An Architecture for Animal Sound Identification based on Multiple Feature Extraction and Classification Algorithms

Leandro Tacioli[1], Luís Felipe Toledo[2], Claudia Bauzer Medeiros[1]

[1]Institute of Computing, Universidade Estadual de Campinas (UNICAMP)
Campinas - São Paulo - Brazil

[2]Institute of Biology, Universidade Estadual de Campinas (UNICAMP)
Campinas - São Paulo - Brazil

leandrotacioli@gmail.com, toledolf2@yahoo.com, cmbm@ic.unicamp.br

Abstract. *Automatic identification of animals is extremely useful for scientists, providing ways to monitor species and changes in ecological communities. The choice of effective audio features and classification techniques is a challenge on any audio recognition system, especially in bioacoustics that commonly uses several algorithms. This paper presents a novel software architecture that supports multiple feature extraction and classification algorithms to help on the identification of animal species from their recorded sounds. This architecture was implemented by the WASIS software, freely available on the Web.*

1. Introduction

Audio recognition systems have been developed to several domains, such as automatic speech recognition [Yu and Deng 2015], music information retrieval [Grosche et al. 2012], and bioacoustics [Aide et al. 2013] - subject of this work. The study of bioacoustics is related to every sound produced by or affecting all kinds of living organisms, although it is a science oriented to animal communication [Schöner et al. 2016]. The vast majority of researchers in this field are specialized in few or only one animal group, hence most of the recognition tools in bioacoustics are designed to meet the needs of the experts in question [Aide et al. 2013]. Algorithms have been created or applied to automate identification of target animal groups, for instance, amphibians [Noda et al. 2016] and birds [Stowell and Plumbley 2014].

Animal identification through their sounds allows, for example, the estimation of population trends of key species in sensitive areas [Bardeli et al. 2010] or provides changes in ecological communities over time [Frommolt and Tauchert 2014]. One advantage of bioacoustics lies in the detection of animal sounds in the absence of an observer [Bardeli et al. 2010]. Moreover, it is a popular non-invasive method to study animal populations, biodiversity, and taxonomy [Frommolt and Tauchert 2014, Köhler et al. 2017].

Primary challenges during the development of sound retrieval systems are the identification of effective audio features and classification techniques. Feature extraction focuses on extracting meaningful information from audio signals, while classification use these extracted data to match against the respective data of samples from a repository. A major concern in audio recognition systems is how feature extraction is coupled to the classification algorithms, preventing the reuse of code in other contexts and limiting

the ability of researchers to exchange features [McEnnis et al. 2005]. Furthermore, researchers demand architectures that allows them to implement new algorithms without major concerns with supporting infrastructure for data manipulation and scheme evaluation [Hall et al. 2009].

Given this motivating and challenging scenario, the main contribution of this paper is a novel architecture that supports multiple feature extraction and classification algorithms to identify animals based on their sounds. The architecture is extensible and can accommodate a number of new algorithms. It has been implemented in the WASIS[1] software, also described in this paper.

2. Related Work

2.1. Audio Features

Audio features represent the way in which meaningful information is analyzed and extracted from audio signals to obtain highly reduced and expressive data that are suitable for computer processing [Schuller 2013]. Note that the amount of data in raw audio files would be too big for their direct processing; moreover, considerable information (e.g., frequency variation and timbre) would not be perceptible in their signal waveforms, often inappropriate for audio retrieval [Mitrovic et al. 2010].

The feature extraction process generates output vectors that are normally called feature descriptors. These feature descriptors are the fundamental information that classifiers use. A failure to capture these relevant information of audio signals will result in poor performance, no matter how good the classifier is [McEnnis et al. 2005].

The performance of audio features may be affected by a series of factors in animal identification systems, such as the presence of background noise and the duration of animal calls [Xie et al. 2016]. Feature fusion is a technique that is able to combine two or more audio features and attenuate their disadvantages, as reported by [Noda et al. 2016].

2.2. Audio Classification

Audio classification is the process by which an individual audio sample is assigned to a class, based on its characteristics [Liu and Wan 2001]. These characteristics are the *feature descriptors* of the audio sample that will be used on the identification. In animal sound recognition, each species represents one class, usually labelled by its taxonomic information (e.g., family, genus, and specific epithet).

Two classification approaches are found in the literature:

- *Brute Force* - The classification is performed by linearly traversing the entire set of feature descriptors, providing similarity results among several audio segments [Mitrovic et al. 2010]. One statistical algorithm used for this approach is Pearson Correlation Coefficient (PCC);
- *Class Model* - Considered by the literature the main approach for audio classification [Sharan and Moir 2016]. Commonly, it employs supervised machine learning algorithms for animal identification. Popular algorithms using this method are Support Vector Machine (SVM) and Hidden Markov Model (HMM).

[1] WASIS: Wildlife Animal Sound Identification System (Version 1.5.0) http://www.naturalhistory.com.br/wasis.html

2.3. Typical Architectures for Audio Retrieval

The general approach to automatic sound recognition (ASR) is commonly inspired from techniques employed in speech recognition systems, and most of these ASR systems have a model based on three key steps, according to [Sharan and Moir 2016]: (a) *signal pre-processing*, responsible for preparing the audio signal for (b) *feature extraction*, and (c) *classification*. However, this model of a typical architecture considers only machine learning-based algorithms, ignoring other techniques, such as the *Brute Force* approach.

[Mitrovic et al. 2010] described a more detailed architecture based on three components: (a) *Input Module* that performs feature extraction from audio stored in an audio database, and persists the descriptors into a feature database; (b) *Query Module* in which the user provides audio objects of interest for identification and feature extraction is also performed in these objects; and (c) *Retrieval Module* that estimates the similarity among the user's and the feature database's audio objects, returning the most similar objects.

3. Proposed Architecture

This work is focused on a novel architecture to support the identification of animal species based on their sounds. This architecture combines multiple algorithms for audio feature extraction and audio classification to a suite of data repositories. The WASIS software is the first implementation of the proposed architecture - described in Section 4.

3.1. Overview

Figure 1 presents an overview of our architecture. The inputs are *Audio Files*, in which users select *Audio Segments* - also known as regions of interest (ROIs). These ROIs are forwarded to the *Feature Extraction* module (1). Several feature extraction techniques can be performed for each audio segment, as well as the *Fusion* among these feature representations (2). The results of this extraction process (3a; 3b) are the *Feature Descriptors*.

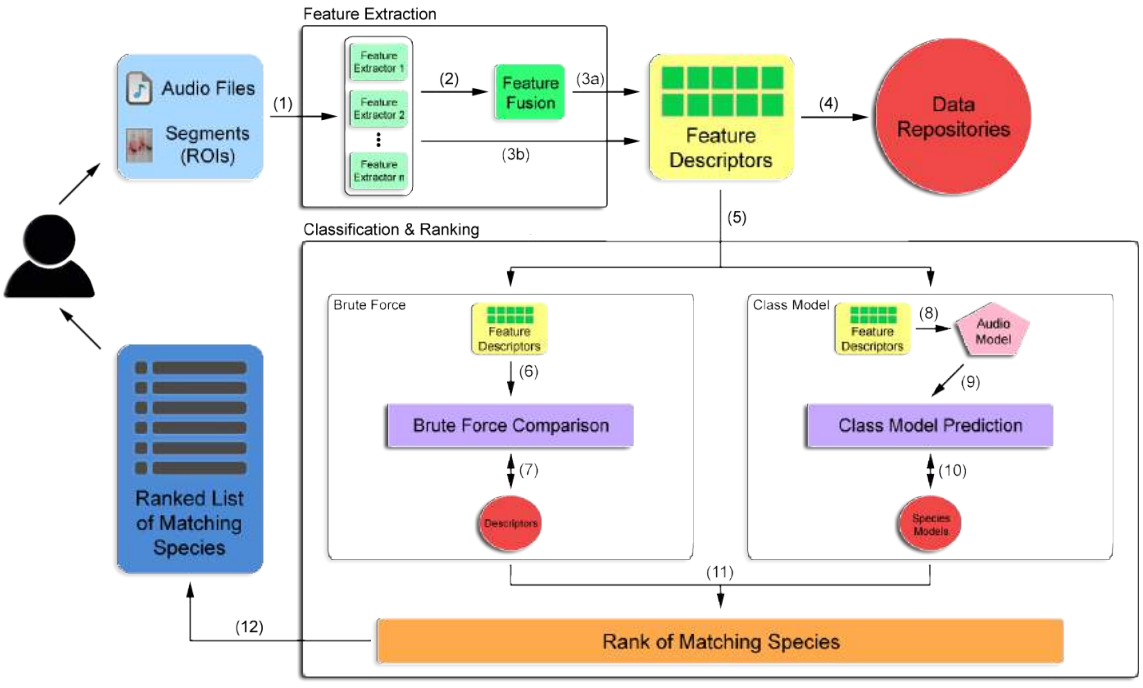

Figure 1. Detailed software architecture.

The *Data Repositories* component represents all the different repositories created/accessed in the architecture. In particular, *Descriptors* and *Species Models* (bottom circles of the figure) belong within the general *Data Repositories* - detailed in Section 3.2.

The *Feature Descriptors* can be either stored into the appropriate data repository with the associated metadata of their audio files (4) or sent directly to the *Classification & Ranking* module (5). The first choice (4) is more suitable for users who want to create their own database for future identification. The second choice (5) is more appropriate for those who just want to identify the animal species from the sound samples.

The *Classification & Ranking* module classifies the input ROIs. It receives *Feature Descriptors* as inputs (5). For the *Brute Force* approach, the *Brute Force Comparison* module calculates the similarities among the *Feature Descriptors* (6) and the descriptors of audio segments previously stored in their appropriate repository (7). In the *Class Model* approach, an *Audio Model* is created from the *Feature Descriptors* based on a machine learning algorithm (8). Then, the *Class Model Prediction* module estimates the similarity degrees among the *Audio Model* (9) and the *Species Models* stored in their repository (10).

Note that both *Brute Force* and *Class Model* approaches are processed totally apart. There is no combination of their results, though both kinds of results are independently ranked by the *Rank of Matching Species* (11). The final output shows a ranked list of matching species (12).

3.2. Data Repositories

Figure 2 details our data repositories and highlights which components of the architecture are responsible for processing, retrieving and persisting information to these data repositories. These are the repositories mentioned in the architecture overview (Figure 1).

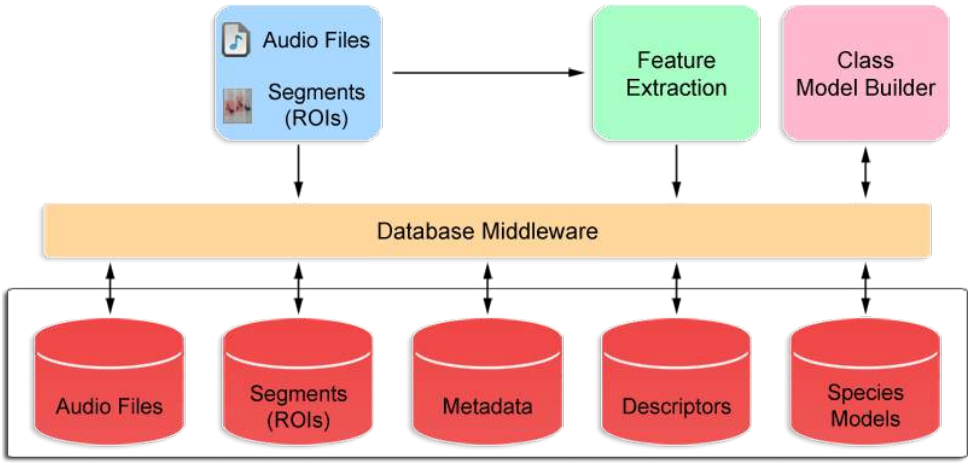

Figure 2. Structure of the data repositories.

Each data repository stores different information from particular modules:

- *Audio Files* - Raw audio files for processing;
- *Segments (ROIs)* - Regions of interest where the audio signals will be used to identification;

- *Metadata* - Information used to identify, describe and organize the audio files. In animal sound recognition, the most important information is scientific classification, followed by recording location, date and time;
- *Descriptors* - The outputs of the *Feature Extraction* module;
- *Species Models* - Particularly used in machine learning-based classifiers, models of animal species are trained from their respective feature descriptors to predict whether an audio segment belongs to a specific species.

The *Database Middleware* provides a bridge between the modules of the architecture and the data repositories. This access granted by the *Database Middleware* allows the modules of the architecture to retrieve or persist information into the data repositories for any desired module. Moreover, if new feature extraction techniques are implemented, the *Feature Extraction* module is able to process the audio files and their ROIs already stored in the data repository and generate its own *Descriptors*. The same goes for newly implemented classifiers that can invoke the *Class Model Builder* module to generate their own *Species Models*.

3.3. Class Model Builder

The architecture also provides the *Class Model Builder* (Figure 3), which requests metadata and feature descriptors of the audio files stored in the data repositories, to create models that are able to identify animal species.

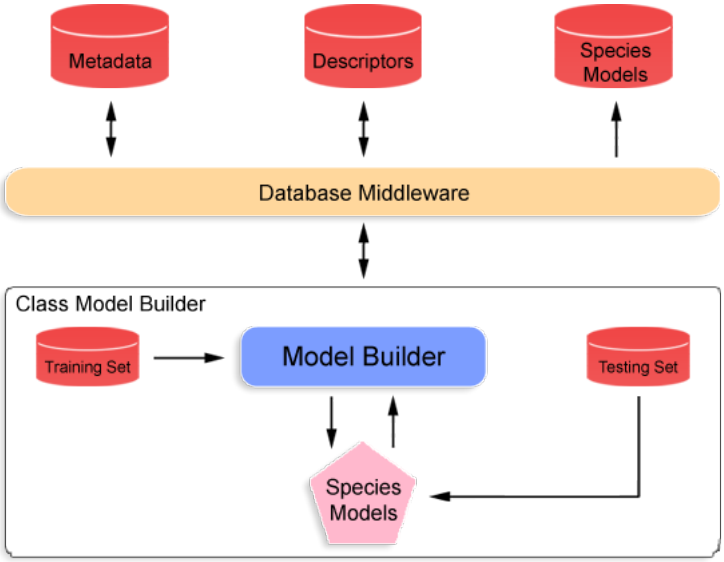

Figure 3. Design of the *Class Model Builder*.

The *Class Model Builder* sets up two datasets with the metadata and features descriptors. The *Training Set* is responsible for providing feature descriptors to the machine learning algorithm that will create the *Species Models*. Using different data from those used by the *Training Set*, the *Testing Set* is set up for the purpose of estimating how well the models were trained and optimize the parameters of the models. Lastly, the final task of the *Class Model Builder* is persisting the trained and optimized *Species Models* to the appropriate data repository.

4. Implementation Aspects

The first prototype developed, WASIS, is based on *Power Spectrum* feature representation and *Pearson Correlation Coefficient*, restricting the present prototype to the *Brute Force* classification approach. *Power Spectrum* describes the distribution of audio signal's maximum power over given frequency bins. *Pearson Correlation Coefficient* is a measure of the strength of the association between two variables.

The prototype was implemented in Java platform, using MySQL and H2 database technologies. Currently, the sound database contains sound samples from amphibians, birds, and primates. Such samples were selected from Fonoteca Neotropical Jacques Vielliard (FNJV)[2], considered one of the ten largest animal sound libraries in the world.

At present, additional features extraction algorithms are being implemented, such as Mel Frequency Cepstral Coefficents (MFCC), Linear Predictive Coding (LPC), Linear Prediction Cepstral Coefficients (LPCC) and Perceptual Linear Predictive (PLP), as well as the fusion among these feature representations. In addition, machine learning algorithms for the *Class Model* approach are being implemented to perform audio classification, such as Hidden Markov Model (HMM) and Support Vector Machine (SVM).

4.1. Case Study

Let us consider the following case study: a scientist has recorded a given bird species and wants to check its identification using WASIS. Initially, the scientist has to select audio segments (ROIs) that contain the bird vocalizations to be identified. Figure 4 illustrates a screen copy of WASIS interface which shows in red squares, the audio segments selected.

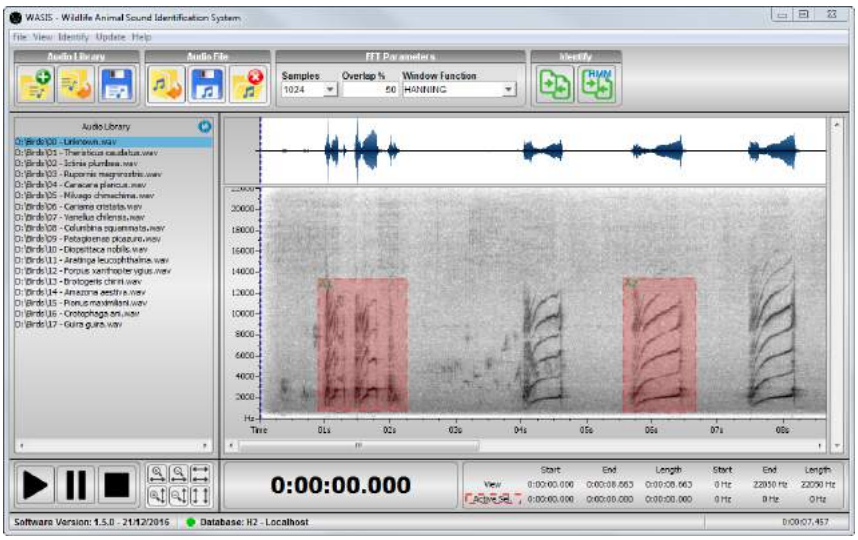

Figure 4. WASIS interface with audio segments to be identified.

Figure 5A shows a screen copy of the results of an audio segment comparison. The prototype performs the comparison according to the architecture flow. Initially, the module extracts feature of the audio segment requested by the scientist, returning the descriptors necessary to the classification. Then, these descriptors are matched against

[2] Fonoteca Neotropical Jacques Vielliard (FNJV), UNICAMP, Brazil - http://www2.ib.unicamp.br/fnjv/

data contained in the *Descriptors* repository using the *Brute Force* approach. A ranked list of matching species is returned. The higher the correlation coefficient between two audio segments, the higher the probability of a species being classified correctly. In this example, the prototype indicates that the audio segment selected by the scientist belongs to a Smooth-billed Ani (*Crotophaga ani*).

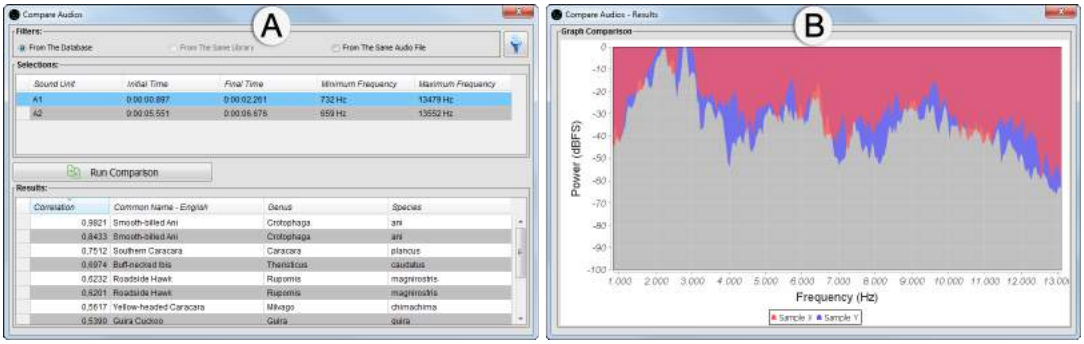

Figure 5. (A) shows the screen for audio comparison with its results, while (B) shows the data of the scientist segment (magenta) against the data of a sample from the *Descriptors* repository (blue).

The prototype also provides detailed information about the audio comparison. Figure 5B illustrates a visual comparison between audio segments, providing more information about the features extracted. The *Power Spectrum* feature extraction employed in the prototype shows the signal's maximum power (vertical axis) over the frequency bins (horizontal axis).

5. Conclusions and Future Work

The ability to identify animal species based on their sounds is extremely useful for scientists. This work presents a software architecture for bioacoustics that supports multiple audio feature extraction, feature fusion, and classification algorithms and is capable of performing the identification of animal based on their sounds. A prototype was implemented with one feature/classifier set for animal sound identification, and a case study explained how a scientist can use the prototype.

We are implementing several feature extraction and classification algorithms for sound recognition. Our purpose is to create a repository for storing these algorithms, avoiding implementation errors of those who want to reuse these techniques.

In the future, we plan to make a comparative study providing recommended sets of features/classifiers for animal identification, exploring sounds of several animal groups. Other enhancements might be related to audio segmentation to support scientists on long-duration recording analysis, and inclusion of techniques other than acoustic, such as semantic features.

6. Acknowledgements

Work partially financed by CNPq (132849/2015-1; 300896/2016-6; 305110/2016-0), FAPESP (2013/02219-0), FAPESP CCES (2013/08293-7), CNPq/INCT in Web Science (557128/2009-9) and FAPESP-PRONEX (eScience project).

References

Aide, T. M., Corrada-Bravo, C., Campos-Cerqueira, M., Milan, C., Vega, G., and Alvarez, R. (2013). Real-time bioacoustics monitoring and automated species identification. *PeerJ*, 1:e103.

Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K.-H., and Frommolt, K.-H. (2010). Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters*, 31:1524–1534.

Frommolt, K.-H. and Tauchert, K.-H. (2014). Applying bioacoustic methods for long-term monitoring of a nocturnal wetland bird. *Ecological Informatics*, 21:4–12.

Grosche, P., Müller, M., and Serrà, J. (2012). Audio content-based music retrieval. In *Multimodal Music Processing*, pages 157–174. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11:10–18.

Köhler, J., Jansen, M., Rodríguez, A., Kok, P. J. R., Toledo, L. F., Emmrich, M., Glaw, F., Haddad, C. F. B., Rödel, M.-O., and Vences, M. (2017). The use of bioacoustics in anuran taxonomy: theory, terminology, methods and recommendations for best practice. *Zootaxa*, 4251(1):1–124.

Liu, M. and Wan, C. (2001). A study on content-based classification and retrieval of audio database. In *International Symposium on Database Engineering and Applications*.

McEnnis, D., Mckay, C., Fujinaga, I., and Depalle, P. (2005). jAudio: A feature extraction library. In *International Conference on Music Information Retrieval*.

Mitrovic, D., Zeppelzauer, M., and Breiteneder, C. (2010). Features for content-based audio retrieval. *Advances in Computers*, 78:71–150.

Noda, J. J., Travieso, C. M., and Sánchez-Rodríguez, D. (2016). Methodology for automatic bioacoustic classification of anurans based on feature fusion. *Expert Systems With Applications*, 50:100–106.

Schöner, M. G., Simon, R., and Schöner, C. R. (2016). Acoustic communication in plant–animal interactions. *Current Opinion in Plant Biology*, 32:88–95.

Schuller, B. (2013). Audio features. In *Intelligent Audio Analysis*, pages 41–97. Springer Berlin Heidelberg.

Sharan, R. V. and Moir, T. J. (2016). An overview of applications and advancements in automatic sound recognition. *Neurocomputing*, 200:22–34.

Stowell, D. and Plumbley, M. D. (2014). Automatic large-scale classifcation of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, 2:e488.

Xie, J., Towsey, M., Zhang, J., and Roe, P. (2016). Acoustic classification of Australian frogs based on enhanced features and machine learning algorithms. *Applied Acoustics*, 113:193–201.

Yu, D. and Deng, L. (2015). *Automatic Speech Recognition: A Deep Learning Approach*. Springer-Verlag London.