Furnariidae species recognition using speech-related features and machine learning

Leandro D. Vignolo[1⋆], Juan A. Sarquis[2], Evelina Leon[2] and Enrique M. Albornoz[1]

[1] Research Institute for Signals, Systems and Computational Intelligence, sinc(*i*), UNL-CONICET. Ciudad Universitaria, Paraje El Pozo, S3000, Santa Fe, Argentina.
[2] National Institute of Limnology, INALI, CONICET
Ciudad Universitaria, Paraje El Pozo, S3000, Santa Fe, Argentina.

Abstract. The automatic classification of calling bird species is important to achieve more exhaustive environmental monitoring and to manage natural resources. Bird vocalizations allow to recognise new species, their natural history and macro-systematic relations, while automatic systems can speed up and improve all the process. In this work, we use state-of-art features designed for speech and speaker state recognition to classify 25 species of Furnariidae family. Since Furnariidae species inhabit the Litoral Paranaense region of Argentina (South America), this work could promote further research on the topic and the implementation of in-situ monitoring systems. Our analysis includes two widely-known classification techniques: random forest an support vector machines. The results are promising, near 86%, and were validated in a cross-validation scheme.

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

1 Introduction

Vocalisations are often the most noticeable manifestations about the presence of avian species in different habitats [22]. Concurrently, birds have been widely used to indicate biodiversity due to 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) [16]. The automatic acoustic tools are useful to collect data from several patterns of bird populations, however, they have experimented several problems as poor sample representation in remote regions, observer bias, defective monitoring and, expensive costs of sampling on large spatial and temporal scales, among others [14,5]. Bird vocalisations have become an important research field, influencing ethology, taxonomy and the evolutionary biology [11,25,21]. In ecosystems monitoring, the technologies are important to record and process the registers and to improve the data collection in large and disjoint areas [31].

⋆ Autor corresponsal: ldvignolo@sinc.unl.edu.ar

Recognition of individuals in passerine species is a challenging task due to they have complex songs and can adapt their content over time. The Furnariidae family have 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 from the same family can 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 vocalizations more difficult. Previous works demonstrate that males and females have differences in tone and among note intervals. Although some works describe the vocalisation changes in some Furnariidae species [32,4,23], no one evaluate several vocalisations of Furnariidae species simultaneously in South America. In this work we analyse vocalisations belonging to 25 Furnariidae species which are distributed in the Litoral Paranaense region. This Region is formed by the Mesopotamia Argentina (Misiones, Corrientes and Entre Rios provinces) plus Chaco, Formosa and Santa Fe provinces, and it is lapped by great rivers of the Plata basin. In the last years, this region became an interesting place to study bird vocalisations [4,15]. In addition, expert people and Furnariidae species availability would allow us to record and analyse these species in real-life conditions in future works.

The access to multimedia data has promoted an interdisciplinary and collaborative science in order to analyse the environment. Although, Human experts (sufficiently trained) can recognise bioacoustic events with a high performance, this is a laborious and expensive process. The full-automatic methods for vocalisations recognition are currently novel and they can be examined from some recent works [24,9]. Regarding to feature extraction, time and frequency based information have been employed [22,13]. In addition, characteristics originally developed for speech analysis are widespread in the context of bird-call recognition. Some of the widely-known features used in literature are Mel Frequency Cepstral Coefficients (MFCCs), Linear Frequency Cepstral Coefficients (LFCCs), and standard functionals (mean, standard deviation, kurtosis, etc.) computed over these [24,8]. A broad range of classifiers has been applied to bird-call classification: Gaussian Mixture Model (GMM), Gaussian Mixture Model-Universal Background Model (GMM-UBM), Support Vector Machine (SVM), Random Forest, among others [24,9,13]. However, none of these works have addressed the vocalisation recognition of species belonging to the same family, which present similar parameters in their vocalisations. Moreover, only a small part of the state-of-the-art speech features [27] have been employed in bird identification tasks.

In this work, we develop a novel approach to bird-call classification to deal with 25 Furnariidae species. It is known that passerines produce complex vocalizations, and the vocalizations of the species under study are similar. All these make the task at hand an interesting challenge, which most of previous work avoided by considering a pool of diverse species instead. Our model uses state-of-art classifiers with speech-related parameterizations. The model is tested in a cross-validation scheme in all cases. In the next section, the proposed features

and classifiers are introduced. Section 3 deals with the experimental set-up, presents the implementation details, explains the validation scheme and discusses the results. Finally, conclusions and future works are presented in the last section.

2 Proposed features and classifiers

As mentioned, the use of speech based features is known in bird-call analysis, identification and classification. The standard sets have shown good performance [8,24]. An extended state-of-art set of features related to human speech is introduced below.

Speech inspired features In speech processing area, researchers have made a great effort to find the best set of features in order to perform speech recognition, speaker recognition, emotion recognition, illness state detection, etc. [29,28]. In The INTERSPEECH 2013 ComParE Challenge [29], a set of 6373 features was presented and it represents the state-of-art for speech processing area. The feature set is built from 65 low-level descriptors (LLD) as energy, spectral, cepstral (MFCC), voicing related (F_0, shimmer, jitter, etc.), zero crossing rate, logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, psychoacoustic spectral sharpness, and their deltas (meaning their first temporal derivatives). These features are computed on a time frame basis, using a window of 60 ms with 10 ms step for F_0 (pitch) and zero crossing rate. The remaining features are computed using 20 ms window size, and the time contour of each attribute is smoothed by a moving average filter. Then, specific functionals are computed for each LLD set. These include the arithmetic mean, maximum, minimum, standard deviation, skewness, kurtosis, mean of peak distances, among others. Besides the complete feature set obtained combining all LLD and functionals (Full-Set), we also propose a subset consisting on the complete set of functionals computed only from the MFCCs, which results in a set of 531 attributes (MFCC+LLD).

2.1 Classifiers

Several techniques from machine learning and computational intelligence have been used in call bird identification [24]. Therefore, we briefly introduce two widespread techniques: Random Forest and Support Vector Machines.

Random Forest The classification and regression tree (CART) models, so called decision trees, are widely known in machine learning and data mining [18]. Some relevant properties are the robustness to several feature transformations as scaling, the ability to discriminate irrelevant information while producing easily analysable models. These are constructed by recursive partitioning the input space (usually represented by a tree), and then, region-specific models are defined for the resulting scheme [6]. Random Forest (RF) is an ensemble learning

method whose decision is based on the average of multiple CARTs, trained on different parts of the same training set, with the goal of reducing the variance of CART overfitting. The computation can be expressed in terms of the *bagging* technique [18] as $f(\mathbf{x}) = \frac{1}{K}\sum_{k=1}^{K} t_k(\mathbf{x})$, where t_k is the k-th tree.

Support Vector Machine A SVM is a supervised learning method widely used for pattern classification, which has theoretically good generalisation capabilities. Its aim is to find a hyperplane able to separate input patterns in a sufficiently high dimensional space. The distances from the hyperplane to the closest patterns, on each side, is called margin. This margin need to be maximised to reach the best generalisation, and in the binary case, it is done by finding the $\mathbf{w}$ and w_0 parameters. These are usually found by performing a standard quadratic optimisation [2]:

$$\begin{aligned} & \min \frac{1}{2} \parallel \mathbf{w} \parallel^2 \\ \text{subject to} \quad & \\ & r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq +1, \forall t \end{aligned} \tag{1}$$

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 non-linear problem could be solved as a linear problem in a new space by doing a non-linear transformation [2]. Then, the new dimensions are computed using the basis functions by inner product. The *kernel trick* is a method to solve this problem without mapping the features in the new space, therefore, the kernel function is applied in the original space [2].

3 Experiments and results

In this section we describe the experimental framework used in this work. We first discuss why and how we have chosen the bird species from the known databases. Secondly, implementation details of the feature extraction and classifiers are presented. Then we explain the validation scheme and discuss the results.

3.1 Target species and birds call corpus

The Study area is located between $22°25'S$ $62°12'W$ and $38°0'S$ $57°26'W$ that integrates several eco-regions along the Paraná river. These regions are dry-Chaco, Espinal, Pampa, Iberá Wetlands and Paraná Islands and Delta.

We decided to study the Furnariidae family due to it presents diverse vocalisations and even some species can sing in duet, however, the experts can usually identify these reaching a good performance. The vocalisations obtained from species of this family could be similar and then, difficult to classify. The Furnariidae family includes 68 genera integrated for 302 species [7], distributed in South America and in a region of Central America [20] forms one of the most

impressive examples of continental adaptive radiation. This family has probably the highest morpho-ecological diversity in birds, inhabiting diverse habitats as desert or arid regions, rocky coasts, ravines, swamps, grasslands and forests [12].

As mentioned in [30], new frontiers has been opened in ecology beyond the analysis performed by expert ecologists due to the propagation of projects as *Xeno-canto*[3] and *EcoGrid*[4]. There, the scientific community evaluates and shares photographies, audio-video recordings, annotations and records about the geographical distribution of birds, as well as other taxonomic groups. We decided to use the *Xeno-canto* project because several state-of-art works recommend its usage [30,22]. Additional records were taken from the *Birds of Argentina & Uruguay: A Field Guide Total Edition* [19], which has almost all the vocalisations of birds that inhabit in Argentina and Uruguay. This database is widely used by the research works performed on this region [3,15]. This combination of data involves an additional complexity that the model should be able to handle.

3.2 Feature extraction and classifiers

There is not a suitable baseline model available, to our knowledge, that allows compare the performance of our proposal. In order to create the baseline, we used the classifiers and the feature set proposed for the task of bird song identification in [8]. The features are functionals computed over the MFCCs (delta coefficients, acceleration coefficients, mean and variance), and then a 102-dimensional vector is obtained for each recording.

Previous to feature extraction, we implemented two processes: a Wiener-based noise filter to reduce noise in the recordings[5] and a detector of acoustic activity using a voice activity detector (VAD) based on Rabiner and Schafer method [10]. Then, the *OpenSMILE* toolkit[6] was used to extract the state-of-art features [29]. The WEKA and Scikit-Neuralnetwork[7] libraries were used to apply RF and SVM classifiers. The SVMs were trained using the Sequential Minimal Optimisation algorithm, considering the polynomial kernel. The RF was implemented following [6], using 10 and 100 trees with unlimited depth. Note that the parameters for all the classifiers where determined based on the results obtained for the baseline features in preliminary experiments, and where not optimized for the proposed feature sets.

3.3 Results and discussion

For all the experiments, the feature vectors were normalised using maximum and minimum values (for each dimension) from the training set. To avoid estimation biases, a cross-validation with the k-fold method was performed [17]. For each

[3] http://www.xeno-canto.org/.
[4] http://www.aiai.ed.ac.uk/project/ecogrid/.
[5] As all utterances have an initial silence, the noise could be easily modelled.
[6] Software available at `http://www.audeering.com/research/opensmile/`.
[7] Software available at `http://scikit-neuralnetwork.readthedocs.org`

Table 1: Weighted average recall and unweighted average recall [%].

		WAR			**UAR**		
	Dim.	RF10	RF100	SVM	RF10	RF100	SVM
Baseline	102	68.45	80.10	84.95	58.25	67.00	74.07
MFCC+LLD	531	69.42	83.01	**85.92**	58.08	70.43	**75.18**
Full-Set	6373	68.93	80.10	83.50	55.74	65.24	72.46

experiment, the classification accuracy was computed by 10-fold stratified cross-validation (SCV), where each fold had 90% of data for training and the remaining 10% for test. Finally, the results were averaged over the 10 test sets. To evaluate the performance of the models, we computed the weighted average recall (WAR or *accuracy*) as the number of correctly classified instances divided by the total number of instances. As the WAR can be biased when the classes are not well-balanced, we computed the *unweighted average recall* (UAR) that gives a more accurate estimation about the performance [26]:

$$\text{UAR} = \frac{1}{K} \sum_{i=1}^{K} \frac{A_{ii}}{\sum_{j=1}^{K} A_{ij}}, \tag{2}$$

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

The obtained results are presented in Table 1. The WAR rates represents the accuracy, while UAR shows the performances taking into accout the individual hit rates with the same weight for each class. It can be seen that the baseline features provide high accuracy rates for the classification of Furnariidae species. Even, these can not be improved using the full-set (more than 6000 features) which has very interesting information but it is too large for the current task. On the other hand, it is interesting to note that the proposed feature set *MFCC+LLD* performs better than the baseline keeping a low dimensionality in the feature vectors. Regarding the classifiers, the SVM provides better performances than RF for all the cases.

As the reached performance is very satisfactory (near to 86%) and the amount of data is limited, a statistical test for these results would not be relevant. This means that a reasonably larger amount of data would be required for a meaningful statistical test. However, it is important to remark that our results show that $5 - 7$ samples per species are sufficient to model and to predict them. Also note that the small amount of available data was exploited in the best possible way, in order to estimate the classification accuracy, by performing cross validation. Moreover, an important result of our experiments is that not many samples are required to obtain an appropiate model for each species.

The acoustic similarities should be explored to define groups of species without taking into account information from the traditional taxonomy of the bird family. Therefore, a hierarchical scheme of classification could be defined [1].

Which would allow to address the classification erros more efficiently, classifying groups of species at a first stage and then, in a second stage, the more similar species within each group.

4 Conclusions and future work

This work explores the bird-call classification using speech related features, and compares the performance using some classification techniques. We analysed species of Furnariidae family from Litoral Paranaense region, well-known in the community but never analysed in a big group. The results show that if the set is too large the performance is not improved, while if some relevant information is added to the standard set it can be reached better rates, keeping a reduced feature set. Summarising, the speech related features are really promising for the automatic bird-call classification. Moreover, since Furnariidae species belong to a specific region and where not yet subjected to similar studies, this work could promote further research on the topic and the implementation of in-situ monitoring systems.

In future work, the model will be improved to detect more than one species in each audio file, doing the correlation of features calculated on short-time frames. Also, we will investigate the performance of the proposed model using data from diverse bird families.

5 Acknowledgements

The authors would like to thank the *National Agency for Scientific and Technological Promotion* (ANPCyT)(with PICT #2014-1442) and *Universidad Nacional de Litoral* (with PACT 2011 #58, CAI+D 2011 #58-511, CAI+D 2011 #58-525), as well as the *National Scientific and Technical Research Council* (CONICET), from Argentina, for their support.

References

1. Albornoz, E.M., Milone, D.H., Rufiner, H.L.: Spoken emotion recognition using hierarchical classifiers. Computer Speech & Language 25(3), 556–570 (2011)
2. Alpaydin, E.: Introduction to Machine Learning. The MIT Press, 2nd edn. (2010)
3. Areta, J.I., Pearman, M.: 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 (2009)
4. Areta, J.I., Pearman, M.: Species limits and clinal variation in a widespread high andean furnariid: The buff-breasted earthcreeper (upucerthia validirostris). The Condor 115(1), 131–142 (2013)
5. Betts, M., Mitchell, D., Diamond, A., Bêty, J.: Uneven rates of landscape change as a source of bias in roadside wildlife surveys. The Journal of Wildlife Management 71(7), 2266–2273 (2007)
6. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)

7. Clements, J., Schulenberg, T., Iliff, M., Roberson, D., Fredericks, T., Sullivan, B., Wood, C.: The ebird/clements checklist of birds of the world (2015), www.birds.cornell.edu/clementschecklist/
8. Dufour, O., Artieres, T., Glotin, H., Giraudet, P.: Soundscape Semiotics - Localization and Categorization, chap. Clusterized Mel Filter Cepstral Coefficients and Support Vector Machines for Bird Song Identification. InTech Open Book (2014)
9. Ganchev, T.D., Jahn, O., Marques, M.I., de Figueiredo, J.M., Schuchmann, K.L.: Automated acoustic detection of vanellus chilensis lampronotus. Expert Systems with Applications 42(15-16), 6098 –6111 (2015)
10. Giannakopoulos, T., Pikrakis, A.: Introduction to Audio Analysis: A MATLAB® Approach. Academic Press, Oxford, 1st edn. (2014)
11. Hesler, N., Mundry, R., Dabelsteen, T.: Does song repertoire size in common blackbirds play a role in an intra-sexual context? Journal of Ornithology 152(3), 591–601 (2011)
12. Irestedt, M., Fjeldså, J., Dalén, L., Ericson, P.G.: Convergent evolution, habitat shifts and variable diversification rates in the ovenbird-woodcreeper family (furnariidae). BMC evolutionary biology 9(1), 1 (2009)
13. Keen, S., Ross, J.C., Griffiths, E.T., Lanzone, M., Farnsworth, A.: A comparison of similarity-based approaches in the classification of flight calls of four species of north american wood-warblers (parulidae). Ecological Informatics 21, 25–33 (2014)
14. Laje, R., Mindlin, G.B.: Highly structured duets in the song of the south american hornero. Physical review letters 91(25), 258104 (2003)
15. Leon, E.J., Beltzer, A.H., Olguin, P.F., Reales, C.F., Urich, G.V., Alessio, V., Cacciabué, C.G., Quiroga, M.A.: Song structure of the golden-billed saltator (saltator aurantiirostris) in the middle parana river floodplain. Bioacoustics 24(2), 145–152 (2015)
16. Louette, M., Bijnens, L., Upoki Agenong'a, D., Fotso, R.: The utility of birds as bioindicators: case studies in equatorial africa. Belgian Journal of Zoology 125(1), 157–165 (1995)
17. Michie, D., Spiegelhalter, D., Taylor, C.: Machine Learning, Neural and Statistical Classification. Ellis Horwood, University College, London (1994)
18. Murphy, K.P.: Machine learning: a probabilistic perspective. MIT press (2012)
19. Narosky, T., Yzurieta, D.: Aves de Argentina y Uruguay–Birds of Argentina & Uruguay: Guía de Identificación Edición Total–A Field Guide Total Edition. Buenos Aires, 16 edn. (2010)
20. Noriega, J.I.: Un nuevo género de furnariidae (ave: Passeriformes) del pleistoceno inferior-medio de la provincia de buenos aires, argentina. Ameghiniana 28, 317–323 (1991)
21. Päckert, M., Martens, J., Kosuch, J., Nazarenko, A.A., Veith, M.: Phylogenetic signal in the song of crests and kinglets (aves: Regulus). Evolution 57(3), 616–629 (2003)
22. Potamitis, I.: Unsupervised dictionary extraction of bird vocalisations and new tools on assessing and visualising bird activity. Ecological Informatics 26, Part 3, 6–17 (2015)
23. Potamitis, I., Ntalampiras, S., Jahn, O., Riede, K.: Automatic bird sound detection in long real-field recordings: Applications and tools. Applied Acoustics 80, 1–9 (2014)
24. Ptacek, L., Machlica, L., Linhart, P., Jaska, P., Muller, L.: Automatic recognition of bird individuals on an open set using as-is recordings. Bioacoustics 0(0), 1–19 (2015)

25. Raposo, M.A., Höfling, E.: Overestimation of vocal characters in suboscine taxonomy (aves: Passeriformes: Tyranni): causes and implications. Lundiana 4(1), 35–42 (2003)
26. Rosenberg, A.: Classifying skewed data: Importance weighting to optimize average recall. In: INTERSPEECH 2012. Portland, USA (2012)
27. Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., , Zhang, Y.: The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load. Proc. Interspeech, ISCA pp. 427–431 (Sep 2014)
28. Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J.: The INTERSPEECH 2011 Speaker State Challenge. Proc. Interspeech, ISCA pp. 3201–3204 (Aug 2011)
29. 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.: The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. Proc. Interspeech, ISCA pp. 148–152 (Aug 2013)
30. Spampinato, C., Mezaris, V., Huet, B., van Ossenbruggen, J.: Editorial - special issue on multimedia in ecology. Ecological Informatics 23, 1 – 2 (2014), special Issue on Multimedia in Ecology and Environment
31. Towsey, M., Wimmer, J., Williamson, I., Roe, P.: The use of acoustic indices to determine avian species richness in audio-recordings of the environment. Ecological Informatics 21, 110–119 (2014)
32. Zimmer, K.J., Whittaker, A.: The rufous cacholote (furnariidae: Pseudoseisura) is two species. The Condor 102(2), 409–422 (2000)