Expert Systems With Applications 42 (2015) 8463–8471

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Audio parameterization with robust frame selection for improved bird identification

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

[a] National Institute for Science and Technology in Wetlands (INAU), Science without Borders Program (CsF), Federal University of Mato Grosso (UFMT), Av. Fernando Corrêa da Costa 2367, Cuiabá-MT, Brazil
[b] Institute of Computing, Federal University of Mato Grosso, Av. Fernando Corrêa da Costa 2367, Cuiabá-MT, Brazil
[c] Institute of Physics, Federal University of Mato Grosso, Av. Fernando Corrêa da Costa 2367, Cuiabá-MT, Brazil
[d] Department of Electronics, Technical University of Varna, str. Studentska 1, 9010, Varna Bulgaria
[e] Zoological Research Museum A. Koenig, Adenauerallee 160, 53113, Bonn Germany
[f] University of Bonn, Regina-Pacis-Weg 3, D -53113, Bonn Germany
[g] Institute of Biosciences, Federal University of Mato Grosso, Av. Fernando Corrêa da Costa 2367, Cuiabá-MT, Brazil

ARTICLE INFO

Keywords:
Computational bioacoustics
Bird identification
Hidden Markov Model (HMM)
Mel Frequency Cepstral Coefficients (MFCCs)
Robust frame selection

ABSTRACT

A major challenge in the automated acoustic recognition of bird species is the audio segmentation, which aims to select portions of audio that contain meaningful sound events and eliminates segments that contain predominantly background noise or sound events of other origin. Here we report on the development of an audio parameterization method with integrated robust frame selection that makes use of morphological filtering applied on the spectrogram seen as an image. The morphological filtering allows to exclude from further processing certain audio events, which otherwise could cause misclassification errors. The Mel Frequency Cepstral Coefficients (MFCCs) computed for the selected audio frames offer a good representation of the spectral information for dominant vocalizations because the morphological filtering eliminates short bursts of noise and suppresses weak competing signals. Experimental validation of the proposed method on the identification of 40 bird species from Brazil demonstrated superior accuracy and faster operation than three traditional and recent approaches. This is expressed as reduction of the relative error rate by 3.4% and the overall operational time by 7.5% when compared to the second best result. The improved frame selection robustness, precision, and operational speed facilitate applications like multi-species identification of real-field recordings.

1. Introduction

Biodiversity monitoring is a prerequisite for sustainable conservation action and is particularly important in efforts to reduce the loss of species (Pereira et al., 2013). Traditionally, animal species distribution, diversity, and population density are assessed with a variety of survey methods that are costly and limited in space and time (e.g., Bibby, Burgess, Hill, & Mustoe, 2000; Jahn, 2011a, 2011b).

Since many animals, such as grasshoppers, crickets, katydids, cicadas, anurans, birds, and certain mammals are more often heard than seen, one promising non-intrusive method for monitoring their presence and activity is the automated acoustic detection and identification. Remote and autonomous survey methods can provide continuous information on the presence/absence of rare and threatened species as well as on the general status of biodiversity in a cost-effective way (e.g., Aide et al., 2013; Ganchev, Jahn, Marques, de Figueiredo, & Schuchmann, 2015; Potamitis, Ntalampiras, Jahn, & Riede, 2014; Sueur, Pavoine, Hamerlynck, & Duvail, 2008). Thus, the use of new technologies is considered as an opportunity for facilitating biodiversity monitoring efforts in remote and difficult-to-access areas, such as the vast Pantanal wetlands of Brazil (Schuchmann, Marques, Jahn, Ganchev, & Figueiredo, 2014).

Based on soundscapes, it is possible to identify the species that are present in an area. However this is not a simple task, since the amount of data to be analyzed is very large, reaching the order

* Corresponding author at: Department of Electronics, Technical University of Varna, str. Studentska 1, 9010 Varna, Bulgaria. Tel.: +359 888096974.
E-mail addresses: thiago@ic.ufmt.br (T.M. Ventura), allan@ic.ufmt.br (A.G. de Oliveira), tganchev@ieee.org, tganchev@hotmail.com, tganchev@tu-varna.bg (T.D. Ganchev), josiel@ic.ufmt.br (J.M. de Figueiredo), o.jahn@zfmk.de (O. Jahn), marinez@ufmt.br (M.I. Marques), klschuchmann@googlemail.com (Karl-L. Schuchmann).

http://dx.doi.org/10.1016/j.eswa.2015.07.002
0957-4174/

of several terabytes per continuous annual cycle of recordings. Consequently, data processing is lengthy and computationally expensive (Oba, 2004). The principle prerequisites for large-scale application of soundscape analysis methods are an increased species recognition accuracy and reduction of the overall computational demands. For that purpose improvements, in the sense of accuracy and speed, are required in the audio parameterization and the classification methods. In the present work we focus on the audio parameterization.

Nowadays, the statistical machine learning approach dominates the field of bioacoustics. The audio signal is first parameterized and subsequently the statistical distribution of the audio parameters is modeled. The most widely used modeling techniques for acoustic animal identification are based on the Hidden Markov Model (HMM) (Bardeli et al., 2010; Chu & Blumstein, 2011; Potamitis et al., 2014; Trifa, Kirschel, Taylor, & Vallejo, 2008) or its single-state version known as Gaussian Mixture Models (GMMs) (Ganchev et al., 2015; Henríquez et al., 2014). The success of the GMM- and HMM-based recognition method depends on the appropriateness of the audio parameterization process, particularly the segmentation and selection of representative portions of the species-specific sound emissions.

Various strategies for audio parameterization were reported in the literature. Simple solutions, which incorporate energy-based frame selection methods for eliminating silent portions of the signal, do not depend on prior knowledge about the signal and are quite easy to implement (Zhang & Li, 2013). This is the main reason for their widespread use in environmental sound recognition. However their accuracy in low signal-to-noise ratio (SNR) conditions is often unsatisfactory.

In a large-scale experiment on the acoustic identification of 501 bird species, Stowell and Plumbley (2014) applied unsupervised feature learning on raw audio, i.e. without prior segmentation and reported species identification accuracy of 42.9%.

Härmä (2003) proposed a method that extracts syllables from bird vocalizations. Huang, Yang, Yang, and Chen (2009) used this approach to classify frogs by determining three different features from the syllables: spectral centroid, signal bandwidth, and threshold-crossing rate. Lee, Han, and Chuang (2008) applied the same algorithm to identify birds sounds by generating Mel Frequency Cepstral Coefficients (MFCCs) from syllables and Lee, Chou, Han, and Huang (2006) classified animal sounds on the basis of linear discriminant analysis. Other syllabification approaches were studied by Chou, Lee, and Ni (2007), who obtained syllables and clustered them with the fuzzy C-means method whereas Chou and Liu (2009) used wavelet transformations to determine sections in the bird songs.

Juang and Chen (2007) proposed an energy-based method for audio segmentation and subsequent selection of segments with bird song activity. In a related work Acevedo, Corrada-Bravo, Corrada-Bravo, Villanueva-Rivera, and Aide (2009) manually selected portions of interest in the spectrogram, and then compared various machine learning techniques for audio data from frog and bird species. Neal, Briggs, Raich, and Fern (2011) used a Random Forest classifier to implement supervised time and frequency audio segmentation and Evangelista, Priolli, Silla, Angelico, and Kaestner (2014) experimented with sound representation in the frequency domain, energy of the signal, and its spectral centroid to carry out an automatic segmentation of audio.

A more recent approach, based on the idea to treat the sound spectrogram as an image, selects regions of interest in the spectrogram and then extracts their statistical characteristics. The features computed from these regions of interest are used to train machine learning algorithms (Aide et al. 2013; Briggs et al., 2012; Kaewtip, Tan, Alwan, & Taylor, 2013; Potamitis, 2014). Likewise, Bardeli (2009) proposed a method in which the sound spectrogram is processed as an image and subsequently used similarity-search techniques to classify a set of animal sounds. In de Oliveira et al. (2015), morphological filtering was employed for the purpose of bird acoustic activity detection which is part of a species-specific recognizer for automated acoustic recognition of *Vanellus chilensis* vocalizations.

Motivated by previous related work, in Section 2 we present an improved audio parameterization method that incorporates robust audio segmentation based on morphological processing of the sound spectrogram considered as an image. Our work differs from previous related work (Aide et al. 2013; Briggs et al., 2012; de Oliveira et al., 2015; Kaewtip et al., 2013; Potamitis, 2014), where morphological filtering of the spectrogram is only part of noise suppression or acoustic activity detection. By contrast, in the current work it is used as part of the robust frame selection that is integrated in the MFCC feature extraction process. By this the audio parameterization computes MFCCs only for the selected audio segments, which speeds up the operation. In Section 3 we describe the experimental setup, which involves the classification of short audio recordings of 40 bird species from Mato Grosso, Brazil. The results of a comparative evaluation of the proposed method with three other frame selection approaches (Briggs et al., 2012; Härmä, 2003; Sahidullah & Saha, 2012) are presented in Section 4. Finally, in Section 5 we evaluate our work providing a detailed discussion on the advantages and shortcomings of the proposed method and its application area.

2. Method

Parameterization transforms the audio signals so that useful information is presented in a compact way and irrelevant information is eliminated. The audio features computed during parameterization are next fed to the classification stage (Fig. 1). The latter allows for a final decision of the category to which each input audio recording belongs, based on the scores computed from the individual species-specific models.

An effective parameterization is crucial for achieving high recognition accuracy. When the parameterization does not fully convey the useful information or when this information is buried in audio feature variability unrelated to the species-specific traits, the modeling and the identification processes are seriously impeded. In an attempt to improve the bird identification accuracy and to reduce computational demands we propose a parameterization method that preserves audio segments carrying useful information, in our case bird sound events. Likewise, short-term and narrow-frequency bursts of energy are discarded because they are definitely not bird vocalizations. Therefore, audio features are computed only for selected subsets of audio frames, reducing the overall computational demands. Elimination of audio segments not containing bird sounds also means a lower risk of misidentification in the classification stage. In Section 2.1 we outline the robust frame selection incorporated in the MFCC computation and post-processing, and in Section 2.2 we elaborate on the classification process that uses these audio features.

2.1. Audio parameterization

The parameterization procedure consists of the following five steps (Fig. 1). First, the audio recording obtained from the database is subject to preprocessing. This step consists of resampling to a sampling frequency of 24 kHz and high-pass filtering of the signal with a 10th order Butterworth filter. The audio is resampled in order to reduce the computational and memory demands in the following processing steps, whereas the high-pass filter with cutoff frequency 1 kHz reduces the influence of wind noise and other low-frequency interferences from the environment.

Thereafter, the dynamic spectrum of the preprocessed time-domain signal $s(n)$ is computed through the short-time discrete

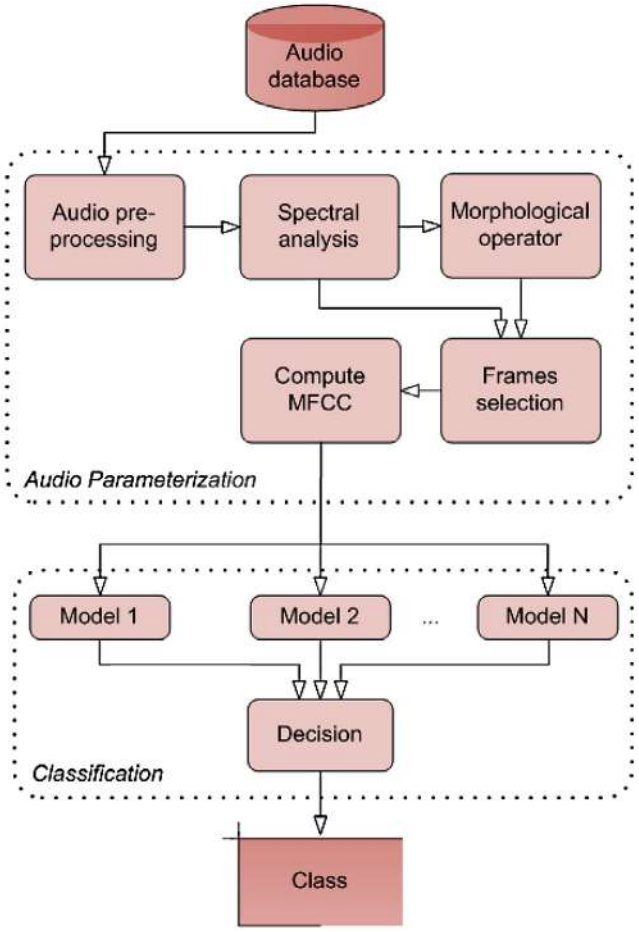

Fig. 1. Overall block diagram of the acoustic bird species identification process, showing the proposed audio parameterization method and the following classification stage.

Fourier transform (STDFT):

$$S(k,l) = \sum_{n=0}^{N-1} s(n) W(n+lL) \exp\left(-\frac{j2\pi nk}{N}\right), 0 \le n,k \le N-1, \quad (1)$$

where n is the index of the time domain samples, k is the index of the Fourier coefficients, and l denotes the relative displacement of the current audio segment in terms of steps of L samples. The Hamming window $W(m)$, defined as

$$W(m) = 0.54 - 0.46\cos\left(\frac{2\pi m}{M}\right), m = 0, 1, \ldots M-1, \quad (2)$$

is applied to reduce the spectral distortions caused by an abrupt change of signal amplitude at the boundary points of each audio segment. Here the STDFT is obtained after applying the discrete Fourier transform for $N = 512$ samples on the zero-padded signal $s(n)$ weighted with a sliding Hamming window of $M = 480$ samples. The Hamming window is sliding with a step of $L = 120$ samples between subsequent segments. The spectrogram $|S(k,l)|$ obtained to this end is seen as an image and is the main source of information for the subsequent processing steps (Fig. 2).

However, the raw image of the spectrogram is not sufficient to perform an accurate selection of the high energy areas, as most recordings contain noise and the competing acoustic activity of other animals. Thus, the direct use of the raw spectrogram increases the probability that sounds originating from different sources are selected as a single acoustic event.

We apply morphological operators on the spectrogram in order to eliminate short-time narrow-band bursts of energy and thus reduce the risk of merging together more informative with less informative sound events. According to Bovik (2005), the morphological operators offer a convenient method of image enhancement by noise suppression and simplification of the spectrogram by retaining only those components that satisfy certain size or contrast criteria. Similarly to Cadore, Gallardo-Antoln, and Pelez-Moreno (2011) and Potamitis (2014) we apply the opening morphological operator, which is described as the *erosion* operator (3) followed by *dilation* (4).

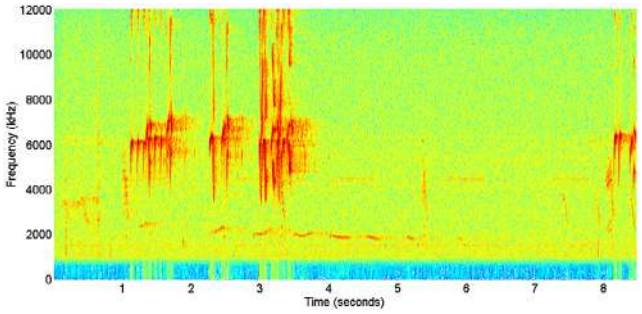

Fig. 2. Spectrogram of a resampled and high-pass filtered audio recording.

$$|S_e(k,l)| = |S(k,l)| \ominus B = \{z : B_z \subseteq |S(k,l)|\}, \quad (3)$$

$$|S_{ed}(k,l)| = |S_e(k,l)| \oplus \hat{B} = \{z : \hat{B}_z \cap |S_e(k,l)| \neq \emptyset\}. \quad (4)$$

Here $|S(k,l)|$, $|S_e(k,l)|$, and $|S_{ed}(k,l)|$ are images corresponding to the spectrogram after applying the erosion operator and after applying both the erosion and dilation operators. B is called the structuring element and has a simple geometrical shape, in our case a rectangle of 40 × 30 pixels. Following erosion, the result will be the set of all points z such that B, translated by z, is contained in the image $|S(k,l)|$. In the dilation operator, $\hat{B}$ is the set of pixel locations z, where the reflected structuring element overlaps with foreground pixels in $|S_e(k,l)|$ when translated to z. Making use of both operators, we represent the opening morphological operator on the spectrogram as:

$$|S_{ed}(k,l)| = (|S(k,l)| \ominus B) \oplus \hat{B}. \quad (5)$$

The operator *erosion* reduces bright regions and enlarges dark regions in the image. By contrast, the operator *dilation* enlarges bright regions and reduces dark image regions. The subsequent application of both operators eliminates certain very small elements in the spectrogram, like short-lived audio signals and weak distortions in the time-frequency space (Fig. 3). The size of the elements that survive this filtering process depends on the size of the structuring elements B and $\hat{B}$.

If we compare the regions of highest energy in Figs. 2 and 3, it is evident that after applying the morphological operators certain elements of the spectrogram are less prominent, e.g. these corresponding to noise and weak vocalizations of competing species.

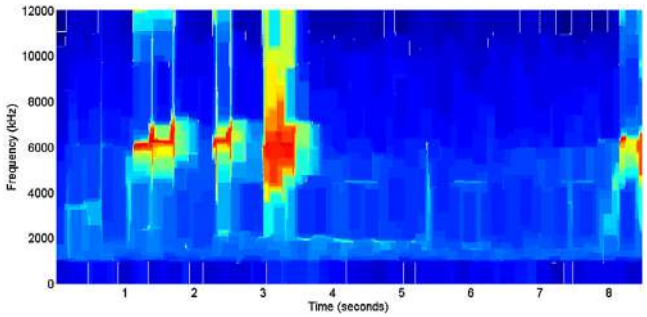

Fig. 3. The spectrogram after applying the opening morphological operators (cf. Fig. 2).

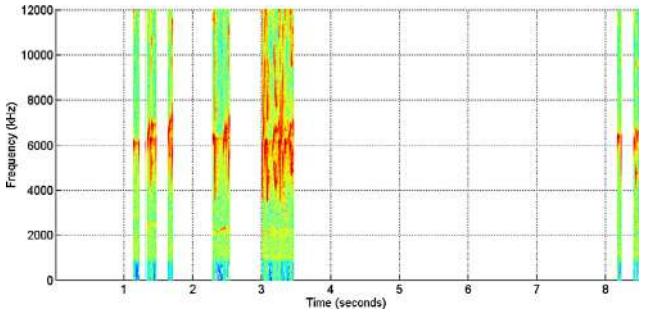

Fig. 4. Selected frames after applying the frame selection algorithm.

The resulting spectrogram appears cleaner and thereby facilitates the more effective and robust selection of the representative audio frames.

Here we incorporate an algorithm for robust frame selection, which adjusts the decision threshold based on the histogram of the sum of amplitudes for all frequency components in each frame l of the processed spectrogram $|S_{ed}(k,l)|$

$$E(l) = \sum_{k=0}^{N-1} |S_{ed}(k,l)|, l = 1, 2, \ldots T, \tag{6}$$

where T is the total number of frames in the specific recording. We select the bin of the histogram that contains less than 30% of the highest bin, considering only those bins that appear after the highest bin. The center of the selected bin specifies the threshold value θ. Each frame l with sum of the amplitude (6) greater than the threshold will be selected for the next processing step:

$$|S_{sel}(k,l_{sel})| = \begin{cases} |S_{ed}(k,l)| \text{ for } E(l) \geq \theta \\ 0 \quad \text{otherwise} \end{cases}. \tag{7}$$

The selected audio frames for the spectrogram in Fig. 2 are presented in Fig. 4. Next, we compute the MFCC for each of the selected frames (7). In order to warp the frequency range [1, 12] kHz according to the Mel-scale, we apply a filter-bank $H_i(k)$ consisting of $K = 22$ Mel-spaced equal-height filters on the power spectrum $|S_{sel}(k, l_{sel})|^2$ for each selected frame l_{sel} and compute the log-energy for the corresponding filter output:

$$S_m(i, l_{sel}) = \ln\left(\sum_{k=0}^{N-1} |S_{sel}(k, l_{sel})|^2 H_i(k)\right), i = 1, 2, \ldots K. \tag{8}$$

Here, each filter in the filter-bank is defined as:

$$H_i(k) = \begin{cases} 0 & \text{for} \quad k < f_{b_{i-1}} \\ \dfrac{\left(k - f_{b_{i-1}}\right)}{\left(f_{b_i} - f_{b_{i-1}}\right)} & \text{for} \quad f_{b_{i-1}} \leq k \leq f_{b_i} \\ \dfrac{\left(f_{b_{i+1}} - k\right)}{\left(f_{b_{i+1}} - f_{b_i}\right)} & \text{for} \quad f_{b_i} \leq k \leq f_{b_{i+1}} \\ 0 & \text{for} \quad k > f_{b_{i+1}} \end{cases}, i = 1, 2, \ldots K, \tag{9}$$

where i stands for the ith filter, f_{b_i} are the boundary points of the individual filters, and $k = 1, 2, \ldots, N$ corresponds to the kth coefficient of the N-point discrete Fourier transform (DFT). The boundary points f_{b_i} are expressed in terms of position, which depends on the sampling frequency $F_s = 24$ kHz and the number of points $N = 512$ in the DFT:

$$f_{b_i} = \left(\frac{N}{F_s}\right) \cdot \hat{f}_{mel}^{-1}\left(\hat{f}_{mel}(f_{low}) + i \cdot \frac{\hat{f}_{mel}(f_{high}) - \hat{f}_{mel}(f_{low})}{K+1}\right), \tag{10}$$

where the function $\hat{f}_{mel}$ stands for the transformation

$$\hat{f}_{mel} = 1127 \cdot \ln\left(1 + \frac{f_{lin}}{700}\right), \tag{11}$$

$f_{low} = 1$ kHz and $f_{hign} = 12$ kHz are respectively the low and high boundary frequencies for the entire filter bank, K is the total number of filters, and $\hat{f}_{mel}^{-1}$ is the inverse transformation to linear frequency scale, defined as:

$$\hat{f}_{mel}^{-1} = f_{lin} = 700\left[\exp\left(\frac{\hat{f}_{mel}}{1127}\right) - 1\right]. \tag{12}$$

Finally, we apply the discrete cosine transform on the result of (8) in order to obtain J+1 MFCC parameters:

$$C_j = \sum_{i=1}^{K} |S_m(i, l_{sel})| \cos\left(j(i+0.5)\frac{\pi}{K}\right), j = 0, 1, \ldots, J. \tag{13}$$

Afterwards, the MFCCs are standardized for zero mean and unit standard deviation. Due to the frame selection process, the MFCCs are computed only for the best frames that represent the dominant sounds, which increase the probability of successful classification in the pattern recognition stage (Fig. 4).

2.2. Classification

The classification stage is fed with the standardized feature vectors obtained in the audio parameterization stage presented in Section 2.1. As we consider a multi-class classification problem, species-specific datasets are required for training each species model. Once all models are trained, the system is ready for classifying recordings of interest.

For each recording to be analyzed, the standardized feature vector obtained from the audio parameterization stage is compared with each model. Based on the similarity scores a final decision is made about the class to which the dominant audio signals belong.

Classification accuracy is affected by the quality of the audio features and is significantly compromised in the presence of noise or feature variability not linked to species-specific traits, and interference from competing species. Therefore, the proposed audio parameterization method aims to select only the best frames of the signal. Likewise, weaker competing sounds, sound events with short duration, and signals that are compact in the frequency domain are eliminated, because they are unrelated to bird vocalizations (Section 4.1).

3. Experimental setup

In the following subsections we briefly outline the common experimental protocol used in the comparative evaluation of the proposed audio parameterization method with other related traditional and recent methods.

3.1. Database

The audio data were downloaded from the xeno-canto archive (www.xeno-canto.org). We made use of a subset consisting of 40 bird species that are present in the State of Mato Grosso, Brazil. It is important to note that these are field recordings and each file potentially contains vocalizations of several animal species and competing noise caused by wind, rain, and anthropogenic interference.

Fifteen audio recordings with an average length of 32.4 s were available for each of the 40 species. We used ten recordings for the model creation and the remaining five for the purpose of performance evaluation. In total there were 400 files for model creation and 200 files for performance evaluation.

3.2. Reference methods

In order to evaluate the practical value of the proposed method, we made a comparative evaluation of traditional MFCCs computed after frame selection with the (1) Gaussian Mixture Model

(GMM)-based energy detector (Alam, Kenny, Ouellet, Stafylakis, & Dumouchel, 2014; Ganchev et al., 2015; Mamiya et al., 2013); (2) "syllables segmentation" approach (Chou, Liu, & Cai, 2008; Härmä, 2003; Lee et al., 2006); and (3) "regions of interest" selection method (Bardeli, 2009; Briggs et al., 2012; Potamitis, 2014). After some experimentation, these reference methods were fine-tuned for the best configuration settings on the current dataset.

3.2.1. Gaussian Mixture Model-based energy detector

The GMM-based energy detector models the distribution of short-term changes in energy and attempts to select only promising frames for further processing (Alam et al., 2014; Bimbot et al., 2004; Mamiya et al., 2013). Typically, the model is designed as a two-component GMM. The first mixture component is fitted to the distribution of the low-energy frames and the second is fitted to the distribution of the high-energy frames. The decision threshold is usually selected as a trade-off at the crossing point of the two Gaussian functions. In our case, the threshold is a function of the mean of the two components. The threshold θ_{trn} for the training dataset was calculated by (14) and θ_{tst} for the test dataset by (15):

$$\theta_{trn} = \frac{\mu_1 + \mu_2}{2} + 0.2|\mu_1 - \mu_2|, \tag{14}$$

$$\theta_{tst} = \max(\mu_1, \mu_2). \tag{15}$$

During the operation of the recognizer, θ_{tst} is selected with a higher value than θ_{trn} because we wished to make a decision based on frames with good signal-to-noise ratio.

3.2.2. Syllable segmentation

In general, bird songs can be broken down to four hierarchical levels: notes, syllables, phrases, and songs. By using the bird song structure, Härmä (2003) proposed an algorithm that extracts syllables from a continuous bird song, whereas Lee et al. (2006) and Chou et al. (2008) used this algorithm to extract syllables and to calculate for each one the MFCC.

One important parameter of this technique is the stopping criterion, which determines the start and the end of the syllables. Here, we set the stopping criterion at -30 dB below the highest amplitude as this guarantees good sensitivity and robustness.

3.2.3. Regions-of-interest-based frame selection

This method aims to select only the interesting regions of the spectrogram from which to extract features. Briggs et al. (2012) manually selected regions of interest (ROIs) for each file. Bardeli (2009) used the Structure Tensor technique to extract the ROIs from spectrograms. By contrast, Potamitis (2014) treated the spectrogram as an image, employed a filter to detach the ROIs, and applied a threshold to obtain a binarized image. Here we make use of the latter procedure to construct an enhanced spectrogram that contains only the ROIs. Subsequently we derive the MFCC for each audio frame.

3.3. Mel Frequency Cepstral Coefficients (MFCCs)

The methods outlined in Section 3.2 aim at selecting only promising portions of the spectrogram for the subsequent feature extraction process. In the present setup we computed the MFCCs as described by Dufour, Artieres, Glotin, and Giraudet (2013). In brief, 16 MFCCs were computed by using a window of 20.0 ms, sliding with a skip step of 5.0 ms. Subsequently, the first and second time-derivatives (delta and delta-delta coefficients) were computed to form the full feature vector of a length of 48 parameters. The feature vectors were normalized to zero mean value and unit standard deviation and then fed to a species-specific Hidden Markov Model (HMM)-based classifier.

3.4. The Hidden Markov Model-based classifier

In the present work we relied on the HMM implementation known as Hidden Markov Model Toolkit (HTK) (Young et al., 2006). Specifically, we built a total of 40 three-state species-specific HMMs. The training for estimating the model parameters was performed by different numbers of iterations of the Baum-Welch algorithm. For the purpose of species identification we made use of a simple grammar that was set to select one species per test trial.

3.5. Experimental protocol and performance metrics

A common experimental protocol was used to compare the proposed method with the three reference methods outlined in Section 3.2. The MFCCs computed from the training dataset described in Section 3.1 were used to train an HMM classifier, and the test dataset was used to evaluate it. The configuration of each HMM classifier is explained in Section 4.

The identification performance was evaluated in terms of percentage correct classifications:

$$\text{Correct} = \frac{H}{N} \times 100[\%], \tag{16}$$

where N is the total number of test trials and H is the number of correctly identified trials.

4. Results

In the following subsections we analyze the experimental results of the bird-identification performance evaluation, which involves different frame selection and audio parameterization methods. These methods are compared in terms of identification accuracy, time needed for training the HMM-based species-specific models, and operational speed.

4.1. Classification performance

In order to guarantee a fair comparison, all methods were adjusted to their best performance. For that purpose we carried out extensive tests with respect to the number of mixture components and the number of training iterations (Table 1).

For the proposed, the GMM-based, and the ROI-based methods the best identification accuracy was achieved with 48 mixture components, reaching 71.5%, 70.0%, and 47.5%, respectively. By contrast, the best result for the syllable-based method was achieved with 80 mixtures (64.5%). For the ROI-based method it was not possible to

Table 1
Performance of four methods, applying different numbers of the mixture components per HMM state and 50 training iterations.

Method	Number of mixtures	Correct (%)
Proposed method	32	70.0
	48	**71.5**
	64	69.5
	80	69.0
GMM-based (Sahidullah & Saha, 2012)	32	69.0
	48	**70.0**
	64	68.5
	80	67.0
Syllable-based (Härmä, 2003)	32	63.0
	48	61.5
	64	64.0
	80	**64.5**
ROI-based (Briggs et al., 2012)	32	42.0
	48	**47.5**
	64	45.5
	80	-

train a model with 80 mixtures because of the insufficient numbers of frames selected by this method. Even so, the results for the ROI method show that the best precision is achieved with 48 mixture components.

For the tests with GMM- and ROI-based methods, the results indicate a possible over-fitting during training, which might be due to the rather small number of training and validation files. With only 20 iterations the GMM-based method gave a precision of 70.5% and the ROI-based method of 48.0%. The proposed method still has the highest precision (71.5%), which corresponds to a relative decrease in the identification error of 3.4%, when compared with the second-best method (GMM).

4.2. Processing time

We estimated the processing time for the two main processing steps: the audio parameterization and the classification stages (Tables 3–5).

The proposed and the GMM-based methods need nearly the same time for the audio parameterization, with a slight advantage for the proposed method. When compared with the GMM approach the proposed method is faster by 4.3%. Such a speed up of computations would make significant difference only when large quantities of audio recordings are processed, which is actually the case in applications used for species recognition in continuous field recordings made with automated recording units. The syllable-based and ROI-based methods are 1.5 and 3.4 times slower when compared with the proposed method.

At the classification stage there are two different aspects: time needed for creating the models (Table 4) and the classification time needed during operation of the recognizer (Table 5). In the training phase the demand for computational resources is high. However this bottleneck can be overcome if the model creation is carried out only once and is performed off-line. In general, the time necessary for processing test files is less than the time needed for creating the statistical models. However, when large amounts of recordings are processed, the detector operation gets time-demanding. Therefore we measure and report these two cases separately.

The proposed method needed less than 21 min for the creation of the 40 species-specific models (Table 4). This is 3 to 6 times faster than the models build for the other three audio parameterization and frame selection methods. Likewise, the proposed method needed less than 65 s processing time for the identification of the 200 test files (Table 5), i.e. on average about 0.33 s per test file. This corresponds to a reduction of the classification time by 20.6% when compared to the second fastest (GMM-based) method, which required 81.29 s.

Combining the results for audio parameterization (Table 3) and classification (Table 5), we can conclude that the proposed method needs 375.84 s to process 200 files, and the second best GMM-based method needs 406.43 s. This is equivalent to a reduction of the overall operational time of 7.5% while reducing the relative error rate by 3.4%.

Table 2
Performance tests with varying numbers of training iterations.

Method	Train iterations	Correct (%)
Proposed method	20	66.5
	50	**71.5**
	100	70.5
GMM-based (Sahidullah & Saha, 2012)	20	**70.5**
	50	70.0
	100	**70.5**
Syllable-based (Härmä, 2003)	20	62.0
	50	**64.5**
	100	63.5
ROI-based (Briggs et al., 2012)	20	**48.0**
	50	47.5
	100	46.5

Table 3
Average time spent per test file and the overall time spent on the audio parameterization (200 files). Time format: hours, minutes, seconds, and milliseconds [hh:mm:ss.ms].

Method	Average [s]	Total time spent [hh:mm:ss.ms]
Proposed method	**1.56**	**00:05:11.28**
GMM-based (Sahidullah & Saha, 2012)	1.62	00:05:25.14
Syllable-based (Härmä, 2003)	2.38	00:07:55.31
ROI-based (Briggs et al., 2012)	5.31	00:17:41.78

Table 4
Time spent to create 40 species-specific HMM-based models for each of the four methods.Time format: hours, minutes, seconds, and milliseconds [hh:mm:ss.ms].

Method	Time spent [hh:mm:ss.ms]
Proposed method	**00:20:46.33**
GMM-based (Sahidullah & Saha, 2012)	01:22:14.85
Syllable-based (Härmä, 2003)	02:15:09.43
ROI-based (Briggs et al., 2012)	01:01:18.45

Table 5
Time spent to process (classify) 200 test files. Time format: hours, minutes, seconds, and milliseconds [hh:mm:ss.ms].

Method	Time spent [hh:mm:ss.ms]
Proposed method	**00:01:04.56**
GMM-based (Sahidullah & Saha, 2012)	00:01:21.29
Syllable-based (Härmä, 2003)	00:02:04.60
ROI-based (Briggs et al., 2012)	00:01:26.26

4.3. Comparison of frame selection methods

For better understanding the four methods evaluated here, we visualized the outcome of frame selection for two different scenarios: (i) “high signal-to-noise ratio (SNR) without competing sound events” and (ii) “low-SNR with competing sound events”. In the case of “high-SNR without competing sound events” all frame selection methods performed similarly well (Fig. 5), whereas the performance was different for recordings with “low-SNR and competing sound events” (Fig. 6).

In the high-SNR case, the ROI-based method selected only the frequency bands with high energy, whereas the other methods produced “cleaned” spectrograms that look very much alike (Fig. 5). By contrast, under noisy ambient conditions with competing sound events the four methods performed differently and the frame selection performance decreased considerably (Fig. 6). The proposed method demonstrated good performance and selected frames only of the dominant species. In the case of the GMM-based method the noise from insects interfered with the frame selection, as the threshold is configured to get the frames with highest energy relative to the average energy. Likewise, with the syllable-based approach the high energy of the insect sounds resulted in the selection of many frames, but the duration of the selected areas was very short in comparison with the other methods. Finally, the ROI-based method selected the insect activity as the main region of interest, getting almost none of the frames representing the dominant bird species.

5. Conclusion

Aiming to improve the audio parameterization process in bird identification tasks, we propose an approach that incorporates robust frame selection based on morphological filtering of the spectrogram

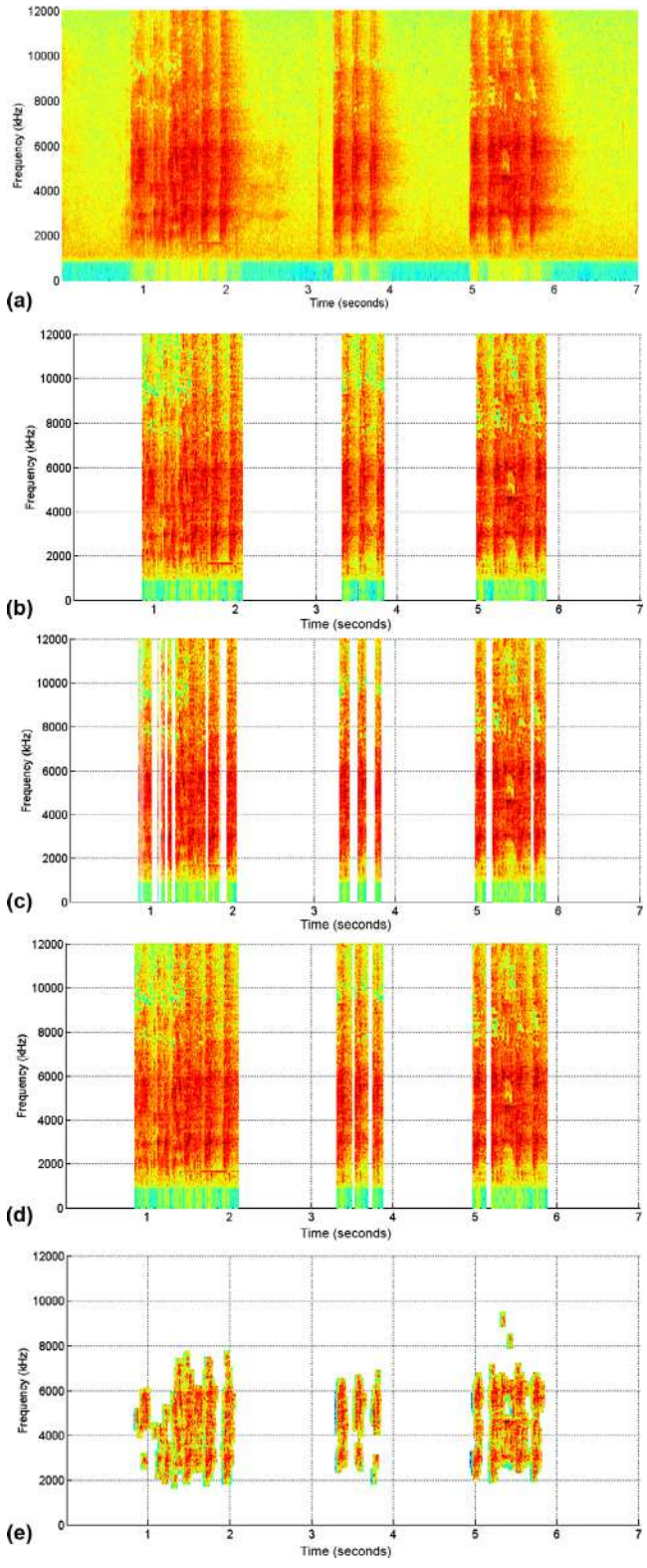

Fig. 5. Comparison of frame selection methods after processing a recording with three short bird call series under "high SNR without competing sound events" conditions: (a) spectrogram of a test audio file that was downsampled to 24 kHz and high-pass filtered; subsequently the recording was processed with (b) the proposed method, (c) GMM-based frame selection, (d) syllable-based approach, or (e) the ROI-based method.

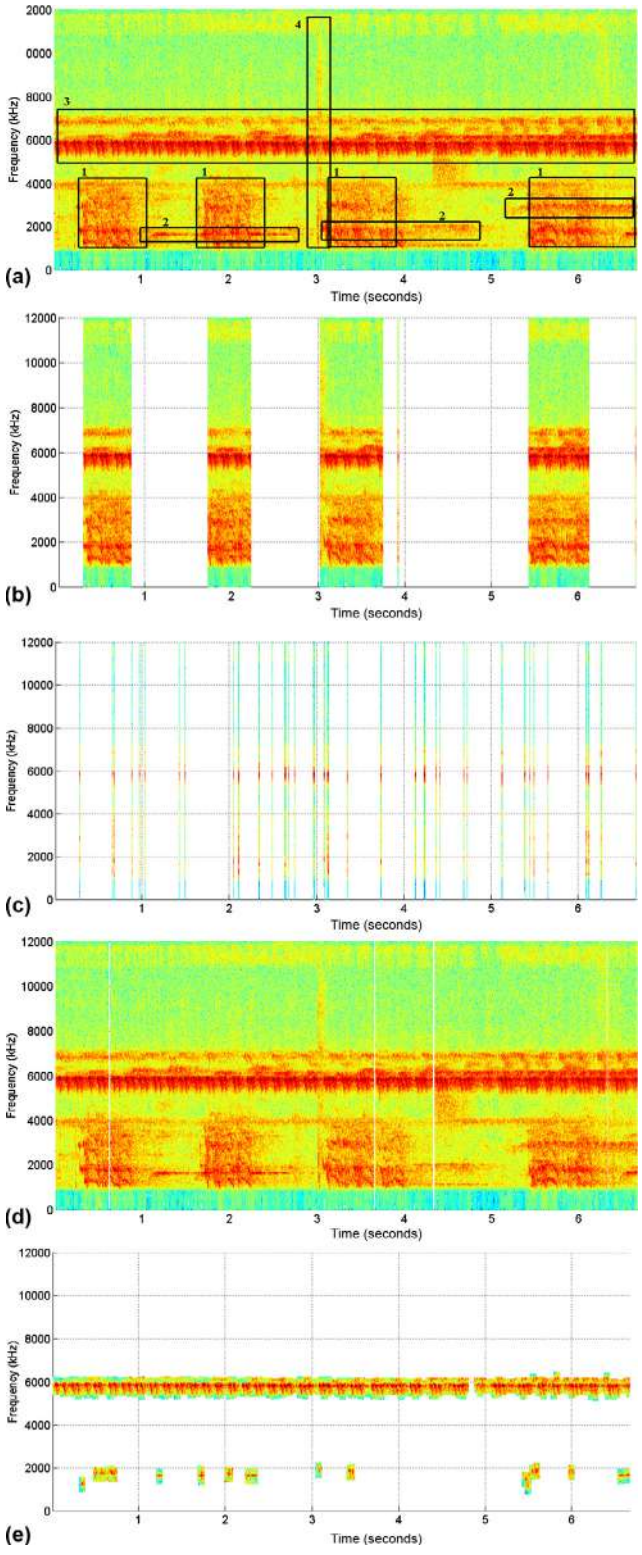

Fig. 6. Comparison of frame selection methods under noisy ambient conditions with competing animal sounds: (a) original spectrogram of a test audio file after downsampling to 24 kHz and high-pass filtering; different sound events were tagged manually, viz. (1) vocalization of dominant species, (2) sounds of several other bird species, (3) acoustic emissions from insects, and (4) unidentified event. Subsequently the preprocessed recording was analyzed with (b) the proposed method, (c) GMM-based frame selection, (d) syllable-based approach, and (e) the ROI-based method.

treated as an image. The robust frame selection shares common processing steps with the MFCC parameter computation so the two algorithms integrate well with only a small overhead. This approach and the fact that MFCC parameters are computed only for a subset of selected frames speeds up the overall computation of audio parameters. Consequently we feed to the classification step only a subset of all audio frames and the relative error rates are reduced. A comparative evaluation based on a common experimental protocol demonstrated the superior accuracy and time-efficiency of the proposed method, when compared to GMM-based frame selection and two other recent methods. The higher accuracy and lower processing time make it more suitable for the needs of automated bird identification tasks than the other methods tested. Other important aspects are its good scalability to very large repositories of audio recordings containing multiple species and especially the advantageous accuracy on real-field recordings captured with different signal-to-noise levels. These characteristics make the proposed method a good start for further technological improvements that soon could facilitate the automated monitoring of species-rich bird communities.

Compared to previous related work (Huang et al., 2009), where syllabification of frog sounds were used for improving recognition accuracy, we provided evidence that a non-syllabification method could perform faster and provide better accuracy on the task of bird identification from noisy field recordings. Furthermore, our audio parameterization with robust frame selection based on the morphological filtering of the spectrogram has the potential to improve recent methods for acoustic recognition of bird species, such as those in Ganchev et al. (2015), where a simple GMM-based acoustic activity detector was used to discard portions of silence in the recordings.

The proposed audio parameterization method computes MFCC parameters for the selected frames of the signal that has been resampled to 24 kHz and high-pass filtered. No special noise reduction is used in the process. Here we make the implicit assumption that since the frame selection is carried out on the spectrogram cleared through morphological filtering, the selected audio frames correspond to audio segments with prominent vocalizations. Evidently, this condition is not always met for field recordings. To overcome this limitation the proposed method could be combined with noise reduction techniques, such as certain subband spectral subtraction-based methods or similar methods affecting the spectral domain. We emphasize that noise reduction in the spectral domain can be integrated in the feature extraction process with only a small increase of the overall computational demand.

Among the potential applications that can benefit of the proposed audio parameterization method are automated tools for indexing bird sound recordings. Given the enormous amount of audio recordings collected by autonomous recording units, e.g. 6 TB per year for recorders operated in 24/7 mode at sampling rates of 48 kHz and a resolution of 16 bits, such a tool would largely speed up the data analysis of biodiversity inventories and monitoring studies. 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 and beyond.

Acknowledgments

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 project support by 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 Border Program of the Brazilian Federal Government (CsF/CNPq) grant from Pesquisador Visitante Especial, Prof. Dr. Karl-L. Schuchmann, process no. 400019/2013-2. Thiago Meirelles Ventura acknowledges his grant from CAPES/PDSE, process no. 14277-13-1. The Fundação de Amparo à Pesquisa do Estado de Mato Grosso (FAPEMAT), Programa de Apoio à Núcleos de Excelência (PRONEX/CNPq), process no. 838265/2009 (Prof. Dr. Marinez Isaac Marques), provided computational equipment. Roseneide Soares kindly provided her immense administrative expertise during our 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. The audio data used here were downloaded from xeno-canto (www.xeno-canto.org) and form part of the recordings provided in the scope of the competition BirdCLEF2014.

References

Acevedo, M. A., Corrada-Bravo, C. J., Corrada-Bravo, H., Villanueva-Rivera, L. J., & Aide, T. M. (2009). Automated classification of bird and amphibian calls using machine learning: a comparison of methods. *Ecological Informatics, 4*(4), 206–214. doi:10.1016/j.ecoinf.2009.06.005.

Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G., & Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. *PeerJ, 1*, e103. doi:10.7717/peerj.103.

Alam, J., Kenny, P., Ouellet, P., Stafylakis, T., & Dumouchel, P. (2014). Supervised/unsupervised voice activity detectors for text-dependent speaker recognition on the rsr2015 corpus. In *Odyssey Speaker and Language Recognition Workshop*.

Bardeli, R. (2009). Similarity search in animal sound databases. *IEEE Transactions on Multimedia, 11*(1), 68–76. doi:10.1109/TMM.2008.2008920.

Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K.-H., & Frommolt, K.-H. (2010). Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters, 31*(12), 1524–1534. doi:10.1016/j.patrec.2009.09.014.

Bibby, C. J., Burgess, N. D., Hill, D. A., & Mustoe, S. H. (2000). *Bird Census Techniques* (2nd ed). London: Academic Press.

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., et al. (2004). A tutorial on text-independent speaker verification. *EURASIP Journal on Advances in Signal Processing, 2004*(4), 430–451. doi:10.1155/S1110865704310024.

Bovik, A. (2005). *Handbook of image and video processing* (2nd ed.). Burlington: Elsevier Academic Press.

Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X. Z., Raich, R., Hadley, S. J. K., et al. (2012). Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. *The Journal of the Acoustical Society of America, 131*(6), 4640–4650. doi:10.1121/1.4707424.

Cadore, J., Gallardo-Antoln, A., & Pelez-Moreno, C. (2011). Morphological processing of spectrograms for speech enhancement. *Advances in Nonlinear Speech Processing, 7015*, 224–231. doi:10.1007/978-3-642-25020-0_29.

Chou, C-H., Lee, C-H., & Ni, H-W. (2007). Bird species recognition by comparing the HMMs of syllabes. In *2nd International Conference on Innovative Computing, Information and Control (ICICIC'07)* (pp. 143–147). doi:10.1109/ICICIC.2007.199.

Chou, C-H., & Liu, P-H. (2009). Bird species recognition by wavelet transformation of a section of birdsong. In *Proceedings of the Symposia and Workshop on Ubiquitous, Autonomic and Trusted Computation* (pp. 189–193). doi:10.1109/UIC-ATC.2009.85.

Chou, C.-H., Liu, P.-H., & Cai, B. (2008). On the studies of syllable segmentation and improving MFCCs for automatic birdsong recognition. In *Asia-Pacic Services Computing Conference* (pp. 745–750). doi:10.1109/APSCC.2008.6.

Chu, W., & Blumstein, D. T. (2011). Noise robust bird song detection using syllable pattern-based hidden Markov models. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 345–348). doi:10.1109/ICASSP.2011.5946411.

Dufour, O., Artieres, T., Glotin, H., & Giraudet, P. (2013). Clusterized mel filter cepstral coefficients and support vector machines for bird song identification. In *1st Workshop on Machine Learning for Bioacoustics*. doi:10.5772/56872.

Evangelista, T. L. F., Priolli, T. M., Silla, C. N., Jr., Angelico, B. A., & Kaestner, C. A. A. (2014). Automatic Segmentation of Audio Signals for Bird Species Identification. In *IEEE International Symposium on Multimedia* (pp. 223–228). doi:10.1109/ISM.2014.46.

Ganchev, T., Jahn, O., Marques, M. I., de Figueiredo, J. M., & Schuchmann, K.-L. (2015). Automated acoustic detection of *Vanellus chilensis lampronotus*. *Expert Systems with Applications, 42*(15–16), 6098–6111 ISSN 0957-4174 http://dx.doi.org/10.1016/j.eswa.2015.03.036 .

Härmä, A. (2003). Automatic identification of bird species based on sinusoidal modeling of syllables. *Acoustics, Speech, and Signal Processing, 5*, 545–548 V. doi:10.1109/ICASSP.2003.1200027.

Henríquez, A., Alonso, J. B., Travieso, C. M., Herrera, B. R., Bolaños, F., Alpízar, P., et al. (2014). An automatic acoustic bat identification system based on the audible spectrum. *Expert Systems with Applications, 41*(11), 5451–5465 ISSN 0957-4174. doi:10.1016/j.eswa.2014.02.021.

Huang, C.-J., Yang, Y. J., Yang, D.-X., & Chen, Y.-J. (2009). Frog classification using machine learning techniques. *Expert Systems with Applications, 36*(2), 3737–3743. doi:10.1016/j.eswa.2008.02.059.

Jahn, O. (2011a). Surveying tropical bird communities: in search of an appropriate rapid assessment method. In K.-L. Schuchmann (Ed.), *Bird Communities of the Ecuadorian Chocó: A Case Study in Conservation* (pp. 63–107). Bonn, Germany: A. Koenig (ZFMK) Bonner zoologische Monographien 56. Zoological Research Museum URL http://zoologicalbulletin.de/BzB_Volumes/BzM_56/BZM_56.pdf last accessed 25 June 2015.

Jahn, O. (2011b). Structure and organization of the bird community. In K.-L. Schuchmann (Ed.), *Bird Communities of the Ecuadorian Chocó: A Case Study in Conservation* (pp. 109–178). Bonn, Germany: A. Koenig (ZFMK) Bonner zoologische Monographien 56. Zoological Research Museum URL http://zoologicalbulletin.de/BzB_Volumes/BzM_56/BZM_56.pdf last accessed 25 June 2015.

Juang, C.-F., & Chen, T.-M. (2007). Birdsong recognition using prediction-based recurrent neural fuzzy networks. *Neurocomputing, 71*(1–3), 121–130. doi:10.1016/j.neucom.2007.08.011.

Kaewtip, K., Tan, L. N., Alwan, A., & Taylor, C. E. (2013). A robust automatic bird phrase classifier using dynamic time-warping with prominent region identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 768–772). doi:10.1109/ICASSP.2013.6637752.

Lee, C.-H., Chou, C.-H., Han, C.-C., & Huang, R.-Z. (2006). Automatic recognition of animal vocalizations using averaged {MFCC} and linear discriminant analysis. *Pattern Recognition Letters, 27*(2), 93–101. doi:10.1016/j.patrec.2005.07.004.

Lee, C-H., Han, C-C., & Chuang, C-C. (2008). Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(8), 1541–1550. doi:10.1109/TASL.2008.2005345.

Mamiya, Y., Yamagishi, J., Watts, O., Clark, R., King, S., & Stan, A. (2013). Lightly supervised GMM VAD to use audiobook for speech synthesiser. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7987–7991). doi:10.1109/ICASSP.2013.6639220.

Neal, L., Briggs, F., Raich, R., & Fern, X. Z. (2011). Time-frequency segmentation of bird song in noisy acoustic environments. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2012–2015). doi:10.1109/ICASSP.2011.5946906.

Oba, T. (2004). Application of automated bioacoustic identification in environmental education and assessment. *Anais da Academia Brasileira de Ciências, 76*(2), 446–451. doi:10.1590/S0001-37652004000200039.

de Oliveira, A. G., Ventura, T. M., Ganchev, T. D., de Figueiredo, J. M., Jahn, O., Marques, M. I., et al. (2015). Bird acoustic activity detection based on morphological filtering of the spectrogram. *Applied Acoustics, 98*(2015), 34–42 ISSN 0003-682X http://dx.doi.org/10.1016/j.apacoust.2015.04.014 .

Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H. G., Scholes, R. J., et al. (2013). Essential biodiversity variables. *Science, 339*(6117), 277–278. doi:10.1126/science.1229931.

Potamitis, I. (2014). Automatic classification of a taxon-rich community recorded in the wild. *PLoS ONE, 9*(5). doi:10.1371/journal.pone.0096936.

Potamitis, I., Ntalampiras, N., Jahn, O., & Riede, K. (2014). Automatic bird sound detection in long real-field recordings: applications and tools. *Applied Acoustics, 80*, 1–9. doi:10.1016/j.apacoust.2014.01.001.

Sahidullah, M., & Saha, G. (2012). Comparison of speech activity detection techniques for speaker recognition. arXiv preprint, 1–7.

Schuchmann, K.-L., Marques, M. I., Jahn, O., Ganchev, T., & Figueiredo, J. M. (2014). Os Sons do Pantanal: Um projeto de monitoramento acústico automatizado da biodiversidade. *Boletim Informativo Sociedade Brasileira de Zoologia, 108*, 11–12.

Stowell, D., & Plumbley, M. D. (2014). Audio-only bird classification using unsupervised feature learning. In *Working Notes for CLEF 2014 Conference* (pp. 673–684).

Sueur, J., Pavoine, S., Hamerlynck, O., & Duvail, S. (2008). Rapid acoustic survey for biodiversity appraisal. *PLoS ONE, 3*(12). doi:10.1371/journal.pone.0004065.

Trifa, V. M., Kirschel, A. N., Taylor, C. E., & Vallejo, E. E. (2008). Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. *The Journal of the Acoustical Society of America, 123*(4), 2424–2431. doi:10.1121/1.2839017.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., et al. (2006). *The HTK book (for HTK Version 3.4)*. Cambridge University Engineering Department.

Zhang, X., & Li, Y. (2013). Environmental sound recognition using double-level energy detection. *Journal of Signal and Information Processing, 4*(3B), 19–24. doi:10.4236/jsip.2013.43B004.