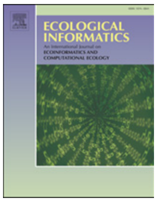
Ecological Informatics 38 (2017) 39–49

Contents lists available at ScienceDirect

Ecological Informatics

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



Automatic classification of Furnariidae species from the Paranaense Littoral region using speech-related features and machine learning

Enrique M. Albornoz[a,*], Leandro D. Vignolo[a], Juan A. Sarquis[b], Evelina Leon[b]

[a] *Research Institute for Signals, Systems and Computational Intelligence, sinc(i), UNL-CONICET, Argentina*
[b] *National Institute of Limnology, INALI, UNL-CONICET, Argentina*

ARTICLE INFO

Article history:
Received 26 May 2016
Received in revised form 25 January 2017
Accepted 26 January 2017
Available online 31 January 2017

Keywords:
Bird sound classification
Computational bioacoustics
Machine learning
Speech-related features
Furnariidae

ABSTRACT

Over the last years, researchers have addressed the automatic classification of calling bird species. This is important for achieving more exhaustive environmental monitoring and for managing natural resources. Vocalisations help to identify new species, their natural history and macro-systematic relations, while computer systems allow the bird recognition process to be sped up and improved. In this study, an approach that uses state-of-the-art features designed for speech and speaker state recognition is presented. A method for voice activity detection was employed previous to feature extraction. Our analysis includes several classification techniques (multilayer perceptrons, support vector machines and random forest) and compares their performance using different configurations to define the best classification method. The experimental results were validated in a cross-validation scheme, using 25 species of the family Furnariidae that inhabit the Paranaense Littoral region of Argentina (South America). The results show that a high classification rate, close to 90%, is obtained for this family in this Furnariidae group using the proposed features and classifiers.

1. Introduction

Vocalisations are often the most noticeable manifestations of the presence of avian species in different habitats (Potamitis, 2015). Birds have been widely used to indicate biodiversity since they provide critical ecosystem services, respond quickly to changes, are relatively easy to detect and may reflect changes at lower trophic levels (e.g. insects, plants) (Burkart et al., 1999; Louette et al., 1995). Technological tools (such as photographic cameras, video cameras, microphones, and mass storage disks) are useful for collecting data about several patterns of bird populations. However, there are a number of problems associated with them, such as poor sample representation in remote regions, observer bias (Laje and Mindlin, 2003), defective monitoring (Betts et al., 2007), and high costs of sampling on large spatial and temporal scales, among others.

Bird vocalisations have become an important research field, influencing ethology (Hesler et al., 2011; Seddon and Tobias, 2010), taxonomy (Bergmann and Schottler, 2001; Raposo and Höfling, 2003; Schottler) and evolutionary biology (Lynch and Baker, 1994; Päckert et al., 2003). One of the main activities that benefits from vocalisation identification is ecosystems monitoring, where the technological advances allow registering and processing the recordings, and improving the data collection in the field (Towsey et al., 2014). This makes it possible to gather data in large and disjoint areas, which is essential for conducting reliable studies.

Although some works describe vocalisation changes in certain Furnariidae species (Areta and Pearman, 2009, 2013; MacKenzie et al., 2002; Potamitis et al., 2014; Zimmer and Whittaker, 2000), none of them simultaneously evaluates several vocalisations of Furnariidae species from South America. In this study, vocalisations belonging to 25 Furnariidae species that are distributed in the Paranaense Littoral region (see Fig. 1) are analysed. This region comprises the Argentinean Mesopotamia (Misiones, Corrientes and Entre Ríos provinces) along with the provinces of Chaco, Formosa and Santa Fe, and it is lapped by great rivers of the Plata basin (Arzamendia and Giraudo, 2009). Over the last years, this region has become an interesting place for studying bird vocalisations (Areta and Pearman, 2009, 2013; León et al., 2014; Leon et al., 2015). In addition, the work of researchers from the National Institute of Limnology (INALI) along with the availability of Furnariidae species would allow us to record and analyse these species in real-life conditions in future studies. Recently, some authors have researched

* Corresponding author at: Research Institute for Signals, Systems and Computational Intelligence, sinc(*i*), FICH-UNL/CONICET, Ciudad Universitaria, Paraje El Pozo, Santa Fe S3000, Argentina.
E-mail addresses: emalbornoz@sinc.unl.edu.ar (E.M. Albornoz), ldvignolo@sinc.unl.edu.ar (L.D. Vignolo), juansarquis@conicet.gov.ar (J.A. Sarquis), evelinaleon@conicet.gov.ar (E. Leon).
URL: http://sinc.unl.edu.ar/.

http://dx.doi.org/10.1016/j.ecoinf.2017.01.004
1574-9541/© 2017 Published by Elsevier B.V.

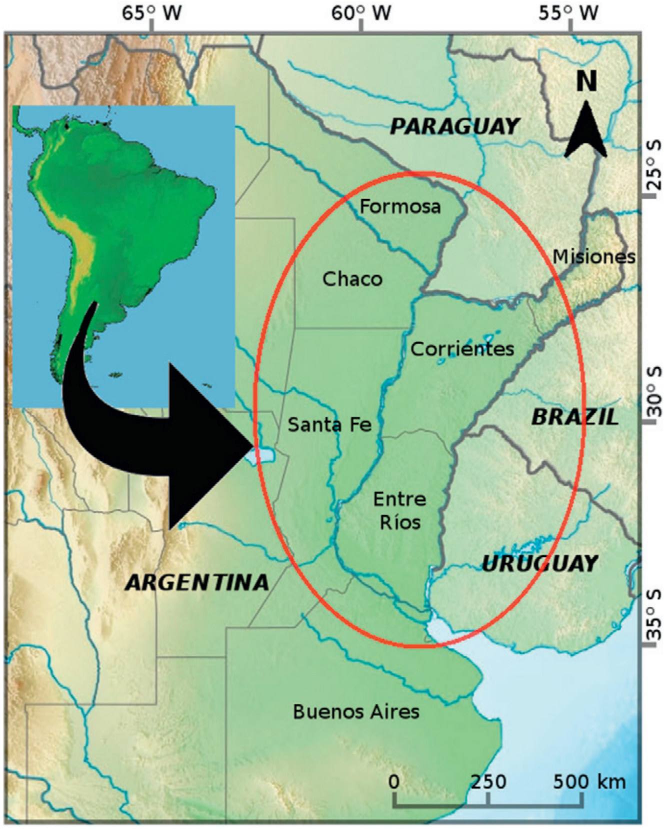


Fig. 1. Paranaense Littoral region (Argentina).

the vocalisations and the natural history of Zimmer and Whittaker (2000) used morphometric analysis, behavioural data and vocalisations to analyse the *Pseudoseisura cristata*. The role of several habitats as well as natural history, taxonomy, morphology, vocalisations and evolution for the *Upucerthia saturatior* was studied in (Areta and Pearman, 2009, 2013).

Recognition of species in passeriformes is a challenging task because to they produce complex songs and can adapt their content over time. It is interesting to note that the song content can be changed depending on the audience, for example, when the receiver is male or female (Byers, 1996), or in order to match it with that of their neighbours (Payne, 1996). Furthermore, they can take possession of new songs or syllables during their lifetime (Marler, 1997). The family Furnariidae produces several songs and some species manifest these as duets. It represents a synchronisation of physiological rhythms in a natural behaviour, which adds more complexity to the analysis. In addition, some species of the same family show similar structures in their songs. These similarities are manifested in introductory syllables or in the trill format, while the complexity of duets within the family makes the analysis and classification of vocalisations more difficult. Previous studies demonstrated that there are differences in tone and note intervals between males and females (Areta and Pearman, 2009, 2013; Pacheco and Gonzaga, 2013; Zimmer and Whittaker, 2000). For this family, the complexity of vocalisations was proved by means of playback experiments. These showed that the different taxa express dissimilar responses to similar patterns.

It should be noted that environmental conditions (humidity, wind, temperature, etc.) may alter the recording process, modifying the features that are present in the structure of songs and in the calls (e.g. frequency, duration, and amplitude) (Harris, 1963, 1966; Zollinger and Brumm, 2015). Since these conditions may lead to errors and distort subsequent analyses and results, researchers usually use recordings from known databases. Even though these registrations can be also affected by environmental issues, their attributes and labels are validated by the scientific community and consequently, they are more reliable than "homemade" records.

As mentioned in Spampinato et al. (2014) , new frontiers have been opened in ecology (besides the analysis performed by expert ecologists) due to the propagation of projects like *Xeno-canto*[1] and *EcoGrid*.[2] The access to multimedia data has promoted an interdisciplinary and collaborative science for analysing the environment. Although human experts (who are sufficiently trained) can recognise bioacoustic events with a high performance, this is a laborious and expensive process that would be more efficient if they had the technical support of a semi-automatic tool (Truskinger et al., 2015). Finally, the goal pursued is the development of an automatic classifier that provide a high accuracy and involve the expert only for evaluating the results. Automatic tools allow simultaneous studies to be conducted and diverse bird communities to be monitored in several areas at the same time, in order to identify when and how the species vocalise. In addition, said tools could be used to create complete inventories of bird communities in unknown or restricted areas, which are essential for conservation or management plans.

In particular, the bird call identification task can be used in two ways (Dong et al., 2015): call retrieval (detection) and call classification. In the call retrieval task, the objective is to identify one or more calls in an audio recording, which can contain multiple calls of different species overlapped or at different times. In the classification task, a set of call classes must be defined and the classifier will be trained to recognise this fixed set. In this way, every input audio (expected to contain only one call) will be classified to one of those classes. A classification scheme can be defined as a pipeline of three modules: preprocessing, feature extraction and classification. The first one depends strongly on the recording process and involves filtering, segmentation and enhancement of audio signals. Furthermore, automatic methods for voice activity detection (VAD) have been recently incorporated (Ptacek et al., 2015). Regarding feature extraction, time- and frequency-based information was employed (Cramer, 2013; Keen et al., 2014; Potamitis, 2015; Truskinger et al., 2015). In addition, characteristics that were originally developed for speech analysis are used in the context of bird call recognition. Some of the features present in the literature are mel frequency cepstral coefficients (MFCCs) (Molau et al., 2001), linear frequency cepstral coefficients (LFCCs) (Zhou et al., 2011), and standard functionals (mean, standard deviation, kurtosis, etc.) computed over these (Briggs et al., 2012; Dufour et al., 2014; Ptacek et al., 2015). Various techniques have been applied to bird call classification: Gaussian mixture model (GMM) (Roch et al., 2007), Gaussian mixture model-universal background model (GMM-UBM) (Xiong et al., 2006), support vector machines (SVM) (Hearst et al., 1998), random forest (RF) (Breiman, 2001), among others. In Ptacek et al. (2015), LFCC features were used along with GMM-UBM to identify some subjects from the same bird species.

A similar approach was proposed in Ganchev et al. (2015) for recognising a single bird species using MFCCs. An interesting strategy based on the pairwise similarity measurements, computed on bird-call spectrograms, was evaluated in Keen et al. (2014), where the authors used different classifiers to recognise four species. In Dufour et al. (2014), thirty-five species were classified using a SVM classifier and six functionals were obtained from each MFCC. A different approach was proposed in Ventura et al. (2015), where a classifier based on hidden Markov models (HMMs) was used to recognise bird calls through their temporal dynamics. Previous works developing

[1] http://www.xeno-canto.org/.

[2] http://www.aiai.ed.ac.uk/project/ecogrid/.

full-automatic methods for vocalisation recognition can be examined in Giannoulis et al., ICML, Stowell and Plumbley (2013a,b), and the current relevance of this topic is shown in some recent works (Ganchev et al., 2015; Ptacek et al., 2015). However, none of these works has addressed the vocalisation recognition of species belonging to the Furnariidae family, which present similar parameters in their vocalisations. Moreover, only a small part of the state-of-the-art speech features have been employed in bird classification tasks. In Schuller et al. (2014), a large set of state-of-the-art speech features is described, comprising more than 6000 features, and many of these are considered within this task for the first time in this work.

This study proposes the development of a bird call recognition model for dealing with the family Furnariidae from the Paranaense Littoral region of Argentina, which is the first approach for these species. Our model is designed to use state-of-the-art classifiers with speech-related parameterisations, and some feature selection techniques are used to reduce dimensionality while maximising accuracy. As a first step, a method for performing the VAD is included. The model is tested in a cross-validation scheme in all cases. Furthermore, the best results are discussed, and the confusion matrix is analysed to introduce the misclassification and how some similarities among some species could be addressed in order to improve the performance.

The following section introduces the proposed features and classifiers. Section 3 deals with the experimental setup, presents the implementation details and describes the validation scheme. The results are presented and discussed in Section 4. In addition, the implementation of a web-demo and an android application for testing the model is explained. Finally, conclusions are summarised and future work is commented in the last section.

2. Proposed features and classifiers

This section introduces the feature extraction process, two different feature selection techniques and the classifier models.

2.1. Feature extraction

As mentioned above, the use of speech-based features is known in bird call analysis, identification and classification. For these tasks, the LFCCs and MFCCs sets (standards in speech recognition) showed good performances (Dufour et al., 2014; Ptacek et al., 2015). An extended state-of-the-art set of features related to human speech is introduced below.

2.1.1. Speech inspired features

In the speech processing area, researchers have made a great effort to find the best set of features for speech recognition, speaker recognition, emotion recognition, illness state detection, etc. (Schuller et al., 2009, 2011, 2013). In the INTERSPEECH 2013 ComParE Challenge (Schuller et al., 2013), a set of 6373 features was presented which is considered the state-of-the-art in speech processing. The feature set is built from 65 low-level descriptors (LLDs) such as energy, spectral, cepstral (MFCC), voicing-related characteristics (F_0, shimmer, jitter, etc.), zero crossing rate, logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, psychoacoustic spectral sharpness, and their deltas (i.e. their first temporal derivatives). These features are computed on a time frame basis, using a 60-ms window with 10-ms step for F_0 (pitch) and zero crossing rate. The remaining features are computed using a window size of 20 ms and the time contour of each attribute is smoothed by a moving average filter. Specific functionals are then computed for each LLD set. These include the arithmetic mean, maximum, minimum, standard deviation, skewness, kurtosis, mean of peak distances, among others. Tables 1 and 2 provide an exhaustive enumeration of all the LLDs and functionals used to constitute the complete feature vector. In addition to the complete feature set obtained by combining all LLDs and functionals (Full-Set), this work also proposes a subset consisting of the complete set of functionals computed only from the MFCCs, which results in a set of 531 attributes (MFCC + Fun).

Table 1
Low-level descriptors (LLDs) (Schuller et al., 2011) . $+\Delta$ means that the first derivative is computed and appended, to the feature vector computed for each analysis frame.

Low-level descriptors
Sum of auditory spectrum (loudness) $+\Delta$
Sum of RASTA-style filtered auditory spectrum $+\Delta$
RMS energy $+\Delta$
harmonic-to-noise ratio (HNR) $+\Delta$
Zero-crossing rate $+\Delta$
RASTA-style filtering. Bands 1–26 (0–8 kHz) $+\Delta$
MFCC 1-14 $+\Delta$
Spectral energy 25–650 Hz, 1 k–4 kHz $+\Delta$
Spectral roll off point 0.25, 0.50, 0.75, 0.90 $+\Delta$
Spectral flux, entropy, variance $+\Delta$
Skewness, kurtosis, slope $+\Delta$
F0, probability of voicing $+\Delta$
Jitter (local, delta) $+\Delta$
Shimmer (local) $+\Delta$

To the best of our knowledge, no suitable baseline models are available for comparing the performance of our proposal. In order to create the baseline, previous works (Dufour et al., 2014; Fagerlund, 2007) were considered to define the classifiers and feature sets for the bird song identification task. The first 17 MFCCs, their deltas and acceleration coefficients were computed using overlapped frames. Then, the mean and variance for each feature (over the entire song) were calculated, which resulted in a 102-dimensional vector for each recording.

2.1.2. Feature selection

Feature selection techniques were defined in order to reduce the dimensionality of data while keeping the most relevant information. This allows less complex models to be generated, which reduces the number of parameters to estimate in the model and the computing cost, and provides a similar or even improved performance. Feature or attribute selection is commonly carried out by searching the space of feature subsets, and each candidate subset is evaluated according to some criteria (Hall, 1998).

In this study, the performance of two well-known attribute selection methods is compared: best first (BF) (Xu et al., 1988) and linear forward selection (LFS) (Gütlein et al., 2009). The BF method performs a greedy hill climbing using backtracking, which means that it can search forward through a specified number of non-improving nodes before the algorithm goes back. This algorithm has proven to guarantee the best global subset without exhaustive enumeration, given that the criterion used satisfies monotonicity. The LFS algorithm is an extension of BF, which aims to reduce the number of evaluations performed during the search process. The number of attribute expansions is limited in each forward selection step, which drastically improves the runtime performance of the algorithm (Gütlein et al., 2009). Both feature selection methods need a criterion to evaluate each considered subset; therefore, the correlation-based feature subset evaluation (CFS) method (Hall, 1998) was applied. This method assesses the predictive ability of each attribute in the subset, and also considers the redundancy among them. Finally, the method picks up the subsets whose attributes are highly correlated within the class and have low intercorrelation among classes. Both feature selection methods were implemented using WEKA library[3] (Hall et al., 2009).

[3] Software available at http://www.cs.waikato.ac.nz/ml/weka/.

Table 2
Functionals applied to LLDs (Schuller et al., 2011) .

Base functionals
Quartiles 1–3
3 inter-quartile ranges
1 % percentile (≈min), 99 % percentile (≈max)
Percentile range 1 %–99 %
Arithmetic mean, standard deviation
Skewness, kurtosis
Mean of peak distances
Standard deviation of peak distances
Mean value of peaks
Mean value of peaks−arithmetic mean
Linear regression slope and quadratic error
Quadratic regression a and b and quadratic error
Simple moving average
Contour centroid
Duration signal is below 25 % range
Duration signal is above 90 % range
Duration signal is rising/falling
Gain of linear prediction (LP)
Linear prediction coefficients 1–5
F0 functionals
Percentage of non-zero frames
Mean, max, min, std. dev. of segment length
Input duration in seconds

2.2. Classifiers

Several techniques from machine learning and computational intelligence have been used in bird call identification (Ptacek et al., 2015). Based on previous studies, the analysis in this work was focused on some of the most commonly used classification algorithms. The following subsections briefly introduce three techniques: multilayer perceptron, random forest and support vector machines. WEKA and Scikit-Neuralnetwork[4] libraries were employed to apply these classifiers.

2.2.1. Multilayer perceptron

A multilayer perceptron (MLP) is a class of artificial neural network that consists of a set of process units (simple perceptrons or neurons) arranged in layers. In the MLP, the nodes are fully connected between layers without connections between units in the same layer (Fig. 2). The input of the MLP is the feature vector (**x**), which feeds each of the neurons of the first layer, the outputs of this layer feed into each of the second layer neurons, and so on (Haykin, 1998). The output of a neuron is the weighted sum of its inputs plus the bias term, and its activation is a function (linear or nonlinear) as

$$y = \mathcal{F}\left(\sum_{i=1}^{n} \omega_i x_i + \theta\right). \tag{1}$$

The output of the MLP (i.e. the output of the neurons in the last layer) is decoded to provide the predicted label for a given input example. The backpropagation method (Haykin, 1998) is commonly used to obtain the synaptic weights for the connections in the network (ω_i). This method computes the gradient of a loss function, with respect to all network weights. The weights are then updated according to the gradient, with the aim of minimising the loss function (usually the mean square error). Since the method requires a desired output for each training input in order to calculate the error, it is considered as a supervised learning technique.

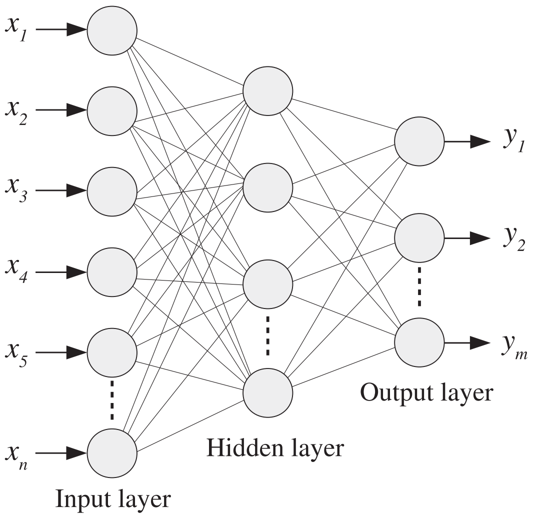


Fig. 2. Example of a MLP network model.

In this work, three architectures were considered: one hidden layer with the number of neurons set as $(Num.of\,inputs + Num.of\,outputs)/2$ (MLP1), one hidden layer with the number of neurons set to the number of inputs (MLP2), and two hidden layers set as in MLP2 and MLP1, respectively (MLP3).

2.2.2. Random forest

Classification and regression tree (CART) models, the so-called decision trees, are widely known in machine learning and data mining (Murphy, 2012). Some relevant properties include their robustness to different feature transformations, such as scaling, and their ability to discriminate irrelevant information while producing easily analysable models. These models are constructed by recursive partitioning the input space and region-specific models are then defined for the resulting scheme (Breiman, 2001). This can be represented with a tree, where the nodes indicate the decision functions and each leaf stands for a region (Fig. 3).

Random forest (RF) is an ensemble learning method whose decision is based on the average of multiple CARTs, which are trained on different parts of the same training set, with the aim of reducing the variance of CART overfitting. The computation can be expressed in terms of the *bagging* technique (Murphy, 2012) as

$$f(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} t_k(\mathbf{x}) \tag{2}$$

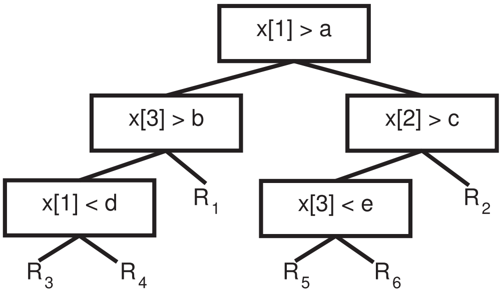


Fig. 3. Example of CART using feature vector $\in R^3$.

[4] Software available at http://scikit-neuralnetwork.readthedocs.org.

where t_k is the k-th tree. Here, the RF was implemented following (Breiman, 2001), considering 10 and 100 trees with unlimited depth.

2.2.3. Support vector machine

A support vector machine (SVM) is a supervised learning method that is widely used for pattern classification and is supposed to have good generalisation capabilities (Vapnik and Cortes, 1995). Its aim is to find a hyperplane that can separate input patterns in a sufficiently high dimensional space. The distances from the hyperplane to the patterns that are closest to it, on each side, is called a *margin*. This margin needs to be maximised to reach the best generalisation. In the binary case, this is done finding the $\mathbf{w}$ and w_0 parameters by means of a standard quadratic optimisation (Alpaydin, 2010; Vapnik and Cortes, 1995):

$$\begin{aligned} &\min \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{subject to} \qquad (3) \\ &r^t\left(\mathbf{w}^T\mathbf{x}^t + w_0\right) \geq +1, \forall t \end{aligned}$$

where $\{\mathbf{x}^t, r^t\}$ is a pattern with $r^t = -1$ if $\mathbf{x}^t$ is class #1, or $r^t = +1$ in the other case.

It is known that a nonlinear problem could be solved as a linear problem in a new space by making a nonlinear transformation (Alpaydin, 2010). The new dimensions are then computed using the basis functions by inner product. The *kernel trick* is a method that solves this problem without mapping the features in the new space; therefore, the kernel function is applied to the original space (Alpaydin, 2010). Some of the more popular kernels used in SVMs are the polynomial of degree q:

$$K\left(\mathbf{x}^t, \mathbf{x}\right) = \left(\mathbf{x}^T\mathbf{x}^t + 1\right)^q \qquad (4)$$

and radial-basis functions:

$$K(\mathbf{x}^t, \mathbf{x}) = \exp\left[-\frac{\mathcal{D}(\mathbf{x}^t, \mathbf{x})}{2s^2}\right] \qquad (5)$$

where x^t is the centre, s is the radius and $\mathcal{D}(\mathbf{x}^t, \mathbf{x})$ is a distance function. In our experiments, the SVMs were trained using the sequential minimal optimisation algorithm and considering the polynomial kernel.

3. Experiments

This section describes the experimental framework used in this study. First, a discussion on why and how the bird species were selected from the known databases. Then, the implementation details of the feature extraction and classifiers are presented. Finally, the validation scheme used to evaluate the models is explained. A general scheme of the whole process for the experiments is shown in Fig. 4.

3.1. Study area and target species

The study area is located between 22°25′ S62°12′ W and 38°0′ S57°26′ W(Fig. 1), and comprises several ecoregions along the Paraná River. These regions are Dry Chaco, Espinal, Pampa, Iberá Wetlands, and Delta and Islands of the Paraná River (Burkart et al., 1999). The family Furnariidae presents diverse vocalisations and some species can even sing male-female duets. In spite of that, the experts are usually able to identify them, reaching a good performance. The vocalisations obtained from species of this family might be similar and thus difficult to classify. In addition, the vocalisations from one species can change depending on its geographical location.

The family Furnariidae includes 68 genera composed of 302 species (Clements et al., 2015). Being distributed in South America and in a region of Central America (Noriega, 1991), it is one of the most impressive examples of continental adaptive radiation. This family has probably the highest morpho-ecological diversity in birds, living in diverse habitats such as desert or arid regions, rocky coasts, ravines, swamps, grasslands and forests (Irestedt et al., 2009; Vuilleumier and Vaurie, 1980). The characteristics described above plus the large number of studies about its taxonomy, the biological and natural history (Areta and Pearman, 2013; Fjeldså et al., 2005; Garcıá-Moreno et al., 1999; Irestedt et al., 2009; Olson et al., 2005; Vuilleumier and Vaurie, 1980) and our own experience at INALI make the family Furnariidae an interesting and open challenge to study. Fig. 5 shows the tree structure of the 25 studied Furnariidae species/genera.

3.2. Bird call corpus

To obtain a suitable number of vocalisations for training the classifiers and evaluating the performance, records from two well-known databases were selected, obtaining a total of 206 recordings. From these, 90 recordings were selected from the *Xeno-canto*[5] database (Joly et al., 2014; Planqué and Vellinga; Potamitis, 2015) and 116 recordings were taken from the *Birds of Argentina & Uruguay: A Field Guide Total Edition* corpus (Contreras et al., 2014; Leon et al., 2015; Narosky and Yzurieta, 2010). This combination of different data sources involves an additional complexity that the model should be able to handle.[6]

3.3. Feature extraction

As mentioned earlier, the step prior to feature extraction is usually the preprocessing and it is carried out to standardise the audio signals. A Wiener-based noise filter (Plapous et al., 2006) was applied to the audio signals to reduce noise in the recordings. As all of the utterances have an initial silence, the noise could be modelled.

The acoustic activity detection (where the information is contained) is an active area of research (de Oliveira et al., 2015). In this work, the endpoints of acoustic activity were computed using a voice activity detector (VAD) based on Rabiner and Schafer's method (Giannakopoulos and Pikrakis, 2014).

The *openSMILE* toolkit (Eyben et al., 2013) was used to extract the state-of-the-art features (Schuller et al., 2013) mentioned in the previous section. This is a feature extraction tool that allows a large set of audio features to be extracted, and it is distributed free of charge for research and personal use.[7]

3.4. Validation

Coefficients in vectors were normalised using the maximum and minimum values (for each dimension) in the training set as follows:

$$C_{i,j}^{norm} = \frac{(C_{i,j} - C_{min,j})}{(C_{max,j} - C_{min,j})}, \qquad (6)$$

where $C_{i,j}^{norm}$ is the normalised coefficient j from recording i, $C_{i,j}$ represents the original value, while $C_{min,j}$ and $C_{max,j}$ represent the minimum and maximum values of coefficient j from all the training recordings.

[5] http://www.xeno-canto.org/.

[6] The list of audio files used in this work was included as Supplementary material.

[7] Software available athttp://www.audeering.com/research/opensmile/.

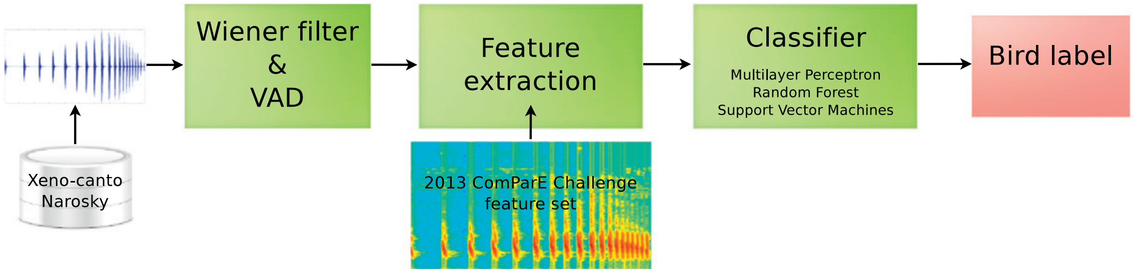


Fig. 4. Conceptual flowchart of the general whole process for the experiments.

The recognition rate estimation may be biased if only one training partition and one test partition are used. To avoid these estimation biases, a cross-validation was performed with the *k*-fold method (Michie et al., 1994). For each experiment the classification results by 10-fold stratified cross-validation (SCV) were computed, where each fold was composed of 90% of data for training and the remaining 10% was used for testing. Finally, the results were computed and averaged over the 10 test sets.

Several classification measures were computed for accurately visualising the performance of the models. The *weighted average recall* or *accuracy* (ACC) is the number of correctly classified instances divided by the total number of instances. Although this measure is widely used, it can be biased when the classes are not balanced. If the classes (species) are unbalanced, the *unweighted average recall* (UAR) gives a more accurate estimation of the performance (Rosenberg, 2012). The UAR was computed as the average of all class accuracies as:

$$\mathrm{UAR} = \frac{1}{K}\sum_{i=1}^{K}\frac{A_{ii}}{\sum_{j=1}^{K}A_{ij}}, \qquad (7)$$

where K is the number of classes and A_{ij} is the number of instances belonging to class i that are classified as j.

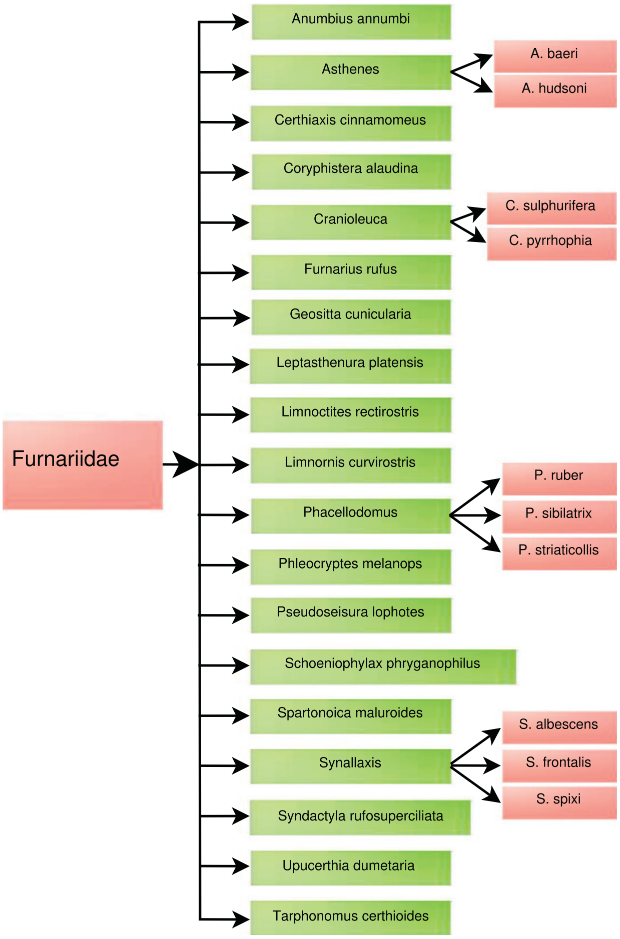


Fig. 5. Tree structure of the 25 studied Furnariidae species.

4. Results and discussion

The baseline feature set and the proposed feature sets were evaluated using all the classifiers described in Section 2.2, considering the normalised attributes explained in Section 3.4. Also, LFS and BF feature selection methods were used to reduce the size of the Full-Set (6373 features), maximising accuracy while keeping the most relevant information. Tables 3 and 4 present the results obtained in terms of accuracy and UAR, respectively.[6]Table 3 shows that the baseline set (102 features) provides high accuracy rates while the proposed sets improve these results, and the best results are close to 90%. However, the performance is lower when the Full-Set is used because the models cannot be properly trained. This means that the complexity of the classifiers is increased due to the high number of inputs (especially in the case of MLP), and the small amount of data available is not enough for appropriately training them, which causes poor performance.

In order to assess how the imbalance of classes affects the results, the UAR values should be analysed, taking into account the hit rates for each class (Table 4). This table presents similar results, where

Table 3
Weighted average recall (accuracy) [%].

Feature vector	MLP1	MLP2	MLP3	RF10	RF100	SVM
Baseline	85.92	**86.89**	78.64	68.45	80.10	84.95
MFCC + Fun	**89.32**	88.83	79.61	69.42	83.01	85.92
Full-Set	74.27	65.05	08.25	68.93	80.10	**83.50**
Full-Set + LFS	**89.32**	86.89	80.58	76.70	86.41	87.38
Full-Set + BF	**89.32**	**89.32**	80.58	76.70	86.41	87.38

Table 4
Unweighted average recall (UAR) [%].

Feature vector	MLP1	MLP2	MLP3	RF10	RF100	SVM
Baseline	77.24	**79.21**	72.06	58.25	67.00	74.07
MFCC + Fun	79.96	**80.85**	69.16	58.08	70.43	75.18
Full-Set	61.90	53.55	05.06	55.74	65.24	**72.46**
Full-Set + LFS	**82.21**	78.74	68.65	64.20	73.65	77.82
Full-Set + BF	**82.10**	80.25	70.42	64.20	73.35	77.82

the proposed feature sets improve the baseline performance. The MFCC + Fun set (531 features) performs better than the baseline for almost all classifiers, whereas both feature selection methods applied over the Full-Set achieve the best performances. It is interesting to note that MLPs and SVMs produce better results than RF for all the feature sets. Finally, one can be conclude that the best performance is obtained using the multilayer perceptron (MLP1) and applying the LFS method over the Full-Set.

The dimension of the best feature set is 153, thus the system has kept a very low dimensionality in addition to achieving the best rates. The retained features include mostly spectral and cepstral coefficients as described next. Thirty-six features were computed based on the MFCC coefficients and some functionals (quartiles, percentiles and mean, among others). Eleven features obtained from the first derivative of MFCC (delta MFCC) (Zheng et al., 2001) and the same functionals. Twenty-four spectral features were selected, including roll-off (percentile of the power spectral distribution), slope (which describes how rapidly the amplitudes of successive component change), harmonicity (which evaluates the total strength of harmonic structure) and flux (a measure that indicates how quickly the power spectrum of a signal is changing) (Eyben et al., 2013). Twelve features computed as functionals from frequency band energies, particularly in bands of 250–650 Hz and 1000–4000 Hz. Forty-four features obtained by applying functionals to 26 spectral bands filtered with RASTA (RASTA uses bandpass filtering in the log spectral domain to remove slow channel variations) (Hermansky and Morgan, 1994). Eleven features computed from the auditory spectrum, which is inspired by psychoacoustic studies on human primary auditory cortex and produces a time-frequency representation. Five features computed as functionals from the auditory spectrum filtered with RASTA (Hermansky and Morgan, 1994). Twelve features computed from the root mean square energy, voicing, harmonic-to-voice ratio, jitter and zero crossing rate.

As the performance obtained is highly satisfactory (close to 90%) and the amount of data is limited, a test of statistical significance like the paired *T*-test (Demšar, 2006) is not relevant. However, our results suggest that 5 samples per species are required for properly training the model (see Table 5). Evidently, patterns from the same species present some differences, therefore analyses where only one sample is used to represent the species (as in Dufour et al., 2014) could be not very reliable. Furthermore, confusions may be explained by certain similarities in vocalisations, such as waveform shapes, harmonic content, placement and separation of syllables, among others. These should be deeply explored in future analyses and modelled in order to improve the results.

Since the limited amount of data might make the result obtained through 10-fold cross-validation unstable, the performance using leave-one-out cross-validation(LOOCV) was also evaluated. LOOCV was performed for the alternative with the best performance (Full-Set + LFS features with MLP1 classifier) and the baseline with best performance (baseline features with MLP2 classifier). As a result, UARs of 85.09% and 80.18% were obtained for the proposed features and the baseline, respectively. The accuracy achieved was 91.75% and 88.35% for the proposed features and the baseline, respectively. Therefore, the results obtained with LOOCV show an even better improvement (almost 5% for UAR) of the proposed approach over the baseline. Moreover, the performances for both alternatives were improved comparing the results obtained with LOOCV and 10-fold cross-validation. Given the small amount of data available, it is reasonable that the higher number of training examples used in each LOOCV iteration[8] helps the classifier to provide a better performance. These results suggest that the overall performance could be further improved if more data was available for training the classifiers.

The results can be further analysed by using confusion matrices. Confusion matrices give a good representation of the results per each class, which allows making a detailed analysis of performance and finding the main classification errors. The confusion matrix (adding all partitions) of our best model (MLP1 and Full-Set + LFS) is shown in Table 5. The rows correspond to the actual class labels, the columns show the predicted labels of bird species, and the main diagonal indicates the species that are correctly recognised. In this matrix, there are no-major errors and the unbalance between the number of examples per species can be noticed. Some confusions (underlined numbers) might be due to the small amount of available patterns for these species when the model is trained (see FuR, GeC, PhS and PhSt in Table 5). The remaining confusions may be explained by the acoustic likeness between species. By contrast, species of the same genus are not confused. Nevertheless, a deeper acoustic analysis would be very useful to define these “similarities”. The acoustic similarities could be exploited to define groups of species without taking into account information from the traditional taxonomy of the bird family. Therefore, a hierarchical classification scheme could be defined (Albornoz et al., 2011; Lee et al., 2009), which allows the mistakes to be addressed more efficiently, classifying these groups at a first stage and then, the more confusing species within the groups.

Fig. 6 shows spectrograms of vocalisation segments from species *Limnoctites rectirostris* (LiR), *Phleocryptes melanops* (PhM), *Upucerthia dumetaria* (UpD), and *Phacellodomus sibilatrix* (PhS). Examples from these species were selected because they are highly confused by the model, as presented in Table 5. The spectral characteristics of all the four vocalisations are very similar. For example, they show successive high energy peaks, which are regular in time and centred around 5000 Hz. Similarly, all the spectrograms present some weaker energy peaks around 10 kHz, which are also regular in time. Since most of the features we considered are based on the spectrum, the auditory spectrum and the spectrogram, it is reasonable that these species be misclassified. Therefore, in order to obtain high a performance for these four species, it would probably be appropriate to include some features based on temporal dynamics of the vocalisations, or to consider a dynamics models for the classification, like hidden Markov models (Ephraim, 2013).

5. Conclusions and future work

The identification of bird species is of increasing importance for ecologists in order to monitor the terrestrial environment, as it reflects important ecosystem processes and human activities. This study explores the bird call classification using speech-related features, and compares the performance using different classification techniques and configurations. Species from the family Furnariidae in the Paranaense Littoral region were analysed, which are well-known in the community but were never studied considering a big group. In addition, our work was motivated by the hypothesis that an extended state-of-the-art feature set, defined for speech-related tasks, would obtain a better performance than the feature set used at present.

The research demonstrated that the baseline results can be improved using additional LLDs, keeping low-dimensional data. The results were poorer when the Full-Set was used, which is expectable due to the high dimensionality of data and the number of samples used to train the multi-class models. This means that the large number of inputs makes the model more complex, and the scarce number of examples available is not enough for appropriately training it. Finally, the best performances (ACC and UAR) were obtained,

[8] It is compared to the number of training examples in each fold for 10-fold cross-validation.

Table 5
Confusion matrix for the MLP1 and Full-Set+LFS. References for the classes are included in the Supplementary material.

Species	CeC	CoA	CrP	CrS	FuR	GeC	LiC	LiR	PhM	SpM	TaC	UpD	AnA	AsB	AsH	LeP	PhR	PhS	PhSt	PsL	ScP	SyA	SyF	SyR	SyS	#
CeC	**11**																									11
CoA		**13**																								13
CrP			**4**													1										5
CrS				**6**																						6
FuR	1				**0**												1									2
GeC						**2**																			1	3
LiC							**4**						1	1	1											7
LiR								**9**					1													10
PhM								1	**3**							1								1		6
SpM										**5**																5
TaC											**11**								1							12
UpD												**8**														8
AnA							1						**14**													15
AsB		1												**7**												8
AsH															**7**											7
LeP			1													**7**										8
PhR																	**9**				1					10
PhS								1	1			1						**1**								4
PhSt									1										**2**							3
PsL																				**4**						4
ScP																					**9**					9
SyA												1										**18**				19
SyF																							**15**			15
SyR																			1					**3**		4
SyS																									**12**	12
#																										206

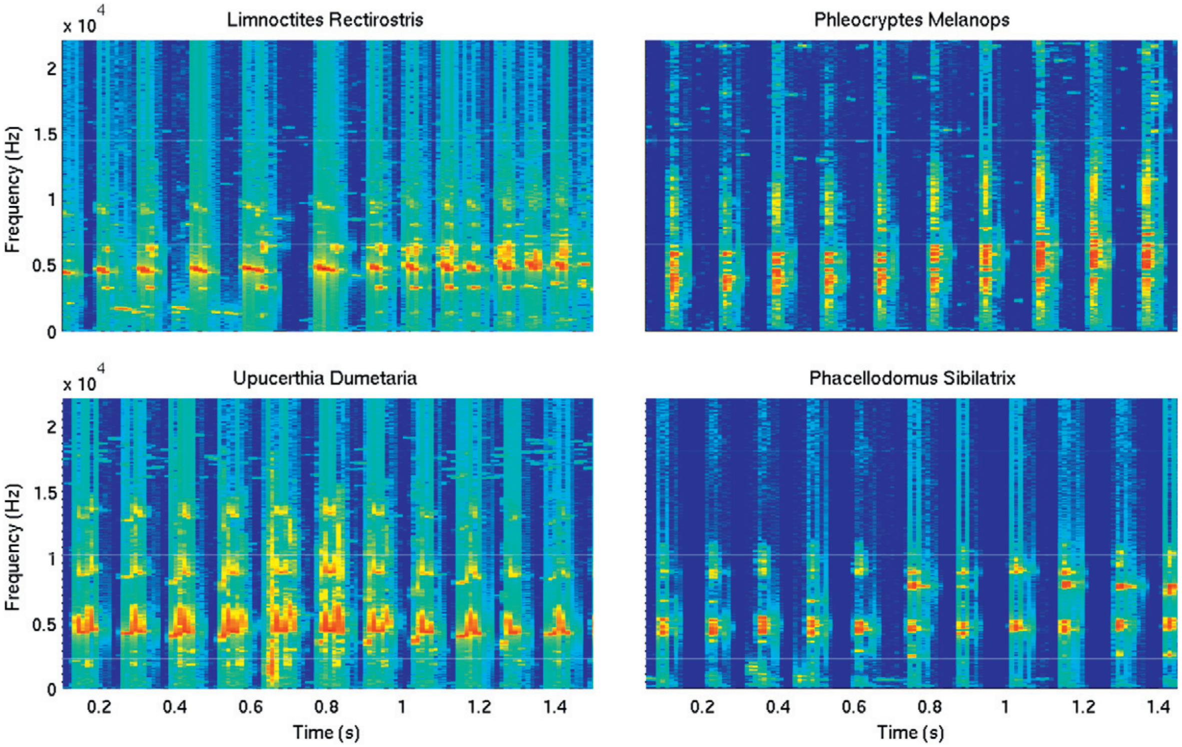


Fig. 6. Spectrograms of vocalisation segments from species *Limnoctites rectirostris* (LiR), *Phleocryptes melanops* (PhM), *Upucerthia dumetaria* (UpD) and *Phacellodomus sibilatrix* (PhS).

keeping a low dimensionality, when feature selection techniques were used. This indicates that said techniques are appropriate for extracting the more discriminative information from the full set of features, and exhibit a good behaviour with unbalanced data. Particularly, the best result is reached using a MLP classifier and the LFS technique. From an ecological monitoring and management point of view, our approach would be useful for developing autonomous tools that allow ornithologists to know which species are present in particular areas. Specifically, it could reduce the effort of manually reviewing recordings of Furnariidae species for labelling. Moreover, it would enable ornithologists to perform remote and simultaneous monitoring in different areas.

In future research, the model will be improved to detect more than one species in each audio file, performing a dynamic analysis of the vocalisations, i.e. frame by frame instead of using static (averaged) features. This could be achieved by matching every frame with short "templates" (Wachter et al., 2007) that should be first obtained for the species. Said matching could be done in terms of cross correlation (Tsai and Lin, 2003) or dynamic time warping (Müller, 2007). Then, a "dictionary" should be built including several templates that capture the characteristics of each species. In addition, it would be interesting to extend this research to perform the classification considering a large number of families with all the genus and species included. A hierarchical classification scheme could also be used, in which the first step would classify bird families, the second step would classify genus and the last step would determine the species. This means that the first classifier would focus on families only. The second step would consist of a set of different classifiers, each of which would be trained to recognise the genus of a particular family, which would be determined in the previous step. Finally, the last step would consist of a classifier for each of the genus under study, which would determine the species given the genus predicted in the previous step. The possibility of developing a semi-automatic tool to provide a list of the most probable species could be also considered. Ornithologists could then select the correct species from the list provided, based on their expertise.

6. Web-demo for reproducible research

A web interface was implemented using the web-demo tool (Stegmayer et al., 2016) in order to obtain further details and test our proposal with some experimental setups. This web interface is available at http://fich.unl.edu.ar/sinc/blog/web-demo/furnariidae/. Also, an android application with the same functionalities was developed, which can be downloaded from the mentioned web page. The system can be tested using an example register or uploading a register. The preprocessing can be set to use or not to use Wiener-based filter and acoustic activity detector. Then, after the feature extraction process, the sample is classified by the best model trained using all the reported data. Moreover, the graphical results of the audio file preprocessing, the features file (arff format), the trained model and the recognised species are freely available for download.

Acknowledgments

The authors would like to thank the National Agency for Scientific and Technological Promotion (ANPCyT)(with PICT #2014-1442 and PICT-2015-977), Universidad Nacional del Litoral (withPACT 2011 #58, CAI+D 2011 #58-511, CAI+D 2011 #58-525), Herpetology Laboratory from INALI (CONICET), and the National Scientific and Technical Research Council (CONICET), from Argentina, for their support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ecoinf.2017.01.004.

References

Albornoz, E.M., Milone, D.H., Rufiner, H.L., 2011. Spoken emotion recognition using hierarchical classifiers. Comput. Speech Lang. 25 (3), 556–570.

Alpaydin, E., 2010. Introduction to Machine Learning. 2nd, The MIT Press.

Areta, J.I., Pearman, M., 2009. Natural history, morphology, evolution, and taxonomic status of the earthcreeper *Upucerthia saturatior* (Furnariidae) from the Patagonian forests of South America. The Condor 111 (1), 135–149.

Areta, J.I., Pearman, M., 2013. Species limits and clinal variation in a widespread high Andean furnariid: the buff-breasted earthcreeper (*Upucerthia validirostris*). The Condor 115 (1), 131–142.

Arzamendia, V., Giraudo, A.R., 2009. Influence of large South American rivers of the Plata basin on distributional patterns of tropical snakes: a panbiogeographical analysis. J. Biogeogr. 36 (9), 1739–1749.

Bergmann, H., Schottler, B., 2001. Tenerife robin *Erithacus* (*Rubecula*) *superbus*—a species of its own. Dutch Birding 23, 140–146.

Betts, M., Mitchell, D., Diamond, A., Bêty, J., 2007. Uneven rates of landscape change as a source of bias in roadside wildlife surveys. J. Wildl. Manag. 71 (7), 2266–2273.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X.Z., Raich, R., Hadley, S.J., Hadley, A.S., Betts, M.G., 2012. Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. J. Acoust. Soc. Am. 131 (6), 4640–4650.

Burkart, R., Bárbaro, N., Sánchez, R., Gómez, D., 1999. Eco-Regiones de la Argentina. Administración de Parques Nacionales (APN). Secretaría de Recursos Naturales y Desarrollo Sostenible, Presidencia de la Nación Argentina.

Byers, B.E., 1996. Geographic variation of song form within and among chestnut-sided warbler populations. The Auk 288–299.

Clements, J., Schulenberg, T., Iliff, M., Roberson, D., Fredericks, T., Sullivan, B., Wood, C., 2015. The eBird/Clements checklist of birds of the world.

Ríos., E. E., Contreras, J.R., Agnolin, F., Davies, Y.E., Godoy, I., Giacchino, A., 2014. Atlas ornitogeográfico de la provincia de Formosa. Vazquez Mazzini.

Cramer, E.R., 2013. Measuring consistency: spectrogram cross-correlation versus targeted acoustic parameters. Bioacoustics 22 (3), 247–257.

de Oliveira, A.G., Ventura, T.M., Ganchev, T.D., de Figueiredo, J.M., Jahn, O., Marques, M.I., Schuchmann, K.-L., 2015. Bird acoustic activity detection based on morphological filtering of the spectrogram. Appl. Acoust. 98, 34–42.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7 (Jan), 1–30.

Dong, X., Towsey, M., Truskinger, A., Cottman-Fields, M., Zhang, J., Roe, P., 2015. Similarity-based birdcall retrieval from environmental audio. Eco. Inform. 29, Part 1, 66–76.

Dufour, O., Artieres, T., Glotin, H., Giraudet, P., 2014. Soundscape semiotics — localization and categorization. Ch. Clusterized Mel Filter Cepstral Coefficients and Support Vector Machines for Bird Song Identification. InTech Open Book.

Ephraim, Y., 2013. Hidden Markov models. Encycl. Oper. Res. Manag. Sci. 704–708.

Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013, October. Recent developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. 21st ACM International Conference on Multimedia, Barcelona, Spain. pp. 835–838.

Fagerlund, S., 2007. Bird species recognition using support vector machines. EURASIP J. Appl. Signal Process. 2007 (1), 64-64

Fjeldså, J., Irestedt, M., Ericson, P.G., 2005. Molecular data reveal some major adaptational shifts in the early evolution of the most diverse avian family, the Furnariidae. J. Ornithol. 146 (1), 1–13.

Ganchev, T.D., Jahn, O., Marques, M.I., de Figueiredo, J.M., Schuchmann, K.-L., 2015. Automated acoustic detection of *Vanellus chilensis lampronotus*. Exp. Syst. Appl. 42 (15-16), 6098–6111.

Garcıá-Moreno, J., Arctander, P., Fjeldså, J., 1999. A case of rapid diversification in the neotropics: phylogenetic relationships among *Cranioleuca spinetails* (Aves, Furnariidae). Molecular phylogenetics and evolution 12 (3), 273–281.

Giannakopoulos, T., Pikrakis, A., 2014. Introduction to Audio Analysis: A MATLAB® Approach. first, Academic Press, Oxford.

Giannoulis, D., Benetos, E., Stowell, D., Rossignol, M., Lagrange, M., Plumbley, M. D., Detection and classification of acoustic scenes and events: an IEEE AASP challenge. In: Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).

Glotin, H., Clark, C., LeCun, Y., Dugan, P., Halkias, X., Sueur, J., 2013. Bioacoustic challenges in icml4b. Proc. of 1st workshop on Machine Learning for Bioacoustics. No. USA. ISSN 979-10-90821-02-6.

Gütlein, M., Frank, E., Hall, M., Karwath, A., 2009. Large-scale attribute selection using wrappers. IEEE Symposium on Computational Intelligence and Data Mining, 2009. CIDM'09. pp. 332–339.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. SIGKDD Explor. 11 (1), 10–18.

Hall, M.A., 1998. Correlation-based Feature Subset Selection for Machine Learning. Ph.D. thesis. University of Waikato, Hamilton, New Zealand.

Harris, C.M., 1963. Absorption of sound in air in the audio-frequency range. J. Acoust. Soc. Am. 35 (1), 11–17.

Harris, C.M., 1966. Absorption of sound in air versus humidity and temperature. J. Acoust. Soc. Am. 40 (1), 148–159.

Haykin, S., 1998, Jul. Neural Networks: A Comprehensive Foundation. 2nd, Prentice Hall.

Hearst, M.A., Dumais, S.T., Osman, E., Platt, J., Scholkopf, B., 1998. Support vector machines. IEEE Intell. Syst. Appl. 13 (4), 18–28.

Hermansky, H., Morgan, N., 1994. Rasta processing of speech. IEEE transact. speech audio process. 2 (4), 578–589.

Hesler, N., Mundry, R., Dabelsteen, T., 2011. Does song repertoire size in common blackbirds play a role in an intra-sexual context? J. Ornithol. 152 (3), 591–601.

Irestedt, M., Fjeldså, J., Dalén, L., Ericson, P.G., 2009. Convergent evolution, habitat shifts and variable diversification rates in the ovenbird-woodcreeper family (Furnariidae). BMC Evol. Biol. 9 (1), 1.

Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.-P., Planque, R., Rauber, A., Fisher, R., Müller, H., 2014. Lifeclef 2014: Multimedia life species identification challenges. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms (Eds.), E., Information Access Evaluation. Multilinguality, Multimodality, and Interaction. Lecture Notes in Computer Science 8685. Springer International Publishing., pp. 229–249.

Keen, S., Ross, J.C., Griffiths, E.T., Lanzone, M., Farnsworth, A., 2014. A comparison of similarity-based approaches in the classification of flight calls of four species of North American wood-warblers (Parulidae). Eco. Inform. 21, 25–33.

Laje, R., Mindlin, G.B., 2003. Highly structured duets in the song of the South American hornero. Phys. Rev. Lett. 91 (25), 258104.

Lee, C., Mower, E., Busso, C., Lee, S., Narayanan, S., 2009. Emotion recognition using a hierarchical binary decision tree approach. Proc. Interspeech, ISCA. pp. 320–323.

León, E., Beltzer, A., Quiroga, M., 2014. El jilguero dorado (*Sicalis flaveola*) modifica la estructura de sus vocalizaciones para adaptarse a hábitats urbanos [the saffron finch (*Sicalis flaveola*) modifies its vocalizations to adapt to urban habitats]. Revista mexicana de biodiversidad 85 (2), 546–552.

Leon, E.J., Beltzer, A.H., Olguin, P.F., Reales, C.F., Urich, G.V., Alessio, V., Cacciabué, C.G., Quiroga, M.A., 2015. Song structure of the golden-billed saltator (*Saltator aurantiirostris*) in the middle Parana river floodplain. Bioacoustics 24 (2), 145–152.

Louette, M., Bijnens, L., Upoki Agenong'a, D., Fotso, R., 1995. The utility of birds as bioindicators: case studies in Equatorial Africa. Belg. J. Zool. 125 (1), 157–165.

Lynch, A., Baker, A.J., 1994. A population memetics approach to cultural evolution in chaffinch song: differentiation among populations. Evolution 351–359.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Andrew Royle, J., Langtimm, C.A., 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology 83 (8), 2248–2255.

Marler, P., 1997. Three models of song learning: evidence from behavior. J. Neurobiol. 33 (5), 501–516.

Michie, D., Spiegelhalter, D., Taylor, C., 1994. Machine Learning, Neural and Statistical Classification. Ellis Horwood, University College, London.

Molau, S., Pitz, M., Schluter, R., Ney, H., 2001. Computing mel-frequency cepstral coefficients on the power spectrum. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001. vol. 1. pp. 73–76.

Müller, M., 2007. Dynamic time warping. Inf. retr. music. motion 69–84.

Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT press.

Narosky, T., Yzurieta, D., 2010. Aves de Argentina y Uruguay-Birds of Argentina & Uruguay: Guía de Identificación Edición Total-A Field Guide Total Edition. 16th, Buenos Aires., pp. 432.

Noriega, J.I., 1991. Un nuevo género de Furnariidae (ave: Passeriformes) del pleistoceno inferior-medio de la provincia de Buenos Aires, Argentina. Ameghiniana 28, 317–323.

Olson, S.L., Irestedt, M., Ericson, P.G., Fjeldså, J., 2005. Independent evolution of two Darwinian marsh-dwelling ovenbirds (Furnariidae: Limnornis, Limnoctites). Ornitologia Neotropical 16, 347–359.

Pacheco, J.F., Gonzaga, L.P., 2013. A new species of synallaxis of the Ruficapilla/Infuscata complex from eastern Brazil (Passeriformes: Furnariidae). Revista Brasileira de Ornitologia-Brazilian J. Ornithol. 3 (3), 10.

Päckert, M., Martens, J., Kosuch, J., Nazarenko, A.A., Veith, M., 2003. Phylogenetic signal in the song of crests and kinglets (Aves: Regulus). Evolution 57 (3), 616–629.

Payne, R.B., 1996. Song traditions in indigo buntings: origin, improvisation, dispersal, and extinction in cultural evolution. Ecology and evolution of acoustic communication in birds, , pp. 198–220.

Planqué, B., Vellinga, W.-P. Xeno-cano.org. Accessed: 2015-07-10, URL http://www.xeno-canto.org.

Plapous, C., Marro, C., Scalart, P., 2006. Improved signal-to-noise ratio estimation for speech enhancement. IEEE Transact. Audio, Speech, Lang. Process. 14 (6), 2098–2108.

Potamitis, I., 2015. Unsupervised dictionary extraction of bird vocalisations and new tools on assessing and visualising bird activity. Eco. Inform. 26, Part 3, 6–17.

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

Ptacek, L., Machlica, L., Linhart, P., Jaska, P., Muller, L., 2015. Automatic recognition of bird individuals on an open set using as-is recordings. Bioacoustics 25 (1), 1–19.

Raposo, M.A., Höfling, E., 2003. Overestimation of vocal characters in suboscine taxonomy (Aves: Passeriformes: Tyranni): causes and implications. Lundiana 4 (1), 35–42.

Roch, M.A., Soldevilla, M.S., Burtenshaw, J.C., Henderson, E.E., Hildebrand, J.A., 2007. Gaussian mixture model classification of Odontocetes in the Southern California Bight and the Gulf of California. J. Acoust. Soc. Am. 121 (3), 1737–1748.

Rosenberg, A., 2012. Classifying skewed data: importance weighting to optimize average recall. INTERSPEECH 2012, Portland, USA.

Schottler, B., Canary Islands blue tits (Parus caeruleus ssp.)—differences and variation in territorial song: preliminary results, 1993.

Schuller, B., Steidl, S., Batliner, A., 2009. The INTERSPEECH 2009 emotion challenge. Proc. Interspeech, ISCA 312–315.

Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., Zhang, Y., 2014. The INTERSPEECH 2014 computational paralinguistics challenge: cognitive & physical load. Proc. Interspeech, ISCA 427–431.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J., 2011. The INTERSPEECH 2011 Speaker State Challenge. Proc. Interspeech, ISCA 3201–3204.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., Kim, S., 2013. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. Proc. Interspeech, ISCA 148–152.

Seddon, N., Tobias, J.A., 2010. Character displacement from the receiver's perspective: species and mate recognition despite convergent signals in suboscine birds. Proc. R. Soc. Lond. B Biol. Sci. 1–9.

Spampinato, C., Mezaris, V., Huet, B., van Ossenbruggen, J., 2014. Editorial — special issue on multimedia in ecology. Eco. Inform. 23, 1–2. Special Issue on Multimedia in Ecology and Environment.

Stegmayer, G., Pividori, M., Milone, D.H., 2016. A very simple and fast way to access and validate algorithms in reproducible research. Brief. Bioinform. 17 (1), 180–183.

Stowell, D., Plumbley, M.D., 2013a. Feature design for multilabel bird song classification in noise. NIPS4B 2013 Bird Challenge.

Stowell, D., Plumbley, M.D., 2013b. Segregating event streams and noise with a Markov renewal process model. J. Mach. Learn. Res. 14, 1891–1916.

Towsey, M., Wimmer, J., Williamson, I., Roe, P., 2014. The use of acoustic indices to determine avian species richness in audio-recordings of the environment. Eco. Inform. 21, 110–119.

Truskinger, A., Towsey, M., Roe, P., 2015. Decision support for the efficient annotation of bioacoustic events. Eco. Inform. 25, 14–21.

Tsai, D.-M., Lin, C.-T., 2003. Fast normalized cross correlation for defect detection. Pattern Recogn. Lett. 24 (15), 2625–2631. http://dx.doi.org/10.1016/S0167-8655(03)00106-5.

Vapnik, V., Cortes, C., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.

Ventura, T.M., de Oliveira, A.G., Ganchev, T.D., de Figueiredo, J.M., Jahn, O., Marques, M.I., Schuchmann, K.-L., 2015. Audio parameterization with robust frame selection for improved bird identification. Exp. Syst. Appl. 42 (22), 8463–8471.

Vuilleumier, F., Vaurie, C., 1980. Taxonomy and geographical distribution of the Furnariidae (Aves, Passeriformes). Bull. Am. Mus. Nat. Hist. 166, 1–357.

Wachter, M.D., Matton, M., Demuynck, K., Wambacq, P., Cools, R., Compernolle, D.V., 2007, May. Template-based continuous speech recognition. IEEE Transact. Audio, Speech, Lang. Process. 15 (4), 1377–1390. http://dx.doi.org/10.1109/TASL.2007.894524.

Xiong, Z., Zheng, T.F., Song, Z., Soong, F., Wu, W., 2006. A tree-based kernel selection approach to efficient Gaussian mixture model-universal background model based speaker identification. Speech Comm. 48 (10), 1273–1282.

Xu, L., Yan, P., Chang, T., 1988, Nov. Best first strategy for feature selection. 9th International Conference on Pattern Recognition. vol. 2. pp. 706–708.

Zheng, F., Zhang, G., Song, Z., 2001. Comparison of different implementations of mfcc. J. Comput. Sci. Technol. 16 (6), 582–589.

Zhou, X., Garcia-Romero, D., Duraiswami, R., Espy-Wilson, C., Shamma, S., 2011. Linear versus mel frequency cepstral coefficients for speaker recognition. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2011. pp. 559–564.

Zimmer, K.J., Whittaker, A., 2000. The rufous cacholote (Furnariidae: Pseudoseisura) is two species. The Condor 102 (2), 409–422.

Zollinger, S.A., Brumm, H., 2015. Why birds sing loud songs and why they sometimes don't. Anim. Behav. 105, 289–295.