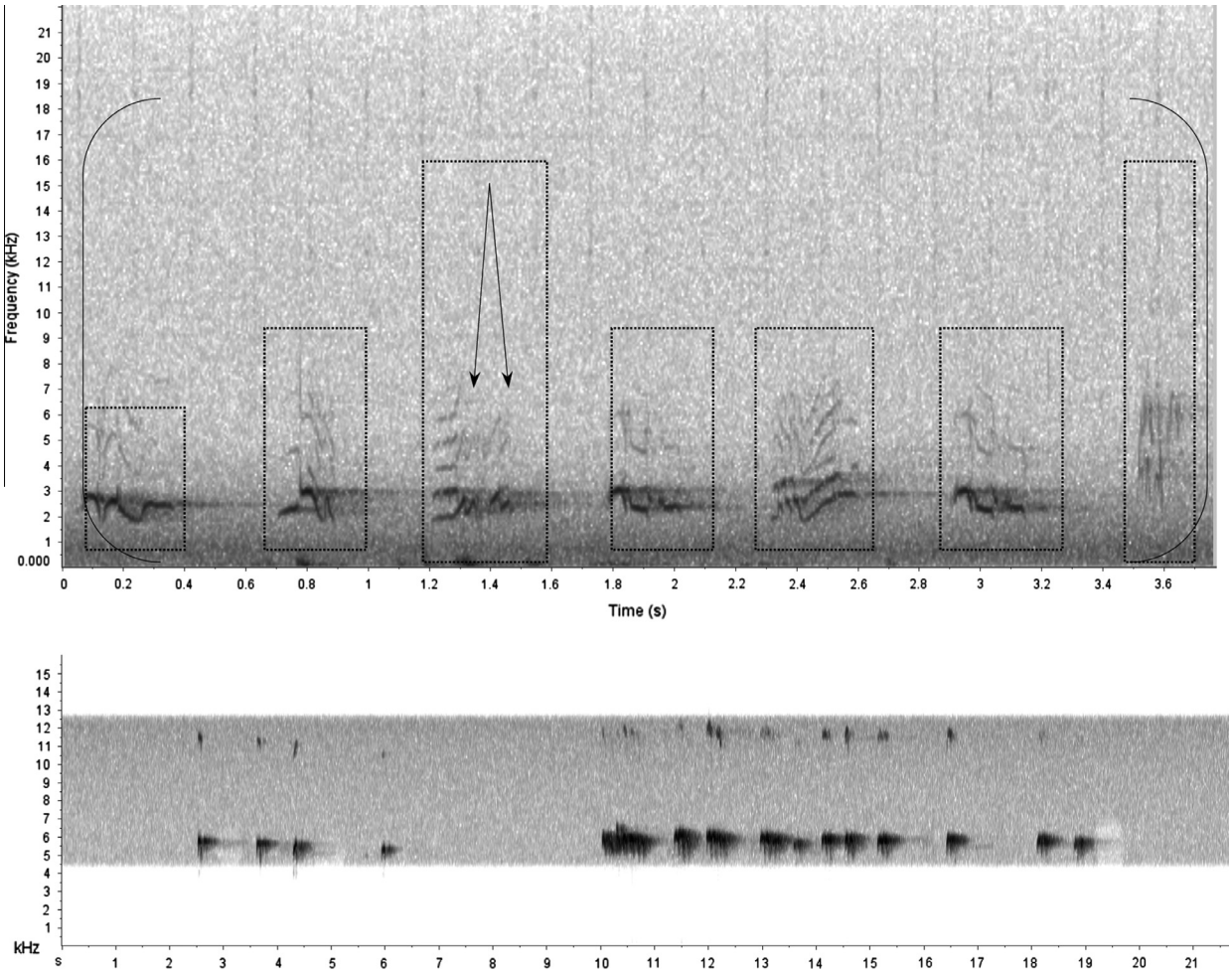

Fig. 1. Spectrograms of target species. Top row: song of American Robin – *Turdus migratorius*, elements (arrows), syllables (boxes), song (parenthesis). Second row: Kingfisher calls.

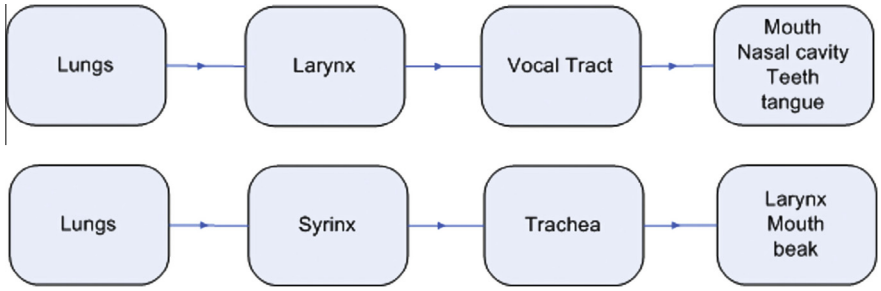

Fig. 2. Human (top) and avian (bottom) sound production model.

independently and are thus capable to sing in two pitches simultaneously [34]. By following this reasonable model we are led naturally to the feature extraction process described in Section 4.

2.1. Databases

In order to move beyond laboratory tests and face the challenges of real-world applications one needs databases from real field recordings that are also annotated at least to the species level. The database is a crucial component as it is needed to provide the ground truth for assessing the performance of different approaches under a common corpus. Unlike the case of speech databases that are now available and annotated for all major languages only a few annotated databases on bird species vocalizations do exist. The annotation of data from the real field is a task that requires an enormous effort. Firstly, the annotator must be an expert in bird vocalizations and, secondly, he has to assess an abundance of interfering audio sources in real environments that must be marked and time-stamped. Visual inspection of spectrograms does not offer sufficient clue as to the identity of calls and songs as more often than not the call repertoire of one species is very similar to that of another species (e.g., different taxa of *Parus*-tits). As the emphasis of this work is in the evaluation under real-field conditions we made use of two large corpora. The recordings in both corpora were *not* screened in any way. Therefore they include all kinds of acoustic degradations encountered in nature. The more important obstacles we faced during our experiments with real data listed in priority order are:

(a) Non-target bird species can and often do overlap both in time and frequency with the target species.
(b) Frequency selective attenuation due to distance. As the target bird can be at any distance from the microphone, many weak calls are picked up. Interestingly, some of the distant calls that are still recognizable by the human ear are already too week for a machine-based recognizer. Moreover, high frequency attenuation due to distance fades out the high frequencies of a target bird forcing its spectrum to potentially resemble a competing species; therefore increasing both misses and false alarms rates.
(c) Acoustic reflections produce a smearing in the frequency spectrum.
(d) Heavy wind, rain, and low battery power can reduce the quality of recordings.

The first corpus, the so called "Vouliagmeni corpus" was recorded using automatic recording units (Song Meter SM2, Wildlife Acoustics®) placed at Lake Vouliagmeni (37°48′28″N, 23°47′08″E; c. 10 m a.s.l.), in the Natura 2000 area "Hymettus – Kaisariani – Lake Vouliagmeni" (GR3000006) at the eastern periphery of the Greek capital Athens. Between 14.12.2010 and 20.12.2010 we made 10,000 audio recordings of 15 s, 48 kHz, 16 bit stereo, corresponding to one recording per minute. In 2011 we partially annotated the "Vouliagmeni corpus" and found the Common Kingfisher *Alcedo atthis* to be one of the more conspicuous sound-emitting birds in the recordings. At Lake Vouliagmeni the species was not previously reported by other observers, which demonstrates the power of ARUs in inventorying and documenting the wildlife of an area. The Common Kingfisher feeds mainly on fish and aquatic arthropods, which is why populations that breed in regions were rivers and lakes freeze in winter perform a southward migration in late autumn. The species is listed in Annex 1 of the European Birds Directive [8], making it an ideal study object for our purpose. As the Kingfisher only has a limited call repertoire, our corpus allows testing the ability of a detector to discern a small number of target vocalizations in a large number of files not containing the target species. To be exact, one call of Kingfisher bears great similarity to a call of Dunnock (*Prunella modularis*). Therefore, hereinafter the results refer to Kingfisher and the specific Dunnock call.

The second corpus on which we tried our approach was the Rocky Mountain Biological Laboratory American Robin database made available from UCLA dept. of Biology [35]. The dataset is 78.3 min long recorded using automatic recording units (Song Meter SM1, Wildlife Acoustics®) at 44.1 kHz, 16 bit and is manually annotated by experts both on the syllable level as well as on the song level with time-boundaries included. Each song usually consists of 2–10 syllables. The database is split into a non-overlapping training and test corpus. We have partitioned the database to training and test corpus according to the setting reported in [35] so that our detection results are directly comparable.

Both corpora are accompanied by a large number of expert-verified calls and can, therefore, serve as ground-truth for training and assessing different feature extraction techniques and classifiers.

3. Forming the target and background training set: Parsing of calls and activity detection

In our approach, the construction of a bird detector starts with an expert familiar with bird vocalizations delivering a small corpus of ~50 signals (calls/songs) of the target species of interest. Ideally these recordings were made in the same area where the ARUs are placed, because many avian species and subspecies have distinct local dialects. This is particularly true for oscine passerines. Hence, the suitable examples provided by the expert for training of a recognizer are usually very few.

The construction of the background set of recordings also is a critical issue. Extensive experimentation on field data has shown that the main source of false alarms is due to competing species vocalizing on the same bandwidth as the target species. We have inspected hundreds of recordings and verified that species vocalizing on the same bandwidth is a very common case indeed. Hereinafter, we present a principled way to form the target and background training set of the recognizer instead of randomly

checking recordings from a pool of tens of thousands of recordings provided by ARUs in long-term monitoring programs.

The procedure is depicted in Fig. 3 and is explained analytically in the paragraphs "*Signal-Preprocessing*" and "*Hilbert Follower*" within the present chapter. In brief, the examples provided by the expert are automatically segmented into syllables from which rough descriptive statistics in terms of spectral energy distribution and call duration are derived. Once the descriptive statistics are derived from the small set of example data, the procedure is repeated for a much larger sample of ARU recordings (parsing of all field recordings). All audio signals falling in the range of the descriptive statistics are copied to a development folder for manual inspection. The folder is examined by the expert, who partitions the data into a target and background set. The gain of this procedure is twofold: (a) the background set consists of signals that are more probable to confuse the detector as they share some statistical characteristics with the target signal and, (b) the training set consists of training data matched to the operational environment.

In the case of Kingfisher, calls range from [150 to 600 ms] and the bandwidth 4.5–6 kHz holds more than 92% of the total energy of the signal in all example calls processed. From the 10,000 initial files of the Vouliagmeni corpus only 623 files had a segment that matched this profile, corresponding to a reduction of 93.77% in search space for the human observer. From these 623 files the 121 were found to contain Kingfisher calls and the rest were moved to the background folder. The large number of files moved to the background folder shows that segment duration and bandwidth are not enough clues to precisely detect the target species; however, they are good enough for reducing the search space for the human observer, who otherwise has to search at random all ARU recordings, which might be of an unmanageable size for finding proper training data for the target and background species.

We proceed into describing the pre-processing and sound activity detection steps that take place during the derivation of the target and background training sets. Note that this procedure is generic and can be applied to any target species. We have applied it only to the Vouliagmeni corpus because the American Robin database is released already annotated. In the later case syllables and inter-syllable gaps of background noise are parsed and extracted from the training part of the database and form the training corpora for the syllable model and background model respectively.

3.1. Signal pre-processing

Due to the fact that different species vocalize in different frequencies it is not optimal to follow the same pre-processing procedure in all species in contrast to the case of speech signal. Therefore, we initially downsample as low as the highest frequency of the target signal allows. In the case of the Kingfisher this is 32 kHz as the species' frequency range is between 4.5 and 12.5 kHz. For the robin it is 16 kHz as its spectrum is constrained between 1 and 5 kHz. Subsequently we apply band-pass filtering retaining only the frequency range of the aforementioned target signals. Band-pass filtering is a crucial step as it is able to reduce wind and interference of competing species. Subsequently, a signal enhancement stage follows that deals with the remaining noise inside the band-pass limits [36].

3.2. Hilbert follower

The Hilbert follower serves as an audio activity detector and segments the recording by following the characteristic shape of the syllable envelopes of bird vocalizations. It is used in the procedure described in Fig. 3 and in training the recognizer. We briefly describe its derivation and function:

Let $x(n)$ denote the discrete time-domain signal holding the original recording, where n is the discrete-time index. In the next steps we outline the manner in which $x(n)$ is transformed into a set of low dimensional descriptors that subsequently are fed to the pattern recognition stage.

Let, $x_h(n) = Hilbert(x(n))$ return a complex sequence called the *analytic signal* of $x(n)$. The analytic signal $x_h(n) = x(n) + jx_i(n)$ has a real part $x(n)$ which is the original data, and an imaginary part, $x_i(n)$, which contains the Hilbert transform of $x(n)$. The envelope $y(n)$ of the sampled time-domain recording is calculated as:

$$y(n) = \left(X_h(n) \otimes \widehat{X}_h(n)\right)^{1/2} \quad (1)$$

where $\widehat{X}_h(n)$ stands for the conjugate of $X_h(n)$ and $\otimes$ for component wise multiplication.

The envelope in Eq. (1) is compared against a threshold θ. When $y(n) > \theta$ the sample $x(n)$ is classified as belonging to the *activity* class otherwise to the *non-activity* class (see Fig. 4). The segmentation procedure imposes time-stamps for the boxed segments detected to have audio activity and, therefore the time-tagged recording can serve for accurate initialization of the HMMs avoiding the need for human annotation of syllables (e.g. by using the Praat software).

4. Acoustic parameters for bird classification

In order to build an efficient bird detector one must discard any information not useful to the detection task. Bird vocalizations may have pure sinusoidal, harmonic content demonstrating modulation in amplitude and frequency or even at both levels and non-harmonic and noisy structure. Fundamental frequency ranges from 100 Hz to 8 kHz [34] while the frequency content spreads at least up to 20 kHz.

The audio feature vector comprises a summarization of the useful information (from the perspective of pattern classification) hidden in the sound signal. The ability to carry out species independent detection and recognition lies in the selection of distinguishing acoustic features that remain relatively invariant regardless of the vocalizations produced and are mostly dependent on the characteristics of the source-filter model (i.e. the vocal tract resonances contain the most individually specific information).

Section 3 described how the target and background folders training folders are composed. Subsequently, the audio signals in these folders are transformed to the corresponding feature vectors

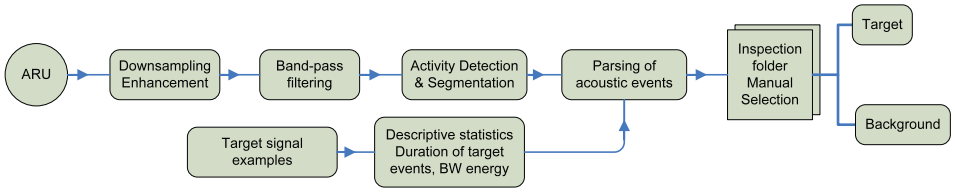

Fig. 3. Derivation of the target and background set from field data. The recordings of the automatic recording unit are pre-processed; the detected audio segments are compared against the target signal profile. Matching recordings are directed to the human observer for final classification. Note, that 90–95% of the ARU data do not reach the observer as they are automatically rejected for not meeting the target profile.

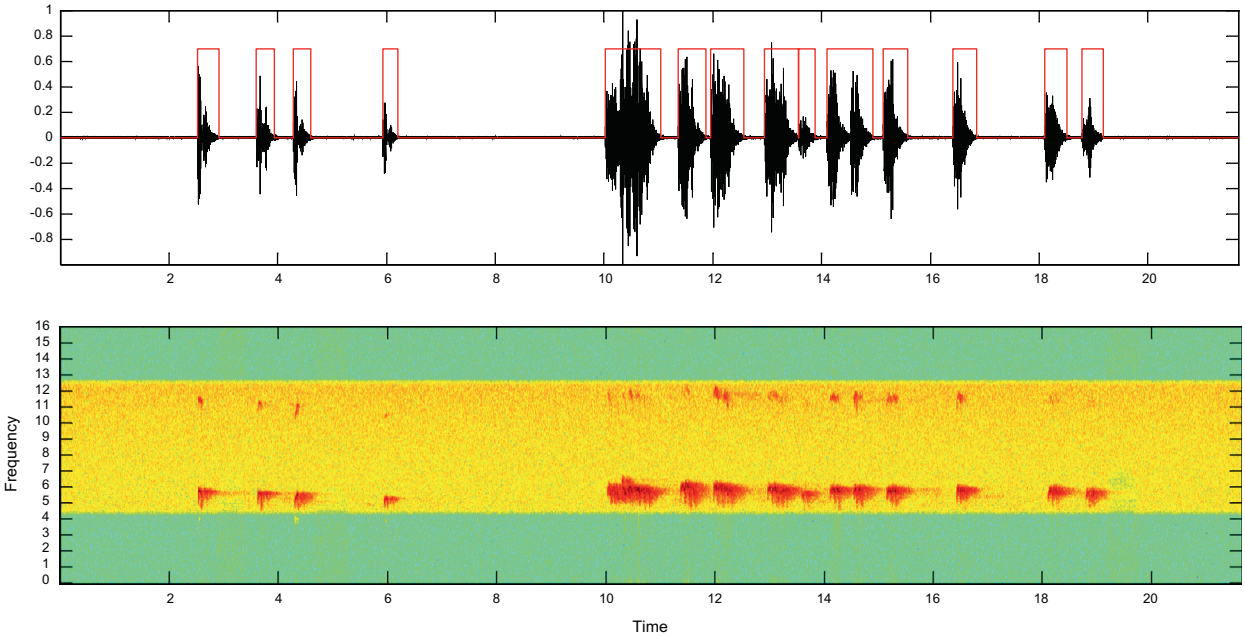

Fig. 4. The Hilbert follower as an activity detector and the segmentation of bird calls. Call sequence of Common Kingfisher shown.

for training the classifier. For extracting HTK [38] compatible Perceptual Linear Prediction cepstral coefficients (PLP-CC) we first follow the following procedure: The recording is frame-blocked into small, overlapping chunks. The frame size is set to 10 ms at a rate of 5 ms. A Hamming function is used to window the frames and let the signal in a frame be denoted by $s(n)$, $n = 0,\ldots N-1$. Each frame has to be multiplied with a hamming window in order to keep the continuity of the ending samples in the frame so that subsequent application of FFT does not produce spikes. The signal after Hamming windowing is $s_2(n) = s(n)w(n)$, where $w(n)$ is the Hamming window defined by:

$$w(n) = 1/2 - 1/2\cos(2\pi n/(N-1)), \quad 0 \leqq n \leqq N-1 \tag{2}$$

The signal $s_2(n)$ is sent to a high-pass filter to partially compensate attenuation of high-frequencies due to long distance from the microphone:

$$s_2(n) = s(n) - ks(n-1) \tag{3}$$

where a pre-emphasis with $k = 0.97$ is applied.

Spectral analysis follows by applying the FFT and by extracting the power in order to get an estimation of how the energy of the chunk is distributed over frequencies and an estimation of the different timbres of the vocalizations. We then multiply the magnitude frequency response by a set of 41 triangular bandpass filters to get the energy of each triangular bandpass filter. The positions of these filters are equally spaced along the Mel frequency, which is related to the common linear frequency f by the following equation:

$$\text{mel}(f) = 1125 * \ln(1 + f/700) \tag{4}$$

At this point we depart from the subsequent steps of extracting Mel-frequency Cepstral Coefficients (MFCC) and instead of applying the log to the filterbank energies as is the case of MFCCs we apply the cube root and subsequently we approximate the power spectrum by applying an all-pole model to the 41 Mel-frequency bands in order to get 16 PLP (0–15) coefficients. The power of the signal segment and the linear predictive coding coefficients are linked through the following equation:

$$P(\omega) = \frac{G^2}{\left|1 + \sum_{k=1}^{p} a_k e^{-jk\omega}\right|^2} \tag{5}$$

where a_k are the coefficients of the resulting pth order polynomial. The derivation of cepstral coefficients from a given set of LPC is simple because there exists direct transformation from LPC to cepstral coefficients [33].

Subsequently, cepstral liftering applies a beneficial weighting to the cepstral features.

$$w_c = 1 + L/2 * \sin([1 : (\text{ncep} - 1)] * \pi/L) \tag{6}$$

where $L = 22$ and $\text{ncep} = 16$ holds the number of cepstral coefficients.

The log-energy is appended to the PLP 0–15 instead of the 0th PLP. The features are mean normalized with respect to the full recording.

Delta and acceleration coefficients (deltas of deltas) are appended to the PLP-CC. The equations to compute these features are:

$$dt = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \tag{7}$$

where d_t is a delta coefficient, from frame t computed in terms of the static coefficients c_{t+N} to c_{t-N}. A typical value for N is 2. Acceleration coefficients follow (7), but they are calculated from the deltas, not the static coefficients.

To sum up, we employed the feature set named PLP_E_D_A_Z which consists of the first 16 perceptual linear prediction (PLP), log energy (E), and their respective delta (D) and double delta (A), while cepstrum mean subtraction is applied (Z) (i.e., 51 dimensions, see Table 1). PLP_E_D_A_Z are almost the same to the popular Mel frequency cepstral coefficients but demonstrated a small consistent gain, which also has been observed in state of the art speech recognition [39], and therefore were the feature extraction process of choice. The Opensmile feature extraction framework [37] which is hard-coded and therefore very fast was used to carry out the feature extraction.

5. Pattern recognition using HMMs and associated techniques

The bird verification process based on its vocalization is a typical two-class decision problem ending with a binary decision of

Table 1
The parameters of the feature extraction front end.

	Common Kingfisher, American Robin
Sampling frequency	32 kHz (Kingfisher), 16 kHz (robin)
High-pass filtering	1–5 kHz (Kingfisher), 4.5–12.5 kHz (robin)
Pre-emphasis	0.97
Window-length	10 ms
Overlap	5 ms
DFT size	256
Filterbank	Mel
Filterbank channels	41
Ceplifters	22
Channel compensation	Cepstral mean subtraction
No. of Coefs	(16 PLP)x3 + 3(E,D,DD) = 51

existence about the species that produced a specific sound signal. In addition to the correct identification, the bird detection task has to detect the time boundaries of the vocalizations in order to allow for subsequent analysis of activity counts. We apply the framework of hidden Markov models (HMMs) as it is able to deal with audio patterns, which vary both in time and frequency, and is able to incorporate a language model into the decoding procedure that we intend to use in order to incorporate context information. We employed the hidden Markov model toolkit (HTK) [38,39], which has been deeply explored by the speech processing community mainly for speech and speaker recognition. HTK, is a compilation of optimized functions for training and decoding using probabilistic models and is fast enough to accommodate applications that require real-time performance. It is not a stand-alone recognizer and its performance depends greatly on the knowledge and experience of the user in pipelining sophisticated tools. We focus on presenting some constructive observations after thoroughly testing recognizers in very large corpora for a number of species.

1. The amount of suitable training data for target species is often very small, especially for species that are rare, mostly silent, occupy large territories or home ranges, or live in environments that do not allow to make large corpora of recordings. Therefore, flat-start of model training using the standard Baum-Welch training procedure is not recommended as it leads to poor-models. The training should be based on training initial models using the Viterbi algorithm on bootstrap data provided by the audio activity detector and then elaborating on this model based on successive passes of Baum-Welch. To make this procedure feasible, the target as well the background folder must be annotated with their time boundaries. This can be done manually – in theory as only the background model has thousands of audio segments – or as in our case automatically using the segmentation results of Hilbert follower as described in Section 3.
2. When band-pass filtering is applied, one should not tailor the band-pass filters only according to the target species. Another species may have part of its spectrum exactly in the same frequency range as the target species. In such case the recognizer will have not enough information to discriminate between classes. Band-pass filtering should allow enough spectral band space for non-target species so that the background models are trained as accurately as possible.
3. The number of mixtures should be small in order not over-optimize on the training data (3 emitting states and 4 mixtures in our experiments).
4. Human experts identify bird sounds by combining many cues of information including higher-level syntactical units (e.g., repetition of syllables and phrases). The higher-level information is partially incorporated into the decoding process through a lattice model (see Fig. 5). Note that the decoding procedure of HMMs integrates language or lattice model information and provides a probability score, which is independent from that coming from the features [38,39]. A degree of context information is also incorporated through the use of deltas and double-deltas in the feature extraction stage.
5. The construction of different kinds of background models is beneficial. In this work we have separate models for target and non-target avian background activity, long pauses, and inter-syllable pauses. The distinct background models allow for more precise segmentation results during the decoding process.
6. The training of a background model with recordings containing vocalization of non-target species that vocalize in the same bandwidth as the target taxon, allows the detector to achieve better discrimination performance and it is found to effectively reduce the number of false alarms.
7. The output results of the detector are parsed and detections with duration outside the time interval of the target signal are rejected. We recommend that the minimum permissible duration of a species' sound signal should be determined by automatic segmentation of the initial training data set and by parsing their corresponding durations. Less reliable is to derive the values from the corresponding literature, due to the aforementioned regional differences of bird sounds within a species. The detections that pass the first parsing stage are further examined against their energy. Detections corresponding to signal segments having energy below −8 dB are also rejected as they correspond to very faint calls, which have lost most part of their high frequency spectrum due to frequency-selective attenuation and are dubious even for an experienced listener.

6. Experimental set-up and analysis of results

The Kingfisher corpus is assessed on file basis and on call basis. Results on file basis answer the question: "From all recordings of the ARU, select all those that have at least one call of the target species in the recording". This helps as to assess how much is the search space (i.e. initially 10,000 files) for human observers reduced by the detector. Results on per detection basis report on the number of detected calls and exact location inside each file.

As regards the American Robin Database syllables as well as songs are manually annotated and time-stamped. Therefore, they offer a convenient and reliable test bed for comparing feature extraction and pattern recognition algorithms. As regards validation results using the robin database (see Table 2) we report scores both on the syllable level and on the song level. Results on the syllable level are based on training with syllables of the training corpus and test with syllables of the test corpus. For recognition on the song level we convert the recognition results of the syllable level to song-level results and then compare them with the manually annotated song data of the test corpus. The conversion is possible as the duration between syllables in a robin song is less than 0.5 s and, therefore syllables with less than 0.5 s inter-syllable distance are grouped into the same song. The HTK is configured to word-spotting mode where each recognition label is compared with the reference transcriptions. If the start and end times lie on either side of the mid-point of an identical label in the reference, then the recognizer label represents a hit; otherwise it is a false alarm.

We evaluate the performance of the detection framework in terms of precision (P) and recall (R) which are defined as:

$$P = \frac{N_{hits}}{N_{hits} + N_{FA}}, \quad R = \frac{N_{hits}}{N_{actual}} \tag{8}$$

where $N_{hits} + N_{FA}$ is the number of detected events and N_{actual} is the number of ground truth (syllables and songs respectively). We measure P and R independently on the song (Songs P, R) and

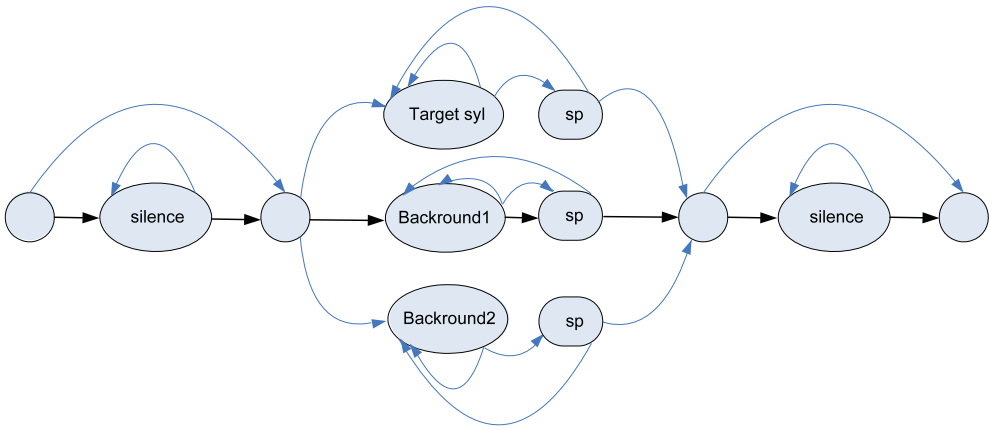

Fig. 5. Lattice network associated with the detection framework: the illustrated modeling scheme includes target syllables, different background models, short pauses (sp), and beginning/ending silence regions.

Table 2
Precision and recall scores for detection counts on the level of a file and call respectively for the Common Kingfisher and for the American Robin.

Species	Precision (%)		Recall (%)		*F*-score	
	Recordings	Calls	Recordings	Calls	Recordings	Calls
Common Kingfisher	88.1[a]	84.9	93.3[a]	85.2	0.916	0.85
American Robin	Syllables	Songs	Syllables	Songs	Syllables	Songs
This work	71.2	85.1	91.3	77.0	0.840	0.79
W. Chu method [28]	N/A	75.2	N/A	76.0	N/A	0.75

[a] *Note*: The 'recordings' and 'call' rate refers to the total number of kingfisher calls within the detected recordings.

syllable level (Syllables *P*, *R*). The effort in detection systems in general is to design ones that can both increase precision and recall. The trade-off between the recall rate and precision rate is well-known and therefore, the F-score is usually reported which is a weighted combination of these two error metrics. The *F*-score is defined as:

$$F = \frac{(1+\beta^2)PR}{\beta^2 P + R} \tag{9}$$

We compare our approach with a state of the art bird detection approach. The Chu approach [35] is based on HMMs and a lattice model that captures the permissible transitions between the syllables of the songs belonging to a target species. Since bird song databases are not provided with annotations to the syllable level it is necessary to infer the numbers of the common syllable patterns from the training set, and then train acoustic models for those patterns. The syllable patterns are inferred by performing a hierarchical clustering analysis. Following the Chu and Blumstein [35] settings, we use $\beta = 1.5$ because in our detection task the primary aim is not to miss calls from the target species, and therefore recall is more important than precision.

In the case of Kingfisher out of 10,000 recordings of 15 s each only 152 recordings were detected by the system to contain Kingfisher vocalizations. After inspection 126 were confirmed to indeed contain one. The total number of Kingfisher calls reported by the human observer was 384 out of which 317 were detected by the system. In the case of Robin, 1071 syllables are correctly discovered out of 1173 tagged by the human annotator and 177 songs are correctly detected out of the 230 on the total corpus of 78.3 min (see also Table 2 for analytic results). One should note that for the experimental set-up the 78.3 min of the Robin database are split into a 45.5 min training corpus and a 32.8 min of non-overlapping corpus as reported in the corresponding reference [35].

7. Discussion

We conclude this work by reporting on its practical achievements and its limitations. To our point of view, the practical gains are:

1. Automatic species recognition can considerably reduce the search space for a human observer (in the case of the Kingfisher the reduction was 98.48%). By itself this is a valuable contribution as biologists are commonly facing the situation of a deluge of audio data from long-term recording programs. Without the automated approach the task is not manageable. Consequently, under the latter condition the data can be analyzed only partially and randomly by an expert, on the basis of visual inspection of the spectrogram.
2. Considering that computer-based detection of species and the time-stamping of the corresponding signals is fast, new options for analysing long-term audio data arise: for instance, it is possible to query (a) the vocal activity patterns of certain species on various time scales, viz. daily activity, seasonality, and arrival/departure dates of migratory species, and to compare the results between years or in dependence of other factors, such as climate data; (b) the effectiveness of habitat restoration efforts [3]; and (3) the presence of rare and cryptic species.
3. Our long practical experience with real-field recordings of calls and song leads us to state that the technology is *not yet* mature enough to completely automate decision making on critical biodiversity issues. Although the precision and recall scores are quite high, a detection system that would work on a continuous basis would still produce a large net number of false alarms and missed positives. It will require a large number of deployed ARUs, larger computational power and a 20% increase in the precision and recall rates, before expert confirmation would be dispensable in for technology applications. However, we

anticipate that ARUs will become smaller, cheaper, and more energy-efficient in the near future and that detection technologies have not used all of the sophisticated methods that are currently used in advanced forensic applications for detecting identity from voice in real-life tasks [40].

8. Conclusion

A detection task has to discern a target signal against anything that is not target (e.g., species vocalizing in the same frequency range and environmental noise). In order to confirm the presence of a species in a large corpus of audio recordings within a reasonable time period it is necessary to use taxon-specific detectors of calls and songs on the basis of statistical models.

With the aim of promoting the widespread use of digital autonomous recording units (ARUs) and species recognition technologies we provide our processing code and a large corpus of audio recordings in order to enable other researchers to perform and assess comparative experiments on real-field data. Ideally, future improvements of these and other tools for computational ecology will lead to the development of proper conservation measures for threatened wildlife.

Acknowledgments

We acknowledge Wei Chu from UCLA for providing the American Robin corpus of recordings and annotations. The authors would like to thank P. Petrakis and S. Kouzoupis for preparing the training data. This work was supported by the European Community LIFE + Program AMIBIO "Automatic Acoustic Monitoring and Inventorying of Biodiversity", Grant: LIFE08 NAT/GR/000539.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.apacoust.2014.01.001.

References

[1] Wiggins S, Hildebrand J. High-frequency Acoustic Recording Package (HARP) for broad-band, long-term marine mammal monitoring. In: International symposium on underwater technology 2007 and international workshop on scientific use of submarine cables & related technologies 2007. Institute of Electrical and Electronics Engineers, Tokyo, Japan; 2007. p. 551–7.
[2] Swiston K, Mennill D. 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 2009;80:42–50.
[3] Buxton R, Jones I. Measuring nocturnal seabird activity and status using acoustic recording devices: applications for island restoration. J Field Ornithol 2012;83(1):47–60.
[4] Slimani T, Beltrán Gala J, El Mouden E, Radi M, Marquez R. Recent discoveries and monitoring populations of the Moroccan Toad (*Alytes maurus* Pasteur & Bons, 1962). In: Proceedings of the XI Congresso Luso-Espanhol de Herpetología/Sevilha; 2010. p. 148.
[5] Riede K. Monitoring biodiversity: analysis of Amazonian rainforest sounds. Ambio 1993;22:546–8.
[6] Huntley B, Green R, Collingham Y, Willis SG. A climatic atlas of European breeding birds. Lynx Edicions; 2008. p. 521.
[7] IUCN. IUCN Red List of threatened species, version 2011.2. <http://www.iucnredlist.org> [last accessed 15.08.13].
[8] EC. Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora. <http://ec.europa.eu/environment/nature/legislation/habitatsdirec-tive /index_en.htm> [last accessed 15/08/2013].
[9] EC. Directive 2009/147/EC of the European Parliament and of the Council of 30 November 2009 on the conservation of wild birds. <http://ec.europa.eu/environment/nature/legislation/birds-directive/index_en.htm> [last accessed 15/08/2013].
[10] Blumstein D, Mennill D, Clemins P, Girod L, Yao K, Patricelli G, et al. Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. J Appl Ecol 2011;48:758–67.
[11] Alldredge M, Pollock K, Simons T. Estimating detection probabilities from multiple-observer point counts. Auk 2006;123(4):1172–82.
[12] Computational bioacoustics for assessing biodiversity. In: Frommolt Karl-Heinz, Bardeli Rolf, Clausen Michael, editors. Proceedings of the international expert meeting on IT-based detection of bioacoustical patterns. International Academy for Nature Conservation (INA), Isle of Vilm, Germany; 2008.
[13] ICML 2013. In: Glotin H, Clark C, LeCun Y, Dugan P, Halkias X, Sueur J, editors. Workshop on machine learning for bioacoustics, June 16–21, 2013 – Atlanta, Georgia, United States.
[14] WWF and ZSL editors. Living Planet Index, Factsheet 1.2.1. Biodiversity Indicator Partnership 2010. <http://static.zsl.org/files/1-2-1-living-planet-index-1062.pdf> [last accessed 15.08.13].
[15] Clemins P, Trawicki M, Adi K, Tao J, Johnson M. Generalized perceptual features for vocalization analysis across multiple species. In: IEEE International conference on acoustics, speech and signal processing; 2006. p. 253–6.
[16] Somervuo P, Harma A, Fagerlund S. Parametric representations of bird sounds for automatic species recognition. IEEE Trans ASSP 2006;14(6):2252–63.
[17] Selin A, Turunen J, Tanttu JT. Wavelets in recognition of bird sounds. EURASIP J Adv Signal Process 2007 [Article ID 51806].
[18] Brandes T. Feature vector selection and use with hidden Markov models to identify frequency-modulated bioacoustic signals amidst noise. IEEE Trans Audio Speech Lang Process 2008;16:1173–80.
[19] Elizabeth J. A new perspective on acoustic individual recognition in animals with limited call sharing or changing repertoires. Anim Behav 2008;75:1187–94.
[20] Jancovic P, Kokuer M. Automatic detection and recognition of tonal bird sounds in noisy environments. J Adv Signal Process 2011;2011:1–10.
[21] Cai J, Ee D, Pham B, Roe P, Zhang J. Sensor network for the monitoring of ecosystem: bird species recognition. In: 3rd International conference on intelligent sensors, sensor networks and information; 2008. p. 293–8.
[22] Kogan J, Margoliash D. Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: a comparative study. J Acoust Soc Am 1998;103(4):2185–96.
[23] Kwan C, Ho K, Mei G, Li Y, Ren Z, Xu R, et al. An automated acoustic system to monitor and classify birds. EURASIP J Appl Signal Process 2006 [Article ID 96706].
[24] Fagerlund S. Bird species recognition using support vector machines. EURASIP J Appl Signal Process 2007 [Article ID 38637].
[25] Trifa V, Kirschel A, Taylor CE, Vallejo EE. Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. J Acoust Soc Am 2008;123(4):2424–31.
[26] Bardeli R. Similarity search in animal sound databases. IEEE Trans Multimed 2009;11(1):68–76.
[27] Ren Y, Johnson M, Clemins P, Darre M, Glaeser S, Osiejuk T. A framework for bioacoustic vocalization analysis using hidden Markov models. Algorithms 2009;2(4):1410–28.
[28] Briggs F, Fern X, Irvine J. Multi-label classifier chains for bird sound. In: Proceedings of the 30th international conference on machine learning. Atlanta, Georgia, USA; 2013 [JMLR, W&CP volume 28].
[29] Briggs F, Lakshminarayanan B, Neal L, Fern X, Raich R, et al. Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. J Acoust Soc Am 2012;131:4640.
[30] Bardeli D, Wolff, Kurth F, Koch M, Tauchert K, Frommolt K. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. Pattern Recogn Lett 2009;31:1524–34.
[31] Marler P. Bird calls: a cornucopia for communication. In: Marler P, Slabbekoorn H, editors. Nature's music: the science of *birdsong*. New York (NY): Elsevier Academic Press; 2004. p. 132–77 [chapter 5].
[32] Doupe A, Kuhl P. Birdsong and human speech: common themes and mechanisms. Annu Rev Neurosci 1999;22:567–631.
[33] Huang X, Acero A, Hon H. Spoken language processing. Prentice Hall; 2001. 1–1008.
[34] Baptista L, Kroodsma D. Avian bioacoustics, handbook of the birds of the world. In: del Hoyo J, Ellio A, Sargatal J, editors. Mousebirds to hornbills. Lynx Editions, Barcelona, Spain, vol. 6; 2001. p. 11–52.
[35] Chu W, Blumstein D. Noise robust bird song detection using syllable pattern-based hidden Markov models. In: ICASSP, IEEE international conference on acoustic speech and signal processing; 2011. p. 345–8.
[36] Cohen I. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans Speech Audio Process 2003;11(5):466–75.
[37] Eyben F, Wöllmer M, Schuller B. OpenSMILE – the munich versatile and fast open-source audio feature extractor. In: Proc ACM multimedia (MM). Firenze, Italy: ACM; 2010.
[38] Hidden Markov model Toolkit. <http://htk.eng.cam.ac.uk/> [date last viewed 15.08. 13].
[39] Gales M, Young S. The application of hidden Markov models in speech recognition. In: Foundations and trends in signal processing, vol. 1, no. 3; 2007. p. 195–304 [c 2008 M. Gales and S. Young, doi: 10.1561/2000000004].
[40] Kinnunen T, Li H. An overview of text-independent speaker recognition: from features to supervectors. Speech Commun 2010;52(1):12–40.
[41] Catchpole C, Slater P. Bird songs: biological themes and variations. Cambridge; 2008.