Expert Systems With Applications xxx (2015) xxx–xxx

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

Q1 Thiago M. Ventura[b,c], Allan G. de Oliveira[a,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 crickets, 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., Sueur et al., 2008, Aide et al., 2013, Potamitis, Ntalampiras, Jahn, & Riede, 2014; Ganchev, Jahn, Marques, de Figueiredo, & Schuchmann, 2015). Q3 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 of several terabytes per continuous annual cycle of recordings.

Q2 * Corresponding author at: Department of Electronics, Technical University of Varna, str Studentska 19010, 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/© 2015 Elsevier Ltd. All rights reserved.

Please cite this article as: T.M. Ventura et al., Audio parameterization with robust frame selection for improved bird identification, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.07.002

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 are 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 (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 spectrogram of the preprocessed time-domain signal $s(n)$ is computed through the short-time discrete Fourier

Please cite this article as: T.M. Ventura et al., Audio parameterization with robust frame selection for improved bird identification, Expert Systems With Applications (2015), http://dx.doi.org/10.1016/j.eswa.2015.07.002