Applied Acoustics 98 (2015) 34–42

Contents lists available at ScienceDirect

Applied Acoustics

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

Bird acoustic activity detection based on morphological filtering of the spectrogram

Allan G. de Oliveira [a,b,c], Thiago M. Ventura [a,b,c], Todor D. Ganchev [a,d,*], Josiel M. de Figueiredo [a,b,c], Olaf Jahn [a,e], Marinez I. Marques [a,f], Karl-L. Schuchmann [a,e,f]

[a] *National Institute for Science and Technology in Wetlands (INAU), Science Without Borders Program, Federal University of Mato Grosso (UFMT), Cuiabá, MT, Brazil*
[b] *Institute of Computing, Federal University of Mato Grosso, Cuiabá, MT, Brazil*
[c] *Institute of Physics, Federal University of Mato Grosso, Cuiabá, MT, Brazil*
[d] *Department of Electronics, Technical University of Varna, Varna, Bulgaria*
[e] *Zoological Research Museum A. Koenig (ZFMK), Bonn, Germany*
[f] *Institute of Biosciences, Federal University of Mato Grosso, Cuiabá, MT, Brazil*

ARTICLE INFO

Article history:
Received 19 November 2014
Received in revised form 15 April 2015
Accepted 21 April 2015
Available online 16 May 2015

Keywords:
Computational bioacoustics
Acoustic activity detection
Feature extraction
Bird sound recognition
Southern Lapwing *Vanellus chilensis*

ABSTRACT

Audio event recognition methods based on the Hidden Markov Model / Gaussian Mixture Model (HMM/GMM) often depend on a large number of mixture components or multi-stage models that require significant computational and memory resources during their operation. A widely used approach for coping with complexity is employing an acoustic activity detector, which selects for further processing only those portions of the audio that are considered promising. As a result, the audio feature extraction and the subsequent pattern recognition stages will process only a subset of the original audio stream, helping to reduce the misclassification rates while lowering the computational demands. In the present work we propose a method for bird acoustic activity detection, based on morphological filtering of the spectrogram seen as an image. The practical significance of the proposed method is validated on the automated acoustic recognition of Southern Lapwing *Vanellus chilensis*, a common Neotropical bird species. Compared with other methods of acoustic activity detection it demonstrates advantageous performance.

1. Introduction

Recent technology developments already offer considerable support to biodiversity monitoring. Automated audio recording devices, for instance, are widely used for scheduled and continuous collection of soundscapes. Several international projects, such as ARBIMON[1], AMIBIO[2], and INAU 3.14[3], have gathered audio data in the range of hundreds of Terabytes. As the number of recordings increases by the minute, handling such a huge amount of data imposes great challenges in terms of storage capacity and data management tools.

Nowadays, the content extraction tools are most often based on statistical machine learning methods, such as the HMM/GMM likelihood estimators. However these likelihood estimators often depend on a large number of mixture components or on multi-stage models that require significant computational and memory resources. A widely used approach for coping with complexity is employing an acoustic activity detector, which selects for further processing only those portions of the audio that are considered promising. As a result, the audio feature extraction and pattern recognition stages that follow only process a subset of the original audio stream, helping to reduce the misclassification rates and computational demands. In Section 2 we briefly discuss the common technological framework of acoustic species recognition and the role of the acoustic activity detector.

Various energy-based methods for acoustic activity detection have been studied in search of a simple solution for eliminating silent portions of the signal. Their operation is quite simple – basically the short-term energy of the signal is computed with a sliding short-time window function and afterward it is compared to a threshold. Typically, the threshold is adjusted to select for further processing only the audio frames with high energy. These methods

* Corresponding author at: Department of Electronics, Technical University of Varna, Varna, Bulgaria. Tel.: +359 888 096974.

E-mail addresses: allan@ic.ufmt.br (A.G. de Oliveira), thiago@ic.ufmt.br (T.M. Ventura), tganchev@tu-varna.bg (T.D. Ganchev), josiel@ic.ufmt.br (J.M. de Figueiredo), o.jahn@zfmk.de (O. Jahn), marinez513@gmail.com (M.I. Marques), klschuchmann@googlemail.com (K.-L. Schuchmann).

[1] ARBIMON project: http://arbimon.com.

[2] AmiBio project: www.amibio-project.eu/en.

[3] INAU project 3.14: http://www.inau.org.br/laboratorios/?LaboratorioCod=12&ProjetoCod=94&ProjetoCod=94, CO.BRA: http://cobra.ic.ufmt.br/.

http://dx.doi.org/10.1016/j.apacoust.2015.04.014
0003-682X/© 2015 Elsevier Ltd. All rights reserved.

do not depend on prior knowledge about the signal and are quite easy to implement, which is the main reason for their widespread use in environmental sound recognition [1] and other application domains, such as bioacoustics as well as signal, speech, and audio processing. More sophisticated methods for acoustic activity detection are based on modeling the distribution of short-term energy with a Gaussian Mixture Model (GMM) [2–4]. These methods rely on a bi-Gaussian Mixture Model, where the first Gaussian component is fitted to the distribution of the low-energy frames and the second is fitted to the distribution of the high-energy frames. In most cases the decision threshold is selected as a trade-off at the crossing point of the two Gaussian functions. The bi-Gaussian method requires training datasets for adjusting the parameters of the GMM, but offers a constructive way of selecting the threshold, and makes it convenient for practical use. An overview and comparative evaluation of various energy-based acoustic activity detection methods is available in Sahidullah and Saha [5]. A common drawback of all energy-based methods is their limited capacity to distinguish between informative sound events and background acoustic activity, which is mainly due to the limited descriptive power of the energy feature. For instance, the acoustic activity detector may miss target sound events with low amplitude and select a certain number of audio segments containing bursts of background noise. This weakness can be compensated by combining the energy feature with other audio descriptors, at the cost of higher computational demands and loss of simplicity.

A HMM-based recognizer for automated detection of American Robin *Turdus migratorius* and Common Kingfisher *Alcedo atthis* is presented in Potamitis et al. [6]. A system for semiautomatic recognition of the nocturnal activity of Eurasian Bittern *Botaurus stellaris* was studied by Frommolt and Tauchert [7]. Its functionality included detecting the direction of the sound source, estimating the number of calling animals, identification of the bird species, and determination of the timestamps. A comparison of various machine learning approaches for the classification of bird and amphibian calls was provided by Acevedo et al. [8]; whereas a comprehensive review of methods is presented in Stowell and Plumbley [9].

In the past few years the research community shifted attention to recordings collected in the wild [10,11,15,34] and to large-scale bird classification tasks [12,13]. Research on acoustic recognition of bird species was fostered by technology evaluation challenges [14–16].

Recently, image processing techniques operating on the spectrogram were employed on the task of automated acoustic species recognition [17–19]. Bardeli [17] used Structure Tensors in order to identify regions of the spectrum, which are suitable for feature extraction. Briggs et al. [18] and Potamitis [19] focused on multi-label classification for identifying multiple species co-occurring in the same audio recording. These approaches treat spectrograms as an image and find regions of interest that are suitable for extracting statistical features.

Other methods for selecting only the crucial portions of audio recordings are based on the structure of bird vocalizations. To that end the elements of bird vocalizations are assigned to four hierarchical levels: notes, syllables, phrases, and song. Härmä [20] proposed an algorithm that extracts syllables from continuous bird songs. This approach was further studied by Lee et al. [21] and Chou et al. [22]. Lee et al. [23] used the same idea, but the birdsong syllables were manually labeled.

Based on these recent studies we propose a new method for acoustic activity detection that uses morphological filtering of the spectrogram treated as an image. The morphological filtering aims to remove high-amplitude events of compact time–frequency distribution, which are not likely to be part of the target signal. In this way we obtain a cleaned spectrogram, where the promising regions are highlighted. Next we compute the histogram, using the sums of spectral magnitude values of each frame. Finally we determine the decision threshold for selecting promising portions of the audio. The proposed method is presented in Section 3.

For the comparative assessment of several acoustic activity detection methods we used a common experimental setup (Section 4). The experimental results demonstrate that the performance of the present approach is advantageous when compared with two other energy-based acoustic activity detection methods (Section 5). Further discussion on the limitations of morphological filtering is provided in Section 6. Finally, Section 7 concludes this work.

2. Technological framework

The technological framework of the Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) methods for automated acoustic bird recognition implements the following steps (Fig. 1):

(1) We use an expert-annotated training dataset to build a species-specific model for the target species (*Vanellus chilensis*). (2) Likewise, we use a representative but target-species-free sample of environment sounds to build a general acoustic background model. These models are most often implemented as GMM or HMM and aim to accurately represent the acoustic variability of the target species and the time-varying acoustic conditions of the environment. For that reason it is quite common that the GMM/HMM models are built from large amounts of audio recordings and either have large numbers of mixture components (GMM) or rely on a multi-state modeling (HMM). On the acoustic bird species classification task, for instance, Graciarena et al. [24] used 1024-component species-specific GMM and a 1024-component acoustic background model.

The use of large numbers of mixture components and multi-stage models increases the computational demands, memory requirements, and energy consumption during operation. Therefore, developers often trade off complexity against accuracy, by reducing the number of states, mixture components, and audio feature parameters. Another way to cope with complexity is to optimize the audio feature extraction stage, which typically claims a significant portion of the computational and memory resources.

Acoustic activity detectors based on the short-term energy are known to be a simple and economical method for implementing the strategy shown in Fig. 1. An overview of various energy-based acoustic activity detection methods is provided in Sahidullah and Saha [5]. Section 3 presents details on the proposed method for acoustic activity detection and Section 5 offers a comparative analysis of a traditional GMM-based energy detector [5], the syllable-based approach of Härmä [20], and our approach.

3. Proposed method

In the present work we seek to reduce the computational demands of bird recognition by pre-scanning the audio stream with an acoustic activity detector (Fig. 2). Promising audio segments are labeled for further processing and parameterization is carried out only for promising high-energy portions of the signal. In the pattern recognition stage the recognizer decides whether a specific audio segment corresponds to a target vocalization or not. If the computation time of the present detector is less than that of a GMM/HMM recognizer operating on continuous audio recordings, we would have accomplished our aim to significantly reduce the overall computational demands of the system. Furthermore, we assume that by excluding portions of audio with predominant environmental noise and by selecting only the most representative audio segments, we can increase the accuracy of our species-specific recognizer. Consequently we focus our

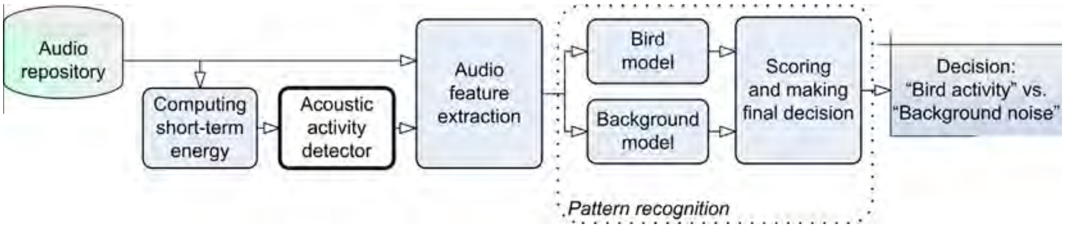
Audio repository
Computing short-term energy
Acoustic activity detector
Audio feature extraction
Bird model
Background model
Scoring and making final decision
Decision: "Bird activity" vs. "Background noise"
Pattern recognition

Fig. 1. Overall block diagram of statistical methods for automated acoustic recognition of bird species.

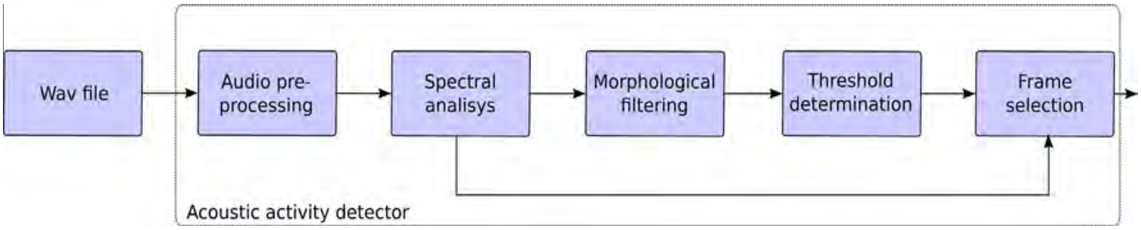

Fig. 2. Audio-processing steps of the proposed method for acoustic activity detection.

attention on the audio segments with high signal-to-noise ratio (SNR) and process only the high-energy portions of the signal. These signals have amplitude above −30 dB and correspond to sounds emitted by birds within a range of few meters to several dozen meters from the microphone. These sound levels correspond to signal strengths that ornithologists can identify with certainty when using traditional audiovisual survey methods [25,26] and thus facilitate the interpretation of the observed acoustic activity patterns.

The individual audio processing steps are discussed in the following subsections.

3.1. Audio pre-processing

Two main actions are performed during the audio pre-processing: resampling and high-pass filtering. The audio is downsampled to 24 kHz, which preserves the frequency range where most of the energy of bird vocalizations is located while saving computational and memory resources. In addition, a high-pass filter, with a cut-off frequency 1 kHz, is applied to reduce the low-frequency noise caused by wind and vibrations of mechanical origin.

3.2. Spectral analysis

Long recordings are divided into non-overlapping pieces of one minute and the spectrogram is generated for each portion of audio. The spectrogram is computed with a Hamming window of 480 samples, sliding with an overlap of 360 samples and a length of the discrete Fourier transform (DFT) of 512 samples (Fig. 3). For an audio signal sampled at 24 kHz, the window size of 480 samples corresponds to 20 ms and the overlap of 360 samples lets the window shift in 5-ms steps. The window shift determines the time resolution of the frame selection method, i.e. affects the number of pixels in a row of the spectrogram. By contrast, the choice of frequency resolution, that is the ratio between sampling rate in Hz and the DFT size, bounds the resolution of the individual pixels. Likewise, the number of pixels in one column of the spectrogram depends on the DFT size. Therefore the spectrogram resolution and the image size in terms of number of pixels depend on the recording duration and on the spectral analysis settings.

The window size (20 ms) and shift (5 ms) settings were found to provide a good trade-off between frequency and temporal resolution in the audio spectrogram. Other researchers used the same window size and shift step in related work on acoustic bird recognition [27,28]. Somervuo et al. [29] reduced the computational demand with a window size of 20 ms and shift step of 10 ms.

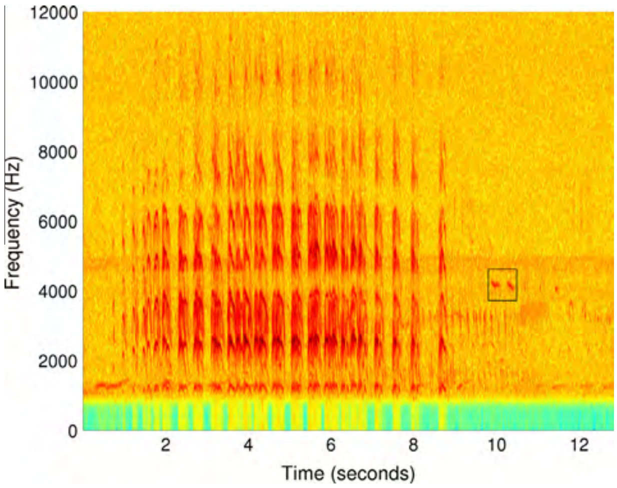

Fig. 3. Spectrogram of a 13-s recording, obtained after the audio pre-processing phase. The segment contains *Vanellus chilensis* calls and faint calls of other species. Two of the competing signals to be removed are marked with a rectangular box (cf. Fig. 4).

In the following we denote the spectrogram by $S(k, l)$, where k is the index of the kth frequency component and l is the frame number index.

3.3. Morphological opening of the spectrogram

Like Cadore et al. [30] and Potamitis [19] we apply the morphological opening operator on the spectrogram $S(k, l)$, which is considered as an image. The morphological opening [31] is defined as the operation erosion followed by the operation dilation using the same structuring element. Erosion reduces the bright regions by removing bright pixels from the object boundaries and thus enlarges the dark regions. By contrast, dilation enlarges bright regions by adding bright pixels to the object boundaries and thus reduces the dark regions. Therefore, erosion removes noise whereas dilation produces an amplification of shapes and fills gaps.

Consequently remaining objects in the image are smoothed versions of the original objects [30].

The result of applying the aforementioned erosion and dilation operators is the processed spectrogram $S_{ed}(k,l)$ (Fig. 4). Following Evans et al. [32] we use a rectangular shape for the structuring element, which we defined as a mask of 40-by-30 pixels. Therefore the morphological filtering eliminates acoustic events with high amplitude and compact localization in time and frequency. However events with a longer duration remain intact, such as *V. chilensis* vocalizations. This leads to lower numbers of audio events to be processed in the GMM/HMM-based pattern recognition stage and reduces the probability of false alarms due to short energy bursts.

3.4. Threshold estimation

The frame selection step begins with the estimation of the decision threshold. For that purpose, we first calculate the sum of magnitudes W_l of all frequency components in the morphologically opened spectrogram $S_{ed}(k,l)$. For each frame we compute

$$W_l = \sum_{k=1}^{K} S_{ed}(k,l), \quad l = 1,2,\ldots,T, \tag{1}$$

and subsequently generate the histogram of the W_l values. Here $K = 256$ is half of the DFT size, and T is the total number of frames in the specific recording. The W_l values computed for the signal in Fig. 4 are shown in Fig. 5 and the corresponding histogram in Fig. 6.

The threshold θ is estimated based on the histogram of W_l (Fig. 7). Firstly we set the selection criterion C, which specifies how strict the frame selection process will be. A lower value of C means a more restrictive criterion is applied, and in the extreme case when $C = 0$ no frames are selected. By contrast, $C = 1$ means that all frames are selected. In the following, we set $C = 0.3$ as this value provides selection of nearly all frames with amplitude over −30 dB. Secondly, the bin B with the highest number of counts X_{max} in the histogram is identified. Next, the right-hand bin $b > B$ with count number x, is compared against X_{max}. If the ratio x/X_{max} is not smaller than C the next bin on the right-hand is evaluated. The process is terminated when the condition $x/X_{max} < C$ is fulfilled. Finally, the threshold value θ is set equal to the center of the bin b for which the condition $x/X_{max} < C$ was fulfilled.

The highest number of frames in the histogram is in the first bin and thus all other bins have to be inspected as well (Fig. 6). With $C = 0.3$ the second bin is selected because its value (85) is less than 30% of the highest bin value (1500). Therefore, the threshold θ is set to 30, which is the center of the second bin.

3.5. Frame selection

In the final step we select for further processing the frames $S_2(k,l)$ based on the threshold θ

$$S_2(k,l) = \begin{cases} S(k,l), & \text{if } W_l \geq \theta \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

According to (1) and (5), the segmentation is carried out based on the value θ, i.e. on the morphologically filtered spectrogram $S_{ed}(k,l)$ and the threshold computed from the W_l values (Section 3.4). However as defined in (5), the frames for further processing are taken directly from the pre-processed spectrogram $S(k,l)$ to avoid the loss of resolution and the nonlinear distortions caused by the morphological filtering. The outcome of the frame selection process applied to the spectrogram in Fig. 3 is shown in Fig. 8.

The computational demand of the proposed algorithm is proportional to the number of pixels in the spectrogram, due to the morphological opening operator which is applied to each pixel. The other calculations are much less demanding – the histogram generation depends mostly on the window shift step (here 5 ms), whereas the threshold estimation loop (Fig. 7) depends on the number of bins in the spectrogram (typically 10) and the data distribution, but not on the number of pixels in the image.

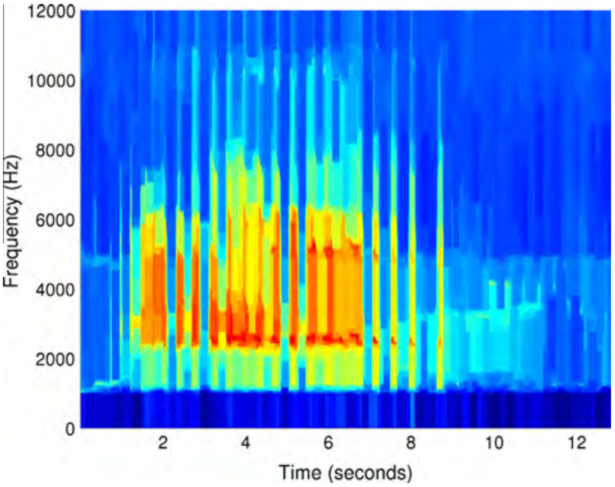

Fig. 4. Spectrogram after applying the operator morphological opening on the spectrogram in Fig. 3. Nearly all non-target sound events with compact time–frequency occurrence were removed.

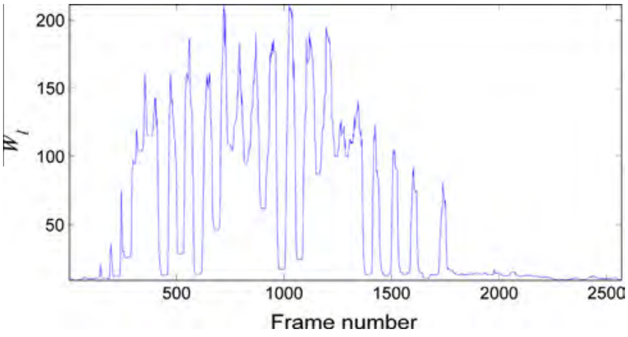

Fig. 5. The sum of magnitude values per frame W_l for the spectrogram in Fig. 4.

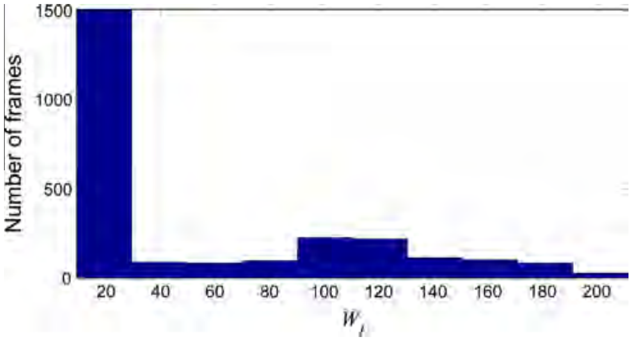

Fig. 6. Histogram of the sum of magnitudes for the spectrogram shown in Fig. 4.

4. Experimental setup

In the following subsections we present the datasets employed in the experimental validation of the proposed method (subsection 4.1) and shortly outline two widely used methods for acoustic

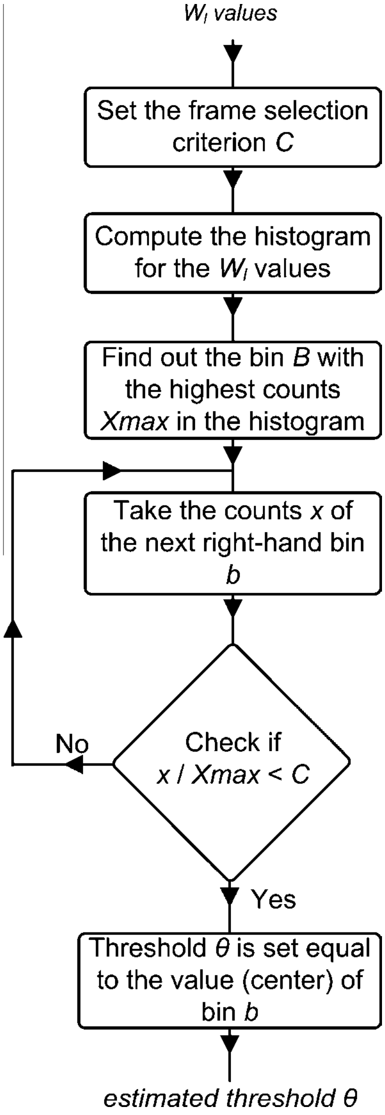
W_l values
Set the frame selection criterion C
Compute the histogram for the W_l values
Find out the bin B with the highest counts $Xmax$ in the histogram
Take the counts x of the next right-hand bin b
Check if $x / Xmax < C$
No
Yes
Threshold θ is set equal to the value (center) of bin b
estimated threshold θ

Fig. 7. The threshold estimation algorithm.

activity detection, namely Sahidulla and Saha [5] and Härmä [20]. Next, we describe the common experimental protocol and performance metrics (subsection 4.4) used in the comparison between the proposed method (Section 3) and the aforementioned approaches [5,20]. The practical significance of these three methods is assessed in terms of recognition performance and speed with the help of a species-specific HMM-based recognizer of *V. chilensis* vocalizations.

4.1. The *V. chilensis lampronotus* datasets[4]

4.1.1. Study area

The study was carried out in the northern Pantanal region, municipality of Poconé, Mato Grosso, Brazil. Between July 2012 and October 2014 we used Song Meter SM2+ recorders (Wildlife Acoustics[5]) for soundscape collection in 24/7 mode, in total ca. 90 TB of audio recordings with a file duration of 14–30 min, 48-kHz sampling rate, and 16-bit resolution [34]. We completed one annual cycle of recordings each at Fazenda Pouso Alegre (−16.50303 S, −56.74533 W; 115–126 m a.s.l.; c. 110 km^2) and SESC Pantanal Private Natural Heritage Reserve (−16.49879 S, −56.41309 W; 119–131 m a.s.l.; 878.7 km^2). These activities were carried out by the Computational Bioacoustics Research Unit[6] within the scope of the INAU 3.14 Project "Monitoring Bioindicators and Migratory Birds in the Pantanal" of the National Institute for Science and Technology in Wetlands[7], aiming at the promotion of Applied Acoustomics as a tool for bio-sustainability assessment [33].

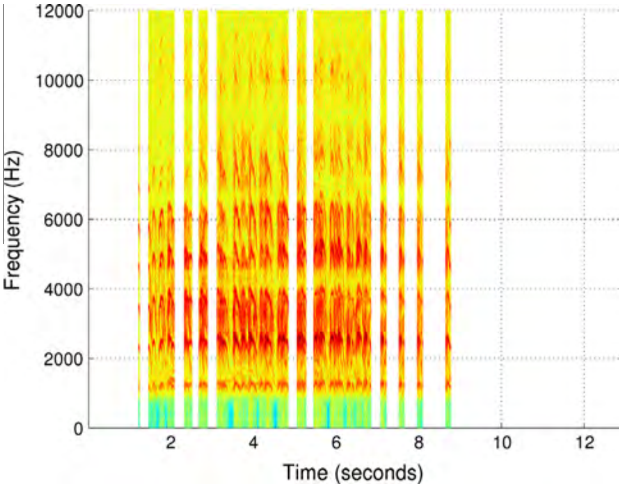

Fig. 8. Selected frames for the spectrogram shown in Fig. 3.

The Southern Lapwing *V. chilensis* is a common and widespread Neotropical waterbird. The ground-dwelling species inhabits open areas and muddy lake shores. The resident subspecies in our central Brazilian study area is *V. c. lampronotus (Wagler, 1827).*

All datasets used here for technology development and evaluation were extracted from the aforementioned Pantanal soundscape collection; cf. Ganchev et al. [34] for further details on the dataset.

4.1.2. Acoustic background dataset

For the creation of a balanced acoustic model of the environmental noise we selected approximately 27 h of representative Pantanal audio recordings: (i) fifty-four 14-min Fazenda Pouso Alegre soundscape recordings with a total duration of over 12 h and (ii) thirty-two 30-min SESC Pantanal soundscape recordings with a total duration of 15 h. These recordings are regarded as *V. chilensis* free, although on occasions some vocalizing lapwings may have flown over the SESC recording station.

4.1.3. *V. chilensis* training dataset

The training dataset consisted of 93 recordings, in total containing ~45 min of sounds from the target species. Among these were 38 recordings that are virtually clean of competing interference (~18 min) and 52 recordings (~24 min) that also contain non-target sounds overlapping with the target signals. Selective filtering was carried out manually to discard the strongest non-target calls from the training dataset. Finally, three recordings of the *V. chilensis* training dataset, representing choruses of lapwing flocks (~2 min), were not used in this work.

4.1.4. *V. chilensis* validation dataset

The validation dataset consisted of fourteen soundscape recordings with duration of 14 min, which were not processed or edited

[4] According to the principles of reproducible research, we will provide all datasets described here through the website of the Computational Bioacoustics Research Unit (CO.BRA; http://cobra.ic.ufmt.br/) of the Federal University of Mato Grosso (UFMT).

[5] Wildlife Acoustics: www.wildlifeacoustics.com.

[6] CO.BRA: http://cobra.ic.ufmt.br/.

[7] INAU: www.inau.org.br.

manually, and thus contain competing sounds from multiple species and certain interferences of abiotic origin. Each recording was time-stamped and tagged on the level of *V. chilensis* call series by a bird-sound expert [34]. In brief, the annotation procedure can be summarized as follows:

(a) Only for the purpose of tagging, we made use of a graphic equalizer to reduce competing noises such as wind and insects by −48 dB in the frequency ranges [0, 630] Hz and [12.5, 24] kHz. However, the acoustic activity detector works directly on the original audio.
(b) The filtered recordings were screened using a headset with integrated manual volume control but without additional software amplification, and simultaneously through visual inspection of the spectrograms in Adobe Audition.
(c) The bird-sound expert tagged the start and end times of each acoustic event; however, the number of calls per vocalization event was not counted.
(d) The following rules were applied to separate call events: (i) a pause of at least one second between target signals was used to separate call events (isolated single calls or call series) of similar sound pressure levels; (ii) abrupt changes in sound pressure levels were used to separate different call series even if there was no pause between the signals; (iii) however, call sequences that varied greatly in sound pressure levels over time were not divided into different call series when the calling birds seemed continuously move (fly) around the recording stations.
(e) In addition, we noted the compound dB-values of the loudest call of each vocalization event; that is, the volume of the target signal plus the volume of the remaining background noise.
(f) The manual annotation of the timestamps was repeated twice. Depending of the complexity of the soundscape, the time effort per 14-min sound file was between about 4 and 16 h.
(g) Some audio signals could not be identified with certainty as target or non-target, either because they were too faint or because background noise levels were too high. Such call series were time-stamped and excluded from further analysis.

The validation dataset was divided into two non-overlapping parts:

(1) A subset of four recordings, containing 80 confirmed *V. chilensis* call series. This subset is referred to as dataset VL01 and was used for the purpose of technology optimization and tuning the adjustable parameters of the *V. chilensis* acoustic activity detector.
(2) A subset of ten recordings, containing 337 confirmed *V. chilensis* call series. This subset is referred to as dataset VL02 and was used in the evaluation of the performance of the acoustic activity detectors and the species-specific *V. chilensis* recognizer.

In fact, evaluation dataset VL02 contains a total of 474 *V. chilensis* tags, split in two groups. The first group of 337 tags was confirmed by a bird-sound expert as originating from the target species *V. chilensis.* The second group of 137 tags, heard and/or seen in the spectrogram, allegedly originated from the target species too, but the sounds could not be identified with certainty, usually due to a high noise floor or the great distance of the calling bird from the microphone. To the latter tags we refer to as *not sure*. Hence, the 137 events of the second group (*not sure*) were excluded from further analysis.

4.2. Syllable-based acoustic activity detection

The method proposed by Härmä [20] segments the acoustic signal as a set of N syllables. The syllables are found from the value of the signal amplitude in the spectrogram. The result is a set of brief frequency and amplitude modulated sinusoidal pulses.

In brief, Härmä [20] proposed the following algorithm:

1. Compute the spectrogram of the sound using short-time Fourier transform (STFT). The spectrogram is a matrix $S(f, t)$ where f represents frequency index and t is the frame index.
2. Repeat steps 3–7 for $n = 0, 1, \ldots, N - 1$.
3. Find f_n and t_n such that $|S(f_n, t_n)|$ is the maximum value in the spectrogram. This position represents the maximum amplitude position of nth sinusoidal syllable.
4. Store frequency parameter $\omega_n(0) = f_n$ and amplitude $a_n(0) = 20\ \log_n 10|S(f_n, t_n)|$ [dB].
5. Starting from $|S(f_n, t_n)|$, trace the maximum peak of $S(f, t)$ for $t > t_0$ and for $t < t_0$ until $a_n(t - t_0) < a_n(0) - T$ dB, were the stopping criteria T is typically 30 dB. It will determine how the sinusoidal syllable starts and ends at times t_s and t_e, respectively around the amplitude maximum t_0.
6. Store obtained frequency and amplitude trajectories corresponding to the nth syllable in functions $\omega_n(\tau)$ and $a_n(\tau)$, where $\tau = t_0 - t_s, \ldots, t_0 + t_e$.
7. Set $S(f, [t_s, t_{s+1}, \ldots, t_e]) = 0$ to delete the area of nth syllable.

The value of T is very important because it determines the search stop criterion for new syllables and also the size of each syllable.

4.3. Energy-based acoustic activity detection with GMM

A widely used energy-based acoustic activity detector implemented with Gaussian Mixture Models (GMM) is described in Sahidullah and Saha [5]. The most frequent implementation uses a two-component GMM in which the first component is fitted to the distribution of the low-energy frames and the second one is fitted to the distribution of the high-energy frames. Subsequently the most promising frames are selected using a threshold learned from the distribution of the data. The threshold is usually selected at the crossing point of the two Gaussian functions.

In the present work we set the threshold depending on the means of the two Gaussian components. The threshold θ_{trn} for the training dataset was calculated by (3) and θ_{tst} for the test dataset by (4):

$$\theta_{trn} = \frac{\mu_1 + \mu_2}{2} + |\mu_1 - \mu_2| \times 0.2, \tag{3}$$

$$\theta_{tst} = \max(\mu_1, \mu_2). \tag{4}$$

During the operation of the recognizer, θ_{tst} is selected with a higher value than θ_{trn} because we wish to make a decision based on frames with signal amplitude over −30 dB.

4.4. Experimental protocol and performance metrics

In all experiments, we followed a common experimental protocol that makes use of the datasets described in Section 4.1. We used the training dataset for adjusting the thresholds 5, 6 and 4 of the acoustic activity detectors. The three acoustic activity detectors are evaluated by means of recognition accuracy after the HMM-based species-specific recognizer.

The training dataset was reused for training a single-state HMM model with 48 mixture components for the species-specific *V. chilensis* recognizer. The acoustic background model, created as a single-state HMM model with the same number of mixture components, was built from the acoustic background dataset. All HMMs were trained by 100 iterations of the Baum-Welch algorithm.

The adjustable parameters of the *V. chilensis* recognizers were tuned by using the four audio recordings, described as dataset VL01. Following this we evaluated the recognizer performance with the test dataset VL02. The results were evaluated in terms of two performance metrics – *correct* and *accuracy*, measured in percentages:

$$\text{Correct} = \frac{H}{N}100, [\%], \tag{5}$$

$$\text{Accuracy} = \frac{H-I}{N}100, [\%], \tag{6}$$

where H is the number of hits, I denotes the number of insertions, and N is the total number of target events according to the annotations of the test dataset VL02. Here, a hit indicates that the *V. chilensis* recognizer correctly detected a target sound event (either single call or call series). Insertions are false positives, i.e. the *V. chilensis* recognizer labeled a sound event as *V. chilensis* call or call series although this sound event has a different origin.

For the purpose of technology evaluation we made use of two subsets of labeled data. The first consists of vocalizations with maximum amplitude in the range [0, −20 dB] (Table 1). In this case, there are $N = 44$ target events which are considered for counting the number of hits (H) and misses ($N–H$) of the *V. chilensis* recognizer. The second subset consists of vocalizations with maximum amplitude in the range [0, −30 dB]. There are $N = 111$ target events in the second subset, with respect to which the number of hits (H) and misses ($N–H$) are counted. In both subsets the remainder of the 337 target tags, together with the 137 of the group *not sure*, were not counted as insertions I. Therefore, for insertions we count only detections which do not coincide with the extended set of 474 *V. chilensis* tags.

5. Results

We present experimental results for the proposed method of acoustic activity detection and for two other energy-based detection methods (i) the traditional GMM-based energy detector [5] and (ii) the syllables-based detector [20]. In order to make a fair comparison among these methods, their settings were adjusted to obtain a similar percentage of selected audio frames on the training dataset. For the GMM-based energy detection, this adjustment was made by setting the threshold equal to the higher mean of the mixtures components. In the case of the syllables-based algorithm we made use of stopping criterion 30 dB as in Härmä [20]. For the method proposed in Section 3, we set the frame selection criterion $C = 0.3$.

All acoustic activity detection methods selected between 90% and 99% of the audio frames containing target vocalizations with signal strengths higher than −30 dB (Table 2). In the following we investigate how these acoustic activity detection methods facilitate the subsequent species recognition step. Therefore, the performance of the species-specific recognizer of *V. chilensis* call series was evaluated in terms of the percentage metrics *correct* and *accuracy* on the subsets VL02 [0, −30 dB] and VL02 [0, −20 dB] (Table 3).

For subset VL02 [0, −20 dB] we observe reasonable *percentage correct* values with the proposed acoustic activity detection method and the GMM-based recognizer (Table 3). For subset VL02 [0, −30 dB] the proposed method returned higher *percentage correct* values when compared with the results of the two other recognizers. Furthermore, the proposed method outperformed the other two methods on subset VL02 [0, −30 dB] with respect to both measures: correct and accuracy. This advantage is due to the morphological filtering of the spectrogram that eliminates high-energy sound events, compactly localized in time and frequency, which otherwise possess sufficient energy to be selected by the acoustic activity detection methods.

Table 1
Subsets of the evaluation dataset VL02.

Amplitude range	# target tags	# not sure tags	Total # of tags
[0, −20 dB]	44	0	44
[0, −30 dB]	111	0	111
[0, −40 dB]	240	2	242
[0, −50 dB]	337	137	474

Table 2
Percentage of selected audio frames after applying the three acoustic activity detectors on the validation dataset VL02.

Method	Ground truth timestamps and tags	
	Subset VL02 [0, −30 dB] (%)	Subset VL02 [0, −20 dB] (%)
Proposed method	89.8	99.2
GMM-based method [5]	91.6	97.8
Syllable-based method [20]	98.8	98.9

Finally, we analyzed the average computation time needed for each of the three methods to process a single 14-min audio recording (Table 4). All tests were carried out on a personal computer with operational memory 8 GBs, using a single core of the Intel Xeon E3-1220 processor operating at 3.1 GHz. The times under *feature extraction* refer to the overall time needed for frame selection (acoustic activity detector) and for computing the audio features (species-specific recognizer).

The average time spent for processing a 14-min audio recording in the *V. chilensis* recognizer is between 13 and 29 s, depending on the acoustic activity detection method. In particular, the HMM-based recognizer needs 1.100 s and 2.590 s on average to carry out the recognition with the proposed method and with the GMM-based acoustic activity detector, respectively. This difference is due to the number of frames selected by each method, specifically, in comparison with the GMM-based approach the proposed method selects more carefully the candidate frames, which results in a smaller number of frames, faster computational times for the HMM recognizer, and lower misclassification rates. By contrast, the shorter time for frame selection in the GMM-based acoustic activity detection method is voided by the longer processing time in the HMM recognizer. Thus, with the proposed method the total processing time of the *V. chilensis* recognizer is 52% shorter when compared to the syllable-based method and only by 8% longer when compared to the GMM-based approach.

In conclusion, when compared to the syllable-based acoustic activity detector the proposed method demonstrates better performance on the subset VL02 [0, −30 dB] in terms of the percentage metrics *correct* and *accuracy* as well as *speed*. The same holds true when the proposed approach is compared with the GMM-based method, except for the minor increase of computational demands.

6. Discussion

The idea of processing the audio spectrogram as an image was exploited in earlier studies [17–19]. The main difference with respect to related work is that here we made use of morphological filtering for the purpose of robust acoustic activity detection.

Table 3
Performance of the *V. chilensis* recognizer for three acoustic activity detection methods.

Method	Subset VL02 [0, −30 dB]		Subset VL02 [0, −20 dB]	
	Correct (%)	Accuracy (%)	Correct (%)	Accuracy (%)
Proposed method	88.3	56.4	93.9	41.3
GMM-based method [5]	74.3	55.2	90.1	48.0
Syllable-based method [20]	49.0	44.1	73.0	62.8

Table 4
Overall computation time for processing a single 14-min audio recording by the *V. chilensis* recognizer for three acoustic activity detection methods.

Method	Feature extraction [ss.ms]	HMM recognizer [ss.ms]	Total time spent [ss.ms]
Proposed method	12.550	01.100	13.650
GMM-based method [5]	10.050	02.590	12.640
Syllable-based method [20]	21.670	06.800	28.470

Potential misclassification is avoided by removing sound events that are localized in a narrow time–frequency range, but otherwise still have considerable energy levels. The experimental results confirm the advantage of the proposed method.

In our case study the proposed method for acoustic activity detection of *V. chilensis* leads to relatively higher precision when compared with the traditional GMM-based [5] and syllable-based [20] methods. This is particularly true when the *V. chilensis* recognizer operates on the noisy signals [0, −30 dB] (Table 3). However, there is potential for further improvements of the proposed method, especially with respect to the detection of weaker audio signals.

One limitation of our approach is that the selection of a certain audio frame always means that the entire frequency bandwidth is selected for further processing, e.g. for computation of Mel Frequency Cepstral Coefficients (MFCC) and subsequent processing steps. In other words, if no special action for noise suppression is foreseen, there is a high risk that signals of more than one species are represented in the feature vector. During training this will interfere with the quality of the target model. In the cases when there is no overlap between target and non-target signals in the frequency domain this drawback can be circumvented by retrieving only regions of interest, as suggested in Briggs et al. [18]. Thus audio frames can easily adapt to the variable bandwidths and lengths of specific call notes, and thereby reflect the true nature of animal vocalizations much better than frames covering the entire frequency range. However, the selection of variable bandwidth regions might not allow straightforward use of MFCC parameters and would require new image-based audio descriptors, such as those used in Potamitis [19].

7. Conclusion

The morphology-based acoustic activity detection method described here was found to outperform the traditional GMM-based [5] and syllable-based [20] methods. The practical significance of our approach was evaluated on the task of automated acoustic species recognition, where we focused on *V. chilensis* vocalizations in real-field recordings. The advantageous performance of the proposed method facilitates the incremental improvement of technology with the medium-term goal of automated monitoring of indicator species and migratory birds in the Brazilian Pantanal.

On a personal computer with processor Intel Xeon E3-1220 at 3.1 GHz and operating system Linux Ubuntu 12.04, the proposed acoustic activity detector operates approximately 62 times faster than real-time. We foresee the integration of the acoustic activity detector in the large Pantanal Database Repository established by the Computational Bioacoustics Research Unit (CO.BRA) at the UFMT's Institute of Computation.

Acknowledgements

The authors acknowledge the financial support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), the financial and logistic support of the National Institute for Science and Technology in Wetlands (INAU/UFMT), the Brehm Foundation for International Bird Conservation, Germany, the project OP "Competitiveness" BG161PO003-1.2.04-0044-C0001 financed by the Structural Funds of the European Union, and the project ISP1 financed by the Technical University of Varna, Bulgaria. Allan Goncalves de Oliveira, M.Sc., visited the Technical University of Varna with a Sandwich Grant from the Science without Borders Program of the Brazilian Federal Government, CAPES / CNPq / FAPs, Bolsa Pesquisador Visitante Especial, Prof. Dr. Karl-L. Schuchmann, Process No. 400019/2013-2. Luiz Vicente da Silva Campos Filho, Fazenda Pouso Alegre, and Director Waldir Wolfgang Valutky, Estância Ecológica SESC Pantanal: Reserva Particular do Patrimônio Natural, Baía das Pedras, generously allowed us to conduct our studies on their habitat-rich lands. Roseneide Soares, INAU, kindly provided her immense administrative expertise during the project implementation phase. Last but not least, special thanks go to the INAU directors Prof. Dr. Wolfgang Junk and Prof. Dr. Paulo Teixeira for their continuous support and encouragement throughout our field and laboratory work.

References

[1] Zhang X, Li Y. Environmental sound recognition using double-level energy detection. J Signal Inform Process 2013;4(3B):19–24.

[2] Bimbot F, Bonastre J-F, Fredouille C, Gravier G, Magrin-Chagnolleau I, Meignier S, et al. A tutorial on text-independent speaker verification. EURASIP J Adv Signal Process 2004;2004(4):430–51.

[3] Mamiya Y, Yamagishi J, Watts O, Clark R, King S, Stan A. Lightly supervised GMM VAD to use audiobook for speech synthesiser. In: Proc of the 2013 IEEE international conference acoustics, speech and signal processing (ICASSP '13); 2013. p. 7987–91.

[4] Alam J, Kenny P, Ouellet P, Stafylakis T, Dumouchel P. Supervised/unsupervised voice activity detectors for text-dependent speaker recognition on the RSR2015 corpus. In: Proc of the Odyssey speaker and language recognition workshop, (Odyssey '14); 2014. p. 123–30.

[5] Sahidullah M, Saha G. Comparison of speech activity detection techniques for speaker recognition. Cornell University Library 2012. arXiv preprint arXiv:1210.0297; 2012.

[6] Potamitis I, Ntalampiras S, Jahn O, Riede K. Automatic bird sound detection in long real-field recordings: applications and tools. Appl Acoust 2014;80:1–9.

[7] Frommolt K-H, Tauchert K-H. Applying bioacoustic methods for long-term monitoring of a nocturnal wetland bird. Ecol Informat 2014;21:4–12.

[8] Acevedo MA, Corrada-Bravo CJ, Corrada-Bravo H, Villanueva-Rivera LJ, Aide TM. Automated classification of bird and amphibian calls using machine learning: a comparison of methods. Ecol Informat 2009;4(4):206–14.

[9] Stowell D, Plumbley MD. Birdsong and C4DM: a survey of UK birdsong and machine recognition for music researchers. Technical report C4DM-TR-09-12. Centre for Digital Music, Queen Mary University of London; 2010.

[10] Mporas I, Ganchev T, Kocsis O, Fakotakis N, Jahn O, Riede K. Integration of temporal contextual information for robust acoustic recognition of bird species from real-field data. Int J Intell Syst Appl 2013;5(7):9–15. ISSN: 2074-904X.

[11] Ganchev T, Mporas I, Jahn O, Riede K, Schuchmann K-L, Fakotakis N. Acoustic bird activity detection on real-field data. In: Maglogiannis I, Plagianakos V, Vlahavas I, editors. SETN 2012, LNAI 7297, Springer-Verlag Berlin Heidelberg; 2012. p. 190–7.

[12] Stowell D, Plumbley MD. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. PeerJ 2014;2:e488. http://dx.doi.org/10.7717/peerj.488.

[13] Stowell D, Plumbley MD. Large-scale analysis of frequency modulation in birdsong databases. Methods Ecol Evol 2014;2014. http://dx.doi.org/10.1111/2041-210X.12223.

[14] Göeau H, Glotin H, Vellinga W-P, Rauber A. LifeCLEF bird identification task 2014. In: CLEF working notes 2014; 2014.

[15] Fodor G. The ninth annual MLSP competition: first place. In: Proc of the international conference on machine learning for signal processing (MLSP 2013); IEEE, 2013; 2013. p. 2. 10.1109/MLSP.2013.6661932.

[16] Glotin H, LeCun Y, Artieres T, Mallat S, Tchernichovski O, Halkias X. Neural information processing scaled for bioacoustics, from neurons to big data. In: Proc of NIPS4B, international workshop joint to NIPS, USA, 2013; 2013.

[17] Bardeli R. Similarity search in animal sound databases. IEEE Trans Multimedia 2009;11(1):68–76.

[18] Briggs F, Lakshminarayanan B, Neal L, Fern XZ, Raich R, Hadley SJK, et al. Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. J Acoust Soc Am 2012;131(6):4640–50.

[19] Potamitis I. Automatic classification of a taxon-rich community recorded in the wild. PLoS One 2014;9(5):e96936.

[20] Härmä A. Automatic identification of bird species based on sinusoidal modeling of syllables. In: Proc of the 2003 IEEE international conference on acoustics, speech, and signal processing, (ICASSP '03), vol. 5; 2003. p. 545–8.

[21] Lee C-H, Chou C-H, Han C-C, Huang R-Z. Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis. Pattern Recogn Lett 2006;27(2):93–101.

[22] Chou C-H, Liu P-H, Cai B. On the studies of syllable segmentation and improving MFCCs for automatic birdsong recognition. In: Proc of the IEEE Asia-Pacific services computing conference, (APSCC '08); 2008. p. 745–50.

[23] Lee C-H, Han C-C, Chuang C-C. Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients. IEEE Trans Audio, Speech, Lang Process 2008;16:1541–50.

[24] Graciarena M, Delplanche M, Shriberg E, Stolcke A, Ferrer L. Acoustic front-end optimization for bird species recognition. In: Proc of the IEEE international conference on acoustics speech and signal processing, (ICASSP'10); 2010. p. 293–6.

[25] Jahn O. Surveying tropical bird communities: in search of an appropriate rapid assessment method. In: Schuchmann K-L, editor. Bird communities of the Ecuadorian Chocó: a case study in conservation; 2011. p. 63–107. Bonner zoologische Monographien 56. Zoological Research Museum A. Koenig (ZFMK), Bonn, Germany. URL: <http://zoologicalbulletin.de/BzB_Volumes/BzM_56/BZM_56.pdf>; (last accessed 06.04.14).

[26] Jahn O. Structure and organization of the bird community. In: Schuchmann K-L, editor. Bird communities of the Ecuadorian Chocó: a case study in conservation; 2011. p. 109–78. Bonner zoologische Monographien 56. Zoological Research Museum A. Koenig (ZFMK), Bonn, Germany. URL: <http://zoologicalbulletin.de/BzB_Volumes/BzM_56/BZM_56.pdf>; (last accessed 06.04.14).

[27] Wei C, Blumstein DT. Noise robust bird song detection using syllable pattern-based hidden Markov models. In: Proc of the 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP-2011); 2011. p. 345–8. 10.1109/ICASSP.2011.5946411.

[28] Tan LN, Alwan A, Kossan G, Cody ML, Taylor CE. Dynamic time warping and sparse representation classification for birdsong phrase classification using limited training data. J Acoust Soc Am 2015;137:1069–80. http://dx.doi.org/10.1121/1.4906168.

[29] Somervuo P, Härmä A, Fagerlund S. Parametric representations of bird sounds for automatic species recognition. IEEE Trans Audio, Speech, Lang Process November 2006;14(6):2252–63. http://dx.doi.org/10.1109/TASL.2006.872624.

[30] Cadore J, Gallardo-Antolan A, Pelaez-Moreno C. Morphological processing of spectrograms for speech enhancement. In: Advances in nonlinear speech processing. Lecture notes in computer science, vol. 7015. Springer, Berlin Heidelberg; 2011. p. 224–31.

[31] Bovik A. Handbook of image and video processing. 2nd ed. Burlington: Elsevier Academic Press; 2005.

[32] Evans NWD, Mason JS, Roach MJ. Noise compensation using spectrogram morphological filtering. In: Proc of the 4th IASTED international con. on signal and image processing; 2002. p. 157–61.

[33] Schuchmann K-L, Marques MI, Jahn O, Ganchev T, de Figueiredo JM. Os sons do Pantanal: um projeto de monitoramento acústico automatizado da biodiversidade. O Biólogo, Revista do Conselho Regional de Biologia 2014;2014(29):12–5.

[34] Ganchev TD, Jahn O, Marques MI, de Figueiredo JM, Schuchmann K-L. Automated acoustic detection of Vanellus chilensis lampronotus. Expert systems with applications; 2015. Accepted manuscript, March 2015. 10.1016/j.eswa.2015.03.036.