ISEIS

Journal of Environmental Informatics Letters 2(1) 19-27 (2019)

Journal of Environmental Informatics Letters

www.iseis.org/jeil

Machine Learning for Identifying an Endangered Brazilian Psittacidae Species

B. T. Padovese[1 *], L. R. Padovese[2]

[1] *Rua Gomes de Medeiros, 190, Sao Paulo 05447-030, Brazil*
[2] *Mechanical Engineering Department, Polytechnic School, University of Sao Paulo, Sao Paulo 05508-030, Brazil*

Received 1 June 2019; revised 15 July 2019; accepted 22 July 2019; published online 30 September 2019

ABSTRACT. Bird population census is an important indicator in conservation programs. However, the process of detecting and identifying particular species is time-consuming and challenging, often being conducted in remote locations. In this scenario, the development of automated acoustic systems for bird monitoring is crucial. In this study, we propose a simple but effective 3-step approach for identifying the *Amazona rhodocorytha*, an endangered Brazilian parrot, among 4 other species belonging to the same family. This approach consists of a pre-processing step, a feature extraction step using the MFCC algorithm and a classification step by employing an Artificial Neural Network. Results show that the proposed approach is both suitable and robust for this type of application, achieving excellent classification results of up to 98% accuracy.

Keywords: machine learning, neural networks, bird detection, MFCC, bird identification

1. Introduction

As the impact of anthropogenic activities on the environment continue to grow, so does scientific interest in detecting and measuring trends in nature conservation. Intensive research and huge amounts of investments in several fields have been applied in order to mitigate, study, and gauge further changes (Rosenzweig and Parry, 2004; Edenhofer, 2015; Simmonds et al., 2007; Stephens et al., 2016). As shown by several studies (Koskimies, 1988; Gregory, 2009; Stephens et al., 2016), changes in bird population size are one indication of the impact of human activity in the environment.

Several animal species are very dependent on sound for their activities. Among others, one can cite bats (Moss and Schnitzler, 1989), birds (Johnson et al., 2002; Selin et al., 2006; Brandes, 2008; Dawson et al., 2009; Lopes et al., 2011), whales (Payne et al., 1971) and dolphins (Au, 2012). Birds, particularly, can often be characterized and identified by their high-energy and diverse vocalizations. Hence, by recording and analyzing bird acoustic signals (Brandes, 2008; Dawson and Efford, 2009; Lopes et al., 2011) in predetermined areas, the identification of species and their population sizes can be evaluated. However, apart from the difficulty of determining population size, the identification of species that are endemic to a certain region faces several further challenges.

Identifying a specific bird song among a collection of other vocalizations and acoustic signals is a very difficult task, even for a trained human operator. Furthermore, if the population density is low, as in an endangered species case, detecting its presence is a labor-intensive and time-consuming process (Bardeli et al., 2010; Sebastián-González, 2015). In addition, these processes often occur in remote areas where supporting resources are limited, further complicating the process.

Development of autonomous low-cost recorder units, capable of passive acoustic monitoring (PAM), has aided researchers by providing an effective way to monitor remote areas (Johnson et al., 2002; Blumstein, 2011). These recorders allow the establishment of large monitoring areas for long periods of time, without the need of constant human interaction. Thus, up to years of recording uninterrupted acoustic data can be collected. However, the massive amount of data collected this way makes it nearly impossible to study and analyze these recordings (Sebastián-González, 2015) effectively and reliably without the aid of artificial intelligence procedures.

Apart from the volume of data to be analyzed, the main recurring challenge, reported by most studies in the field of bioacoustics (Johnson, 2002; Selin et al., 2006; Brandes, 2008; Jaffar et al., 2013), also includes the chaotic nature of the data and the access to remote locations, where resources and manpower are scarce. Methodologies of study often focus on a solution that is not only both robust and flexible, but also simple to be implemented in these kinds of scenarios.

Therefore, the incorporation of automated methods to process, classify and extract valuable information from huge collections of acoustic datasets becomes necessary. Along the past decades, enormous progress in the field of pattern recognition, signal processing and machine learning have enabled efficient analysis of big datasets with methods that are widely used in

* Corresponding author. Tel.: +1 (902) 402 8326.
E-mail address: bpadovese@gmail.com (B. T. Padovese).

ISSN: 2663-6859 print/2663-6867 online
 doi:10.3808/jeil.201900013

both the scientific sector and the industry (Cortes and Vapnik, 1995; Yegnanarayana, 2009). Particularly, for animal detection and identification, machine-learning models have been trained in order to detect and identify specific species such as amphibians (Jaffar et al., 2013), birds (Johnson, 2002; Selin, 2006; Brandes, 2008; Lopes et al., 2011) and whales (Pace, 2008).

In this context of big data, machine learning models can be employed alongside autonomous recording units, creating fully automated systems. These systems are able to support researchers and Non-Governmental Organizations (NGOs) in the detection and identification of different animal species and population census.

For example, concerning birds, the study presents an overview of automated sound recordings (Brandes, 2008). The authors expose a list of challenges and motivations to perform bird survey analysis, as well as a description of several hardware designs and concepts in order to conduct PAM. This study includes a discussion about possible microphones arrays settings, timers for scheduling purposes, embedded systems, and even smart phones. Lastly, an analysis of automated methods for processing the collected recordings is presented. The analysis also mentions a list of the most used types of features extracted to conduct a classification task, as well as the classifiers used along with it, such as Multilayer Perceptron and Bayesian classifiers.

In contrast, Jaffar et al. (2013) pursue the identification of frog vocalization by applying an automatic syllable segmentation method along with the k-nearest neighbors (kNN) classifier. The methodology can be described in 4 steps. In the first step, the input acoustic signal is segmented by the proposed method into a set of syllables. In the next step, each syllable is subjected to pre-emphasizing, framing and windowing. Then, features are extracted by the Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) algorithms. Finally, in the last step a model is trained with the kNN classifier.

Johnson et al. (2002) discusses an approach for monitoring nocturnal avian vocalization, such as for owls. While diurnal survey already presents its series of challenges, the reduced visibility makes it even harder for manual survey. Therefore, the authors discuss a technique for optimizing both time and resources for nocturnal monitoring, while also presenting challenges that still need to be taken into account, such as hardware failure in a remote location.

In another study, Pace (2008) compare feature extraction methods such as MFCC, LPC and real cepstrum coefficients, for humpback whale song classification. These feature extraction methods are then combined with a clustering algorithm, namely k-means clustering, and Artificial Neural Networks (ANN).

The work presented by Lopes et al. (2011) focuses on the automatic identification of numerous species from the Southern Atlantic Brazilian Coast. The recordings used in this study were collected from two Datasets and passed through a series of preprocessing techniques to enhance the quality of the recordings. Next, the authors employ a feature extraction step using the MARSYAS framework and compare different classification techniques. These are the kNN, MLP, Naïve Bayes, and SVM. Obtained results showed a clear advantage of the MLP and SVM methods for the proposed scenario.

In its turn, Selin et al. (2006) describes an approach using wavelets decomposition as the feature extraction method. As with many of the other methods already discussed, a first step of pre-processing, consisting of noise reduction and segmentation, was carried out. Then, the proposed wavelet approach of feature extraction was conducted, where four features were extracted and fed to two types of Neural Networks, the unsupervised self-organizing map (SOM) and the MLP. The results show that the MLP achieved up to 96% accuracy while the SOM 78%.

Priyadarshani et al. (2018) presents a literature review of the state-of-the-art in birdsong recognizers, summarizing and discussing currently available methods, as well as available software. Furthermore, the authors discuss and review studies in all stages of a birdsong classifier such as signal segmentation, call detection, noise reduction methods, feature extraction steps and classification methods and software. Additionally, performance measures for all methods are given in the form of accuracy, precision, recall, and F-score.

A fully automated real-time bird sound recognition system is proposed by Küçüktopcu et al. (2019) using a low-cost, low-level microcontroller capable of simultaneously on-board recording and signal processing. Due to the limited processing power and memory of the hardware, the most intensive processing tasks such as training and cross-validating a classification algorithm are perfomed off-line, that is, apart from the actual system implementation. The system implementation has 6 stages, comprised of sampling and recording; noise removal; detection of sound parts in segments; feature extraction; classification; and storage of the classification results. Results achieved up to 83% accuracy with the minimum distance classifier, however, the authors also experimented off-line with multilayer neural networks and convolutional neural networks (CNN) with up to 93% and 96% accuracy respectively. It is worth noting that recent released libraries by ARM allows the use of these more complex algorithms in some of their microcontrollers.

To sum up, across the literature, in what concerns automated bird recognition, it is possible to identify two major challenges that are still relevant today: the difficulty to access remote areas, in order to conduct the necessary survey, and the huge datasets to analyze and classify. These challenges are usually overcome by using a combination of remote sensors, to automatically collect vocalizations, and an automatic classification model to detect and classify the collected data. In this context, several types of feature extraction methods and classification methods are described. A clear pattern that can be observed is the use of MFCC as a feature extraction method and the use of a couple of classification methods, particularly the MLP. The constant use of this method is due to the excellent results it provides, often better than other methods, as well as to its simple implementation in a multitude of different scenarios.

Finally, adding another contribution to the subject of environment conservation, the present work proposes the use of a

simple, but effective machine learning method, in order to identify the *Amazona rhodocorytha*, an endangered species endemic of Brazil. To our knowledge, there are very few studies tackling automatic identification of endangered Brazilian bird species, and no papers concerning the Psittacidea family specifically, of which several endangered sub-species exist. This work provides a first look into this identification task. Although the present study focuses on a specific Brazilian bird, the proposed methodology could be applied in recognizing other species. Additionally, we collected and worked on labeled data from 2 public avian datasets, the international Xeno-canto and Brazilian Wikiaves datasets. To our knowledge, no work of this context exists for the Wikiaves dataset, and few for the Xeno-canto collection (Vellinga, 2015).

The red-browed parrot (*Amazona rhodocorytha*), from the Psittacidae family, is a species of parrot that can be found in the Atlantic Forest in Brazil. This species is a prime example where anthropogenic activity is affecting its population, since less than 1% of its habitat still remains due to deforestation (BirdLife International, 2017). Nowadays, even though the parrot's remaining habitats are conservation units, this offers little protection from illegal poaching. Indeed, with less than 10,000 estimated individuals remaining, the *Amazona rhodocorytha* is classified as "endangered" by the IUCN (International Union for Conservation of Nature) in its Red List of Threatened Species (BirdLife International, 2017). Besides the difficulties already listed, due to being present in an ecosystem with very high biodiversity, recordings will often be subjected to a high volume of environmental noise present in the forest. Furthermore, this type of bird often shares the same ecosystem as others from the same family, further impeding detection and identification due to their similar vocalizations.

The present study proposes the utilization of a Neural Network for the detection and identification of the *Amazona rhodocorytha* under a real-world soundscape scenario. This will be tackled in combination with a pre-processing step composed of signal segmentation, filtering and feature extraction with the Mel-Frequency Cepstrum Coefficients (MFCC). Additionally, 4 other bird species from the Psittacidae family are considered in this study as to verify the methods capability to handle the difficult task described in the last paragraph.

In summary, this paper describes a method for the detection and classification of an endangered Brazilian bird species. This method is also capable of being easily implemented in systems that are deployed in remote locations. Furthermore, this paper aids researchers and NGOs in handling the difficult task of differentiating a particular species among several other similar acoustic vocalizations.

2. Methodology

In this section, we describe the methodology followed in the present work. Section 2.1 gives an overview of the proposed methodology. Section 2.2 presents the pre-processing steps conducted before a feature extraction step in section 2.3 and finally, Section 2.4 describes the Neural Network Multilayer Perceptron used to train the model.

2.1. Overview

Figure 1 illustrates the overall architecture of the proposed approach. First, a 2-step pre-processing phase is carried out by segmenting the recordings and applying a band-pass filter. Next, a feature extraction step is performed aiming at capturing relevant information to the classification problem. Finally, the resulting data is fed to an Artificial Neural Network to train a classification model for *the Amazona rhodocorytha*. It is worth noting that the dataset used for these steps is described in Section 4.1.

2.2. Pre-Processing

As shown in Figure 2, the collected recordings contain a wide array of acoustic events, some of which are not only not relevant to the problem in question, but may also yield undesirable results when training a classification algorithm. Therefore, we conducted a segmentation process to isolate and retrieve only the time periods where a probable acoustic event from one of the Psittacidae was present. This segmentation process was done in a supervised way, using the Audacity (Audacity, 1999 ~ 2018) software and spectrogram tool for signal visualization. With it, it was possible to recognize the Psittacidae patterns from noise, such as human speech, other animals' vocalizations, or instrumentation handling. It is worth noting that, when there was over-lap of vocalization and noise, we selected only the cases where it was still possible to easily identify the desired signal.

Another problem that is evident from the collected samples is constant background noise, such as cricket sounds shown in Figure 2. These types of noise are present in mostly all the recordings, often overlap with the desired signal, and are detrimental to the classification process. Our proposed solution uses a Butterworth bandpass filter to emphasize only the frequencies that contain the *Amazona rhodocorytha* vocalizations. A band-pass filter can be defined as a filter that isolates the data in a certain frequency band of a time series (Christiano et al., 2003).

In this work, the selected band, from 900 to 5000 Hz, was chosen after closely analyzing the *Amazona rhodocorytha* vocalization pattern. Furthermore, this filter was applied to all segmented recordings, even the ones where a vocalization from another Psittacidae was recorded, since this paper focuses solely on the detection and identification of the *Amazona rhodocorytha*.

2.3. MFCC Feature Extraction

Choosing an appropriate feature to describe each signal is a defining step when building an intelligent acoustic model. The Mel-frequency cepstral coefficients is described as a robust and reliable method, suitable in applications where a lot of noise can be expected (Foote, 1997). Originally developed and used in speech and speaker recognition systems, extensive research has been conducted in a wide range of acoustic applica-

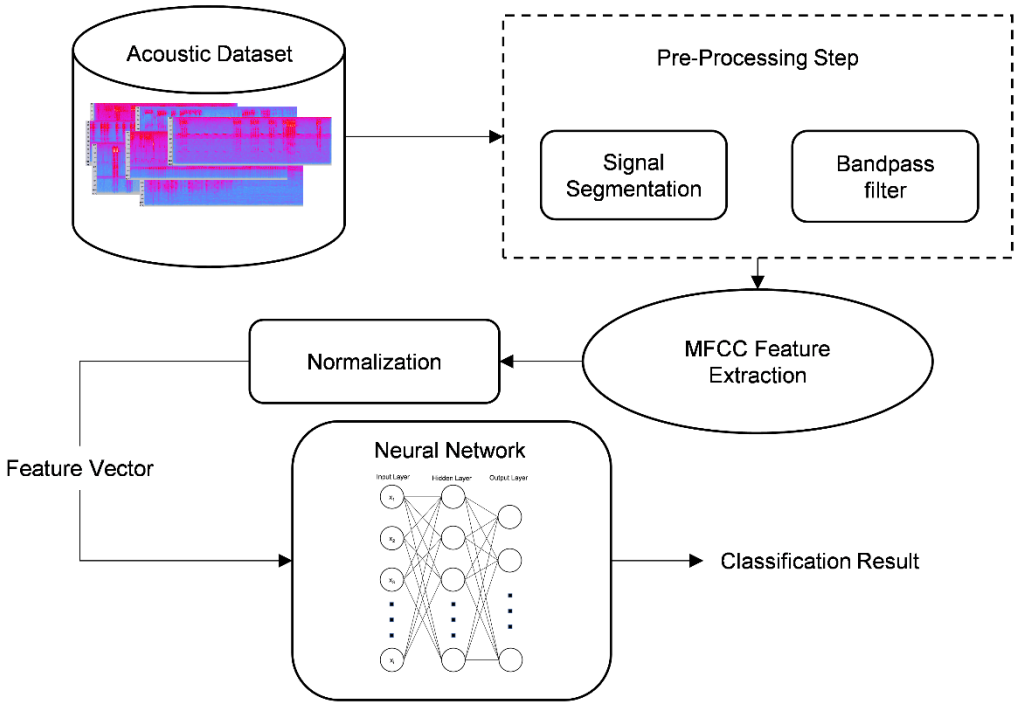

Figure 1. *Amazona rhodocorytha* identification methodology.

tions, including animal vocalization classification (Johnson et al., 2002; Selin et al., 2006; Pace, 2008; Brandes, 2008; Jaffar et al., 2009, 2013; Lopes et al., 2011).

In the present work, due to its flexibility, robustness and ability to be easily implemented in a remote environment, the MFCC feature extraction method was considered in building the feature vector that will feed the MLP.

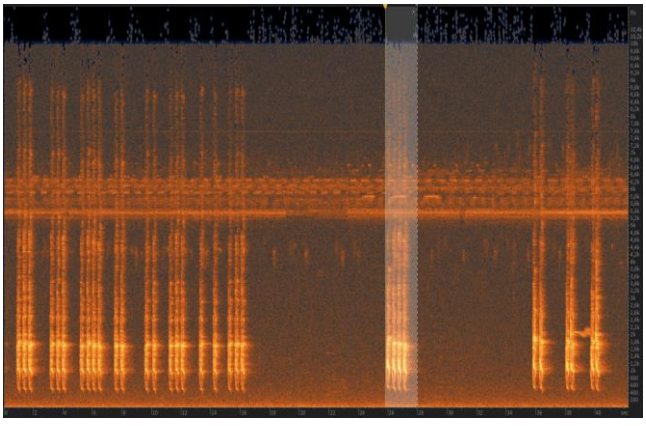

Figure 2. Spectrogram of *Amazona rhodocorytha* vocalization and an example of a selection.

The Mel-frequency scale is used to represent the perceived pitch frequency in a non-linear way, which closely models the human auditory perception. It considers properties such as power spectrum and critical-band frequencies. Let f_{mel} be the perceived pitch, Mel frequency can be defined as follows:

$$f_{mel} = 2595\log_{10}(1+\frac{f_{Hz}}{700}) \quad (1)$$

where f_{Hz} is the real linear frequency to be converted.

The Mel frequency cepstral coefficients can be defined in a series of steps (Sigurdsson et al., 2006), as follows. Firstly, we compute the Discrete Fourier Transform (DFT) of every frame in the signal. The DFT can be defined as:

$$X(k) = \sum_{n=0}^{N-1} W(n)x(n)e^{-j\frac{2\pi}{N}nk} \quad (2)$$

where $x(n)$ is the discrete-time signal with length N, k = 0, 1, …, N-1 corresponds to the frequency $f(k) = kf_s/N$, f_s, is the sampling frequency in Hertz (*Hz*), and $w(n)$ is a time window (Hanning in this work).

Next, we compute the Mel filter bank, which is a series of overlapping triangular shaped bandpass filters, uniformly distributed on the Mel scale and spanning the entire signal bandwidth. Usually, this is a set of 20 ~ 40 filters which produces an equal number of same length vectors. These vectors are mostly composed of zeros and are non-zero for the section of the spectrum each filter band comprises. Next, each filter output is calculated by multiplying each filter bank with its power spectrum and accumulating the results. This gives a number that represents how much energy each filter bank contained. It is worth noting that the center frequency of each band n is $f_{mel}(n)$ and has max amplitude, starts at amplitude 0 at $f_{mel}(n - 1)$ and decays again to 0 at $f_{mel}(n + 1)$.

The Mel filter bank can be written as:

$$H(k,m)=\begin{cases}0 & \text{for } f(k)<f_c(m-1)\\ \dfrac{f(k)-f_c(m-1)}{f_c(m)-f_c(m+1)} & \text{for } f_c(m-1)\le f(k)<f_c(m)\\ \dfrac{f(k)-f_c(m+1)}{f_c(m)-f_c(m+1)} & \text{for } f_c(m)\le f(k)<f_c(m+1)\\ 0 & \text{for } f(k)\ge f_c(m+1)\end{cases} \quad (3)$$

and $f_c(m)$ is the center frequency.

In the next step the logarithm X' of each filter output is calculated, and the Discrete Cosine Transform (DCT):

$$c_n=\sum_{m=1}^{M} X'(m)\cos(n(m\frac{1}{2})\frac{\pi}{M}) \quad (4)$$

For n = 1, 2, ..., M, where M is the number of filter banks. The resulting c_n feature vectors are *Mel* frequency cepstral coefficients. This process results in a matrix for each input signal.

As each sample signal may generate MFCC matrixes with different lengths, each resulting matrix was flattened into a vector and zero padded as to have the same length. The resulting matrix is the extracted MFCC features from every sample.

Finally, before being fed to the Neural Network the training data is normalized, as to eliminate scale factors that might exist between variables of the data. This can be done by making all variables have similar weight. One possible formula to achieve this is:

$$z_i=\frac{x_i-\min(x)}{\max(x)-\min(x)} \quad (5)$$

where z_i is the i^{th} normalized data.

2.4. Multilayer Perceptron

There are several types of Artificial Neural Networks (ANN), and we chose the Multilayer Perceptron algorithm for its simplicity and usage in a wide range of applications. The MLP is a type of feedforward Neural Network (Gardner et al., 1998).

The MLP consists of a series of interconnected neurons (often called nodes or units), that can be arranged in 3 layered structures, the Input Layer, one or more Hidden Layers, and an Output Layer as shown in Figure 3. Each unit is connected to every other unit of the adjacent layers by axons, composed of weights and output signals. Basically, this consists of the sum of all the input units multiplied by the weights and then modified by a sigmoid activation function expressed as (Gardner et al., 1998):

$$a_i=\left(1/1+e^{-x}\right) \quad (6)$$

where a_i is the activation of unit i.

As with other supervised methods, the MLP learns through training with a series of labeled data. In the training process, the MLP maps the input data to the output data by adjusting the weights of the axons. In this process, errors are determined as the difference between the target output and the obtained output. Concretely, what the training process does is try to minimize this error, by changing the value of the weights, through an optimization algorithm (such as gradient descent, LBFGS (Nocedal et al., 2006) and ADAM (Kingma et al., 2014)).

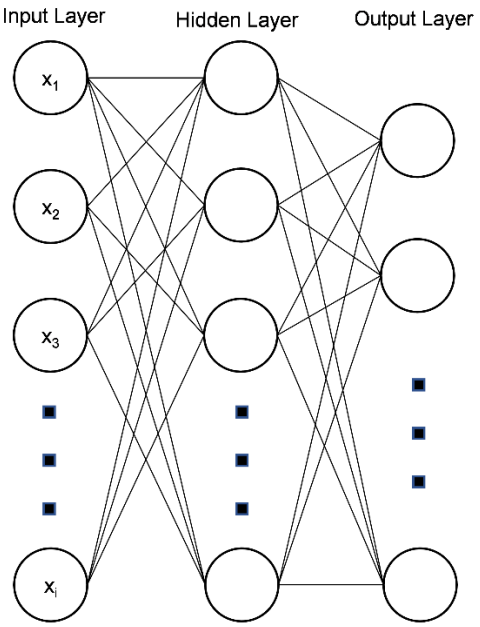

Figure 3. Schematic representation of a Multilayer Perceptron. Each circle represents a unit in its respective layers, and x is the input vector.

In this context, the most widely used algorithm that envelops this training process is the backpropagation algorithm (Rumelhart et al., 1985). The backpropagation algorithm can be summarized in a series of steps:

- Initialize weights.
- Feed the input vector.
- Forward propagate the input vector through the Neural Network.
- Calculate the error by comparing the target vector with the current output.
- Back propagate the error through the Neural Network.
- Apply the optimization algorithm to minimize error and adjust the weights.
- Repeat from step 2.
- A formal definition of the backpropagation algorithm can be found in Bishop, 1995.
- Finally, with the MLP model trained, a new acoustic signal can be fed to the network, and it will be classified as pertaining to the class *Amazona rhodocorytha* or not.

Furthermore, it is worth noting, that new acoustic signals should pass through the same pre-processing steps conducted

in the model generation. This is because the trained model works on the representation of the data given by its input and output vectors and the pre-processing steps are part of that representation.

3. Results and Discussion

This section presents the results obtained by the MLP classifier. Section 3.1 describes the dataset used for the training and testing phase, Section 3.2 discusses the experimental protocol cases, and Section 3.3 presents study cases and the obtained results.

3.1. Datasets

To emulate the chaotic nature of the data - that the model will be subjected to in a real-world scenario - we considered imperative the usage of unprocessed recordings taken in the field in different circumstances. Thus, the recordings have been taken by means of different microphones, different hardware, and different operators, thus having different quality and varying degrees of noise.

In this context, all the data was collected from the Wikiaves (2018) and Xeno-canto (2018) databases. Wikiaves is a community driven website for sharing information, photos and recordings of Brazilian birds. Similarly, Xenocanto is a website for sharing bird vocalizations from species around the world. Additionally, these databases provided us data in ideal settings, as all recordings were taken in different circumstances. Furthermore, it is worth noting that as there is a reasonable amount of redundancy between these datasets, we took the pre-caution of analyzing each recording to check for this issue.

Besides the *Amazona rhodocorytha*, we selected recordings from 4 other birds of the same family, having similar vocalizations, sharing the same ecosystem of the *Amazona rhodocorytha*. Table 1 lists the common name, scientific name and family of all species used in this work.

Table 1. Bird Species

Scientific Name	Common Name	Family
Amazona rhodocorytha	Red-browned Amazon	Psittacidae
Amazona aestiva	Turquoise-fronted Amazon	Psittacidae
Aratinga auricapillus	Golden-capped Parakeet	Psittacidae
Primolius maracana	Blue-winged Macaw	Psittacidae
Triclaria malachitacea	Blue-bellied Parrot	Psittacidae

We collected in total 109 different recordings of which 59 were from the Xeno-canto database, and 50 from the Wikiaves dataset. These 109 recordings can be separated into 5 classes, according to the subspecies of Psittacidae. In this way, 47 recordings were of the *Amazona rhodocorytha* (AR), 10 for the *Amazona aestiva* (AA), 12 classified as *Aratinga auricapillus* (AU), 15 identified as *Primolius maracana* (PM), and 25 of the recordings were of the *Triclaria malachitacea* (TM). It is worth noting that, as most of these recordings are over long periods of time and rich in acoustic vocalizations, the segmentation stage will yield multiple examples from each recording for training the neural network.

3.2. Experimental Protocol

Firstly, it is worth remembering that, for training the model, all 109 recordings were subjected to a pre-processing step. A manual segmentation of each recording file into multiple small segments containing only the relevant vocalization signal of a particular species was conducted. Next, all segments were subjected to a bandpass filter that only passed signals between 900 and 5000 Hz. Afterwards, in the feature extraction step, the MFCC of each segment was extracted with the number of MFCC as n_mfcc = 20. Finally, the resulting matrix was flattened and zeropadded, as to have the same dimensions when training the ANN, and then normalized.

Therefore, after this preprocessing step, the 109 original recordings were processed into 1185 samples that were fed to the ANN. Of the 1185, 424 were examples of *Amazona rhodocorytha*, 176 of *Amazona aestiva*, 179 of *Aratinga auricapillus*, 194 of *Primolius maracana*, and 212 of *Triclaria malachitecea*.

All experiments with the MLP algorithm were conducted using lambda = 0.3 for the regularization parameter, with one hidden layer, and 100 hidden units. Experimentation with higher values for the hidden units and hidden layers did not yield noticeable better results to justify the additional computational overhead. Furthermore, we used the second order LBFGS optimization algorithm since the dataset used to train the network is relatively small. This algorithm also provides the advantage that fewer parameters need to be set, such as the learning rate. However, it is worth noting that tests conducted with the ADAM optimization algorithms displayed nearly the same results.

For evaluation purposes, the main metric used was accuracy, and can be described as how close the result is to the correct value. Let TP be the number of true positives, TN be the number of true negatives and FP and FN be the number of false positives and false negatives respectively, accuracy can be given as:

Accuracy: (TP + TN)/(TP + TN + FN + FP)

Additionally, we also obtained results for 3 other evaluation metrics, they are: precision, recall, and the F1 score. Precision tells how many of the selected objects were relevant (correct), while recall gives a measure of how many of the relevant (correct) items were actually selected. Finally, the F1 score is a function of the precision and recall, giving a balance of both. These 3 metrics can be defined as follows:

Precision: (TP)/(TP + FP)
Recall: (TP)/(TP + FN)

$$\text{F1: } 2 * \frac{precision * recall}{precision + recall}$$

It is worth noting that all metrics used for the evaluation

were obtained using scikit-learn (a machine learning library) evaluation tools and functions, as to maintain consistency (Scikit-learn, 2011).

Furthermore, in order to build a robust model and help avoid overfitting, all experiments were conducted using a 5-fold stratified cross validation step, that is, all data was randomly split 5 times while preserving the proportion of samples for each class. Next, for each of the folds, the model was trained on the remaining data and evaluated with the one left out. With that, 5 results for each experiment were obtained and the mean and standard deviation were calculated. Finally, to estimate a reliable result, we repeated the described cross validation 10 times and calculated the average performance presented in the following section.

3.3. Experimental Results

Several scenarios were considered in the experiments for evaluating the effectiveness of the model in identifying the *Amazona rhodocorytha*. We conducted experiments with different combinations considering the 5 classes (mapped to each of the 5 species of Psittacidae present in the same geographic localizations). Therefore, 5 different binary classification experiments were performed. These combinations were *Amazona rhodocorytha* vs all other classes (ALL), and *Amazona rhodocorytha* vs each other class individually, that is AR vs AA, AR vs AU, AR vs PM and AR vs TM.

Tables 2 and 3 list the classification result of the model under these scenarios, presenting the mean performance and standard deviation respectively.

Table 2. Performance of the Neural Network for MFCC

	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
AR vs ALL	90.54	91.36	90.8	90.85
AR vs AA	89.83	89.96	89.83	89.68
AR vs PM	90.61	90.12	89.96	89.67
AR vs TM	93.7	93.8	93.7	93.7
AR vs AU	98.17	98.21	98.17	98.16

Table 3. Standard Deviation of the ANN for MFCC

	Accuracy	Precision	Recall	F1 Score
AR vs ALL	0.028	0.035	0.041	0.039
AR vs AA	0.024	0.024	0.024	0.023
AR vs PM	0.019	0.019	0.019	0.02
AR vs TM	0.009	0.008	0.009	0.008
AR vs AU	0.009	0.009	0.009	0.009

The results show that even though the vocalizations of these species are similar to each other, the model was capable of differentiating the *Amazona rhodocorytha* from the other Psittacidae with 90% accuracy or more for all the scenarios. It is worth highlighting the results obtained for the AR vs TM and AR vs AU, with 93.7 and 98.17% respectively. Furthermore, as both cases presented a very low standard deviation error (0.009), it is possible to see that the model fit the data very well.

For the AR vs ALL, AR vs AA and AR vs PM even though the results were slightly inferior, the model still showed great performance, especially considering all recordings used were real world data taken in different circumstances. An assumption can be made that these 2 species, the *Amazona aestiva* and *Primolius maracana*, present the most similar vocalization pattern to the *Amazona rhodocorytha*. Furthermore, the slightly higher standard deviation shown in AR vs ALL is due to this experiment having more variation among the samples given to the model to train and cross-validate, therefore increasing the complexity of the task.

Additionally, all other evaluation metrics presented nearly identical performance results and standard deviation to the accuracy, for all cases. This further proves the robustness of the model under different scenarios, with a high precision and high recall.

In Table 4 we break down the results listed in Tables 2 and 3 and present a confusion matrix showing the number of TP, FP, TN and FN across one repetition of each experiment.

Table 4. Confusion Matrix breakdown

	TP	TN	FP	FN
AR vs ALL	372	702	59	52
AR vs AA	401	138	38	23
AR vs PM	409	157	37	15
AR vs TM	401	187	23	25
AR vs AU	422	172	7	2

As expected, the number of True positives and True Negatives was high across all cases as compared to the number of False Positives and False Negatives. Furthermore, we can notice a high capability of the model in correctly identifying the *Amazona rhodocorytha* in all cases, with only few mistakes less than 10% error for this class in the individual scenarios. Furthermore, we can see that except for 2 cases, the AR vs ALL and AR vs TM, the number of FN was significantly lower than the number of FP. An interesting question of whether FP and FN have the same weight can be explored. In these types of application, such as detecting endangered species, we consider that higher number of FP is more desirable than a high number of FN. This is because a human operator will be looking specifically for these species, and it would be better to check something incorrect than to completely miss the presence of the species in question.

Additionally, it is possible to see that the results were better for the class that had more samples. Thus, we believe that some of the error shown is because of class imbalance during training, and the model could have achieved even better results, had the same number of samples been used for each class.

4. Conclusions

In this study, we present an approach for the automatic detection and identification of the *Amazona rhodocorytha*, an endangered bird species endemic to Brazil. The proposed method consists of a 3-step approach, starting with a segmentation and

filtering phase of the original signal, passing through a feature extraction step, and ending with the training of an Artificial Neural Network model for classification tasks.

Being simple to implement and flexible, the approach is capable of being carried out in several environments and with other species of birds. Furthermore, we tested the classification process with a difficult task of identifying the *Amazona rhodocorytha* among 4 other species that have very similar vocalization patterns. As shown by the good results obtained, the methodology proved to be both effective and robust for this type of application.

Finally, as future work, we contemplate the following: (i) implementation of this method in conservation units in Brazil; (ii) apply the method in an extensive real soundscape database that is currently being constructed by continuous recording from multiple sensors.

Acknowledgment. This work was partly supported by CNP_q Grant (303992 / 2017-4).

References

Au, W.W. (2012). *The sonar of dolphins*. Springer Science & Business Media.

Audacity® software is copyright © 1999-2018 Audacity Team. https://audacityteam.org/. It is free software distributed under the terms of the GNU General Public License. The name Audacity® is a registered trademark of Dominic Mazzoni.

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

BirdLife International. (2017). Amazona rhodocorytha. The IUCN Red List of Threatened Species 2017. https://doi.org/10.2305/IUCN.UK.2017-3.RLTS.T22686288A118968809.en.

Bishop, C.M. (1995). *Neural networks for pattern recognition*. Oxford university press, London. https://doi.org/10.1201/9781420050646.ptb6.

Blumstein, D.T., Mennill, D.J., Clemins, P., Girod, L., Yao, K., Patricelli, G., Deppe, J.L., Lrakauer, A.H., Clark, C., Cortopassi, K.A., Hanser, S.F., McCowan, B., Ali, A.M., and Kirschel, A.N.G. (2011) Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus. *Journal of Applied Ecology*, 48(3), 758-767. https://doi.org/10.1111/j.1365-2664.2011.01993.x.

Brandes, T.S. (2008). Automated sound recording and analysis techniques for bird surveys and conservation. *Bird Conservation International*, 18(S1), S163-S173. https://doi.org/10.1017/S0959270908000415.

Christiano, L.J. and Fitzgerald, T.J. (2003). The band pass filter. International *Economic Review*, 44(2), 435-465.https://doi.org/10.1111/1468-2354.t01-1-00076

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297. https://doi.org/10.1007/BF00994018.

Dawson, D.K. and Efford, M.G. (2009). Bird population density estimated from acoustic signals. *Journal of Applied Ecology*, 46(6), 1201-1209. https://doi.org/10.1111/j.1365-2664.2009.01731.x.

Edenhofer, O. (Ed.). (2015). *Climate change 2014: mitigation of climate change (Vol. 3)*. Cambridge University Press, Cambridge.

Foote, J.T. (1997). *Content-based retrieval of music and audio. In Multimedia Storage and Archiving Systems II*. International Society for Optics and Photonics, 3229, 138-148. https://doi.org/10.1117/12.290336.

Gardner, M.W. and Dorling, S.R. (1998). Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14-15), 2627-2636. https://doi.org/10.1016/S1352-2310(97)00447-0.

Gregory, R.D., Willis, S.G., Jiguet, F., Voříšek, P., Klvaňová, A., Strien, A.V., Hauntley, B., Collingham, Y.C., Couvet, D., and Green, R.E. (2009). An indicator of the impact of climatic change on European bird populations. *PLOS ONE*, 4. 3, e4678.https://doi.org/10.1371/journal.pone.0004678.

Jaafar, H. and Ramli, D.A. (2013). Automatic syllables segmentation for frog identification system. *2013 IEEE 9th International Colloquium on Signal Processing and its Applications*, IEEE, pp. 224-228. https://doi.org/10.1109/CSPA.2013.6530046

Johnson, J.B., Saenz, D., Burt, D.B., and Conner, R.N. (2002). *An automated technique for monitoring nocturnal avian vocalizations*. Bulletin of the Texas Ornithological Society, 35 (2), 24-29.

Kingma, D.P. and Ba, J. (2014). Adam: A method for stochastic optimization. *Computer Science*.

Koskimies, P. (1988). Trends in bird populations as environmental indicators. *Statistical Journal of the United Nations Economic Commission for Europe*, 5(3), 231-238.

Küçüktopcu, O., Masazade, E., Ünsalan, C., and Varshney, P.K. (2019). A real-time bird sound recognition system using a low-cost microcontroller. *Applied Acoustics*, 148, 194-201. https://doi.org/10.1016/j.apacoust.2018.12.028.

Lopes, M.T., Gioppo, L.L., Higushi, T.T., Kaestner, C.A.A., Jr, C.N.S., and Koerich, A.L. (2011). Automatic Bird Species Identification for Large Number of Species. *IEEE International Symposium on Multimedia*. IEEE Computer Society. https://doi.org/10.1109/ISM.2011.27.

Moss, C.F. and Schnitzler, H.U. (1989). Accuracy of target ranging in echolocating bats: acoustic information processing. *Journal of Comparative Physiology A*, 165(3), 383-393. https://doi.org/10.1007/BF00619357.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Pace, F. (2008). *Comparison of feature sets for humpback whale song classification*. Ph.D. dissertation, MSc dissertation, University of Southampton, UK.

Payne, R. and Webb, D. (1971). Orientation by means of long range acoustic signaling in baleen whales. *Annals of the New York Academy of Sciences*, 188(1), 110-141. https://doi.org/10.1111/j.1749-6632.1971.tb13093.x

Priyadarshani, N., Marsland, S., and Castro, I. (2018). Automated birdsong recognition in complex acoustic environments: a review. *Journal of Avian Biology*, 49(5), jav-01447. https://doi.org/10.1111/jav.01447

Rosenzweig, C. and Parry, M.L. (1994). Potential impact of climate change on world food supply. *Nature*, 367(6459), 133. https://doi.org/10.1038/367133a0

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1988). Learning internal representations by error propagation. *Cognitive Science*, 399-421. https://doi.org/10.1016/B978-1-4832-1446-7.50035-2.

Swami, A. and Jain, R. (2012). Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12(10), 2825-2830.

Sebastián-González, E., Pang-Ching, J., Barbosa, J.M., and Hart, P. (2015). Bioacoustics for species management: two case studies with a Hawaiian forest bird. *Ecology and evolution*, 5(20), 4696-4705. https://doi.org/10.1002/ece3.1743.

Selin, A., Turunen, J., and Tanttu, J.T. (2006). Wavelets in recognition of bird sounds. *EURASIP Journal on Advances in Signal Processing*, 2007(1), 051806. https://doi.org/10.1155/2007/51806.

Sigurdsson, S., Petersen, K.B., and Lehn-Schiøler, T. (2006). Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music. *ISMIR*, 286-289.

Simmonds, M.P. and Isaac, S.J. (2007). The impacts of climate change on marine mammals: early signs of significant problems. *Oryx*, 41(1), 19-26. https://doi.org/10.1017/S0030605307001524.

Stephens P.A., Mason L.R., Green R.E., Gregory R.D., Sauer J.R., Alison J., Aunins A., Brotons L., Butchart S.H., Campedelli T., Chodkiewicz T., Chylarecki P., Crowe O., Elts J., Escandell V., Foppen R.P., Heldbjerg H., Herrando S., Husby M., Jiguet F., Lehikoinen A., Lindström Å., Noble D.G., Paquet J.Y., Reif J., Sattler T., Szép T., Teufelbauer N., Trautmann S., van Strien A.J., van Turnhout C.A., Vorisek P., and Willis S.G. (2016). Consistent response of bird populations to climate change on two continents. *Science*, 352.6281 84-87. https://doi.org/10.1126/science.aac4858.

Vellinga, Willem-Pier and Robert Planqué. (2015). The Xeno-canto Collection and its Relation to Sound Recognition and Classification. *Working Notes of CLEF-Conference and Labs of the Evaluation forum, 2015*.

WikiAves (2018). Brazillian community foir sharing bird vocalizations. https://www.wikiaves.com.br/.

Xeno-Canto (2018). Brazillian bird vocalization sharing site. www.xeno-canto.org.

Yegnanarayana, B. (2009). *Artificial neural networks*. PHI Learning Pvt. Ltd.