Automatic bird sound detection: Applications and tools

Ilyas Potamitis[1], *Olaf Jahn*[2], *Klaus Riede*[2]

[1] Department of Music Technology and Acoustics, Tech. Educ. Institute of Crete, Crete, Greece
[2] Zoologisches Forschungsmuseum Alexander Koenig, 53113 Bonn, Germany
potamitis@staff.teicrete.gr, {o.jahn.zfmk,k.riede.zfmk}@uni-bonn.de

Abstract

Our primary purpose for pursuing this research is to present a body of disciplines and techniques required to enable reliable automatic bird species identification on the basis of their sound emissions in the field. We propose a practical and complete computer-based framework to detect and time-stamp particular bird species in continuous streams of 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. We evaluate our framework on a large corpus of real field data that we make available along with its verified detections, targeting the Common Kingfisher *Alcedo atthis.*

Index Terms: *Alcedo atthis*, automatic bird detection, bioacoustics, Common Kingfisher, computational ecology.

1. Introduction

In our rapidly changing world the monitoring of animal communities is becoming increasingly important. For instance, global warming is projected to profoundly change the distribution pattern of European birds by shifting their average distributional range nearly 550 km north-east by the end of this century [1]. In the same period, about 75% of the avian species might suffer range declines due to numerous threats such as intensive agriculture and pollution, and the overlap of the current and future distribution might be only 40%. Monitoring of populations is necessary to document, understand, and mitigate the diverse impacts of these global changes.

Reliable estimates of the range, population size, and population trends are critical for assessing the conservation status of species and generate lists according to their global and regional threat status as these species are of special concern [2-4]. Such “Red Lists” can help to implement conservation measures and, hopefully, avoid extinctions. However, the high costs of classical observer-based survey techniques are a major problem for effective wildlife monitoring. A potential solution is the automated acoustic monitoring of sound-producing animals, which is more cost-effective than classical surveys. Furthermore automatic monitoring systems can provide continuous real-time information on the presence/absence of target species.

Consequently, in recent years biologist started to use autonomous recording units (ARUs) to survey different taxonomic groups of sound-producing animals, such as mammals, amphibians, birds, and insects [5-7]. Considering that these ARUs can be operated in 24/7 modus and that several recorders can be operated 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 inspect the complete sample of recordings.

To address this challenge, we propose a workflow for the detection, segmentation, time stamping, and delivery of activity counts of target species in long time series of real field recordings.

Pattern recognition of bird sound has a long history and many feature extraction techniques [8-10] and pattern recognition approaches [11-14] have been applied to the problem of automatic bird detection and identification. However, in the majority of the literature the recognizers are usually trained and tested on pre-segmented data that contain the target signal. This allows the recognizer to take a decision on the majority of the observations in a file (by adding class log-likelihoods of all frames). Pre-segmentation of train and test data is useful for research on features and classifiers, but, in on-line operational condition the recognition process is forced to make decisions using a small fraction of the data, usually on per-segment basis.

We are currently implementing a number of tools in order to provide an explicit, objective framework for the detection of bird species with the ultimate goal to automate decision making in long-term monitoring studies regarding the presence/absence of species, rough estimation of their population status and trends. These tools were designed to solve the following subtasks:

a) assessing syllable segmentation approaches that are used to prepare the training corpora.

b) detection and time- stamping of calls and songs of target bird species in reasonable time for real field recordings of many thousands files.

c) making available a large corpus of recordings along with our detection results and expert confirmations 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.

Our work brings together different research disciplines that have been ignoring each other for a long time; on the one hand bioacoustics and ecology and on the other hand engineering. Engineers working on audio signal processing and pattern recognition as applied to bird vocalizations have much to learn about song and call repertoires and different species having similar sound repertoires in order to fine-tune their approaches based on ground truth that can only be provided by expert zoologists. In exchange, ornithologists obtain advanced tools that reduce their acoustic search space, providing them with an abundance of high quality vocalizations and automated guidance on species presence and population counts. Furthermore they can take advantage of powerful, state-of-the-art speaker and speech recognition technologies, hitherto not yet applied in animal sound recognition.

2. Calls & Songs

In this paper we will focus on sounds produced by the vocal organ of birds, the syrinx. Calls usually refer to simple frequency patterns of short monosyllabic sounds that may have many functions, which are expressed by biologists with a descriptive terminology referring to the behavioral context in which the vocalizations are emitted, e.g., begging calls, warning and alarm calls, territorial calls, and contact calls. While all birds emit calls, although with different variability and frequency, only some birds also produce songs. Songs are longer and more complex than calls, and often have a modular structure. In general, the songs of the non-passerines are less complex than the songs of the passerines (perching birds). The highest complexity of bird songs is found in the oscine passerines (suborder Passeri, songbirds), which can control both branches of the trachea independently and are thus capable to sing in two voices simultaneously.

The basic elementary units of songs are simple non-separable segments of the signal (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. As mentioned above calls have a simple and more or less stereotype pattern while songs resemble the composition of human language, though at a simpler level.

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. 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 is not enough 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). In this work we made use of a real-field corpus recorded at 48 KHz, 16 bit stereo. The "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 sec, corresponding to one recording per minute. The records were *not* screened in order to be of high quality. Therefore they include all kinds of acoustic interference such as wind, rain, audio sources like traffic noise, frequency selective attenuation due to distance and acoustic reflections, and last but not least many other bird species.

In 2011 we partially annotated the "Vouliagmeni corpus" and found the Common Kingfisher *Alcedo atthis* (Linnaeus, 1758) 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 species is listed in Annex 1 of the European Birds Directive [4], 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. The corpus is accompanied by a large number of expert-verified calls and can, therefore, serve as ground truth for training and assessing different feature and classifiers.

3. Sound Activity Detection & Syllables

The construction of a bird detector starts with the compilation of an initially very small corpus of 50-100 signals (calls/songs) of the target species, which are selected by an expert familiar with bird vocalizations. In addition, a small background corpus of non-target recordings from the study site is needed. The processing of the training corpus begins with the sound activity detector (SAD) tagging and time-stamping syllables.

The SAD is a crucial element of the workflow: It functions as an automatic annotator for boot-strapping the training data, and therefore the quality of the models are based on its efficiency. Its function is to detect the vocalizations in the initial target and background corpora and tag them with time boundaries in order to divide the initial recordings into target, non-target, inter-syllable pauses, and into segments of long pauses. In Section 5 we describe analytically why this procedure takes place.

We examined three different SAD approaches: a statistical one, a novel approach based on a Hilbert follower, and a syllabification approach. All three SADs can efficiently detect and time-stamp correctly calls. However, only the third has demonstrated success in segmenting syllables within songs and series of calls under the strong reverberation found in many natural environments. Due to the importance of this stage we report all approaches.

Statistical SAD

The signal received from the microphone is Short Time Fourier Transformed and frame t has D discrete Fourier transform (DFT) coefficients $x(f_k)$ $k=1,..D$ where, $x(f_k,)$ is the DFT coefficient of the received signal at frequency f_k. Under the null hypothesis H_0 an audio event is absent, while under the alternative, H_1, an audio event is present. The log of the probability density function of the received frame under hypothesis H_i (i=0,1) for is given by:

$$L\left(x\left(f_k\right)\right)=\log\frac{p_0\left(x\left(f_k\right)|\,\mathrm{H}_0\right)}{p_1\left(x\left(f_k\right)|\,\mathrm{H}_1\right)}$$

Evaluating the global log-likelihood ratio (GLRT) jointly for all the D DFT bins leads to

$$L\left(x\left(f_k\right)\right)=\sum_{k=1}^{D}\left(\frac{\gamma_k\xi_k}{1+\xi_k}-\log\left(1+\xi_k\right)\right)$$

where ξ_k is the *a-priori* signal-to-noise ratio (SNR) and γ_k is the *a-posteriori* SNR estimated using the Ephraim and Malah minimum mean-square error (MMSE) estimator [15]. The decision for the proposed technique is based on the same rule:

$$\text{if } \sum_{k=1}^{D}L\left(x\left(f_k,t\right)\right)>\sum_{k=1}^{D}\gamma_k \text{ accept } H_0 \text{ otherwise accept } H_1$$

In the GLRT detector we compare $L(x(f_k))$ to a threshold γ_k derived from the initial frames, which are assumed to have only noise and the threshold define the false alarm rate. This SAD is very efficient in detecting calls even in very noisy conditions. However, it often fails to resolve syllables within a song that is recorded under real field conditions. This is due to reverberation

that fills in the inter-syllable gap with energy as the song originates from a distance, and therefore, the activity detector perceives the whole song as one segment (Fig. 1).

Hilbert follower

The Hilbert follower segments the recording by following the characteristic shape of the syllable envelopes of bird vocalizations. Let x_r be the signal in the time domain holding the original recording. Let, `x = hilbert(`x_r`)` return a complex sequence called the *analytic signal* of x_r. The analytic signal $x = x_r + jx_i$ has a real part x_r which is the original data, and an imaginary part, x_i, which contains the Hilbert transform of x_r. After deriving the Hilbert transform we estimate the envelope $y=(x\dot{x})^{1/2}$ where $\dot{x}$ stands for the conjugate of x.

The envelope is compared, in our case, against a threshold of 0.03. The Hilbert follower can resolve syllables to large extent as it is able to follow the characteristic envelope of bird songs. However, it can over-segment or partially segment a song depending on the level of the reverberation, which is not known a-priori to the algorithm.

Syllabification approach [16]

The model performs a spectro-temporal analysis of the signal. The frequency band of the target species is covered by 40 partly overlapping critical band filters. The auditory model is generating a loudness spectrum every 10 ms. The sum of the components of that spectrum at a certain time constitutes the total loudness at that time (the so-called loudness pattern) [16]. It merely identifies the major minima and maxima in this pattern and it considers the intervals surrounding the maxima as the envisaged syllables. The minimum/maximum detection algorithm has two free parameters: alpha and beta. The search for a maximum is stopped as soon as the loudness is below alpha times the last maximum that was detected and the search for a minimum is stopped as soon as the loudness is larger than a factor 1+beta times the last minimum that was detected. Though simple, this technique manages to track efficiently syllables within a song having distortions due to reflections (Fig. 1).

For the task of assessing the segmentation accuracy we used 100 manually tagged syllables of the Eurasian Chaffinch *Fringilla coelebs* Linnaeus, 1758 as it is a "songbird" (oscine passerine), whereas the Common Kingfisher does not "sing". The syllabification outperforms all other techniques being accurate 92%, followed by the Hilbert Follower by 84% and the statistical SAD being 41% accurate.

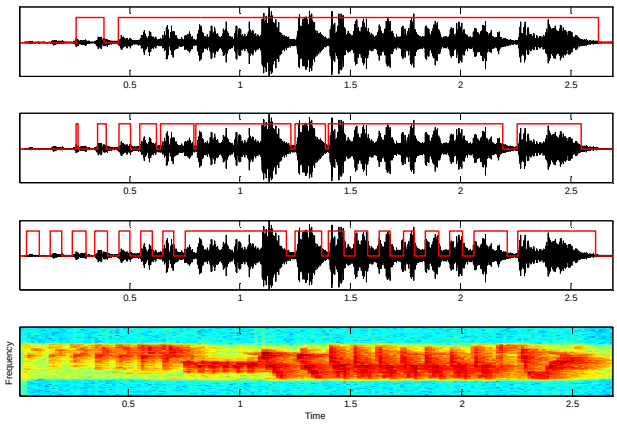

Fig. 1: Chaffinch song, high reverberation. a) Statistical detector, b) Hilbert follower, c) syllabification approach, d) spectrogram

4. Signal Preprocessing & Features

In order to build an efficient bird detector one must discard any information not useful to the detection task. Due to the fact that different species vocalize in different spectral bandwidths ranging from frequencies as low as 100 Hz and reaching over 16 KHz it is not optimal to follow the same pre-processing procedure in all species in contrast to speech signals. Therefore, we initially downsample as low as the highest frequency of the target signal allows (i.e. 32 KHz in the case of the kingfisher). Subsequently we apply band-pass filtering retaining only the frequency range of the target signal (4.5-12.5 KHz are kept for the kingfisher). 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 [15] that deals with the remaining noise inside the band-pass limits. Finally, the audio signals are transformed to feature vectors for training the classifiers. The frame size is 10 ms with frame shift of 5 ms. We employed the feature set named PLP_E_D_A_Z which consists of the first 16 Perceptual Linear Prediction, log energy their respective delta and double delta, while cepstrum mean subtraction is applied (i.e., 51 dimensions).

Due to space limitations of this paper we do not provide a thorough search on features as this is the subject of a forthcoming project.

5. Pattern Recognition Techniques

The bird detection task in very long recordings has to detect those sections of interest, where a specific species vocalizes. It also has to determine the exact boundaries of the audio parts uttered by the target bird species. We apply the framework of hidden Markov models (HMMs) as it is able to deal with patterns that vary both in time and frequency and is able to incorporate a lattice model into the decoding procedure that we intend to use in order to incorporate context information (Fig. 2). We employed the hidden Markov model toolkit (HTK) which has been deeply explored by the speech processing community mainly for speech and speaker recognition [17].

Generally, there are two ways of initializing HMMs models. The first, the so called 'flat start' (`Hcompv` function in HTK) is based on assigning a single tag to the whole recording (e.g. all chaffinch recording have the tag: *Chaffinch*). It is left to the HMM to realign data as a part of its training procedure on pauses, noise and detection events. The second way is based on the availability of data for which the boundaries of the events have also been marked and training uses the fully labeled bootstrap data. The second way is very accurate and is known to result into better detection accuracies. However, as yet recordings in bird databases do not have this level of information tagging. This information can be provided by hand labeling the training data with time boundaries of acoustic events, which is a very laborious procedure. Here we present a novel approach to use the time boundaries that are naturally produced as the sound activity detector is applied on the recordings and, subsequently, use these boundaries for bootstrapping the training data (`Hinit` function in HTK). In the case that the initial available target calls/songs provided by the experts are low (which is the normal case), bootstrapping is the only alternative as the target model is otherwise poorly modeled.

The models that were trained are:

a) a model of the target species (i.e., Common Kingfisher),

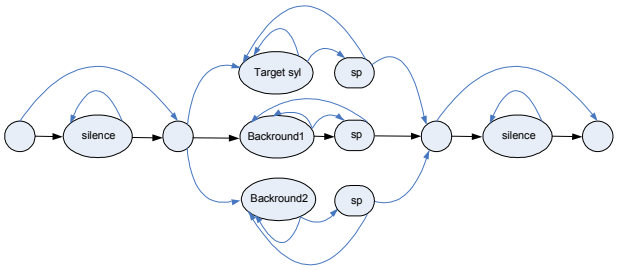

Fig. 2. Lattice network associated with the detection framework: the illustrated modeling scheme includes target syllables, different background models and short-pauses (sp).

b) multiple background models for different aspects of background activity, namely a cohort model trained on the syllables of the class of other bird taxa that resemble the target species and a short-pause model tied to a silence model as well as a long-pause model modeling any other non-target activity.

6. Experiments & Analysis of Results

The confirmation of the presence of a certain bird species with the help of ARUs involves the investigation of audio data by experts. However, for human observers it is neither feasible to hear nor visually inspect (the spectrograms of) several hundred thousand audio files, reaching terabytes of data.

Our approach is to first use the small target data provided by experts (48 recordings containing 110 single Kingfisher calls) to get a large number of detections in the real field recordings. Subsequently, these findings are also tagged by the experts and the larger set of confirmed ID tags form the enhanced training set used in the final run. We assessed the detectors on a large corpus and we report results on per file and on per call basis. Result on per file basis answers the question: If we have 10,000 audio files, to what extent is the search space for human observers reduced by the detector? In the case of the kingfisher the reduction was 98% and 92,96% of the files found verifiably referred to the target species (Table 1). On *per call* basis the precision was 98,4%, whereas the recall rate within the detected files was 70,21%. Results on per detection basis report on the number of detected calls and exact location inside each file. These results can be used to draw statistics on species activity which will be reported elsewhere. Detections below a low energy threshold correspond to very faint calls and are discarded as they have lost a large part of their high frequency spectrum due to frequency selective attenuation.

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

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

where N_{hits} +N_{FA} = Number of detected events and N_{actual} is the number of ground truth. The following table depicts P, R results calculated on per file and on per detected segment respectively.

	P (%)		***R (%)***	
Species	***file***	***call***	***file***	***call***
Kingfisher	*97,06*	*98,4*	*92,96*	*70,21**

Table 1. Precision and Recall scores on per file and per detection basis for the Common Kingfisher. *Note: the recall rate on per file and per call rate refers to the total number of kingfisher calls within the detected recordings .

7. Conclusions

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 bird detectors on the basis of statistical models. Towards this aim we presented some tools and an audio processing framework for detecting and tagging bird vocalizations in massive real-field continuous streams of audio data from different habitats. With the aim of promoting the development of proper conservation measures for threatened wildlife we will continue to present tools for computational ecology by giving emphasis to real-field applications.

8. Acknowledgements

This work was supported by the LIFE+ Program AMIBIO, "Automatic monitoring of biodiversity", NAT/GR/000539.

9. References

[1] B. Huntley, R. Green, Y. Collingham, and S. G. Willis, "A climatic atlas of European breeding birds," Lynx Edicions, 521 pages, 2008.

[2] IUCN, "IUCN Red List of threatened species, version 2011.2," online at http://www.iucnredlist.org, last accessed on 24/3/12.

[3] EC, "Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora," online at http://ec.europa.eu/environment/nature/legislation/habitatsdirective /index_en.htm, last accessed on 24 March 2012.

[4] EC, "Directive 2009/147/EC of the European Parliament and of the Council of 30 November 2009 on the conservation of wild birds. Online at http://ec.europa.eu/environment/nature/legislation/birdsdirective/index_en.htm last accessed on 24 March 2012.

[5] R. T. Buxton, and I. L. Jones, "Measuring nocturnal seabird activity and status using acoustic recording devices: applications for island restoration," J. of Field Ornith. 83 (1), pp. 47-60, 2012.

[6] K. Riede, "Monitoring biodiversity: analysis of Amazonian rainforest sounds," Ambio 22, pp. 546-548, 1993.

[7] P. J. Clemins, et al., "Generalized perceptual features for vocalization analysis across multiple species," in Proceedings IEEE ICASSP, pp. 253-256, 2006.

[8] P. Somervuo, A. Harma, and S. Fagerlund, "Parametric representations of bird sounds for automatic species recognition," IEEE Transactions on ASSP, 14(6), 2252–2263, 2006.

[9] A. Selin, J. Turunen, and J. T. Tanttu, "Wavelets in Recognition of Bird Sounds," EURASIP Journal on Advances in Signal Processing, Article ID 51806, 2007.

[10] V. Trifa, et al., "Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models," Journal of the Acoustical Society of America, 123(4), pp. 2424-2431, 2008.

[11] C. Kwan, et al., "Bird classification algorithms: Theory and experimental results," in Proceedings IEEE ICASSP, pp. 289-292, Montreal, Canada, 2004.

[12] R. Bardeli, "Similarity Search in Animal Sound Databases," IEEE Transactions on Multimedia 11(1), pp. 68-76, 2009.

[13] Y. Ren, et al., "A Framework for Bioacoustic Vocalization Analysis Using HMMs," Algorithms 2(4), pp. 1410-1428, 2009.

[14] S. Fagerlund, "Bird species recognition using support vector machines," EURASIP Journal on Applied Signal Processing, Article ID 38637, 2007.

[15] Y. Ephraim, and D. Malah, "Speech enhancement using a MMSE short-time spectral amplitude estimator," IEEE Transactions on ASSP, 32(6), pp. 1109-1121, 1984.

[16] T. De Mulder, J.P. Martens, M. Lesaffre, M. Leman, B. De Baets, H. De Meyer, "Recent Improvements of an Auditory Model Based Front-end for the Transcription of Vocal Queries", in Proceedings IEEE ICASSP, IV, pp. 257-260, 2004.

[17] Hidden Markov model Toolkit, available at http://htk.eng.cam.ac.uk/.