# How to correctly evaluate an automatic bioacoustics classification method

Juan G. Colonna[1], João Gama[2], and Eduardo F. Nakamura[1]

[1] Federal University of Amazonas (UFAM), Institute of Computing (Icomp),
Avenida General Rodrigo Octávio 6200, Manaus-AM, 69077-000, Brazil
{juancolonna,nakamura}@icomp.ufam.edu.br,
[2] Laboratory of Artificial Intelligence and Decision Support (LIAAD), INESC Tec,
Campus da FEUP, Rua Dr. Roberto Frias, Porto, 4200-465, Portugal
jgama@fep.up.pt

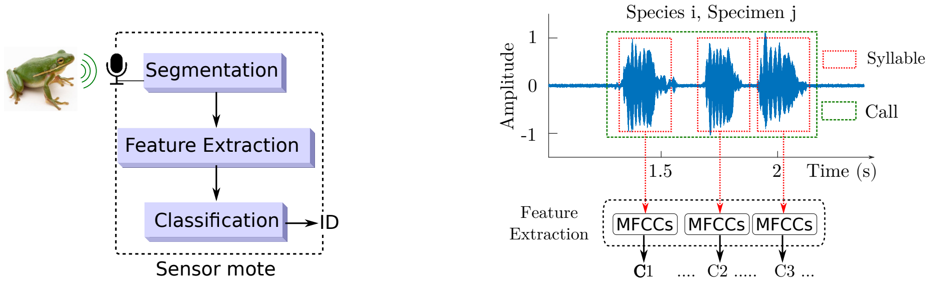
**Abstract.** In this work, we introduce a more appropriate (or alternative) approach to evaluate the performance and the generalization capabilities of a framework for automatic anuran call recognition. We show that, by using the common k-folds Cross-Validation (k-CV) procedure to evaluate the expected error in a syllable-based recognition system the recognition accuracy is overestimated. To overcome this problem, and to provide a fair evaluation, we propose a new CV procedure in which the specimen information is considered during the split step of the k-CV. Therefore, we performed a k-CV by specimens (or individuals) showing that the accuracy of the system decrease considerably. By introducing the specimen information, we are able to answer a more fundamental question: Given a set of syllables that belongs to a specific group of individuals, can we recognize new specimens of the same species? In this article, we go deeper into the reviews and the experimental evaluations to answer this question.

**Keywords:** Automatic anuran call recognition, Cross-Validation, Bioacoustics, One-against-All, One-against-One.

## 1 Introduction

Nowadays Wireless Acoustic Sensor Networks (WASNs) are used in several environmental applications including bioacoustic monitoring programs [1]. These networks are composed by small sensor nodes that can: collect, process and transmit the audio data and correlated environment variables. In this context, the problem of automatic bioacoustic monitoring can be addressed by embedding a Machine Learning (ML) classification technique into the sensor nodes [2, 3]. Thus, by combining ML and WASNs, we can identify different animal calls without human intervention. However, the low cost of the sensor nodes imposes restrictions on the hardware and software, and consequently, affects the classification techniques.

Among all the species commonly used in bioacoustic monitoring programs anuran (frogs and toads) are natural indicators of the environmental health [4].



(a) Automatic Call Recognition System (ACR).

(b) An audio record of the species Adenomera hylaedactyla.

Fig. 1: A framework for automatic frog's calls recognition.

Thus, tracking the variations of frog populations can help to identify ecological problems in early stages [5]. Moreover, with WASNs, we can develop an autonomous system to support researchers in long-term ecological studies.

The general framework for recognizing frog species, based on their calls, is shown in Figure 1(a). This system consists of three main blocks. The first block performs an acoustics signal segmentation that recognize the start and end time where a minor vocalization unit occur, named **syllable** (see Figure 1(b)) [6, 7]. The second block maps each syllable into a set of Low Level acoustic Descriptors (LLDs or feature vector). The last block, is a ML algorithm that makes a pattern matching between the unknown input feature vector and a feature set representing all the species included into the dataset (see Table 3).

In the related literature, presented in Section 3, we found works concerned with the segmentation and pre-processing steps [7], also works mainly concerned with feature analysis and selection [6, 8–10] and, finally, works comparing different ML algorithms for classification [11, 12]. These are examples on how this framework can be flexible. However, most of these systems are based on syllable recognition approaches that use Cross-Validation ($k$-CV) to evaluate the classification performance and the generalization capabilities of the system. In these cases, the $k$-CV procedure splits the dataset in two subsets: one for training and another for testing, ignoring if all the samples chosen (or syllables in this case) belong to the same individual (or specimen). This becomes a problem when syllables of one particular specimen are at the same time in these two subsets. When it happens, we noticed that the accuracy of the classifier increases being over estimated. Our new evaluation and validation proposal incorporates the specimen information as additional label and considers this new information during the $k$-CV split procedure to avoid mixing syllables from the same individual in the training and testing sets at the same time.

To the best of our knowledge, this is the first work proposing a CV strategy dividing the testing and training sets by specimens. We believe that specimen-based cross validation is the best way to test the generalization capabilities of recognition models, without falling in a bias problem and overestimate the final accuracy.

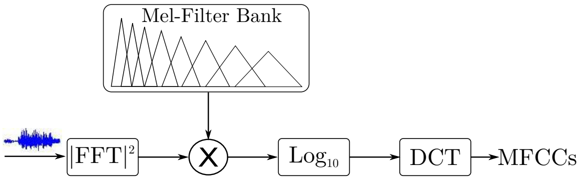


Fig. 2: MFCCs steps. Here, FFT stands for Fast Fourier Transform and DCT for Discrete Cosine Transform.

## 2 Fundamentals

Bioacoustics classification systems are traditionally composed of three main steps with different purposes (see Figure 1(a)). Formally, the input bioacoustic signal $X = \{x_1, x_2, \cdots, x_N\}$ is a time series of length $N$, in which its values represent the acoustics pressure levels (or amplitude). A syllable $\mathbf{x}_k = \{x_t, x_{t+1}, \cdots, x_{t+n}\}$ is a subset of $n$ consecutively signal values. Thus, the pre-processing step segments the signal $X$ by identifying the beginning and the endpoints of $\mathbf{x}_k$.

After the syllable extraction we need to represent each $\mathbf{x}_k$ by a set features, commonly called Low Level Descriptors (LLDs). The most frequent LLDs are the Mel-Frequency Spectral Coefficients (MFCCs). The MFCCs perform a spectral analysis based on a triangular filter-bank logarithmically spaced in the frequency domain (Figure 2) [2, 3, 12]. The feature extraction using the MFCCs allows to represent any syllable by a set of coefficients ($\text{MFCC}(\mathbf{x}_k) \rightarrow \mathbf{c}_k$), i.e.: $X \rightarrow \{(\mathbf{c}_1, y_i), (\mathbf{c}_2, y_i), \ldots, (\mathbf{c}_k, y_i)\}$, where each $\mathbf{c}_k = [c_1, c_2, \ldots, c_l]$ is a feature vector with $l$ coefficients (Figure 1(b)), and $y_i$ is the species name (or label). The representation of $\mathbf{x}_k$ through $\mathbf{c}_k$ is more robust, compact, and simpler for recognizing, compared to raw data.

Finally, the challenge is how to assign the species name to a new syllable by using the MFCC values. This is a supervised classification task and is performed by the last step of the system. For this purpose several ML algorithms could be applied to create and train a model $f(\cdot)$ with capabilities to predict new incoming samples, i.e., given an unknown $\mathbf{c}$ estimates the most probable label by evaluating $f(\mathbf{c}) \rightarrow y_i$, where $S = \{s_1, s_2, \ldots, s_i\}$ is the set of species names. To test how well the model performs and, estimate the expected error, a common choice is the use of stratified $k$-CV. However, there are few related problems to the classical $k$-CV in this type of application (see Section 4). This is the main concern of this work and, therefore, we propose a different Cross-Validation procedure, especially adapted for this task. We present our proposal in the Section 5.

## 3 Related Work

Amphibians are directly affected by environmental changes [4, 5]. This observation has motivated many researchers to develop Automatic Calls Recognition (ACR) systems to monitor anuran populations. Thus, the general idea consists of treating the problem of species recognition as an audio classification task.

Table 1: Summary of few related works. The # stands for the number of different frog species, **ML** for Machine Learning Algorithm, **Acc** for the accuracy, and GMM for Gaussian Mixture Models.

| **Author** | **#** | **ML** | **Acc** |
|---|---|---|---|
| Colonna *et. al.* [12] | 9 | kNN, SVM | 97% |
| Huang *et. al.* [6] | 5 | kNN, SVM | 100% |
| Jaafar *et. al.* [15] | 28 | kNN, SVM | 98% |
| Xie *et. al.* [17] | 4 | GMM | 90%* |
| Dayou *et. al.* [14] | 9 | kNN | 90% |
| Han *et. al.* [8] | 9 | kNN | 100% |
| Vaca-Castaño *et. al.* [16] | 20 | kNN | 91% |
| Yuan *et. al.* [18] | 8 | kNN | 98% |

* identify the F-score measure.

In this context, there are three possible approaches: (1) classify the entire audio recorded without segmentation; (2) use a fixed size segmentation by frames; or (3) classify by syllables [6, 7, 12, 13]. However, the last approach is widely adopted among related works, because signal segments between syllables do not carry useful information about the acoustic frequencies of the species. Therefore, the syllable-based approach achieves better results.

Several comparative studies about ACR can be found in the literature. Table 1 summarizes the related works. Note that commonly these methods achieves high accuracy rates, even with very different features and ML methods. In this paper, we investigate why this happens. One insight about this is the way in which $k$-CV is applied to this task. However, in the majority of the related works the description of the CV procedure adopted is not always explicit, this fact makes the reproduction of the results, and the critical analysis, difficult.

Briggs *et. al.* [19] had used 5-CV to evaluate their method although they express concern about this, pointing that it may be a problem: *"(...) We expect that prediction accuracy would decrease in an experiment where the classifier is applied to individuals that do not appear in the training set (...)"*. Other example can be found in Dong *et. al.* [20]: *"(...) The selection of recordings was made so as to ensure that no two queries within one call class came from the same site on the same day. This was to minimize the probability that calls of the same individual appeared in more than one recording (...)"*. Therefore, the community is concerned about the problem caused when syllables of the same specimen are present in the testing and training sets at the same time. However, there is no consensus on how to evaluate the gains over the recognition rate and which is better suited to this context.

## 4 Problem Description

This section describes the major problem related with the performance validation of the bioacoustic classification approaches used to recognize anuran calls. We called this "the generalization problem", exemplifying the problem as follow. The Figure 1(b) represents an audio signal (or a call) with three syllables of one specimen from the species Adenomera hylaedactyla. Visually these syllables appear slightly different, but in the frequency domain their differences are not very noticeable. Table 2 shows 10-MFCC values extracted from these three syllables.

The last row summarize the mean and the standard deviation (Std) of each column. The low Std indicates that these syllables are very similar. For instance, assuming that we choose a kNN classifier with the Euclidean distance separating the first syllable for testing and the two remaining syllables for training. After running the classifier the dissimilarity score is 0.0545 between the first and second syllables, and 0.0546 between the first and third syllables. This situation is likely to result in a high recognition rate. Technically, this situation may not be considered as overfitting, but as bias. This illustrative example helps understand why some related works achieve almost 100% of accuracy.

Table 2: MFCCs example.

| | **MFCCs** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **syllable$_1$** | 0.00 | 0.14 | 0.42 | 0.69 | 0.81 | 0.90 | 0.93 | 0.97 | 1.00 | 0.91 |
| **syllable$_2$** | 0.00 | 0.13 | 0.42 | 0.72 | 0.84 | 0.93 | 0.94 | 0.96 | 1.00 | 0.92 |
| **syllable$_3$** | 0.00 | 0.14 | 0.44 | 0.72 | 0.83 | 0.91 | 0.92 | 0.95 | 1.00 | 0.92 |
| **Mean** | 0.00 | 0.14 | 0.42 | 0.71 | 0.83 | 0.91 | 0.93 | 0.96 | 1.00 | 0.92 |
| **±Std** | ±0.000 | ±0.005 | ±0.010 | ±0.018 | ±0.017 | ±0.014 | ±0.012 | ±0.008 | ±0.000 | ±0.003 |

Learning the parameters of the classification function and testing it on syllables arising from the same specimen is a methodological misconception. The model would have a perfect score repeating the labels of the samples, but could fail to predict syllables from new specimens. To avoid it, and to increase the generalization capabilities of the system, a common practice when performing a ML experiment is to adopt Cross-Validation. Thus, part of the available examples are separated for testing and other part for training, but in this context, we must avoid generating a random split containing syllables of the same specimen into two subsets. So the classical $k$-CV procedure is not suitable to this application context. Now we can define our research question as: Given a classification's model $f(\cdot)$, trained on a subset of $j$ specimens from the $i$th species, is possible to recognize a new specimen of the same species? Thereby, we want to know how well the trained model generalizes the concept learned for unknown specimens.

We also describe a secondary problem related with the accuracy measure when the dataset is unbalanced. Diverse species of anuran have different syllable's rate (amount of syllables per unit time) in their calls. This is a particular vocalization characteristic of each anuran species and an unequal number of samples could be retrieved from each one [7]. Thus, a classification model that always estimate the most numerous species would have a high accuracy even losing all syllables from the less numerous classes. To overcome this matter we suggest to use the macro-accuracy instead of the traditional micro-accuracy. It means, the final accuracy value is calculated as the average accuracy of each species individually [7, 21].

## 5 Proposed Methodology

Cross-Validation (CV) is used to estimate the expected error in a real situation. With $k$-CV the original dataset is split into $k$ disjoint folds, and for each one the conditional error ($e_k$) is estimated training the model $f(\cdot)$ with $k$-1 folds. Thus, this procedure is repeated $k$ times and the expected generalized error can be obtained as the mean of $e_k$. As mentioned earlier, we might hope that $k$-CV estimates the real error, but when the information of the specimen is omitted we fall in a situation in which the split could leave syllables from one specimen in the testing and training sets.

To address this problem, we propose to consider the specimen information during the $k$-CV splitting, leaving all the syllables that belongs to the same specimen (or individual) together, avoiding mixing them in the testing and training sets. Then, we propose a Leave-one-Out CV (LOOCV) by individuals (or records) for measuring the performance of the classification algorithms, i.e., being $k$ equal to number of different specimens. Therefore, the individuals are separated into two groups, one for testing and the others for training. These steps are repeated until every individual (or record) has been used as test set. In each $k$ step, the predictions for every syllable are saved. After LOOCV is completed, Micro- and Macro-accuracy are calculated using the confusion matrix.

Because we are dealing with a supervised problem, and we want to consider this new information during the LOOCV evaluation, now each syllable must be associated with two labels: one for the specimen ($s_j$) and one for the species ($y_i$). Therefore, an example of dataset could be:

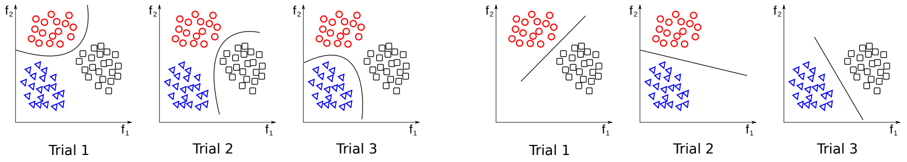
$$\begin{aligned}
\mathbf{c}_1 &= [c_1, c_2, \ldots, c_l],\ s_1,\ y_1 \\
\mathbf{c}_2 &= [c_1, c_2, \ldots, c_l],\ s_1,\ y_1 \\
\mathbf{c}_3 &= [c_1, c_2, \ldots, c_l],\ s_2,\ y_1 \\
& \qquad\quad \vdots \qquad\qquad\;\; \vdots \;\;\; \vdots \\
\mathbf{c}_k &= [c_1, c_2, \ldots, c_l],\ s_j,\ y_i
\end{aligned}$$

in which $i$ is the species ID and $j$ is the specimen ID. In this example, the first two syllables belong to the same specimen and the same species.

This way of splitting shows two main particularities. First, at least two specimens of each species are needed. Second, the number of examples in each fold could be not balanced. On the other hand, we assume that the generalization error will be more realistic, because we are training with one specimen to predict a different one.

Apply this procedure to solve a multiclass problem could be more complex when increasing the number of specimens, but these can be simplified by creating and combining a pool of binary problems. For this purpose there are two well know strategies: One-against-All (1AA) and One-against-One (1A1 or Round Robin) [22]. These approaches are also useful to adapt a binary classifier to a multiclass task. The Figure 3 exemplifies these concept.

The 1AA procedure begins by separating all the syllables of the first specimen in testing set and grouping the remaining syllables of the same species in training set for the target class ("+1"). The syllables of all remaining species, that not belongs to the target species, are grouped in the negative class ("-1"). Then, the model $f(\cdot)$ is trained and applied to estimate the labels of the testing group. In the second round, this procedure is repeated but separating all the



(a) One-against-All decomposition.

(b) One-against-One decomposition.

Fig. 3: Decomposition and simplification of the problem. New classification functions over three rounds of 1AA and 1A1. After that, the majority vote is applied.

syllables of the second specimen for testing. The validation is repeated until all specimens in the dataset are evaluated. It is analogous to perform a Leave-one-Out CV using specimens instead off syllables. With the 1AA and the classical $k$-CV setting the number of rounds are equal to $r = i \times k$, but after incorporating the specimen information the amount of rounds become $r = i \times j$. However, with this decomposition the complexity of trained model $f(\cdot)$ is smaller than in the multiclass configuration (Figure 3(a)).

The second procedure we propose, called 1A1, breaks the original problem in smaller problems than 1AA. The label estimation proceeds similar to the Leave-one-Out by specimen, but with the main difference that the negative class is breaking down into several small groups, considering one group for each species (Figure 3(b)). After that, the result of each sub-problem is combined by using the majority voting rule. Typically, with 1A1 and $k$-CV the number of rounds increases with the rule $\frac{i\cdot(i-1)}{2} \times k$, but, in our case, this rule becomes $\frac{i\cdot(i-1)}{2} \times j$. This decomposition reduces the complexity of each sub-problems, compared to the multiclass approach.

## 6 Experiment Setting and Results

The dataset used in our experiments is summarized in Table 3. It has 10 different species, 55 specimens and 5799 syllables. These samples were collected *in situ* under real noise conditions. Some species are from the Federal University of Amazonas, Brazil$^{*}$, other from Mata Atlântica, Brazil$^{**}$, and the last from Córdoba, Argentina$^{+}$. These recordings were stored in *wav* format with 44.1 kHz of sampling frequency and 32 bit, which allows us to analyze signals up to 22.05 kHz. From each extracted syllable, 24 MFCCs were calculated by using 44 triangular filters. For the segmentation task we based our approach on the work of Colonna *et. al.* [7], but using only the energy of the signal[3]. Finally, the frame size was 0.0464 s with 66% of overlap to obtain a good energy-time resolution.

We compared the results showed in the Tables 4 and 5 by using four classifiers: kNN; Quadratic Discriminant Analysis (QDA); Decision Tree; and Support Vector Machine (SVM) using RBF and polynomial kernels with degrees $p = \{1, 2, 3\}$.

[3] The segmentation code is available at http://goo.gl/vjVQ2c.

Table 3: Species Dataset. The $\boldsymbol{s}$ and the $\boldsymbol{k}$ stands for the amount of specimens and syllables respectively.

| **Species** | $\boldsymbol{s}$ | $\boldsymbol{k}$ | **Species** | $\boldsymbol{s}$ | $\boldsymbol{k}$ |
|---|---|---|---|---|---|
| Adenomera hylaedactyla$^{**}$ | 11 | 3039 | Adenomera andreae$^{*}$ | 8 | 471 |
| Leptodactylus fuscus$^{*}$ | 4 | 222 | Ameerega trivittata$^{**}$ | 5 | 493 |
| Hyla minuta$^{**}$ | 11 | 227 | Hypsiboas cinerascens$^{*}$ | 2 | 361 |
| Hypsiboas cordobae$^{+}$ | 4 | 703 | Osteocephalus oophagus$^{*}$ | 3 | 96 |
| Scinax ruber$^{**}$ | 4 | 77 | Rhinella granulosa$^{*}$ | 3 | 110 |

Table 4: Comparison result using 1AA decomposition.

| | **$k$NN** | **$k$NN** | **$k$NN** | **Tree** | **QDA** | **SVM** | **SVM** | **SVM** | **SVM** |
|---|---|---|---|---|---|---|---|---|---|
| **Species** | $k = 1$ | $k = 3$ | $k = 5$ | | | RBF | $p = 1$ | $p = 2$ | $p = 3$ |
| Adenomera andreae | 33.46 | 32.66 | 34.67 | 30.64 | 86.69 | 72.58 | 59.27 | 74.79 | 72.98 |
| Ameerega trivittata | 89.88 | 89.33 | 88.23 | 42.83 | 88.60 | 67.46 | 57.90 | 64.52 | 70.77 |
| Adenomera hylaedactyla | 98.68 | 99.37 | 99.50 | 94.29 | 98.29 | 99.77 | 99.77 | 99.73 | 99.60 |
| Hyla minuta | 61.57 | 53.71 | 53.27 | 34.49 | 52.40 | 55.02 | 25.32 | 55.02 | 65.06 |
| Hypsiboas cinerascens | 96.39 | 98.06 | 96.95 | 71.74 | 90.02 | 93.35 | 90.02 | 96.67 | 97.50 |
| Hypsiboas cordobae | 100.00 | 100.00 | 100.00 | 98.29 | 95.86 | 98.29 | 96.43 | 97.86 | 98.57 |
| Leptodactylus fuscus | 63.96 | 59.90 | 49.09 | 9.00 | 0.45 | 9.45 | 0.45 | 70.27 | 57.20 |
| Osteocephalus oophagus | 42.70 | 34.37 | 32.29 | 17.70 | 11.45 | 0.00 | 0.00 | 0.00 | 9.37 |
| Rhinella granulosa | 39.84 | 32.81 | 30.46 | 9.37 | 0.78 | 28.12 | 12.50 | 37.50 | 45.31 |
| Scinax ruber | 0.00 | 0.00 | 0.00 | 3.94 | 0.00 | 0.00 | 0.00 | 3.94 | 1.31 |
| **Micro-accuracy** | 86.21 | 85.80 | 85.36 | 73.52 | 85.38 | 84.34 | 80.09 | 86.93 | **87.61** |
| **Macro-accuracy** | **62.65** | **60.02** | **58.45** | 41.23 | **52.45** | **52.40** | **44.16** | **60.03** | **61.77** |
| **Precision** | 0.62 | 0.60 | 0.59 | 0.53 | 0.65 | 0.66 | 0.63 | 0.72 | 0.70 |
| **Recall** | 0.63 | 0.60 | 0.58 | 0.41 | 0.52 | 0.52 | 0.44 | 0.60 | 0.62 |

For each configuration we calculated the micro- and macro-accuracy. The baselines results for comparison are: 52.40% in the case of micro-accuracy and 10% in the case of macro-accuracy, i.e. the baseline value for a classifier, which always chooses the most numerous species is the micro and for a classifier, which randomly chooses one species is the macro. In the Macro-accuracy rows we applied the $t$-Test to compare the means obtained against the best value in the row. Therefore, the boldface values could be considered a tie with confidence level $p = 0.05$. In the last row of each table, we have the standard deviation values of each column.

Among these results, we note that Scinax ruber was the most difficult species. However, Adenomera hylaedactyla and Hypsiboas cordobae appear to be easier to classify. The configuration using polynomial SVM with $p = 3$ and 1A1 is the better option. In general, comparing the standard deviation of the methods we can conclude that 1A1 decrease the variance showing a more uniform accuracy among all the species tested. In the last table of results (6) we compare the macro-accuracy gains of 1A1 against 1AA. This values are presented in percentage. The gains obtained by the Tree classifier and by the linear SVM show that these methods take advantages from the 1A1 decomposition.

Table 5: Comparison result using 1A1 decomposition.

| | *k*NN | *k*NN | *k*NN | **Tree** | **QDA** | **SVM** | **SVM** | **SVM** | **SVM** |
|---|---|---|---|---|---|---|---|---|---|
| **Species** | $k = 1$ | $k = 3$ | $k = 5$ | | | RBF | $p = 1$ | $p = 2$ | $p = 3$ |
| Adenomera andreae | 33.46 | 32.05 | 31.45 | 26.00 | 26.61 | 31.85 | 28.42 | 28.62 | 30.24 |
| Ameerega trivittata | 89.88 | 89.70 | 88.97 | 70.40 | 99.26 | 92.83 | 91.36 | 78.86 | 63.78 |
| Adenomera hylaedactyla | 98.68 | 99.37 | 99.50 | 98.19 | 98.49 | 99.86 | 99.96 | 99.77 | 99.34 |
| Hyla minuta | 61.57 | 53.71 | 53.27 | 58.07 | 84.27 | 61.57 | 62.44 | 66.81 | 68.99 |
| Hypsiboas cinerascens | 96.39 | 98.06 | 97.22 | 88.36 | 88.64 | 97.22 | 96.95 | 96.12 | 94.18 |
| Hypsiboas cordobae | 100.00 | 100.00 | 100.00 | 95.58 | 95.72 | 99.85 | 99.00 | 99.71 | 100.00 |
| Leptodactylus fuscus | 63.96 | 59.90 | 50.90 | 45.49 | 1.351 | 59.00 | 36.93 | 67.56 | 62.16 |
| Osteocephalus oophagus | 42.70 | 36.45 | 34.37 | 20.83 | 15.62 | 6.25 | 1.04 | 14.58 | 36.45 |
| Rhinella granulosa | 39.84 | 33.59 | 33.59 | 17.96 | 1.56 | 31.25 | 28.12 | 32.81 | 46.87 |
| Scinax ruber | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 9.21 | 18.42 | 23.68 | 32.89 |
| **Micro-accuracy** | **86.21** | 85.83 | 85.34 | 80.85 | 82.66 | 86.14 | 84.82 | 85.32 | 84.43 |
| **Macro-accuracy** | **62.65** | **60.28** | **58.93** | 52.09 | **51.15** | **58.89** | **56.26** | **60.85** | **63.49** |
| **Precision** | 0.62 | 0.61 | 0.60 | 0.57 | 0.53 | 0.67 | 0.63 | 0.68 | 0.70 |
| **Recall** | 0.63 | 0.60 | 0.59 | 0.52 | 0.51 | 0.59 | 0.56 | 0.60 | 0.63 |

Table 6: Gains of 1A1 over 1AA.

| | *k*NN | *k*NN | *k*NN | **Tree** | **QDA** | **SVM** | **SVM** | **SVM** | **SVM** |
|---|---|---|---|---|---|---|---|---|---|
| | $k = 1$ | $k = 3$ | $k = 5$ | | | RBF | $p = 1$ | $p = 2$ | $p = 3$ |
| **Gains** | 0.00 | **+0.26** | **+0.48** | **+10.86** | -1.30 | **+6.49** | **+12.10** | **+0.82** | **+1.72** |

For comparison purpose we have tested two additional configuration applying the traditional *k*-CV with ten folds and LOOCV by syllables, i.e., without taking care about individuals information, with *k*NN ($k = 3$ and $k = 1$). In the first case, the Micro- and Macro-accuracy were 99.45% and 99.14%, and in the second case, were 99.66% and 99.53% respectively. These results are equivalent to the approaches described by several authors [6, 8, 12, 14–16, 18], but using our own dataset. Comparing these against the results obtained using our *k*-CV by individuals, showed in last lines of the tables 4 and 5, we realize that when the specimen information is not considered, the accuracy is overestimated due the problem described in Section 4.

## 7 Discussion and Conclusion

In this work we introduced a different *k*-CV procedure to evaluate a bioacoustic recognition framework. The main contribution is the incorporation of the specimens information (or individuals) as an additional label and consider it when performs the *k*-CV. This extra label helps to split the dataset without mixing up syllables from the same specimen into the testing and training groups avoiding an overestimate of the accuracy. Thus, the results are more representative of a real situation, in which different specimens would be found in the rainforest. In addition, we showed a problem simplification using 1A1 and 1AA approaches.

Comparing the related works against our results we notice a considerably difference from similar configurations, showing that not separate the testing by specimens causes a high bias of the accuracy, and consequently, the model has less generalization capabilities. Moreover, the difference between the macro- and micro-accuracy exposes the problem of working with unbalanced datasets as commonly happens in these type works. Inspecting several confusion matrix of our experiments we also note that the information about others individuals was not enough to recognize new ones in some cases, as the Scinax ruber. This may be caused by: (1) the features were insufficient to extract the shared information between specimens of the same species; or (2) the discriminatory power of the MFCCs was very detailed capturing fine-grained differences of the frequencies. Anyway, others LLDs should be investigated and evaluated with our methodology. Finally, we recommend to the authors of future works give more details about the adopted evaluation procedures and the generalization capabilities of the proposed approaches.

## Acknowledges

This work was partly supported by the European Commission through MAESTRA (ICT-2013-612944) and the Project TEC4Growth - Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE-01-0145-FEDER-000020 is financed by the North Portugal Regional Operational Program (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

Juan G. Colonna gratefully acknowledge to National Council of Technological and Scientific Development (CNPq, Brazil) by the PhD fellowship. Eduardo F. Nakamura acknowledge to FAPEAM by the support granted through the Anura Project (FAPEAM/CNPq PRONEX 023/2009). We also thank to professors Marcelo Gordo and the biologist Celeste Salineros for the help with the recordings.

## References

1. Bertrand, A.: Applications and trends in wireless acoustic sensor networks: A signal processing perspective. In: Communications and Vehicular Technology in the Benelux (SCVT), 2011 18th IEEE Symposium on. (Nov 2011) 1–6
2. Ribas, A.D., Colonna, J.G., Figueiredo, C.M.S., Nakamura, E.F.: Similarity clustering for data fusion in wireless sensor networks using k-means. In: International Joint Conference on Neural Networks (IJCNN), IEEE (June 2012) 1–7
3. Colonna, J.G., Cristo, M.A.P., Nakamura, E.F.: A distribute approach for classifying anuran species based on their calls. In: 22nd International Conference on Pattern Recognition. (2014)
4. Cole, E.M., Bustamante, M.R., Reinoso, D.A., Funk, W.C.: Spatial and temporal variation in population dynamics of andean frogs: Effects of forest disturbance and evidence for declines. Global Ecology and Conservation **1**(0) (2014) 60–70
5. Carey, C., Alexander, M.A.: Climate change and amphibian declines: is there a link? Diversity and Distributions **9**(2) (2003) 111–121

6. Huang, C.J., Yang, Y.J., Yang, D.X., Chen, Y.J.: Frog classification using machine learning techniques. Expert Systems with Applications **36**(2) (2009) 3737–3743
7. Colonna, J.G., Cristo, M.A.P., Salvatierra, M., Nakamura, E.F.: An incremental technique for real-time bioacoustic signal segmentation. Expert Systems with Applications **42**(21) (2015) 7367 – 7374
8. Han, N.C., Muniandy, S.V., Dayou, J.: Acoustic classification of australian anurans based on hybrid spectral-entropy approach. Applied Acoustics **72**(9) (2011) 639–645
9. Jaafar, H., Ramli, D., Shahrudin, S.: Mfcc based frog identification system in noisy environment. In: International Conference on Signal and Image Processing Applications (ICSIPA), IEEE. (Oct 2013) 123–127
10. Xie, J., Zhang, J., Roe, P.: Acoustic features for hierarchical classification of australian frog calls. In: In 10th International Conference on Information, Communications and Signal Processing. (2015)
11. Yen, G., Fu, Q.: Automatic frog call monitoring system: a machine learning approach. In: Proceedings of SPIE. Volume 4739., SPIE (2002) 188–199
12. Colonna, J.G., Ribas, A.D., Santos, E.M.d., N., E.F.: Feature subset selection for automatically classifying anuran calls using sensor networks. In: International Joint Conference on Neural Networks (IJCNN), IEEE (June 2012) 1–8
13. Xie, J., Towsey, M., Truskinger, A., Eichinski, P., Zhang, J., Roe, P.: Acoustic classification of australian anurans using syllable features. In: IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP 2015), IEEE (2015)
14. Dayou, J., Han, N.C., Mun, H.C., Ahmad, A.H., Muniandy, S.V., Dalimin, M.N.: Classification and identification of frog sound based on entropy approach. In: International Conference on Life Science and Technology. Volume 3. (2011) 184–187
15. Jaafar, H., Ramli, D.A., Rosdi, B.A.: Comparative study on different classifiers for frog identification system based on bioacoustic signal analysis. In: Proceedings of the 2014 International Conference on Communications, Signal Processing and Computers. (2014)
16. Vaca-Castaño, G., Rodriguez, D.: Using syllabic mel cepstrum features and k-nearest neighbors to identify anurans and birds species. In: Signal Processing Systems (SIPS), 2010 IEEE Workshop on. (2010) 466–471
17. Xie, J., Towsey, M., Yasumiba, K., Zhang, J., Roe, P.: Detection of anuran calling activity in long field recordings for bio-acoustic monitoring. In: IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP 2015), IEEE (2015)
18. Yuan, C.L.T., Ramli, D.A.: Frog sound identification system for frog species recognition. In: Context-Aware Systems and Applications. Springer (2013) 41–50
19. Briggs, F., Lakshminarayanan, B., Neal, L., Fern, X.Z., Raich, R., Hadley, S.J.K., Hadley, A.S., Betts, M.G.: Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. The Journal of the Acoustical Society of America **131**(6) (2012) 4640–4650
20. Dong, X., Towsey, M., Truskinger, A., Cottman-Fields, M., Zhang, J., Roe, P.: Similarity-based birdcall retrieval from environmental audio. Ecological Informatics **29, Part 1** (2015) 66–76
21. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management **45**(4) (2009) 427–437
22. Fürnkranz, J.: Round robin rule learning. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01 (2001) 146–153