Acoustic Bird Activity Detection on Real-Field Data

Todor Ganchev[1], Iosif Mporas[1], Olaf Jahn[2], Klaus Riede[2], Karl-L. Schuchmann[2], and Nikos Fakotakis[1]

[1] Artificial Intelligence Group, Wire Communications Laboratory
Dept. of Electrical and Computer Engineering, University of Patras, 26500 Patras, Greece
[2] Zoologisches Forschungsmuseum Alexander Koenig 53113 Bonn, Germany
tganchev@upatras.gr

Abstract. We report on a research effort aiming at the development of an acoustic bird activity detector (ABAD), which plays an important role for automating traditional biodiversity assessment studies – presently performed by human experts. The proposed on-line ABAD is considered an integral part of an automated system for acoustic identification of bird species, which is currently under development. In particular, taking advantage of real-field audio recordings collected in the Hymettus Mountains east of Athens, we investigate the applicability of various machine learning techniques for the needs of our ABAD, which is intended to run on a mobile device. Performance is reported in terms of recognition accuracy on audio-frame level, due to the restrictions imposed by the requirement of run-time decision making with limited memory and energy resources. We report recognition accuracy of approximately 86% on a frame level, which is quite promising and encourages further research efforts in that direction.

Keywords: acoustic bird activity detection, bioacoustics, biodiversity surveys, real-field data.

1 Introduction

At present biodiversity inventories and monitoring studies are typically performed by expert biologists, who have to visit (periodically) sites and habitats of interest to conduct audiovisual, capture-recapture, or collection surveys. This is a time-consuming and costly task, which, due to multiple reasons, cannot be performed continuously and systematically for extended periods of time. Therefore, even a partial automation of the data collection and analysis procedures are considered to be important for developing future biodiversity assessment approaches.

Birds are an important indicator of the conservation status of habitats and landscapes as well as a proxy for biodiversity patterns. Thus, the detection of the presence and the estimation of population trends and reproductive success of certain bird species groups are of significant importance as they offer a general measurement of the health of an ecosystem [1].

I. Maglogiannis, V. Plagianakos, and I. Vlahavas (Eds.): SETN 2012, LNAI 7297, pp. 190–197, 2012.

As birds are heard more often than seen, one promising non-intrusive method for monitoring their presence and activity is the acoustic detection and identification of avian taxa. In the present work, we focus on investigating the feasibility of automatic acoustic detection of bird vocalizations from real-field audio recordings and evaluate the recognition accuracy of the proposed ABAD, when implemented with different classifiers.

In the present paper, all sounds of non-bird origin, e.g. human- or machine-made sounds, sounds due to natural phenomena (e.g., wind and rain), sounds from other animals or unanimated objects co-existing in that environment, are collectively referred to as *background audio* or *noise*. Next, all sounds of bird origin that can be distinguished from the audio background by a human listener are collectively labeled as *bird vocalizations*, regardless of the coexistence of background interference.

2 Acoustic Bird Activity Detection in Real-Field Environment

The acoustic bird activity detector serves as a gateway, which aims to eliminate from the input audio stream these portions of the signal that correspond to sounds of non-bird origin. Thus, the ABAD excludes from storing or passing to the consequent processing stages, such as species identification, the silence intervals and any non-bird sounds, but passes through unaltered these portions of the audio which were recognized as bird vocalizations.

We aimed at an efficient design with respect to computation and memory, and by using frame-by-frame detections of the presence or absence of bird vocalizations in the input audio stream or with at minimal delay (Fig. 1). The acoustic bird activity detection process consists of three main stages *audio acquisition*, *audio parameterization*, and *pattern recognition*. While the audio parameterization step aims at computing descriptors, which capture the generalized acoustic properties of bird vocalizations, the pattern recognition step categorizes the current input audio frame either as *bird vocalization* or as *background noise*. Depending on the machine learning technique employed, this stage either estimates the degree of match between an unknown input signal and the pre-computed general models for the bird vocalizations and the background acoustic environment or, alternatively, makes decision without using any explicit modeling of the class-specific distributions. Lastly, after some post-processing of the binary decisions (or the scores) obtained for the current audio frame, a final decision is made with respect to a predefined threshold: either the current audio frame contains a bird vocalization or not. In the following, we briefly outline the consequent steps of signal acquisition, pre-processing, parameterization, and classification:

Audio acquisition: Audio is captured by a microphone, next amplified and then sampled at 32 kHz, so that the wide frequency range of bird vocalizations from various species is covered. Precision of 16-bits per sample is used to guarantee sufficient resolution of details for the subsequent processing of the signal.

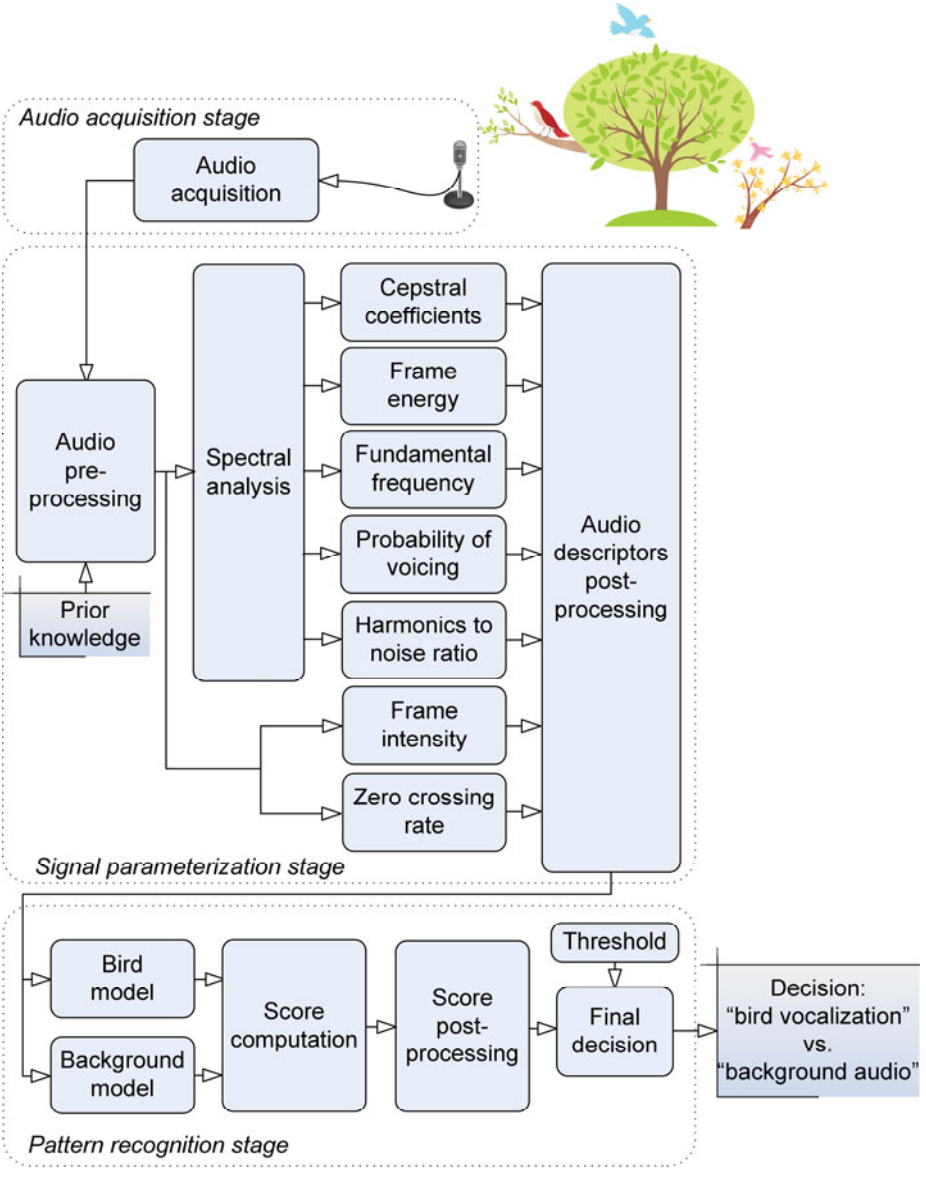

Fig. 1. Block diagram of the acoustic bird activity detection process (see text for details)

Audio Pre-processing: The pre-processing of the input audio stream consists of mean value removal, which is performed on the time domain signal, for eliminating the dc-offset that might have occurred during signal acquisition and amplification. Furthermore, based on prior knowledge we assume that, for most avian species, there is no useful information characterizing the bird vocalizations in the audio signal for the frequency band below 400 Hz. Thus, in order to reduce the influence of any

environmental noise and low-frequency interferences, such as wind, vibrations of nearby objects, traffic noise, a high-pass filtering is applied on the amplified audio signal. The high-pass filtering with a low-order filter was found to provide a reasonable trade-off between computational demands and improvement of recognition accuracy, as it requires significantly fewer computational and memory resources when compared to contemporary noise reduction methods, such as those discussed in [2]. Specifically, it was experimentally found that a Butterworth filter of order 6 with cut-off frequency 400 Hz effectively reduces the low-frequency noise and improves the overall recognition accuracy. The audio pre-processing also includes Hamming windowing of the signal, and thus all consequent processing is performed on a frame level. In the following we assume audio frame size of 20 ms and skip step of 10 ms.

Audio Parameterization: Previous work on acoustic bird activity detection from real-world data reported the importance of tonal audio features [3] and their advantage over the traditional Mel-frequency cepstral coefficients (MFCC) in noisy conditions. In the present work, we make use of a more diverse set of audio parameters that facilitates the robust detection of bird vocalizations in non-stationary noise environments. In particular, each audio segment obtained after the audio pre-processing step is zero-padded to 1024 samples and then becomes subject to the audio parameterization procedure. Specifically, we compute two types of complementary audio descriptors: temporal (zero crossing rate, frame intensity) and spectral (MFCC, frame energy, fundamental frequency, probability of voicing, and the harmonics-to-noise ratio). These audio descriptors have been successfully used in audio processing and sound classification tasks, and offer improved robustness in noisy conditions. In the present work, we computed all audio parameters via the openSMILE acoustic parameterization tool [4]. In particular, for each audio frame we computed 12 MFCCs following the default HTK setup [5], the root mean square energy of the frame (*E*), the voicing probability (*Vp*), the harmonics-to-noise ratio (*HNR*) by autocorrelation function, the dominant frequency (*Fd*) normalized to 500 Hz, the intensity (*Int*) and the zero crossing rate (*ZCR*). Stacking together these audio parameters results in a feature vector of 18 audio features. Post-processing for dynamic-range normalization was applied to all audio features for equalizing the range of their numerical values.

A series of feature-ranking and selection tests have shown that the abovementioned *static* audio features are sufficient to carry out the recognition task and that appending to the feature vector their first and second time derivatives contributes less to improving the overall recognition accuracy. Yet, appending the time derivatives to the static features increases significantly the length of the feature vector, and thus, the demand of training data needed for robust model development, but also the computational demands.

Pattern Recognition Stage: The audio features obtained to this end are fed to a binary classifier which is trained to discriminate between the *bird vocalizations* and the *background* acoustic noise. Depending on the machine learning method employed, the decisions obtained (or alternatively the scores computed) for each audio frame are next post-processed. This post-processing aims at eliminating sporadic erroneous labeling of the current audio frame, e.g., due to momentary burst of interference, and

thus contributes to the improvement of the overall recognition accuracy. A simple and computationally effective rule for post-processing is smoothing each decision (or score) with respect to its closest temporal neighbors. In particular, when the previous N neighbor audio frames and the following N neighbor frames were recognized as bird vocalization then the current frame is also (re)labeled as bird vocalization. Likewise are processed the frames whose neighbors were recognized as category background audio. The length w of the smoothing window is subject to investigation and in the general case is equal to $w = 2N + 1$, where $N \geq 0$. The case $N = 0$ corresponds to eliminating the post-processing of the recognized labels, while the cases $N = 1, 2, 3$ correspond to window size $w = 3, 5, 7$.

Eventually a final decision about the label of the current audio frame (either *bird vocalization* or *background noise*) is made after applying a predefined threshold on the post-processed scores. This threshold controls the 'sensitivity' of the ABAD and allows for some trading-off of false alarm errors vs. target miss errors, and thus allows for fine-tuning the operational mode and the gating properties of the ABAD. Furthermore, the choice of threshold levels directly affects the amount of the audio data fed to the subsequent audio processing steps.

3 Evaluation Setup and Results

In the following subsections, we describe the evaluation dataset, the experimental setup, and discuss the experimental results.

3.1 Real-World Dataset

The dataset used in the present research is a small excerpt from our collection of audio recordings, obtained in the Hymettus Mountains west of Athens, Greece. The recordings have been manually tagged (labels: bird vocalization vs. background audio) by an engineer with considerable experience in the area of audio processing. The training data representing the category *bird vocalizations* consists of approximately 6 minutes of concatenated bird vocalizations (35636 audio frames) extracted from 50 audio files, each with average duration of approximately 30 seconds. The training data for the category *background audio* was represented by a similar amount of audio, extracted from the same 50 files, and contains environmental sounds typical for our study area.

The test dataset consisted of another subset of 150 files (271024 audio frames), which were processed by each of the machine learning algorithms outlined in the next subsection.

3.2 Experimental Setup

We investigated the applicability of various machine-learning techniques, for the implementation of the binary classifier. Classifiers belonging to different categories of algorithms were selected:

- *k*-nearest neighbors classifier with linear search of the nearest neighbor and without weighting of the distance – also known as instance based classifier (*IBk*) [6],
- Bayes network (*BayesNet*) [7], with Simple Estimator (alpha = 0.5) and the K2 search algorithm (maximum number of parents = 1);
- 3-layer Multilayer perceptron (*MLP*) neural network [8] with architecture 18–10–1 neurons (all sigmoid) trained with 50 000 iterations;
- Pruned C4.5 decision tree (*J48*), with 3 folds for pruning and 7 for growing the tree [9];
- Support vector machine with sequential minimal optimization (*SMO*) algorithm [10] and RBF kernel [11].

We made use of the Weka [12] implementations of these algorithms with the default values of all parameters, which are not specified here.

A common experimental protocol was followed during the evaluation of all classifiers. Specifically, the ABAD implemented with different binary classifiers were treated uniformly and were trained with the dataset outlined in Section 3.1. The recognition accuracy of the ABAD was evaluated on audio-frame level in terms of percentages of correct detections on the test dataset specified in Section 3.1.

3.3 Experimental Results

The ranking of the machine-learning methods was made on the basis of their classification accuracy for the case of no post-processing of the frame decisions ($N = 0$) (Table 1). Specifically, the ABAD implementation based on the Multilayer Perceptron (*MLP*) neural network demonstrated the highest recognition accuracy, followed by the Support Vector Machine with sequential minimal optimization (*SMO*), the *k*-nearest neighbors classifier (*IBk*), the decision tree (*J48*), and finally the Bayes Network (*BayesNet*).

Table 1. Recognition accuracy (reported in percentages) for the acoustic bird activity detection implemented with various machine-learning techniques and different length, *w*, of the smoothing window

Binary Classifier	*w*=1 (*N*=0)	*w*=3 (*N*=1)	*w*=5 (*N*=2)	*w*=7 (*N*=3)
MLP	**85.0**	**86.3**	**86.1**	**86.0**
SMO	79.1	81.4	81.0	80.8
IBk	78.5	82.9	82.1	81.4
J48	74.9	79.1	78.0	77.2
BayesNet	64.9	65.8	65.6	65.5

This ranking is not surprising considering the limited amount of data used for training the binary classifiers, as well as the fact that some of the audio features in the feature vector are correlated to a certain degree. For instance the frame energy (*E*) is

correlated with the frame intensity (*Int*), and the voicing probability (*Vp*) is correlated to the harmonics-to-noise ratio (*HNR*). However, it was observed that this redundancy in the feature vector, in fact, contributes to the improvement of the robustness to noise. One explanation for this effect could be that since the audio features in these pairs are computed in dissimilar manner, they are affected in different ways by the interference. Thus, in noisy conditions, depending on the noise type, these audio features complement each other, and thus contribute to the overall improvement of the noise robustness.

We want to emphasize that smoothing the scores (or the decisions) of the binary classifier is beneficial only when the closest temporal neighbors (corresponding to the previous and next audio frame) are used ($N = 1$) but this advantage is reduced when the decisions for the more distant neighbors ($N = 2, 3$) are accounted for.

The highest overall recognition accuracy of 86.3% was obtained for the *MLP* classifier in combination with post-processing with a smoothing window, $w = 3$, i.e., when the scores of the two closest neighbor frames (previous and next) are used. In practice, this smoothing scheme requires delaying the final decisions of the ABAD with one time-step with respect to the decisions made by the binary classifier. In our setup this delay is 10 ms, as the overlapping between two consecutive audio frames is 10 ms.

The relatively low absolute value of the recognition accuracy, 86.3%, can be explained with the non-stationary and fully uncontrolled conditions in our real-field environment (Hymettus Mountains), where interferences resulting from human presence and activities and from natural phenomena, such as wind and rain, are quite common and co-occur in time and space with the bird vocalizations. Nevertheless, when more annotated real-field data become available, we intend to experiment with more advanced statistical modeling techniques which explicitly address the intra-class distribution of the data for each class, which is expected to reduce the false acceptance rates.

In conclusion, it is worth noticing that the ABAD implemented with *MLP*-based binary classifier with architecture 18-10-1 is quite compact. This makes the ABAD computationally inexpensive, and less memory demanding, in contrast to the other implementations, with the instance based *k*-nearest neighbor classifier, or with the *SMO* (with RBF kernel) classifier. All these make the ABAD implementation with the *MLP* classifier followed by smoothing window, $w = 3$, quite suitable for porting as an "*App*" on a contemporary handheld mobile device.

Acknowledgements. The research reported in the present paper was supported by the AmiBio project (LIFE08 NAT/GR/000539), which is implemented with the contribution of the LIFE+ financial instrument of the European Union (project web-site: www.amibio-project.eu).

The authors wish to acknowledge the contribution of Mr. Stavros Ntalampiras, and Mr. Theodoros Kostoulas from the University of Patras and also to the entire team of the Association for the Protection and Development of Hymettus (SPAY), who supported the implementation of the audio data collection campaign in the Hymettus area.

References

1. Dawson, D.K., Efford, M.G.: Bird population density estimated from acoustic signals. Journal of Applied Ecology 46, 1201–1209 (2009)
2. Loizou, P.: Speech Enhancement: Theory and Practice. CRC Press (2007)
3. Jančovič, P., Köküer, M.: Automatic detection and recognition of tonal bird sounds in noisy environments. EURASIP Journal on Advances in Signal Processing 2011, Article ID 982936, 10 (2011), doi:10.1155/2011/982936
4. Eyben, F., Wöllmer, M., Schuller, B.: OpenEAR - introducing the Munich open-source emotion and affect recognition toolkit. In: Proc. of the 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction, ACII 2009 (2009)
5. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK book (for HTK Version 3.4), Cambridge University Engineering Department (2006)
6. Aha, D., Kibler, D.: Instance-based learning algorithms. Machine Learning 6, 37–66 (1991)
7. Bouckaert, R.R.: Bayesian networks in Weka. Technical Report 14/2004. Computer Science Department. University of Waikato (2004)
8. Chester, D.L.: Why two hidden layers are better than one. In: Proc. of the International Joint Conference on Neural Networks, vol. 1, pp. 265–268 (1990)
9. Quinlan, R.: C4.5: Programs for machine learning. Morgan Kaufmann Publishers, San Mateo (1993)
10. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation 13(3), 637–649 (2001)
11. Scholkopf, B., Smola, A.J.: Learning with Kernels. MIT Press (2002)
12. Witten, H.I., Frank, E.: Data Mining: practical machine learning tools and techniques. Morgan Kaufmann Publishing (2005)