Universidade Estadual de Campinas
Instituto de Computação

Leandro Tacioli

WASIS - Bioacoustic Species Identification based on Multiple Feature Extraction and Classification Algorithms

WASIS - Identificação Bioacústica de Espécies baseada em Múltiplos Algoritmos de Extração de Descritores e de Classificação

CAMPINAS
2017

Leandro Tacioli

WASIS - Bioacoustic Species Identification based on Multiple Feature Extraction and Classification Algorithms

WASIS - Identificação Bioacústica de Espécies baseada em Múltiplos Algoritmos de Extração de Descritores e de Classificação

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Thesis presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientadora: Profa. Dra. Claudia Maria Bauzer Medeiros

Este exemplar corresponde à versão final da Dissertação defendida por Leandro Tacioli e orientada pela Profa. Dra. Claudia Maria Bauzer Medeiros.

CAMPINAS
2017

Agência(s) de fomento e nº(s) de processo(s): CNPq, 132849/2015-1

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

T118w

Tacioli, Leandro, 1986-
WASIS - Bioacoustic species identification based on multiple feature extraction and classification algorithms / Leandro Tacioli. – Campinas, SP : [s.n.], 2017.

Orientador: Claudia Maria Bauzer Medeiros.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Computação.

1. Reconhecimento de padrões. 2. Sistemas de computação. 3. Bioacústica. 4. Animais - Identificação. I. Medeiros, Claudia Maria Bauzer,1954-. II. Universidade Estadual de Campinas. Instituto de Computação. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: WASIS - Identificação bioacústica de espécies baseada em múltiplos algoritmos de extração de descritores e de classificação
Palavras-chave em inglês:
Pattern recognition
Computer systems
Bioacoustics
Animals - Identification
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
Claudia Maria Bauzer Medeiros [Orientador]
Lucas Rodrigues Forti
Sandra Eliza Fontes de Avila
Data de defesa: 03-07-2017
Programa de Pós-Graduação: Ciência da Computação

Universidade Estadual de Campinas
Instituto de Computação

Leandro Tacioli

WASIS - Bioacoustic Species Identification based on Multiple Feature Extraction and Classification Algorithms

WASIS - Identificação Bioacústica de Espécies baseada em Múltiplos Algoritmos de Extração de Descritores e de Classificação

Banca Examinadora:

- Profa. Dra. Claudia Maria Bauzer Medeiros (*Orientadora*)
  Instituto de Computação - UNICAMP
- Prof. Dr. Lucas Rodriguez Forti
  Instituto de Biologia - UNICAMP
- Profa. Dra. Sandra Eliza Fontes de Avila
  Instituto de Computação - UNICAMP

A ata da defesa com as respectivas assinaturas dos membros da banca encontra-se no processo de vida acadêmica do aluno.

Campinas, 03 de julho de 2017

Acknowledgements

I would like to express my sincere thanks and gratitude to professor Claudia Bauzer Medeiros, not only for her great work and ability to handle my concerns, but also for sharing her precious knowledge. Thanks to Professor Luís Felipe Toledo for sharing his knowledge in biology, for the opportunity to work together in several projects, especially WASIS, and for opening several doors in my life. Professor André Santanchè and Professor Hélio Pedrini for all the insightful ideas, suggestions and feedback I have received throughout my work.

My parents, José Valdir and Marlene, for their unconditional love, caring and education, for always believing in my potential, and always supporting in troubled moments. My brothers and sisters (actually cousins), Daniel, André, Cintia, Camila and Simone for always being together in special moments. I also thanks my aunts and uncle, Marli, Marlei and Eduardo for always caring and supporting me. Special thanks to Ivie, Eder, Renato and Cleber who together with my brother Daniel and sisters gave me my most precious gifts: Júlio César, Gabriel, Cecília, Marília and Catarina. Thanks to my future sister-in-law Jéssica for always being a good partner to my brother André. My grandmother, Iolanda, who is sadly not with us anymore. Thank you eternally for everything that you did for our family. We all miss you so much.

Thanks to all my friends of LIS: Camilla, Celso, Daniel Cugler, Ewerton, Fabrício, Fagner, Felipe, Flávia, Ivelize, Jacqueline, Jaudete, Joana, João, Juan, Kléber, Luana, Lucas Batista, Lucas Carvalho, Márcio and Ray.

This gratitude extends to all my friends of FNJV and LaHNAB: Simone, Camila, Sandra, Lucas, Guilherme, Roseli, Mariane, Milena, Victor, Patrícia, Alexandre, Carlos Henrique, Carol, Tamilie, Luisa, Mariana, Joice, Anat and Cecília.

Finally, I would like to thank the financial support from funding agencies: CNPq (132849/2015-1), FAPESP (2013/02219-0), the Microsoft Research FAPESP Virtual Institute (NavScales project), FAPESP CCES (2013/08293-7), CNPq/INCT in Web Science (557128/2009-9) and FAPESP-PRONEX (eScience project).

Resumo

A identificação automática de animais por meio de seus sons é um dos meios para realizar pesquisa em bioacústica. Este domínio de pesquisa fornece, por exemplo, métodos para o monitoramento de espécies raras e ameaçadas, análises de mudanças em comunidades ecológicas, ou meios para o estudo da função social de vocalizações no contexto comportamental. Mecanismos de identificação são tipicamente executados em dois estágios: extração de descritores e classificação. Ambos estágios apresentam desafios, tanto em ciência da computação quanto na bioacústica. A escolha de algoritmos de extração de descritores e técnicas de classificação eficientes é um desafio em qualquer sistema de reconhecimento de áudio, especialmente no domínio da bioacústica. Dada a grande variedade de grupos de animais estudados, algoritmos são adaptados a grupos específicos. Técnicas de classificação de áudio também são sensíveis aos descritores extraídos e condições associadas às gravações. Como resultado, muitos sistemas computacionais para bioacústica não são expansíveis, limitando os tipos de experimentos de reconhecimento que possam ser conduzidos. Baseado neste cenário, esta dissertação propõe uma arquitetura de software que acomode múltiplos algoritmos de extração de descritores, fusão entre descritores e algoritmos de classificação para auxiliar cientistas e o grande público na identificação de animais através de seus sons. Esta arquitetura foi implementada no software WASIS, gratuitamente disponível na Internet. Como o WASIS é de código e expansível, especialistas podem realizar experimentos com diversas combinações de pares descritor-classificador para escolher os mais apropriados para a identificação de determinados sub-grupos de animais. Diversos algoritmos foram implementados, servindo como base para um estudo comparativo que recomenda conjuntos de algoritmos de extração de descritores e de classificação para três grupos de animais.

Abstract

Automatic identification of animal species based on their sounds is one of the means to conduct research in bioacoustics. This research domain provides, for instance, ways to monitor rare and endangered species, to analyze changes in ecological communities, or ways to study the social meaning of animal calls in their behavioral contexts. Identification mechanisms are typically executed in two stages: feature extraction and classification. Both stages present challenges, in computer science and in bioacoustics. The choice of effective feature extraction and classification algorithms is a challenge on any audio recognition system, especially in bioacoustics. Considering the wide variety of animal groups studied, algorithms are tailored to specific groups. Audio classification techniques are also sensitive to the extracted features, and conditions surrounding the recordings. As a results, most bioacoustic softwares are not extensible, therefore limiting the kinds of recognition experiments that can be conducted. Given this scenario, this dissertation proposes a software architecture that allows multiple feature extraction, feature fusion and classification algorithms to support scientists and the general public on the identification of animal species through their recorded sounds. This architecture was implemented by the WASIS software, freely available on the Web. Since WASIS is open-source and expansible, experts can perform experiments with many combinations of pairs descriptor-classifier to choose the most appropriate ones for the identification of given animal sub-groups. A number of algorithms were implemented, serving as the basis for a comparative study that recommends sets of feature extraction and classification algorithms for three animal groups.

List of Figures

Contents

Chapter 1

Introduction

The typical scenario in eScience involves collaboration among computer scientists and researchers from other branches of science for the development of their fields [25], as well as empowering scientists to do their research and obtain results in faster, better and different ways [43]. One such example is the work in bioacoustics, in which biologists and computer scientists collaborate in research concerning sounds produced by or affecting living species.

Audio recognition systems have been developed in several domains, such as automatic speech recognition [44], music information retrieval [36], acoustic surveillance [24], and bioacoustics [1] – subject of this work. Primary challenges during the development of these sound retrieval systems are the identification of effective audio features and classification methods [63]. Feature extraction focuses on analyzing and extracting meaningful information from audio signals, while classification use these extracted data to match against the respective data of samples previously stored in a repository.

A major concern in audio recognition systems is how feature extraction is coupled to the classification algorithms. In many cases, poor software design restricts the reuse of code in other contexts and limits the ability of researchers to exchange feature extraction algorithms [59]. In bioacoustics, the vast majority of researchers 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 [1]. On the other hand, researchers demand generic architectures that allow them to implement new algorithms without major concerns with supporting infrastructure for data manipulation [38].

Typical architectures for audio retrieval systems follow general guidelines that consider classification essentially based on machine learning algorithms [81]. However, the architecture of software systems for bioacoustic recognition is seldom configurable or expansible, and lacks information on internals - such as documentation. Thus, it is hard for experts to test different sets of feature extraction and classification algorithms to check for the most appropriate combinations thereof.

Given this scenario, the goal of this dissertation is to design a software architecture that supports multiple feature extraction, feature fusion, and classification algorithms to identify animals based on their sounds. To obtain this goal, we designed and implemented WASIS[1] - a free and extensible software platform that allows scientists to identify animal species based on their recorded sounds. A suite of data repositories that specifies which components are responsible for processing, retrieving and persisting information was integrated to this architecture. To the best of our knowledge, no similar architecture has ever been designed for bioacoustic identification. A previous version of the WASIS software (Version 1.0.0) was implemented for testing the ideas. The results of this implementation raised several open questions, which resulted in the proposed architecture and the implementation of several algorithms for sound identification.

The main contributions of this dissertation are:

- an architecture to help on the identification of animal species from their sounds - this architecture supports multiple feature extraction algorithms, feature fusion and classification algorithms, and facilitates the extension for new techniques;
- a free software that implements the architecture, to be used by scientists/users on the identification of animals based on their sounds;
- a comparative study providing recommended sets of feature extraction/classification algorithms for animal sound identification, exploring different animal groups.

The evaluation and validation of the comparative study was conducted using audio recordings deposited at Fonoteca Neotropical Jacques Vielliard (FNJV), considered one of the ten largest animal sound libraries in the world.

Part of this dissertation produced the paper *An Architecture for Animal Sound Identification based on Multiple Feature Extraction and Classification Algorithms* [86] that was published and presented at the *11º BreSci - Brazilian e-Science Workshop.*

The remainder of this document is structured as follows: Chapter 2 provides an overview of the basic concepts and related work to support and develop this dissertation; Chapter 3 presents details of the proposed architecture. Chapter 4 contains implementation aspects of the architecture, a case study on how a scientist can use the WASIS software, and the comparative study. Finally, Chapter 5 concludes this dissertation and discusses future work.

[1] WASIS: Wildlife Animal Sound Identification System (Version 2.0.0)
http://www.naturalhistory.com.br/wasis.html

Chapter 2

Basic Concepts and Related Work

This chapter presents the main concepts and related work for the development of this research. Section 2.1 introduces the study of bioacoustics and the relevance of animal sound identification. Section 2.2 addresses the concept of Audio Features, while Section 2.3 focuses on Audio Classification, both key elements of audio recognition. At last, Section 2.4 explores typical architectures for audio retrieval systems.

2.1 Bioacoustics

Bioacoustics is a branch of science related to every sound produced by or affecting all kinds of living organisms. Although it is a research line oriented to animal communication, studies have been conducted showing that plants can also emit acoustic signals and communicate through them [33, 34], or even showing an interaction between plants and animals from acoustic communication [79].

Bioacoustics studies sounds of all animal groups. However, the vast majority of researchers in this field are specialized in few or only one specific group. Vocalizations are species-specific for many animal groups, being beneficial by means of identifying species [57]. Algorithms have been created or applied to automate animal identification of amphibians [66, 97], birds [47, 84], insects [17, 20], primates [41, 62] and whales [68, 88], for instance. Thus, most of the recognition tools in bioacoustics are designed to meet the needs of the experts in question. In addition, a considerable number of researchers – who make use of these techniques – do not have the mathematical and programming expertise to develop efficient algorithms [1]. The design and development of new algorithms for analysis and recognition of animal sounds is one of the greatest contributions of the collaboration among computer scientists and bioacoustic researchers. Moreover, the advent of new equipments for sound recording and analysis (e.g., recorders, microphones, sound level meters) made technology essential for the development of the bioacoustics [92].

Developing animal sound recognition techniques is not a trivial task. Firstly, it is necessary to understand the significance, functions and strategies used by animals for the emission of acoustic signals. The main function of sound communication between animals is to attract mates for reproduction and territorial defense [22, 83, 90]. In dangerous situations, animals emit sounds to astonish or threat predators, as well as warn members of their species [56, 90]. Due to large competition in the acoustic space, animals go through periods of evolutionary and ecological adaptations, and strategies are selected to maximize the transmission and reduce the interference of their sounds [16]. One example of an ecological strategy is when animals set the frequencies of their songs [28, 101], which means that various acoustic signals can occur simultaneously in different frequency ranges and still be recognized by individuals of the same species (Figure 2.1).

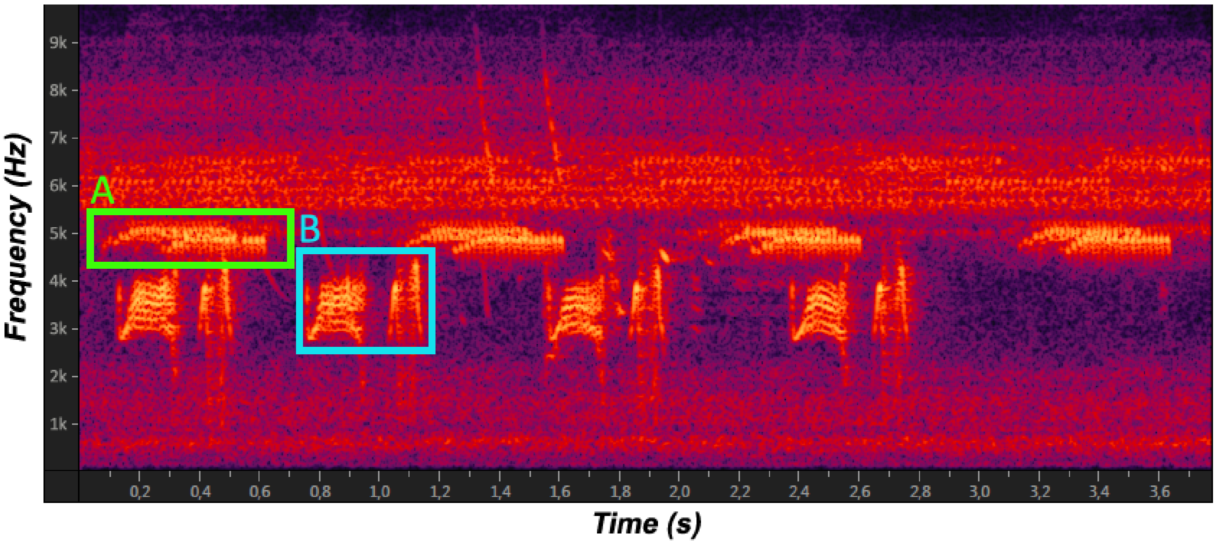


Figure 2.1: Different species sharing the same acoustic space. Species A emits sounds in higher frequencies (4.5-5.3kHz), while species B calls in lower frequencies (2.8-4.5kHz).

Monitoring animal populations is a recurrent subject in bioacoustics. With climate change, habitat loss, and high rates of species decline and losses, monitoring animals is an essential approach to deal with these threats and to manage conservation units [58, 96]. Animal monitoring through their sounds allows the estimation of population trends of key species in sensitive areas [8], provides evidences of changes in ecological communities through time [30] and increases the scale of ecological research from various locations over extended periods [93]. The main advantage of bioacoustic monitoring lies in the detection of animal sounds in the absence of an observer [8], even over larger spatial temporal scales [97]. Moreover, it is a popular non-invasive method to study animal populations, biodiversity and taxonomy [30, 50, 85].

2.2 Audio Features

Defining appropriate audio features is one of the crucial tasks regarding audio retrieval systems. Audio features represent the way in which meaningful information is analyzed and extracted from audio signals in order to obtain highly reduced and expressive data that are suitable for computer processing [2, 63, 80]. 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 [63].

The feature extraction process generates output vectors that are usually called descriptors [65]. These descriptors are the fundamental information that algorithms use to process the original audio files. A failure to capture appropriate feature descriptors of audio signals will result in poor performance, no matter how good the classifier is [59].

There are no optimal feature representations for particular applications, whether directed to an animal sound identification system or an automatic speech recognition application. Nevertheless, it is desirable that the choice of audio features covers the following properties [3]: (a) allows a system to automatically discriminate between different and similar sounds; (b) allows the creation of acoustic models without the need for excessive amount of training data; and (c) exhibits statistics that are largely invariant across the audio source and the environment. In addition, the feature extraction method should describe an audio segment in such a particular way that other similar segments can be grouped together by comparing their feature descriptors [82].

Mitrovic et al. [63] performed an extensive review of feature representations for audio retrieval. Audio features are categorized in different domains that provide information about their extraction process and computational complexity, as well as allowing the interpretation of the data. Most audio features representations belong to the following domains:

- *Temporal domain* – Based on the aspect represented by the audio signal changes over time, such as amplitude and power. This domain is considered the basis for feature extraction. For better audio analysis and identification, the audio signals are often transformed into more expressive domains;

- *Frequency domain* – Represents the spectral distribution of the audio signals, transforming such signals from *Temporal* to *Frequency* domain. A feature representation of this domain is Power Spectrum (PS) that employs Fast Fourier Transform (FFT) algorithm to compute the distribution of signal's power over given frequency bins of an audio file [73, 100];

- *Correlation/Autocorrelation domain* – Represents temporal relationships between audio signals. This domain reveals repeating patterns and their periodicities in a signal.

- *Cepstral domain* – Based on an approximation of the spectral envelope, capturing timbral information.

The following subsections describe common feature representations from the literature:

2.2.1 MFCC (Mel Frequency Cepstral Coefficients)

Firstly introduced by Bridle and Brown [13] and later developed by Mermelstein [61] in the 1970s, the Mel Frequency Cepstral Coefficents (MFCC) are widely used in audio recognition systems due to their abilities to represent the audio spectrum according to a perceptual scale that reflects the human auditory perception [37].

The human auditory system follows a linear scale up to 1kHz and logarithmic scale for frequencies above 1kHz, so humans hear lower frequencies more clearly than higher frequency components [31]. MFCC redistribute the frequency bands across the spectrum on the Mel scale, an approximation of the nonlinear human auditory system's response. MFCC provide a compact representation of the spectral envelope, thus timbre perception. Terasawa et al. [87] compared different Cepstral feature representations and determined that the MFCC representation is a good model for the perceptual timbre space.

Initially and regularly applied for automatic speech recognition [26, 44], MFCC have also had effective use in music information retrieval [31, 55]. In animal sound recognition, MFCC presented significant results in amphibians [10, 66, 97], birds [18, 19, 91], among other animal groups. Jančovič et al. [47] reported significant reduction in the accuracy of the identification when MFCC are applied in noisy environments. The failure of the conventional MFCC lies on the capture of the entire frequency band, which may contain significant background noise and/or presence of other animal sounds simultaneously [47, 48].

As shown in Figure 2.2, MFCC extraction consists of seven steps:

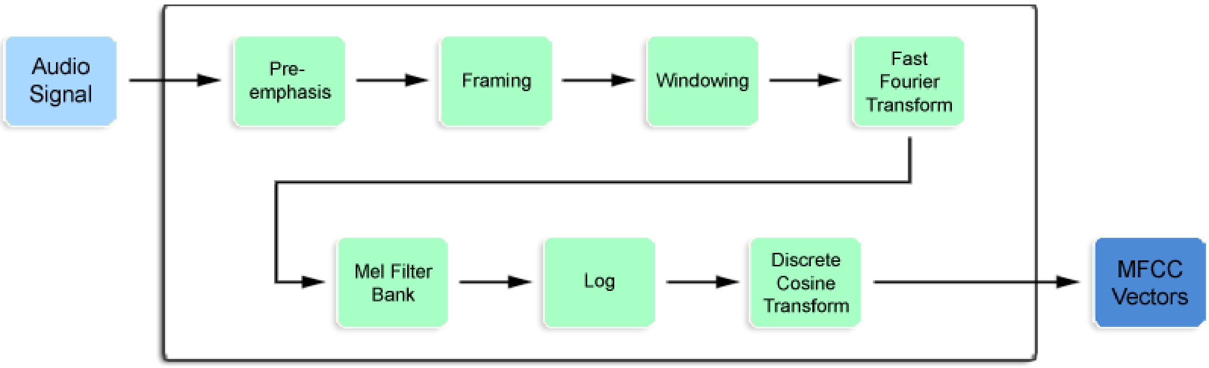


Figure 2.2: Block diagram of the MFCC algorithm.

- *Pre-emphasis* - Passes the audio signal through a filter that equalizes amplitude of high and low frequencies (high frequencies have smaller magnitudes compared to lower frequencies);

- *Framing* - Splits the signal into smaller frames, usually 20ms to 40ms with 50% overlap between consecutive frames (audio signals do not change much over short time scales and further processing across the entire signal would lose frequency contours over time);

- *Windowing* - Applies a window function to reduce discontinuities and smooth the audio signals at the edges of each frame;

- *Fast Fourier Transform (FFT)* - Converts each windowed frame from the time domain into the frequency domain by computing the Discrete Fourier Transform (DFT) and returns the magnitude distribution over different frequency bands;

- *Mel Filter Bank* - Multiplies the frequency magnitudes by a set of filters (the output adapts the magnitude spectrum to the Mel scale which satisfies the properties of the human ears, and reduces the size of the feature);
- *Log* - Computes the logarithm of the Mel Filter Bank output;
- *Discrete Cosine Transform (DCT)* - Converts the Mel Log powers into a time-like domain resulting in the desired set of MFCC.

MFCC extraction creates a vector of coefficients for each frame created in the *Framing* process. These MFCC vectors describe only the spectral envelope, thus they do not provide information about temporal changes in the spectra that also play an important role in human perception [45]. One method to capture this information is to calculate delta coefficients that measure the changes in coefficients over time. Delta MFCC (Differential Coefficients) are extracted from the static MFCC vectors, while Delta-Delta MFCC (Acceleration Coefficients) are extracted from the dynamic Delta MFCC.

There is no common guideline for the number of MFCC coefficients. A large number of coefficients increases the feature dimensionality and may cause data redundancy, which demands more computational resources, while a small number of coefficients may lead to insufficient data which results in low recognition performance [46]. A typical MFCC vector consists of 13 coefficients, but the 0th coefficient is commonly ignored because it is considered as a collection of average energies of the frequency bands [27], resulting on 12 coefficients. Adding 12 Delta coefficients and 12 Delta-Delta coefficients, the final MFCC vector contains 36 coefficients for each frame.

Most of the meaningful information needed for audio recognition is already contained in the 12 static MFCC coefficients, but the inclusion of Delta and Delta-Delta coefficients can significantly reduce the recognition error [45]. The performance of MFCC may also be affected by several factors, such as the number of filters, the shape of filters, and the type of window function [27, 89].

2.2.2 LPC (Linear Predictive Coding)

The basic concept of Linear Predictive Coding (LPC) is that a given audio sample at the current time can be well estimated based on a linear combination of previous sample values [67]. The goal of LPC is to estimate time-varying parameters of speech wave signals, such as the transfer function of the vocal tract and formant[1] frequencies [6].

Rabiner & Juang [74] reviewed the reasons why LPC has been widely used in speech analysis: (a) LPC is a good model of the speech signal, providing a good approximation to the vocal tract envelope shape; (b) LPC leads to a reasonable source-vocal tract separation, resulting on a parsimonious representation of the vocal tract characteristics; (c) the LPC model is mathematically precise, as well as simple and straightforward to implement both in software and hardware; and (d) LPC works well in recognition applications - performance based on LPC front ends is comparable or better that of recognizers based on different front ends.

[1] A formant is a concentration of acoustic energy within a particular frequency region.

LPC also makes a good representation of the spectral envelope, being applied to domains other than speech analysis [63]. Schön et al. [78] classified stress calls of domestic pigs using LPC and most of the unknown calls were correctly assigned. Mitrovic et al. [64] reported that LPC outperformed MFCC results using the popular Support Vector Machine (SVM) classifier when applied to a database with birds and domestic animals. However, the accuracy of LPC was considerably lower compared to other feature representations for frog sound identification, as reported by [97, 99].

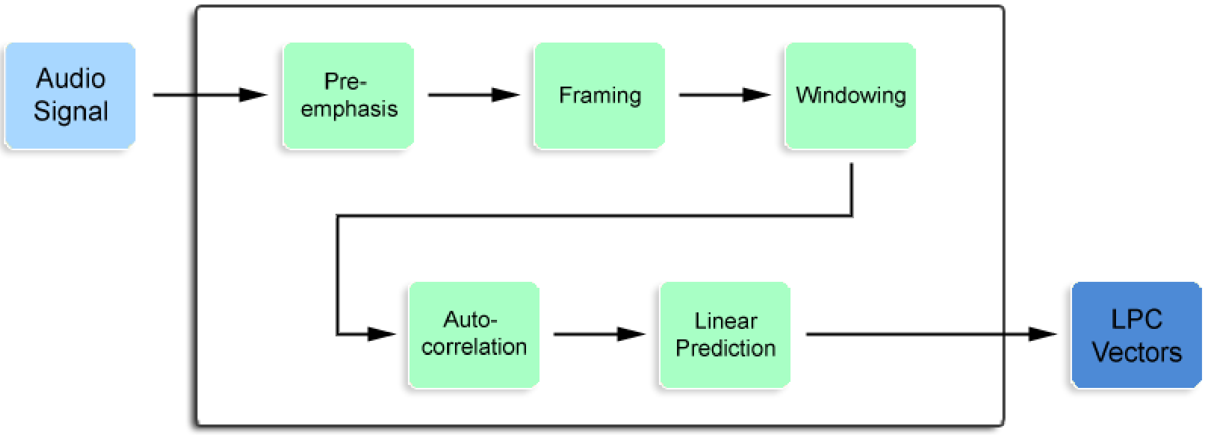


Figure 2.3: Block diagram of the LPC algorithm.

Figure 2.3 illustrates the LPC feature extraction process, which consists on five steps: (1) *Pre-emphasis*, (2) *Framing* and (3) *Windowing* are performed the same way as in the MFCC algorithm; (4) *Autocorrelation Analysis* provides a set of $(N+1)$ coefficients, where N is the order of the LPC analysis; and (5) *Linear Prediction Analysis* computes the final LPC coefficients from the autocorrelated vector using Levinson-Durbin algorithm.

2.2.3 LPCC (Linear Prediction Cepstral Coefficients)

Created as an audio representation in the Cepstral domain [32], Linear Prediction Cepstral Coefficients (LPCC) is an extension of Linear Predictive Coding (LPC). The fundamental idea of LPCC extraction is to apply a recursion technique to the LPC vectors rather than applying Fourier transform to the original audio signals.

One significant drawback in LPCC and LPC is their high sensitivity to noisy environments [102]. LPC components are also highly correlated, but it is desirable to obtain less correlated feature descriptors for acoustic modeling [4]. Thanks to the cepstral analysis, LPCC feature components are decorrelated, which is important to reduce computational complexity for probabilistic modeling [39].

Both LPC and LPCC extraction create vectors of coefficients for each frame created in the *Framing* process, similar to the process performed in the MFFC extraction.

2.2.4 PLP (Perceptual Linear Predictive)

Like many other popular audio feature representations, Perceptual Linear Predictive (PLP) was introduced for automatic speech recognition [42]. PLP produces a better representation of the spectral shape than the conventional linear predictive analysis by approximating three properties from the human hearing: (a) the critical-band spectral resolution, (b) the equal-loudness curve, and (c) the intensity-loudness power law.

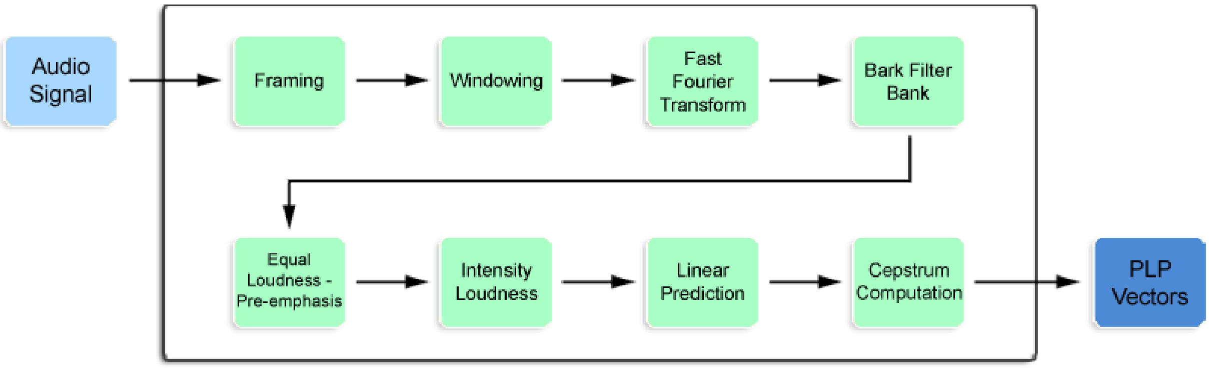


Figure 2.4: Block diagram of the PLP algorithm.

Figure 2.4 shows detailed steps of the PLP computation:

- *Framing*, *Windowing* and *Fast Fourier Transform (FFT)* are performed the same way as explained in previous feature representations;

- *Bark Filter Bank* - Converts the frequency magnitudes to Bark scale (better representation of the human auditory resolution in frequency);

- *Equal Loudness / Pre-emphasis* - Provides an approximation to the non-equal sensitivity of human hearing at different frequencies;

- *Intensity Loudness* - Provides an approximation to the power law of hearing and simulates the non-linear relation between the intensity of sound and its perceived loudness.

- *Linear Prediction* - Computes a linear prediction model (LPC) from the perceptually equally weighted signals;

- *Cepstrum Computation* - Cepstral coefficients (LPCC) are obtained from the predictor model using a recursion technique resulting in the desired set of PLP.

PLP has not been widely used in bioacoustics, despite offering better performance than linear predictive analysis in noisy conditions [51, 102]. Clemins & Johnson [21] created a generalized model (gPLP) that generates perceptual features for animal vocalizations by incorporating information about each species' sound perception. Potamitis et al. [72] applied PLP to automatic bird sound detection in continuous real field recordings with high scores of precision and recall.

2.2.5 Feature Fusion

The performance of audio feature representations may be affected by a series of factors in animal identification systems, such as the presence of background noise [5] and the duration of animal calls [97]. Feature fusion is a technique that is able to combine two or more audio feature representations. Disadvantages of feature extraction methods can be attenuated when several of these techniques are combined, as reported by Noda et al. [66] who merged four different methods. Their resulting feature descriptors held information of lower and higher frequency ranges and time variable characteristics employed separately in previous state-of-the-art work. Their combinations of feature extraction methods have also had better performance than the use of single feature representations for every identification using different retrieval techniques.

Arencibia et al. [5] combined temporal, frequency and cepstral domain features, showing they are more efficient than a single cepstral feature. Xie et al. [97] merged features from different domains (e.g., temporal, cepstral) that were able to better distinguish the content of frog calls. The authors concluded that their enhanced feature representation presented better classification accuracy than non-fused features, as well as good anti-noise ability.

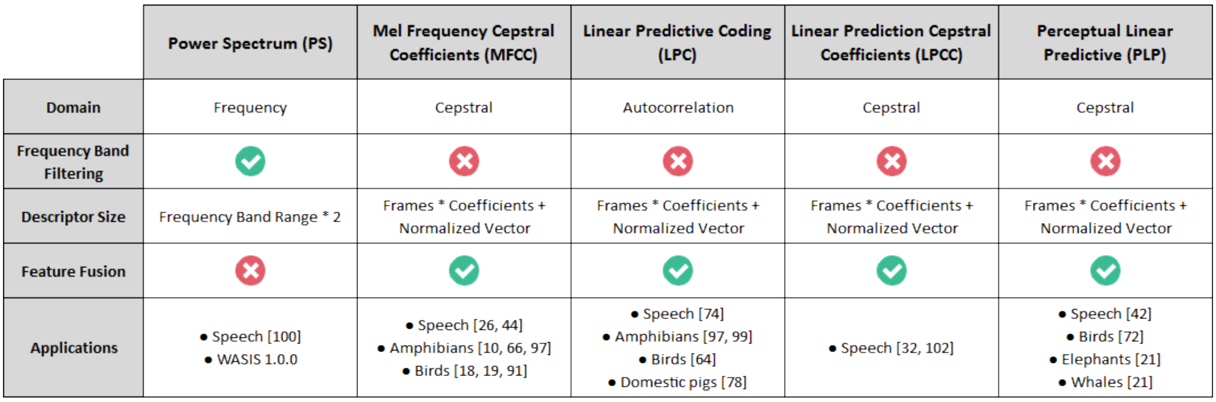
One simple way to implement feature fusion is concatenating the descriptors of different feature representations horizontally, creating a matrix where each row corresponds to an audio frame or audio segment (Figure 2.5), same as employed by Noda et al. [66].

$$\text{FUSION} = \begin{pmatrix} \text{Feature } A_1 & \text{Feature } B_1 & \text{Feature } C_1 & \text{Feature } D_1 & \text{Feature } E_1 \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \text{Feature } A_n & \text{Feature } B_n & \text{Feature } C_n & \text{Feature } D_n & \text{Feature } E_n \end{pmatrix}$$

Figure 2.5: Example of feature fusion matrix with 5 feature representations concatenated.

Table 2.1 shows an overview of the most common feature extraction algorithms presented and highlights some of their characteristics. Most of these feature representations belong to the Cepstral domain, in which timbre properties are extracted. Power Spectrum (PS) is the only algorithm that has the ability to filter the frequency band of audio signals, helpful in situations with significant background noise or even when several animal species calls at the same time. The size of the resulting descriptors may vary according to the information that is being filtering, as well as the number of coefficients of the algorithms. The resulting descriptors may also define whether or not a feature representation is allowed to perform fusion with other representations. MFCC, LPC, LPCC and PLP are allowed to perform fusion due to the *Framing* step in their extraction process. The last line of the table points out some applications, and in which animal groups the algorithms were applied, along with related literature. Note that most of the feature representations were initially designed for speech recognition and later applied to the bioacoustic domain.

Table 2.1: Overview of common feature representations



	Power Spectrum (PS)	Mel Frequency Cepstral Coefficients (MFCC)	Linear Predictive Coding (LPC)	Linear Prediction Cepstral Coefficients (LPCC)	Perceptual Linear Predictive (PLP)
Domain	Frequency	Cepstral	Autocorrelation	Cepstral	Cepstral
Frequency Band Filtering	✓	✗	✗	✗	✗
Descriptor Size	Frequency Band Range * 2	Frames * Coefficients + Normalized Vector	Frames * Coefficients + Normalized Vector	Frames * Coefficients + Normalized Vector	Frames * Coefficients + Normalized Vector
Feature Fusion	✗	✓	✓	✓	✓
Applications	• Speech [100] • WASIS 1.0.0	• Speech [26, 44] • Amphibians [10, 66, 97] • Birds [18, 19, 91]	• Speech [74] • Amphibians [97, 99] • Birds [64] • Domestic pigs [78]	• Speech [32, 102]	• Speech [42] • Birds [72] • Elephants [21] • Whales [21]

2.3 Audio Classification

Audio classification is the process by which an individual audio sample is assigned to a class, based on its characteristics [54, 95]. 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 their taxonomic information (e.g., genus, and specific epithet). Two classification approaches are found in the literature:

- *Brute Force* - The classification is performed by linearly traversing either the normalized or the entire set of feature descriptors, providing similarity results among every possible audio segment [77]. One statistical algorithm used for this approach is Pearson Correlation Coefficient (PCC);

- *Class Model* - Considered by the literature the main approach for audio classification [15, 81]. Commonly, it employs supervised machine learning algorithms for animal sound identification. These algorithms allow the computer to understand a data collection on a semantic level and assign them to previously created categories [53]. One popular techniques using this approach is Hidden Markov Model (HMM).

Standard classification is defined in the current context where each audio instance contains only one label (Single-Label). A sound can also be described by several meaningful tags where the sound may be related to multiple categories. The purpose of this kind of annotation is to label new sounds with all relevant tags [31]. For instance, on the top panel of Figure 2.6, each audio is assigned to a single species class (Single-Label). Occasionally, it is also significant to annotate/identify the animal sound category (e.g., advertisement call, social call, distress call) providing a better understanding of the animal behavior (Multi-Label), illustrated on the bottom panel of Figure 2.6.

In sound processing, the classification performance is mostly evaluated in terms of accuracy and speed [3]. Accuracy may be measured in percentage as the total number of correctly classified samples divided by the number of total samples [81]. Speed may be related to the total time needed on the classification, both in the training phase and in the evaluation phase. Minimum computing times on the evaluation phase makes a classifier more suitable for implementation in portable devices [66].

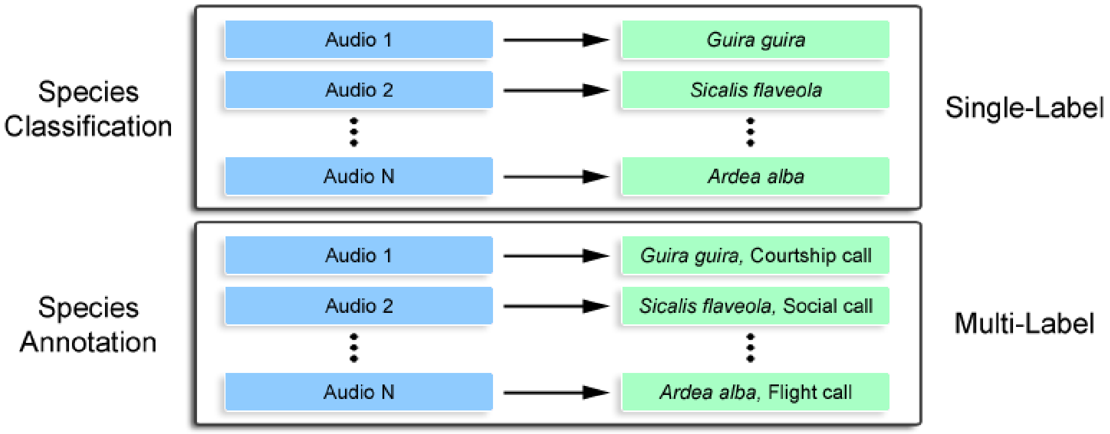


Figure 2.6: Species Classification (Single-Label) versus Species Annotation (Multi-Label).

2.3.1 PCC (Pearson Correlation Coefficient)

Pearson Correlation Coefficient is a measure of the strength of the linear association between two variables [40]. PCC considers the range values between +1 and −1, where +1 is total positive linear correlation, 0 is no linear correlation and −1 is total negative linear correlation.

Figure 2.7 exemplifies four different associations between two variables and their respective correlations. For example, the left top result shows total positive correlation, while right top shows total negative correlation.

PCC has been used in many applications related to audio analysis, such as noise reduction [11] and sound recognition [35]. This technique was also applied to establish a relationship between body measurements and acoustic features in primates [70].

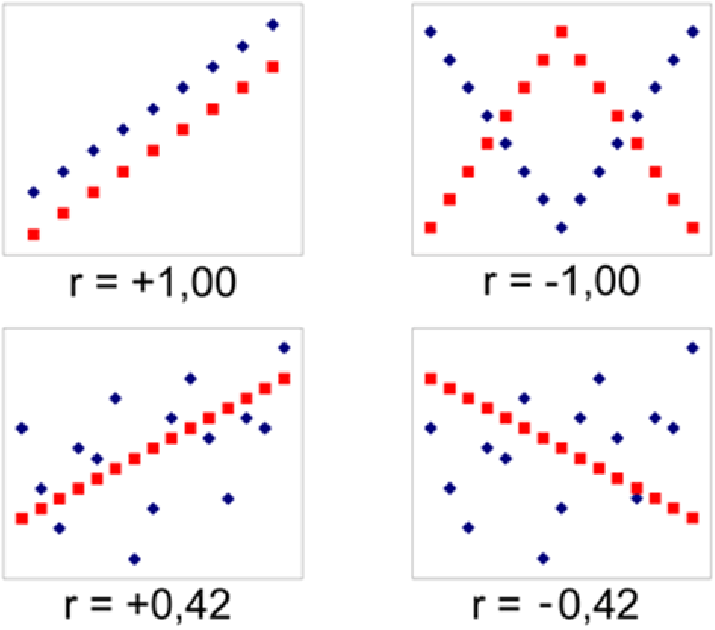


Figure 2.7: Pearson Correlation Coefficient results of different associations.

2.3.2 HMM (Hidden Markov Model)

The basic concept of Hidden Markov Model was introduced by Baum [9] in the 1960s. It is defined as a powerful statistical technique of characterizing observed data samples of a discrete-time series [45]. The main assumption of HMM is that the data samples can be well characterized as a parametric random process, and the parameters of the stochastic process can be estimated in a precise and well-defined manner [75].

The Hidden Markov Model is an extension of the Markov Chain Models, a discrete random process whose probability state at a given time depends solely on the state at the previous time [45]. HMM incorporates an observation which is a probabilistic function of the state. Hence, HMM is a double stochastic process with an underlying stochastic process which is not observable (hidden), but can only be observed through another set of stochastic process that produces the sequence of observations [75].

Figure 2.8 illustrates a Bakis type of HMM, also called left-right model, particularly appropriate for the bioacoustic domain because the transitions between states are produced in a single direction, similar to audio signal properties that change over time [75].

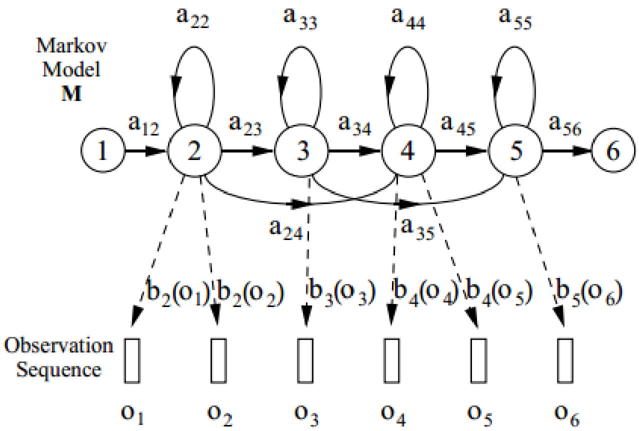


Figure 2.8: Diagram of a Hidden Markov Model, extracted from [98].

HMM have been used for numerous purposes in bioacoustics including species recognition of birds [48, 91], amphibians [1, 66] and whales [68]. Potamitis et al. [72] employed HMM for automatic bird sound detection focusing on small amount of training data and evaluated the proposal in continuous real field recordings with high scores of precision and recall. Pace [68] stated that the performance of HMM is maximized when several samples of various call categories and recordings of different quality are included in the classifier, so that variability amongst calls is taken into account.

2.4 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 & Moir [81]: (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. [63] described an architecture with more detailed database components with three main modules: (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.

Foggia et al. [29] presented an architecture that employs bag-of-audio-words (BoAW) approach between the typical feature extraction and classification steps. The idea of this approach is to create a dictionary of perceptual units of hearing using a clustering process and feed the classifier with a histogram of the occurrences of these perceptual units. However, this method has numerous criticisms, mostly due to information loss in the clustering step [69].

Classical feature extraction-classification architectures may use a variety of techniques for each stage. For instance, Deep Learning has gained significant attention in the classification stage for pattern recognition [44, 81]. In recent years, this technique has been adopted in architectures for animal sound identification based on acoustic and image features [12, 60, 76].

Chapter 3

The WASIS 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 feature extraction and audio classification to a suite of data repositories. The WASIS software is the first implementation of the proposed architecture – described in Chapter 4.

3.1 Overview

Figure 3.1 presents an overview of the WASIS architecture, which uses the classical feature extraction and classification structure. 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*.

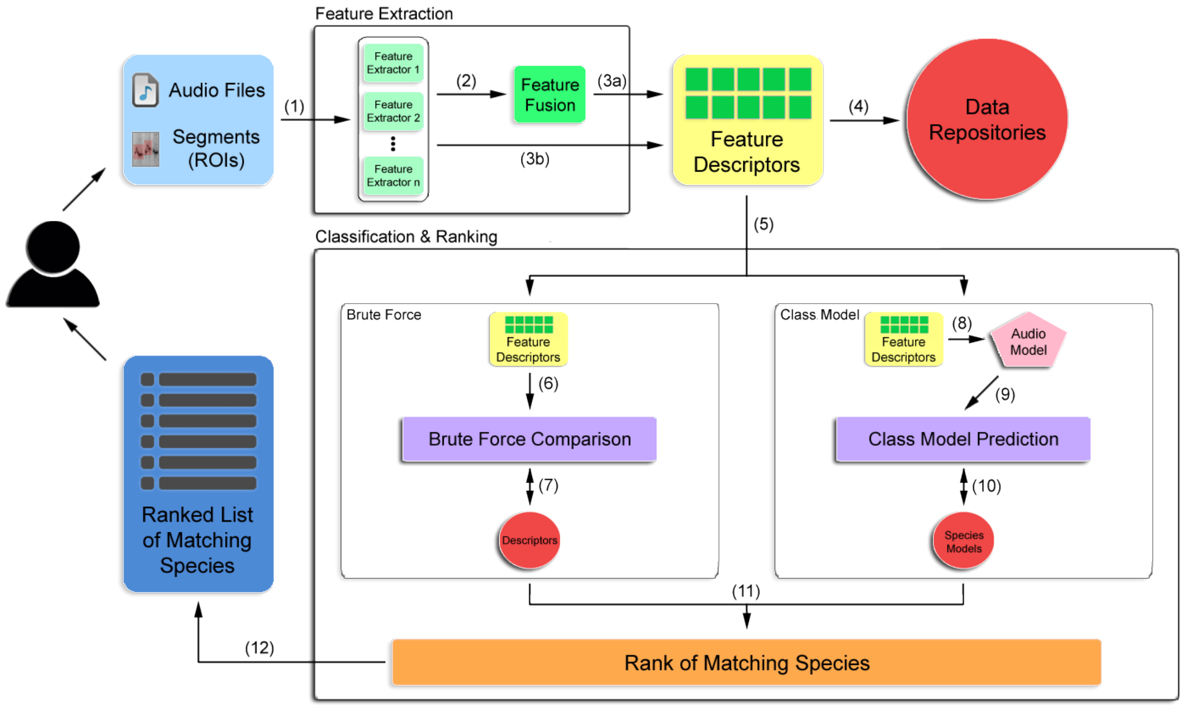


Figure 3.1: Detailed software architecture.

The *Data Repositories* component represents all the different repositories created and 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 sound samples.

The *Classification & Ranking* module classifies the input ROIs, receiving *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 the 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 3.2 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 previously mentioned in the architecture overview (Section 3.1).

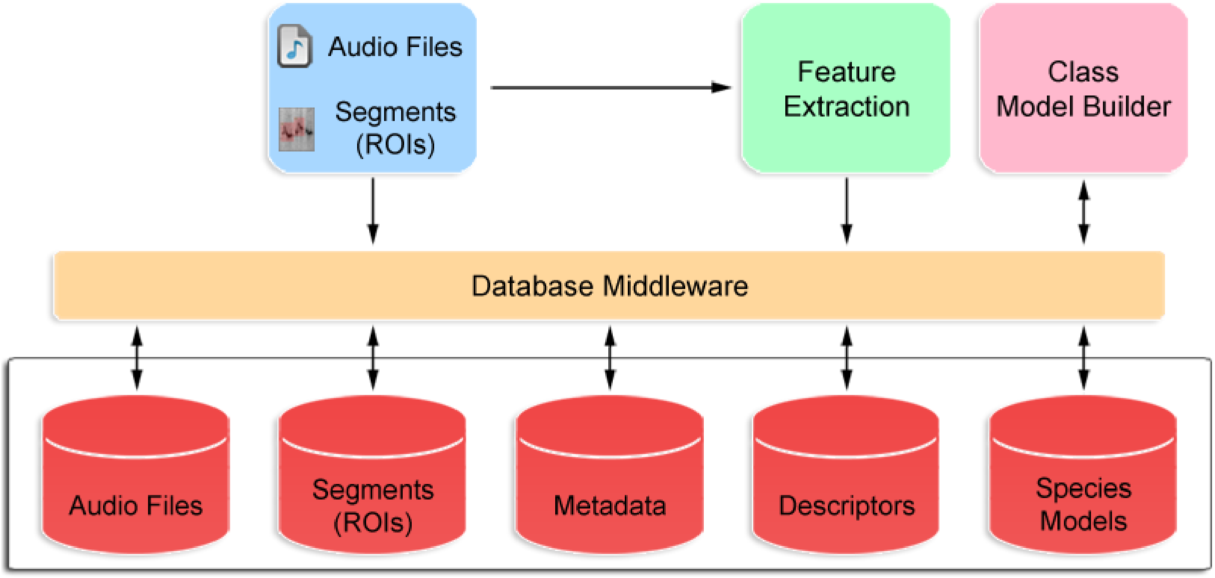


Figure 3.2: Structure of the data repositories.

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*.

Figure 3.3 shows the database schema describing the organization of data and how these data are related in the repositories.

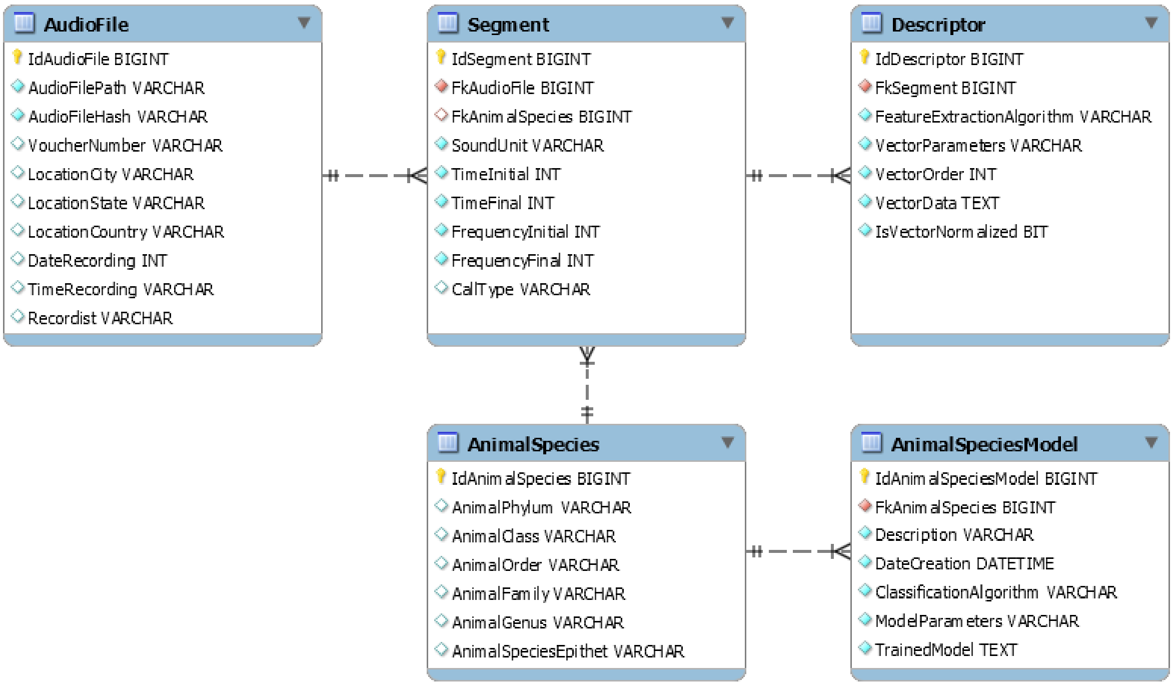


Figure 3.3: Database schema of the data repositories.

The table *AudioFile* contains fields that have information about path locations of the raw audio files, together with some of their related metadata (e.g., location, date, time). In addition to the *AudioFile* table, the table *AnimalSpecies* stores metadata about scientific classification of the species. This separation of metadata between tables allows the inclusion of several ROIs of different species for a same audio file, as seen in table *Segment*, which is related to *AudioFile* and *AnimalSpecies*.

The table *Descriptor* stores the audio segment vectors obtained from the feature extraction process. Note that the identifier of feature extraction algorithm and its parameters are stored as well. Different modules of the architecture must retrieve from this table the data related to the feature extraction algorithm that will be used on training and classification stage. For instance, in a given scenario, the MFCC algorithm is selected to perform the *Brute Force* identification, therefore the *Brute Force Comparison* module must retrieve only descriptors associated with the MFCC feature extraction algorithm.

Lastly, table *AnimalSpeciesModel* contains fields to store information about the trained models for the *Class Model* approach – detailed in Section 3.3.

3.3 Class Model Builder

The architecture provides the *Class Model Builder* (Figure 3.4), 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 through the *Class Model* approach.

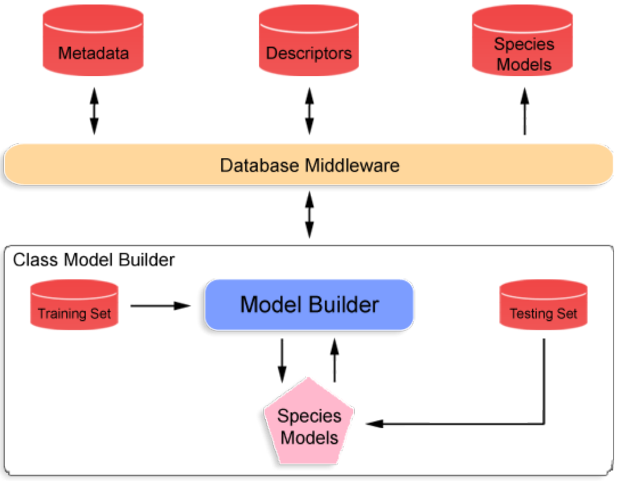


Figure 3.4: Design of the *Class Model Builder*.

The *Class Model Builder* may eventually create two datasets using the metadata and feature descriptors. The *Training Set* is obligatory created to provide feature descriptors to the machine learning algorithm that will create the *Species Models*. The *Testing Set* is created using different data from those used in the *Training Set* 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.

Chapter 4

Implementation Aspects

This chapter presents the implementation aspects and the comparative study of this dissertation. Section 4.1 presents an overview of the software WASIS that implements the proposed architecture and a brief description of the technologies used on this software. Section 4.2 contains examples of WASIS usage. Lastly, Section 4.3 presents the experimental methodology applied and the results obtained from the comparison study of feature representations and classification algorithms.

4.1 WASIS

The first implementation of the architecture produced the second version of the WASIS[1] software, originally created as part of a joint research partnership between Laboratory of Information Systems (LIS) and Fonoteca Neotropical Jacques Vielliard (FNJV) of the University of Campinas (UNICAMP) in 2013. The first version of WASIS was designed to support only Power Spectrum (PS) feature extraction and Pearson Correlation Coefficient (PCC) as the comparison method. In the course of this research, the original architecture of WASIS was replaced by the proposed architecture in this dissertation.

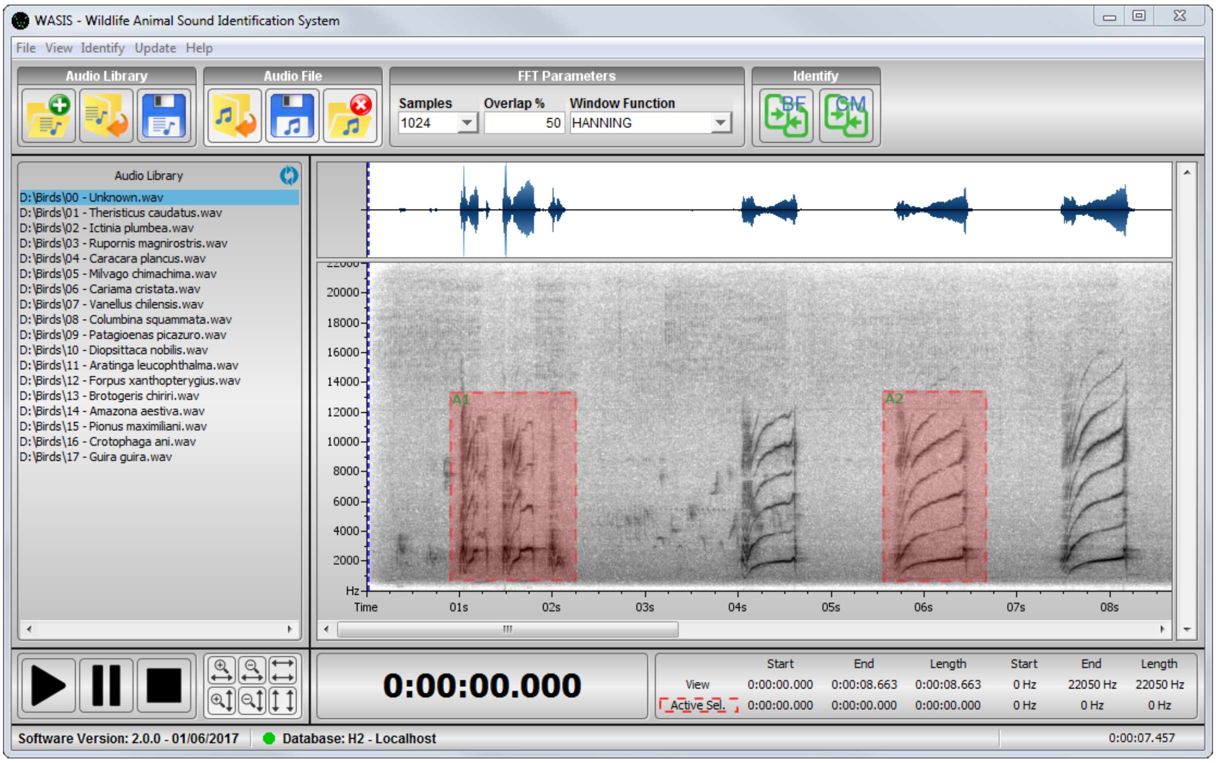
WASIS is implemented in Java due to its cross-platform support that allows running the application on multiple operating systems, such as Windows, Linux and Mac OS. In addition, Java applications may be executed on mobile devices that run Android platform. Since the main focus of this application is bioacoustical identification and there is extensive work in field locations in this domain, the software can be further extended to portable devices, such as smartphones and tablets.

The database needs of WASIS were implemented using MySQL and H2 databases. H2 is more adequate for those who just need the software for sound identification due to its embedded mode that does not require previous installation, and is more flexible for portable device. On the other hand, MySQL is more appropriate for those who want to do research, store and analyze more volume of data.

[1] WASIS: Wildlife Animal Sound Identification System (Version 2.0.0)
http://www.naturalhistory.com.br/wasis.html

4.2 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 4.1 illustrates a screen copy of WASIS interface which shows in red squares the audio segments selected by the scientist.



Samples: 1024
Overlap %: 50
Window Function: HANNING

Audio Library

D:\Birds\00 - Unknown.wav
D:\Birds\01 - Theristicus caudatus.wav
D:\Birds\02 - Ictinia plumbea.wav
D:\Birds\03 - Rupornis magnirostris.wav
D:\Birds\04 - Caracara plancus.wav
D:\Birds\05 - Milvago chimachima.wav
D:\Birds\06 - Cariama cristata.wav
D:\Birds\07 - Vanellus chilensis.wav
D:\Birds\08 - Columbina squammata.wav
D:\Birds\09 - Patagioenas picazuro.wav
D:\Birds\10 - Diopsittaca nobilis.wav
D:\Birds\11 - Aratinga leucophthalma.wav
D:\Birds\12 - Forpus xanthopterygius.wav
D:\Birds\13 - Brotogeris chiriri.wav
D:\Birds\14 - Amazona aestiva.wav
D:\Birds\15 - Pionus maximiliani.wav
D:\Birds\16 - Crotophaga ani.wav
D:\Birds\17 - Guira guira.wav

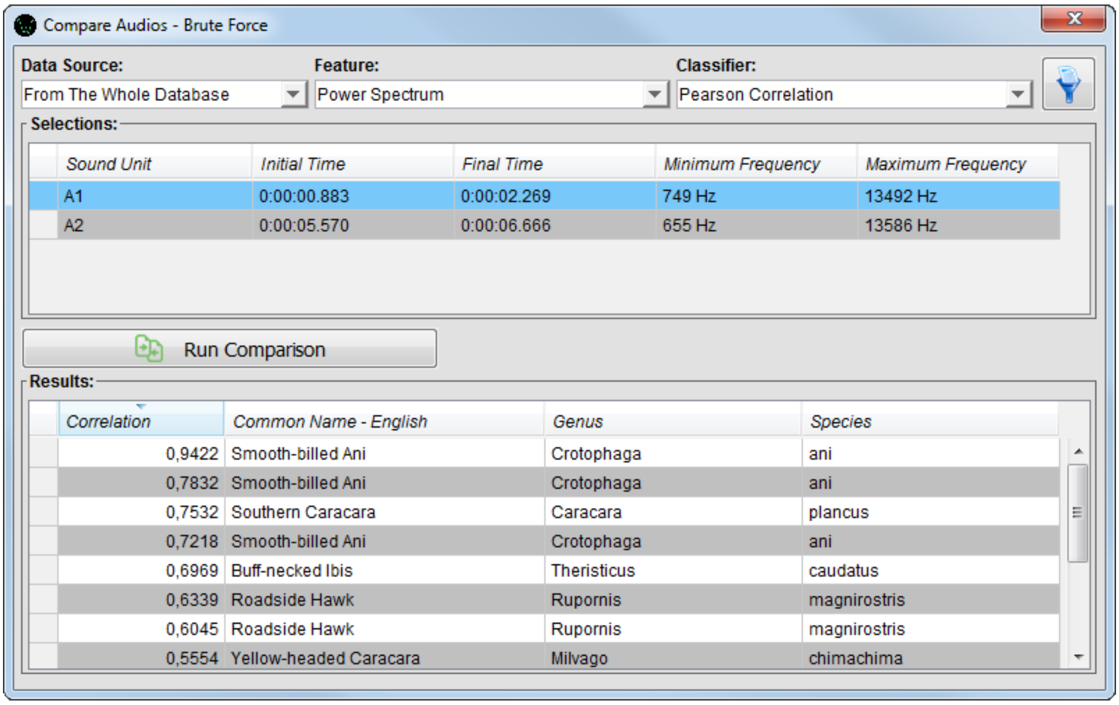
0:00:00.000

	Start	End	Length	Start	End	Length
View	0:00:00.000	0:00:08.663	0:00:08.663	0 Hz	22050 Hz	22050 Hz
Active Sel.	0:00:00.000	0:00:00.000	0:00:00.000	0 Hz	0 Hz	0 Hz

Software Version: 2.0.0 - 01/06/2017 ● Database: H2 - Localhost 0:00:07.457

Figure 4.1: WASIS interface with audio segments to be identified.

Figure 4.2 shows a screen copy of a *Brute Force* classification. The scientist filters the source from where the data will be compared, along with the feature extraction and classifier. The software performs the comparison according to the architecture flow (Section 3.1). Initially, the module extracts features of the audio segment requested by the scientist and returns the descriptors necessary to the classification. Then, these descriptors are matched against other audio files vectors contained in the *Descriptors* repository, returning a ranked list of matching species, using Pearson Correlation Coefficient. The higher the correlation coefficient between two audio segments, the higher the probability of a species being classified correctly. In this example, the software indicates that the audio segment selected by the scientist is more likely to belong to a Smooth-billed Ani (*Crotophaga ani*).



Compare Audios - Brute Force

Data Source: From The Whole Database

Feature: Power Spectrum

Classifier: Pearson Correlation

Selections:

Sound Unit	Initial Time	Final Time	Minimum Frequency	Maximum Frequency
A1	0:00:00.883	0:00:02.269	749 Hz	13492 Hz
A2	0:00:05.570	0:00:06.666	655 Hz	13586 Hz

Results:

Correlation	Common Name - English	Genus	Species
0,9422	Smooth-billed Ani	Crotophaga	ani
0,7832	Smooth-billed Ani	Crotophaga	ani
0,7532	Southern Caracara	Caracara	plancus
0,7218	Smooth-billed Ani	Crotophaga	ani
0,6969	Buff-necked Ibis	Theristicus	caudatus
0,6339	Roadside Hawk	Rupornis	magnirostris
0,6045	Roadside Hawk	Rupornis	magnirostris
0,5554	Yellow-headed Caracara	Milvago	chimachima

Figure 4.2: Screen copy for *Brute Force* audio comparison with its results.

The software also provides detailed information about audio identification. Figure 4.3 illustrates a visual comparison between audio segments, providing more information about the feature descriptors extracted. The *Power Spectrum* feature comparison shows information about the signal's maximum power (vertical axis) through the frequency bins (horizontal axis).

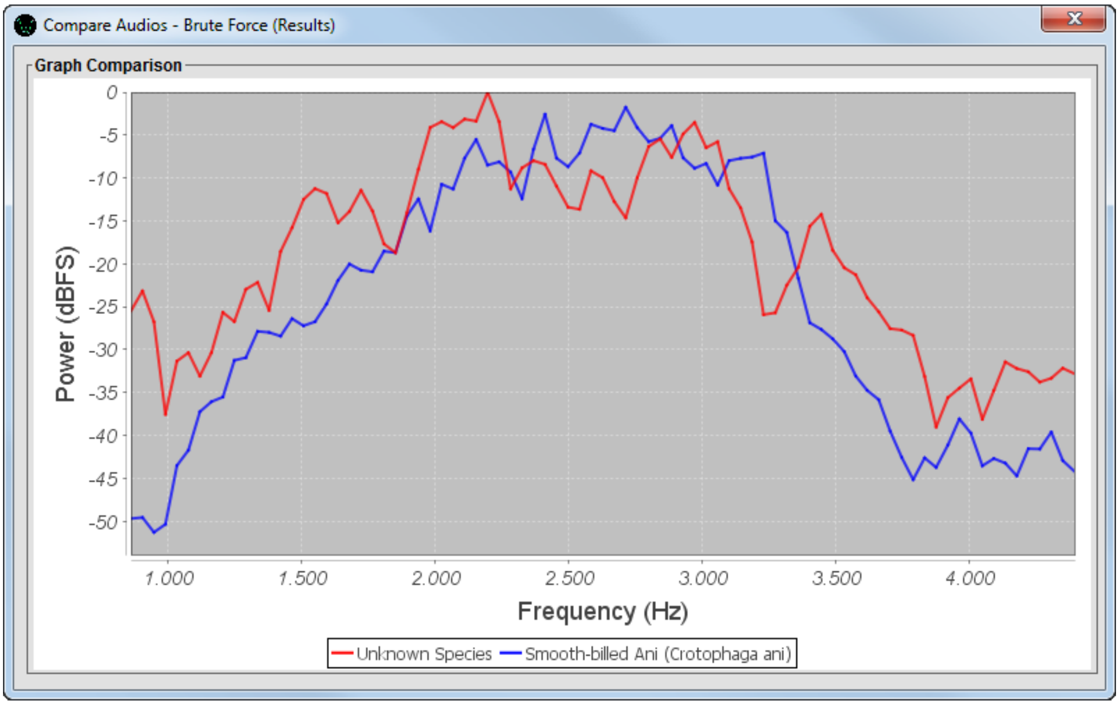
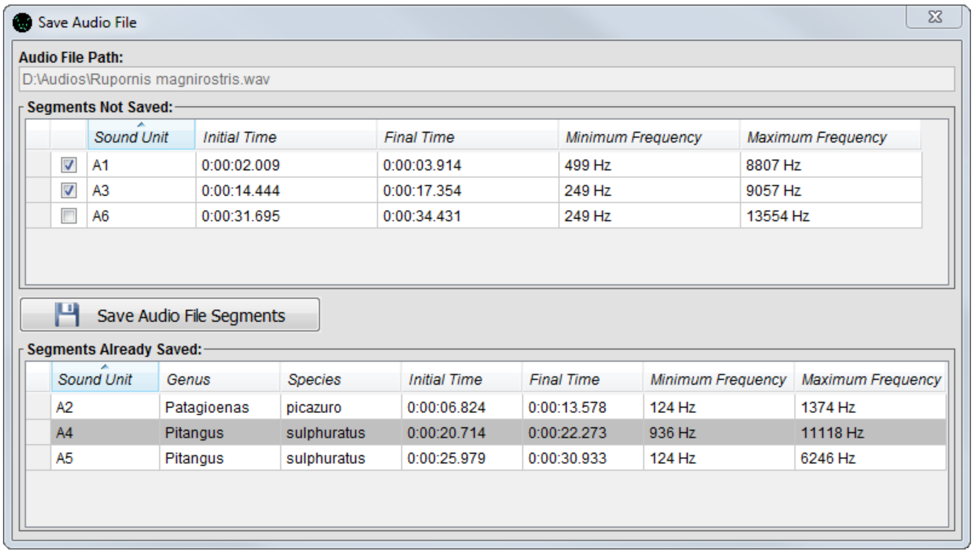


Figure 4.3: Power Spectrum comparison using *Brute Force* shows the data of the scientist audio segment (red) against the data samples from the *Descriptors* repository (blue).

The scientist has the choice of saving the audio segments into the appropriate data repositories and use their information for future identification. Figure 4.4 illustrates a screen with the information of several audio segments selected by the scientists. The top part of the figure shows the audio segments that were not saved into the database. Considering that an audio file may contain calls of several species, the scientist has to select the audio segments of one species and press the button "Save Audio File Segments" to continue the saving process. The bottom part of the same figure shows the audio segments already stored into the database and details to which species they belong.



Save Audio File

Audio File Path:
D:\Audios\Rupornis magnirostris.wav

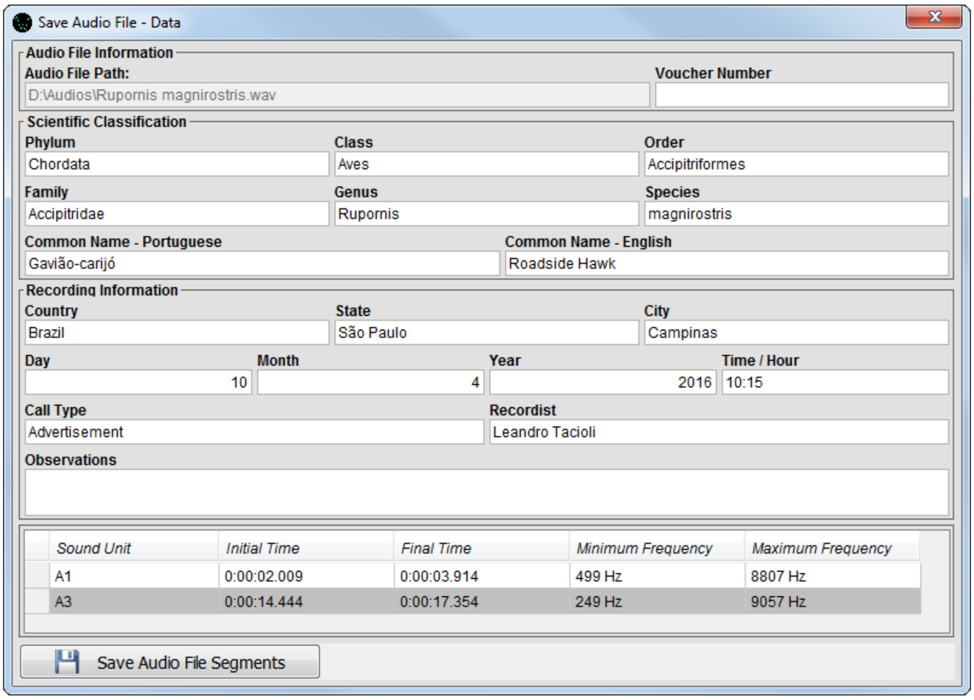
Segments Not Saved:

	Sound Unit	Initial Time	Final Time	Minimum Frequency	Maximum Frequency
☑	A1	0:00:02.009	0:00:03.914	499 Hz	8807 Hz
☑	A3	0:00:14.444	0:00:17.354	249 Hz	9057 Hz
☐	A6	0:00:31.695	0:00:34.431	249 Hz	13554 Hz

Segments Already Saved:

Sound Unit	Genus	Species	Initial Time	Final Time	Minimum Frequency	Maximum Frequency
A2	Patagioenas	picazuro	0:00:06.824	0:00:13.578	124 Hz	1374 Hz
A4	Pitangus	sulphuratus	0:00:20.714	0:00:22.273	936 Hz	11118 Hz
A5	Pitangus	sulphuratus	0:00:25.979	0:00:30.933	124 Hz	6246 Hz

Figure 4.4: Screen copy for selection of audio segments to be saved.



Save Audio File - Data

Audio File Information

Audio File Path:	**Voucher Number**
D:\Audios\Rupornis magnirostris.wav	

Scientific Classification

Phylum	**Class**	**Order**
Chordata	Aves	Accipitriformes
Family	**Genus**	**Species**
Accipitridae	Rupornis	magnirostris
Common Name - Portuguese	**Common Name - English**	
Gavião-carijó	Roadside Hawk	

Recording Information

Country	**State**	**City**	
Brazil	São Paulo	Campinas	
Day	**Month**	**Year**	**Time / Hour**
10	4	2016	10:15
Call Type	**Recordist**		
Advertisement	Leandro Tacioli		

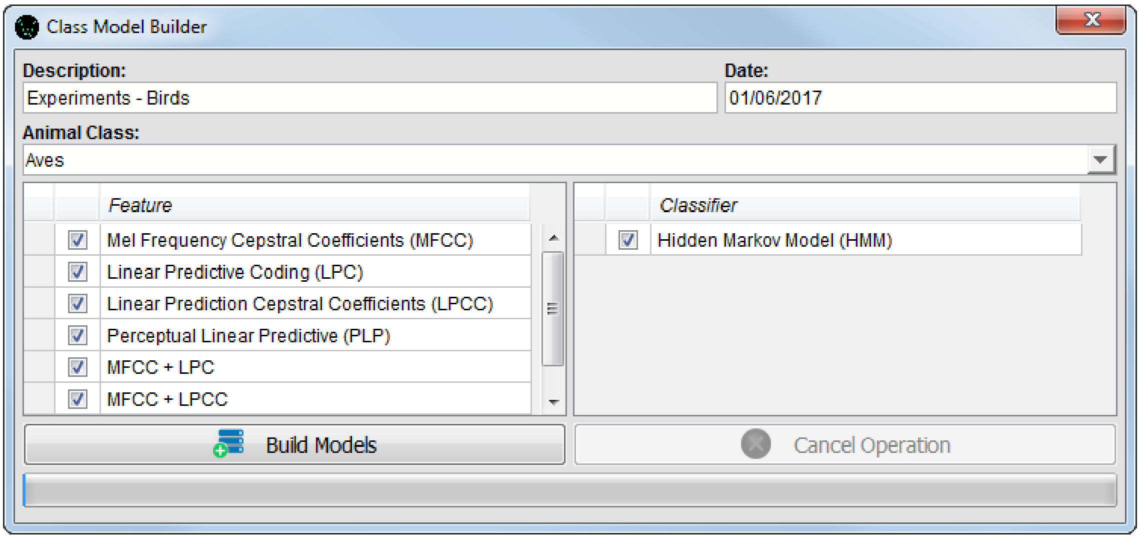
Observations

Sound Unit	Initial Time	Final Time	Minimum Frequency	Maximum Frequency
A1	0:00:02.009	0:00:03.914	499 Hz	8807 Hz
A3	0:00:14.444	0:00:17.354	249 Hz	9057 Hz

Figure 4.5: Form containing metadata and details about audio segments to be stored into the data repositories.

Figure 4.5 shows a screen copy with a form that contains the metadata belonging to the audio segments. When saving these metadata, the software automatically extracts the feature descriptors and stores them into the adequate *Descriptors* repository. For instance, the extracted MFCC coefficients can be stored in a table that contains only MFCC descriptors.

For the *Class Model* approach, the scientist initially needs to train species models prior to the identification. Figure 4.6 shows a screen copy of the *Class Model Builder*. Considering the scientist an expert in a specific field of research (e.g., birds, amphibians), he/she is able to train only the models of the specialized field by filtering the taxonomic data on the top of the screen. Then he/she selects the sets of features and classifiers, and starts building the species models. The time required to train the models may vary depending, mainly, on the number of audio records stored into the repositories, as well as the size of the feature descriptors, the fusion among feature representations, and the classifiers. Lastly, the species models created are stored into their respective *Species Models* repository.



Description: Experiments - Birds

Date: 01/06/2017

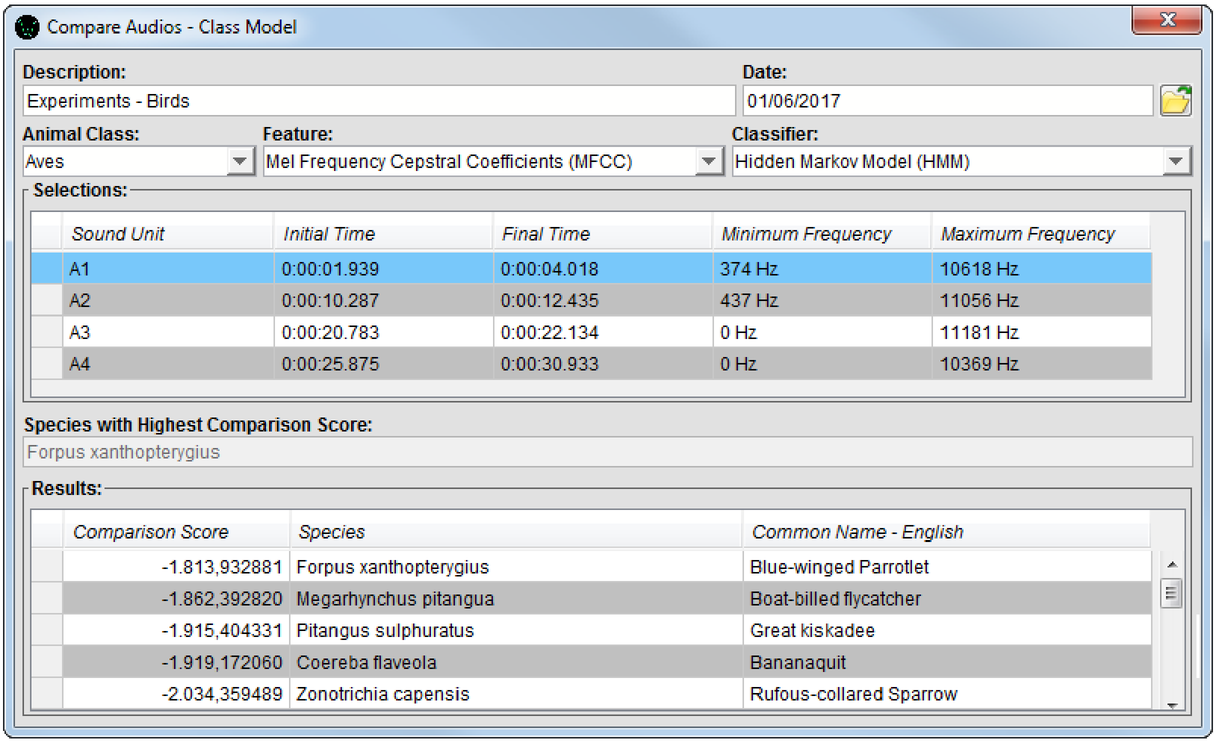
Animal Class: Aves

	Feature
☑	Mel Frequency Cepstral Coefficients (MFCC)
☑	Linear Predictive Coding (LPC)
☑	Linear Prediction Cepstral Coefficients (LPCC)
☑	Perceptual Linear Predictive (PLP)
☑	MFCC + LPC
☑	MFCC + LPCC

	Classifier
☑	Hidden Markov Model (HMM)

Figure 4.6: *Class Model Builder* screen copy.

Figure 4.7 shows a screen copy of *Class Model* classification and comparison. Similar to the *Brute Force*, the scientist chooses the taxonomic data, feature representation and classifier, and the software performs the comparison. Instead of retrieving information from the *Descriptors* repository, the software performs the classification from the *Species Models* repository. In this example, HMM was applied to classify the feature descriptors based on MFCC and the result indicates that the audio segment is more likely to belong to a Blue-winged Parrotlet (*Forpus xanthopterygius*).



Compare Audios - Class Model

Description: Experiments - Birds

Date: 01/06/2017

Animal Class: Aves

Feature: Mel Frequency Cepstral Coefficients (MFCC)

Classifier: Hidden Markov Model (HMM)

Selections:

Sound Unit	Initial Time	Final Time	Minimum Frequency	Maximum Frequency
A1	0:00:01.939	0:00:04.018	374 Hz	10618 Hz
A2	0:00:10.287	0:00:12.435	437 Hz	11056 Hz
A3	0:00:20.783	0:00:22.134	0 Hz	11181 Hz
A4	0:00:25.875	0:00:30.933	0 Hz	10369 Hz

Species with Highest Comparison Score: Forpus xanthopterygius

Results:

Comparison Score	Species	Common Name - English
-1.813,932881	Forpus xanthopterygius	Blue-winged Parrotlet
-1.862,392820	Megarhynchus pitangua	Boat-billed flycatcher
-1.915,404331	Pitangus sulphuratus	Great kiskadee
-1.919,172060	Coereba flaveola	Bananaquit
-2.034,359489	Zonotrichia capensis	Rufous-collared Sparrow

Figure 4.7: Screen copy for *Class Model* audio comparison with its results.

4.3 Comparison Study

The following experiments have been designed to provide recommended sets of feature extraction and classification algorithms for animal identification, exploring sounds of different animal groups. In this comparison study, 30 species of birds, 15 species of amphibians and 5 species of primates were chosen. All recordings were obtained from Fonoteca Neotropical Jacques Vielliard (FNJV)[2], one of the ten largest animal sound libraries in the world. This sound collection has more than 33,000 digitized files - mainly birds. Most of these audio files were recorded in the Neotropical Region (mainly Brazil), but FNJV also possesses files from North America, Europe and Africa.

The audio recordings of FNJV cover a wide distribution of the Neotropical region, mainly Brazil. It is important to note that these are field recordings and each file potentially holds vocalizations of several species and background noise caused by weather or anthropogenic interference. For each animal species, 10 audio files are evaluated and a maximum of 5 audio segments per file were manually selected with various duration ranges, depending on the duration of the vocalizations. A number of 2,019 of segments were selected from the 500 audio files, combining a total of 1 hour, 21 minutes and 29 seconds of recordings to be analyzed. The information of the audio files chosen for the comparison study, along with the information of their selected segments is available at `http://www2.ib.unicamp.br/wasis/Segments.xlsx`.

Four experiments were considered based on the selected animal groups: (1) *Amphibians*, (2) *Birds*, (3) *Primates*, and (4) *All Groups*. A total of 10 sets of data for testing

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

were generated for the experiments, each with 70% of the audio files for training, and the remaining 30% of the recordings for the purpose of evaluation. Note that in case of Brute Force classification, it is not necessary to perform training, but each segment of an evaluation dataset will be matched against each segment of its respective training dataset. The catalog number of the audio files belonging to each dataset is available at `http://www2.ib.unicamp.br/wasis/Experiments.xlsx`.

All experiments were performed in an Intel i7-7700 3.60Ghz computer with 8GB of RAM, non-dedicated Windows 7 64-bit. Before any analysis, the audio files were encoded as 16-bit mono WAV format with a sampling rate of 44.1 kHz. After feature extraction, the descriptors and metadata of all audio files were stored into a MySQL database (Version 5.7.18), and further evaluations were performed by retrieving the stored information.

The following measures were analyzed in this comparison study: (1) time required to extract features of the audio segments; (2) time required to classify and rank the audio segments; and (3) the accuracy of different sets of feature representations and classifiers.

4.3.1 Feature Extraction

A total of five audio feature representations were evaluated in this comparison study: Power Spectrum (PS), Mel Frequency Cepstral Coefficents (MFCC), Linear Predictive Coding (LPC), Linear Prediction Cepstral Coefficients (LPCC) and Perceptual Linear Predictive (PLP). MFCC implementation was based on the algorithm contained in the jAudio[3] library, a framework for feature extraction. LPC, LPCC and PLP implementations were based on the algorithms from CMUSphinx[4], a set of speech recognition libraries.

Power Spectrum is the only feature representation capable of filtering the minimum and maximum frequencies of the selected audio segments. The other representations extract their information based on full individual frames, considering the whole frequency spectrum [14]. Hence, PS generates only a single vector per audio segment with variable size, not suitable to the *Class Model* approach. The Power Spectrum descriptors were extracted with approximately 23ms frames with 50% overlapped Hamming window and FFT size of 1024.

The remaining feature representations (MFCC, LPC, LPCC and PLP) were extracted with approximately 23ms frames with 50% overlapped Hamming window. MFCC were computed with 23 Mel filter bank and a total of 12 static coefficients were generated. Delta and Delta-Delta coefficients were also computed to form the full MFCC vectors with 36 coefficients, considering that these dynamic coefficients can reduce the recognition error [45]. The LPC and LPCC vectors were extracted with 24 coefficients, which is more suitable for the sampling rate in which the audio files were encoded [71]. PLP vectors were computed with 21 filters and a total of 24 coefficients, which had the best recognition results for various numbers of filters and coefficients in [49].

Similar to a time series, MFCC, LPC, LPCC and PLP generate multiple vectors for each audio segment (precisely one vector for each frame extracted). In order to reduce

[3] jAudio, McGill University, Canada - https://sourceforge.net/projects/jaudio/

[4] CMUSphinx, Carnegie Mellon University, USA - https://cmusphinx.github.io/

them to compact feature representations [81, 84], mean and standard deviation were also computed and concatenated forming a new vector per audio segment.

The comparison study also evaluated the fusion among feature representations. Power Spectrum was discarded in this analysis due to its variable vector sizes and its inability to create one vector for each frame of the audio segments. The fusions considered in this study were: MFCC+LPC, MFCC+LPCC, MFCC+PLP and MFCC+LPC+LPCC+PLP.

4.3.2 Classification

Two classifiers were assessed in this comparison study: Pearson Correlation Coefficient (PCC) using the *Brute Force* approach, and Hidden Markov Model (HMM) using the *Class Model* approach. HMM implementation was based on the algorithm contained in the OC Volume[5], a speech recognition engine.

Pearson Correlation Coefficient is able to classify all feature extraction techniques in the experiments. The whole set of descriptors resulting of the PS extraction are processed, while the vectors containing mean and standard deviation are the only ones processed for the MFCC, LPC, LPCC and PLP representations.

Temporal information of audio signals can be well applied to Hidden Markov Model [44], which means that HMM can handle the whole set of vectors of an audio segment (discarding mean and standard deviation vectors). Bakis type of HMM were employed, particularly appropriate for audio analysis [75]. K-means clustering algorithm were applied to generate a codebook with 256-dimension size. Additional parameters include a total of 5 HMM states, Baum-Welch algorithm to estimate the species models, and Viterbi algorithm to calculate the likelihood among observation sequences (audio segments) given the models.

4.3.3 Results and Discussions

Figure 4.8 shows the average time required to extract the feature descritors. Ten repetitions were performed to extract each feature representation from the 2,019 audio segments. There was no need to calculate the time required for feature fusions, since it is a simple concatenation of already computed descriptors.

MFCC, LPC and LPCC had the best results with averages of 203.67s, 202.95s and 206.63s, respectively. PLP took about 54% longer than the previous feature representations, due to its extraction that also includes LPCC calculation. PS extraction was approximately 80% slower than the best results, explained by a flaw in the implementation that performed some unnecessary verifications regarding minimum and maximum frequencies contained in the audio segments. The highest standard deviation observed was 4.89 seconds for LPC, indicating that the results of each feature extraction representation are very close to the average.

[5] OC Volume - https://github.com/dannysu/ocvolume

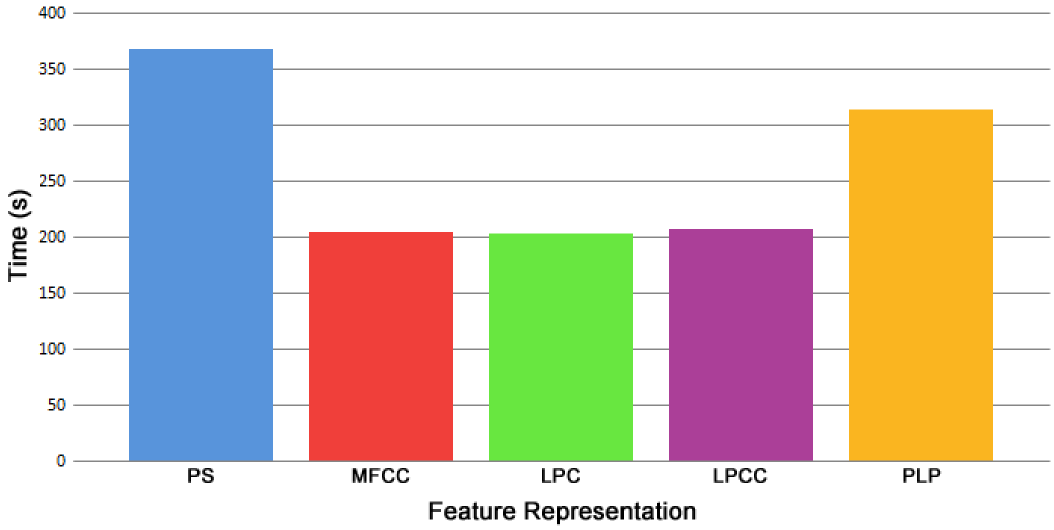


Figure 4.8: Comparison of time required for feature extraction (in seconds).

Figure 4.9 details the average time spent to classify and rank the matching species using Pearson Correlation Coefficient (PCC). For each feature representation, the fastest results were related to the *Primates* experiments, which contains fewer species for identification (only five). As we increase the number of species to be identified, the average time may increase considerably as observed in PS feature representation that use the entire set of descriptors for identification.

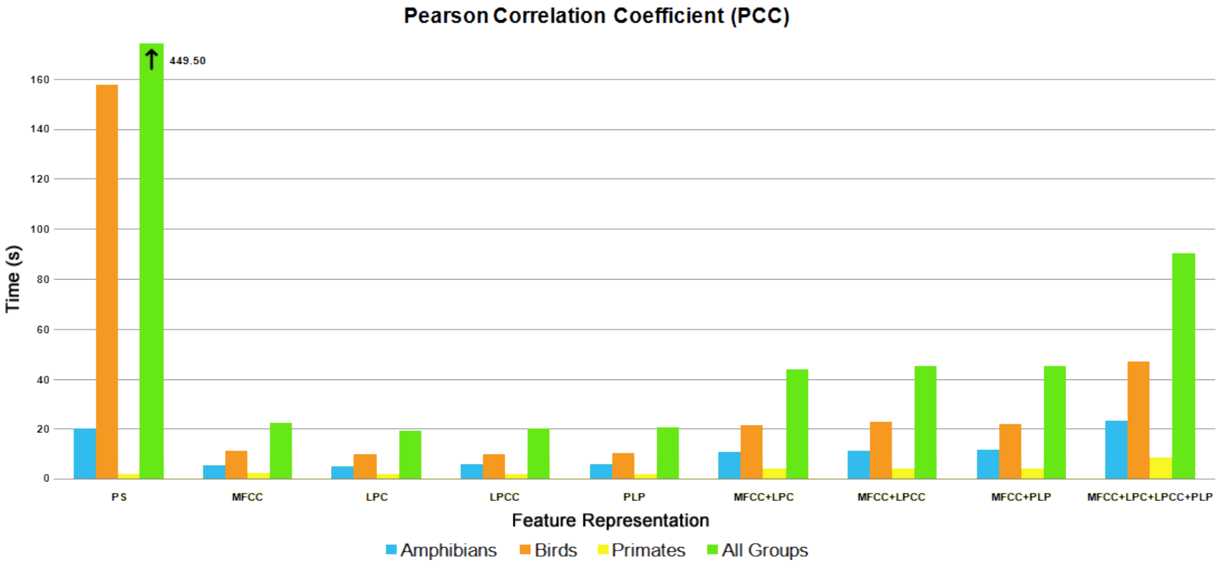


Figure 4.9: Comparison of time required for PCC classification and ranking (in seconds) with different feature representations.

MFCC, LPC, LPCC and PLP have similar number of coefficients and use only the mean and standard deviation vectors for identification in the *Brute Force* approach. Hence, their average time for classification and ranking dropped significantly compared to PS. Considering feature fusion, it is necessary to concatenate the mean and standard deviation vectors, which generally doubles the average time spent.

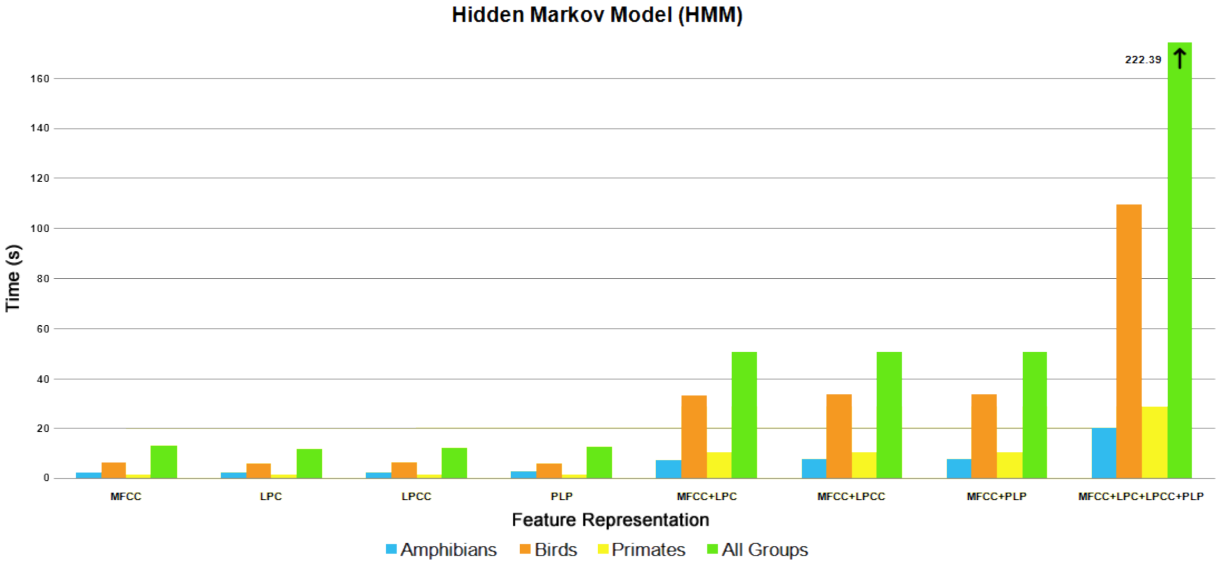


Figure 4.10: Comparison of time required for HMM classification and ranking (in seconds) with different feature representations.

It can be seen from Figure 4.10 that MFCC, LPC, LPCC and PLP together with Hidden Markov Model (HMM) can perform the fastest classification and ranking for all experiments. When combining two feature representations, such as MFCC+LPC, the average time for HMM classification and ranking increases similarly to PCC. HMM uses all descriptors obtained from the feature extraction process for classification, making it very costly to concatenate more than two feature representations, as seen in MFCC+LPC+LPCC+PLP results.

Note that even with a larger number of species of *Amphibians* compared to *Primates*, *Amphibians* had faster classification with HMM. It can be explained by the fact that the audio segments extracted for the *Amphibians* experiments are much shorter, generating less descriptors than the *Primates* experiments. All segments from bird audio files have approximately 53 minutes, while segments of amphibians have 14 minutes, and segments of primates have 14 minutes and 30 seconds.

The following results are related to the recognition accuracy of the feature representation and classifiers. We calculated the mean of the true positive rate (TPR), that measures the proportion of species that were correctly identified. We also calculate the standard deviation to assess the variability of the results around the mean. Figure 4.11 illustrates the results for the Pearson Correlation Coefficient comparison, while Figure 4.12 shows the results for the Hidden Markov Model classification.

The combination of Power Spectrum and Pearson Correlation Coefficient achieved the best performance for the *Amphibians* experiment – 74.38%. In spite of its slow average time of classification and ranking, this combination of PS and PCC was the only one able to reach accuracy rate close to 75% in any of the experiments.

Pearson Correlation Coefficient can be recommended for every animal group due to the highest classification accuracy for all feature representations. Several factors contribute for future improvements of recognition rates of HMM classification, such as setting better initial parameters and estimates, and increasing the size of training data [45].

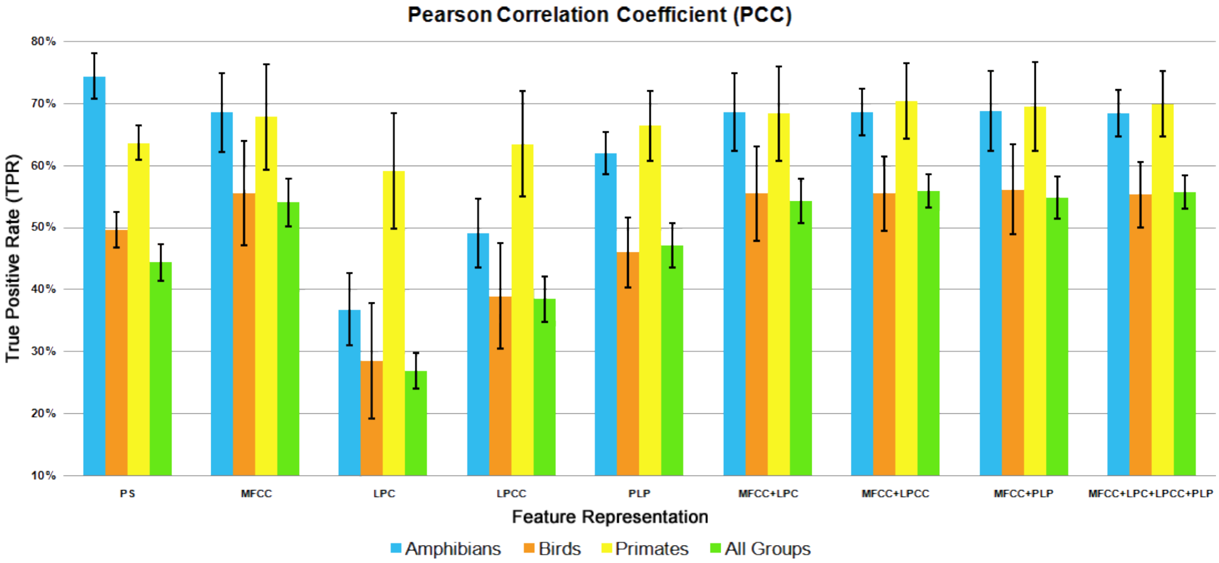


Figure 4.11: Comparison of true positive rate (TPR) for PCC among feature representations.

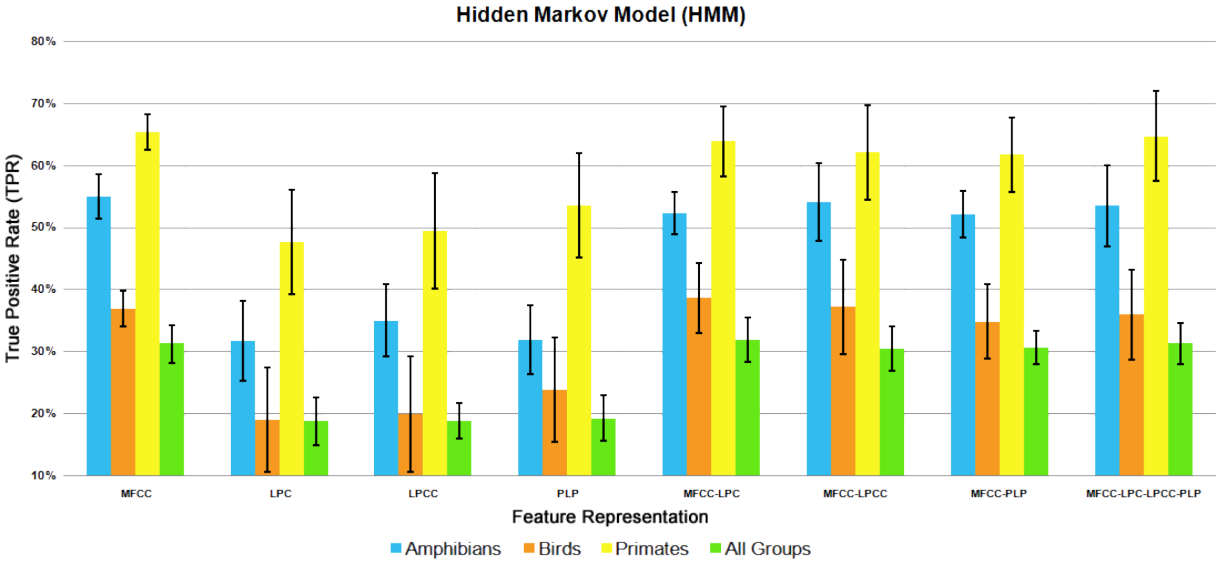


Figure 4.12: Comparison of true positive rate (TPR) for HMM among feature representations.

In most cases, the experiments also confirm that feature fusion slightly enhances the identification rate. For *Birds*, MFCC+PLP achieved slight improvement when compared to other fusions. MFCC+LPCC outperformed all the other feature representations for the *Primates* and *All Groups* experiments.

We can observe that the results involving MFCC and the fusions that contain MFCC had similar performances. Also considering the average time from previous evaluations (feature extraction and classification/ranking), MFCC is a suitable feature representation for implementations in portable field devices.

Last but not least, it is best to perform analysis for close taxonomic groups, and when there are large amount of data. The *All Groups* experiments did not perform well for these two reasons.

Chapter 5

Conclusions and Future Work

The ability to identify animal species based on their sounds is extremely useful for scientists and the general public. Besides the curiosity itself of knowing which species is calling, we can, for instance, identify invasive species in a certain area, estimate population trends of key species in sensitive areas and analyze changes in ecological communities over time.

This dissertation proposed 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 species based on their sounds. Along with the architecture, a conceptual database design described many different entities and their relationships, and illustrates the logical structure of a data repository suite. This architecture also allows the implementation of new algorithms without major concerns with supporting infrastructure.

The software WASIS was the first implementation of the architecture. In addition, several feature extraction and one classification algorithm for each classification approach were implemented in this software, validating the architecture as feasible. A case study was presented showing how scientists and users can use the software.

This dissertation also conducted a comparative study of different sets of feature extraction and classification algorithms for animal sound identification. Four sets of tests to measure accuracy and time needed to execute the experiments were generated. Three of these tests were performed with different animal groups and one with the combination of the groups. The results indicate that a Brute Force comparison technique (Pearson Correlation Coefficient) outperformed a Class Model technique (Hidden Markov Model) in all experiments. The results also showed that feature fusion slightly enhances the recognition rate, even though this combination of feature representations increases the time required for classification and ranking.

There are many possible extensions to this dissertation. Some examples of these extensions are:

- Redesign the implementation database to improve performance and flexibility. For instance, store the raw audio files into the software database instead of just persisting a path from where the files are located;
- Investigate techniques to automatically select audio segments from which to extract feature descriptors. The recommendation of ROIs in an audio file would support

researchers on the analysis of long-duration recordings;

- Correct the flaws on Power Spectrum extraction algorithm and adapt this feature representation to the Class Model approach. Instead of considering the whole information in just one vector, several vectors would be generated, similar to other feature representations;

- Investigate sound recognition techniques other than acoustic-based features. For instance, image shape features extracted from spectrograms [52];

- Reduce the dimensionality of feature vectors. Techniques, like Linear Discriminant Analysis (LDA), aim to perform dimensionality reduction and retain the class discriminatory capacity as much as possible [7]. This reduction would consume significant less amount of memory, together with less computing power;

- Investigate different approaches of fusion. Late-fusion techniques may be applied to the classification stage, improving the recognition accuracy [76];

- Create a bioacoustic repository for storing feature extraction and classification algorithms. Challenges for this domain-specific repository would be implementing it, maintaining it, and gaining adoption;

- Integration of well-known software workbenchs to current implementations. Weka [94] is a collection of machine learning algorithms and has achieved widespread acceptance within academia. The integration of softwares, like Weka, would not require the implementation of complex algorithms;

- Provide more extensive comparative studies in animal sound identification. Inclusion of new feature extraction and classification algorithms (such as Support Vector Machine [23]), setting different number of coefficients/filters for each feature representation, and more training data for the classification algorithms would contribute to the recommendation of appropriate sets of feature extraction/classification techniques for different animal groups;

- Perform comparison studies with other tools for bioacoustics. For instance, Raven[1] is a software specialized in sound analysis with correlation functionality for sound comparison. ARBIMON[2] is specialized in acoustic monitoring and provides tools for species identification.

[1] Raven, Cornell Lab of Ornithology, USA - http://www.birds.cornell.edu/brp/raven/RavenOverview.html

[2] ARBIMON, University of Puerto Rico-Rio Piedras, Puerto Rico/USA - https://www.sieve-analytics.com/arbimon

Bibliography

[1] T. Mitchell Aide, Carlos Corrada-Bravo, Marconi Campos-Cerqueira, Carlos Milan, Giovany Vega, and Rafael Alvarez. Real-time bioacoustics monitoring and automated species identification. *PeerJ*, 1:e103, 2013.

[2] Flora Amato, Luca Greco, Fabio Persia, Silvestro Roberto Poccia, and Aniello De Santo. Content-based multimedia retrieval. In *Data Management in Pervasive Systems*, pages 291–310. Springer International Publishing, 2015.

[3] M. A. Anusuya and S. K. Katti. Speech recognition by machine: A review. *International Journal of Computer Science and Information Security*, 6:181–205, 2009.

[4] M. A. Anusuya and S. K. Katti. Front end analysis of speech recognition: a review. *International Journal of Speech Technology*, 14:99–145, 2011.

[5] Juan J. Noda Arencibia, Carlos M. Travieso, David Sánchez-Rodríguez, Malay Kishore Dutta, and Garima Vyas. Automatic classification of frogs calls based on fusion of features and svm. In *Eighth International Conference on Contemporary Computing*, 2015.

[6] B. S. Atal and Suzanne L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America*, 50:637–655, 1971.

[7] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44:572–587, 2011.

[8] Rolf Bardeli, Daniel Wolff, Frank Kurth, Martina Koch, Klaus-Henry Tauchert, and Karl-Heinz Frommolt. Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters*, 31:1524–1534, 2010.

[9] Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37:1554–1563, 1966.

[10] Carol Bedoya, Claudia Isaza, Juan M. Daza, and José D. López. Automatic recognition of anuran species based on syllable identification. *Ecological Informatics*, 24:200–209, 2014.

[11] Jacob Benesty, Jingdong Chen, and Yiteng Huang. On the importance of the Pearson Correlation Coefficient in noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 16:757–765, 2008.

[12] Amira Boulmaiz, Djemil Messadeg, Noureddine Doghmane, and Abdelmalik Taleb-Ahmed. Robust acoustic bird recognition for habitat monitoring with wireless sensor networks. *International Journal of Speech Technology*, 19:631–645, 2016.

[13] J. S. Bridle and M. D. Brown. An experimental automatic word-recognition system. *JSRU Report No. 1003, Joint Speech Research Unit, Ruislip, England*, 1974.

[14] Forrest Briggs, Balaji Lakshminarayanan, Lawrence Neal, Xiaoli Z. Fern, Raviv Raich, Sarah J. K. Hadley, Adam S. Hadley, and Matthew G. Betts. Acoustic classification of multiple simultaneous bird species: A multi-instance multi-label approach. *The Journal of the Acoustical Society of America*, 131:4640–4650, 2012.

[15] Forrest Briggs, Raviv Raich, and Xiaoli Z. Fern. Audio classification of bird species: a statistical manifold approach. In *Ninth IEEE International Conference on Data Mining*, 2009.

[16] Henrik Brumm. *Animal Communication and Noise*. Springer-Verlag Berlin Heidelberg, 2013.

[17] Yanping Chen, Adena Why, Gustavo Batista, Agenor Mafra-Neto, and Eamonn Keogh. Flying insect classification with inexpensive sensors. *Journal of Insect Behavior*, 27:657–677, 2014.

[18] Jinkui Cheng, Yuehua Sun, and Liqiang Ji. A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines. *Pattern Recognition*, 43:3846–3852, 2010.

[19] Jinkui Cheng, Bengui Xie, Congtian Lin, and Liqiang Ji. A comparative study in birds: call-type-independent species and individual recognition using four machine-learning methods and two acoustic features. *Bioacoustics*, 21:157–171, 2012.

[20] David Chesmore. Automated bioacoustic identification of species. *Anais da Academia Brasileira de Ciências*, 76:435–440, 2004.

[21] Patrick J. Clemins and Michael T. Johnson. Generalized perceptual linear prediction features for animal vocalization analysis. *The Journal of the Acoustical Society of America*, 120:527–534, 2006.

[22] Sarah Collins. Vocal fighting and flirting: the functions of birdsong. In *Nature's Music - The Science of Birdsong*, page 39–79. Elsevier Academic Press, 2004.

[23] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[24] Marco Crocco, Marco Cristani, Andrea Trucco, and Vittorio Murino. Audio surveillance: A systematic review. *ACM Computing Surveys*, 48:52:1–52:46, 2016.

[25] Daniel Cintra Cugler, Claudia Bauzer Medeiros, and Luís Felipe Toledo. Managing animal sounds - some challenges and research directions. In *Proceedings V eScience Workshop - XXXI Brazilian Computer Society Conference*, 2011.

[26] Steven B. Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[27] Zheng Fang, Zhang Guoliang, and Song Zhanjiang. Comparison of different implementations of mfcc. *Journal of Computer Science and Technology*, 16:582–589, 2001.

[28] Neville H. Fletcher. Animal bioacoustics. In *Springer Handbook of Acoustics*, pages 821–841. Springer New York, 2014.

[29] Pasquale Foggia, Nicolai Petkov, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. Reliable detection of audio events in highly noisy environments. *Pattern Recognition Letters*, 65:22–28, 2015.

[30] Karl-Heinz Frommolt and Klaus-Henry Tauchert. Applying bioacoustic methods for long-term monitoring of a nocturnal wetland bird. *Ecological Informatics*, 21:4–12, 2014.

[31] Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang. A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, 13:303–319, 2011.

[32] Sadaoki Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29:254–272, 1981.

[33] Monica Gagliano, Stefano Mancuso, and Daniel Robert. Towards understanding plant bioacoustics. *Trends in Plant Science*, 17:323–325, 2012.

[34] Monica Gagliano, Michael Renton, Nili Duvdevani, Matthew Timmins, and Stefano Mancuso. Out of sight but not out of mind: Alternative means of communication in plants. *PLOS ONE*, 7(5):e37382, 2012.

[35] Mayur R. Gamit and Kinnal Dhameliya. English digits recognition using MFCC, LPC and Pearson's Correlation. *International Journal of Emerging Technology and Advanced Engineering*, 5:364–367, 2015.

[36] Peter Grosche, Meinard Müller, and Joan Serrà. Audio content-based music retrieval. In *Multimodal Music Processing*, pages 157–174. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012.

[37] David K. Grunberg, Alyssa M. Batula, Erik M. Schmidt, and Youngmoo E. Kim. Synthetic emotions for humanoids: Perceptual effects of size and number of robot platforms. *International Journal of Synthetic Emotions*, 3:68–83, 2012.

[38] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11:10–18, 2009.

[39] Asmaa El Hannani, Dijana Petrovska-Delacrétaz, Benoît Fauve, Aurélien Mayoue, John Mason, Jean-François Bonastre, and Gérard Chollet. Text-independent speaker verification. In *Guide to Biometric Reference Systems and Performance Evaluation*, pages 167–211. Springer London, 2009.

[40] Jan Hauke and Tomasz Kossowski. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, 30:87–93, 2011.

[41] Stefanie Heinicke, Ammie K. Kalan, Oliver J. J. Wagner, Roger Mundry, Hanna Lukashevich, and Hjalmar S. Kühl. Assessing the performance of a semi-automated acoustic monitoring system for primates. *Methods in Ecology and Evolution*, 6:753–763, 2015.

[42] Hynek Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87:1738–1752, 1990.

[43] Tony Hey and Jessie Hey. e-science and its implications for the library community. *Library Hi Tech*, 24:515–528, 2006.

[44] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29:82–97, 2012.

[45] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Prentice Hall, 2001.

[46] Inshirah Idris and Md Sah Salam. Improved speech emotion classification from spectral coefficient optimization. In *Advances in Machine Learning and Signal Processing*, pages 247–257. Springer International Publishing, 2016.

[47] Peter Jančovič and Münevver Köküer. Automatic detection and recognition of tonal bird sounds in noisy environments. *EURASIP Journal on Advances in Signal Processing*, 2011:982936, 2011.

[48] Peter Jančovič, Masoud Zakeri, Münevver Köküer, and Martin Russell. Hmm-based modelling of individual syllables for bird species recognition from audio field recordings. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[49] Josef V. Psutka Josef Psutka, Luděk Müller. Comparison of MFCC and PLP parameterizations in the speaker independent continuous speech recognition task. In *INTERSPEECH*, 2001.

[50] Jörn Köhler, Martin Jansen, Ariel Rodríguez, Philippe J. R. Kok, Luís Felipe Toledo, Mike Emmrich, Frank Glaw, Célio F. B. Haddad, Mark-Oliver Rödel, and Miguel Vences. The use of bioacoustics in anuran taxonomy: theory, terminology, methods and recommendations for best practice. *Zootaxa*, 4251(1):1–124, 2017.

[51] Veton Z. Këpuska and Hussien A. Elharati. Robust speech recognition system using conventional and hybrid features of MFCC, LPCC, PLP, RASTA-PLP and Hidden Markov Model classifier in noisy conditions. *Journal of Computer and Communications*, 3:1–9, 2015.

[52] Chang-Hsing Lee, Sheng-Bin Hsu, Jau-Ling Shih, and Chih-Hsun Chou. Continuous Birdsong Recognition Using Gaussian Mixture Modeling of Image Shape Features. *IEEE Transactions on Multimedia*, 15:454–464, 2013.

[53] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2:1–19, 2006.

[54] Mingchun Liu and Chunru Wan. A study on content-based classification and retrieval of audio database. In *International Symposium on Database Engineering and Applications*, 2001.

[55] Beth Logan. Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval*, 2000.

[56] Peter Marler and Hans Slabbekoorn. *Nature's Music - The Science of Birdsong*. Elsevier Academic Press, 2004.

[57] Nicolas Mathevon and Thierry Aubin. Sound-Based Species-Specific Recognition in the Blackcap Sylvia atricapilla Shows High Tolerance to Signal Modifications. *Behaviour*, 138:511–524, 2001.

[58] Brenda McComb, Benjamin Zuckerberg, David Vesely, and Christopher Jordan. *Monitoring Animal Populations and Their Habitats: A Practitioner's Guide*. CRC Press, 2010.

[59] Daniel McEnnis, Cory Mckay, Ichiro Fujinaga, and Philippe Depalle. jaudio: A feature extraction library. In *International Conference on Music Information Retrieval*, 2005.

[60] Ian McLoughlin, Haomin Zhang, Zhipeng Xie, Yan Song, , and Wei Xiao. Robust Sound Event Classification Using Deep Neural Networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23:540–552, 2015.

[61] Paul Mermelstein. Distance measures for speech recognition, psychological and instrumental. In *Pattern Recognition and Artificial Intelligence*, page 374–388. Academic Press, 1976.

[62] Alexander Mielke and Klaus Zuberbühler. A method for automated individual, species and call type recognition in free-ranging animals. *Animal Behaviour*, 86:475–482, 2013.

[63] Dalibor Mitrović, Matthias Zeppelzauer, and Christian Breiteneder. Features for content-based audio retrieval. *Advances in Computers*, 78:71–150, 2010.

[64] Dalibor Mitrovic, Matthias Zeppelzauer, and Christian Breiteneder. Discrimination and retrieval of animal sounds. In *12th International Multi-Media Modelling Conference Proceedings*, 2006.

[65] David Moffat, David Ronan, and Joshua D. Reiss. An evaluation of audio feature extraction toolboxes. In *Proceedings of the 18th International Conference on Digital Audio Effects (DAFx-15)*, 2015.

[66] Juan J. Noda, Carlos M. Travieso, and David Sánchez-Rodríguez. Methodology for automatic bioacoustic classification of anurans based on feature fusion. *Expert Systems With Applications*, 50:100–106, 2016.

[67] Douglas O'Shaughnessy. Linear predictive coding. *IEEE Potentials*, 7:29–32, 1988.

[68] Federica Pace. *Automated classification of humpback whale (Megaptera novaeangliae) songs using Hidden Markov Models.* PhD thesis, University of Southampton, Engineering and the Environment, 2013.

[69] Stephanie Pancoast and Murat Akbacak. Softening quantization in bag-of-audio-words. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[70] Dana Pfefferle and Julia Fischer. Sounds and size: identification of acoustic variables that reflect body size in hamadryas baboons, Papio hamadryas. *Animal Behaviour*, 72:43–51, 2006.

[71] Bartek Plichta. Best practices in the acquisition, processing, and analysis of acoustic speech signals. *University of Pennsylvania Working Papers in Linguistics*, 8:Article 16, 2002.

[72] Ilyas Potamitis, Stavros Ntalampiras, Olaf Jahn, and Klaus Riede. Automatic bird sound detection in long real-field recordings: Applications and tools. *Applied Acoustics*, 80:1–9, 2014.

[73] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in Pascal: The Art of Scientific Computing.* Cambridge University Press, 1989.

[74] Lawrence Rabiner and Biing Hwang Juang. *Fundamentals of Speech Recognition.* Prentice-Hall, Inc, 1993.

[75] Lawrence R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989.

[76] Justin Salamon, Juan Pablo Bello, Andrew Farnsworth, and Steve Kelling. Fusing shallow and deep learning for bioacoustic bird species classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[77] Alan P. Schmidt and Trevor K. M. Stone. Music classification and identification system. *Technical report, Department of Computer Science, University of Colorado, Boulder*, 2002.

[78] Peter-Christian Schön, Birger Puppe, and Gerhard Manteuffel. Linear prediction coding analysis and self-organizing feature map as tools to classify stress calls of domestic pigs (sus scrofa). *The Journal of the Acoustical Society of America*, 10:1425–1431, 2001.

[79] Michael G. Schöner, Ralph Simon, and Caroline R. Schöner. Acoustic communication in plant–animal interactions. *Current Opinion in Plant Biology*, 32:88–95, 2016.

[80] Björn Schuller. Audio features. In *Intelligent Audio Analysis*, pages 41–97. Springer Berlin Heidelberg, 2013.

[81] Roneel V. Sharan and Tom J. Moir. An overview of applications and advancements in automatic sound recognition. *Neurocomputing*, 200:22–34, 2016.

[82] Urmila Shrawankar and Vilas M. Thakare. Techniques for feature extraction in speech recognition system: A comparative study. *International Journal Of Computer Applications In Engineering, Technology and Sciences*, pages 412–418, 2013.

[83] Hans Slabbekoorn and Thomas B. Smith. Bird song, ecology and speciation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 357(1420):493–503, 2002.

[84] Dan Stowell and Mark D. Plumbley. Automatic large-scale classifcation of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, 2:e488, 2014.

[85] Jérôme Sueur and Almo Farina. Ecoacoustics: the ecological investigation and interpretation of environmental sound. *Biosemiotics*, 8:493–502, 2015.

[86] Leandro Tacioli, Luís Felipe Toledo, and Claudia Bauzer Medeiros. An architecture for animal sound identification based on multiple feature extraction and classification algorithms. In *11th BreSci - Brazilian e-Science Workshop.* Sociedade Brasileira de Computação (SBC), 2017.

[87] Hiroko Terasawa, Malcolm Slaney, and Jonathan Berger. Perceptual distance in timbre space. In *Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display*, 2005.

[88] Aaron M. Thode, Katherine H. Kim, Susanna B. Blackwell, Charles R. Greene Jr., Christopher S. Nations, Trent L. McDonald, and A. Michael Macrander. Automated detection and localization of bowhead whale sounds in the presence of seismic airgun surveys. *The Journal of the Acoustical Society of America*, 131:3726, 2012.

[89] Vibha Tiwari. Mfcc and its applications in speaker recognition. *International Journal on Emerging Technologies*, 1:19–22, 2010.

[90] Luís F. Toledo, Itamar A. Martins, Daniel P. Bruschi, Michel A. Passos, César Alexandre, and Célio F. B. Haddad. The anuran calling repertoire in the light of social context. *acta ethologica*, 18:87–99, 2014.

[91] Thiago M. Ventura, Allan G. de Oliveira, Todor D. Ganchev, Josiel M. de Figueiredo, Olaf Jahn, Marinez I. Marques, and Karl-L. Schuchmann. Audio parameterization with robust frame selection for improved bird identification. *Expert Systems with Applications*, 42:8463–8471, 2014.

[92] Jacques Marie Edme Vielliard and Maria Luisa da Silva. Bioacústica: Bases teóricas e regras práticas de uso em ornitologia. In *Ornitologia e Conservação: Ciência Aplicada, Técnicas de Pesquisa e Levantamento*, pages 313–326. Technical Books, 2010.

[93] Jason Wimmer, Michael Towsey, Birgit Planitz, Ian Williamson, and Paul Roe. Analysing environmental acoustic data through collaboration and automation. *Future Generation Computer Systems*, 29:560–568, 2013.

[94] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2015.

[95] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaton. Content-based classification, search, and retrieval of audio. *IEEE MultiMedia*, 3:27–36, 1996.

[96] Peter H. Wrege, Elizabeth D. Rowland, Sara Keen, and Yu Shiu. Acoustic monitoring for conservation in tropical forests: examples from forest elephants. *Methods in Ecology and Evolution*, 2017.

[97] Jie Xie, Michael Towsey, Jinglan Zhang, and Paul Roe. Acoustic classification of Australian frogs based on enhanced features and machine learning algorithms. *Applied Acoustics*, 113:193–201, 2016.

[98] Steve Young, Gunnar Evermann, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. *The HTK Book (for HTK version 3.4).* Cambridge University Engineering Department, 2006.

[99] Clifford Loh Ting Yuan and Dzati Athiar Ramli. Frog sound identification system for frog species recognition. In *Context-Aware Systems and Applications*, pages 41–50. Springer Berlin Heidelberg, 2013.

[100] Donglai Zhu and K. K. Paliwal. Product of power spectrum and group delay function for speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004.

[101] Sue Anne Zollinger, Jeffrey Podos, Erwin Nemeth, Franz Goller, and Henrik Brumm. On the relationship between, and measurement of, amplitude and frequency in birdsong. *Animal Behaviour*, 84:e1–e9, 2012.

[102] Youssef Zouhir and Kaïs Ouni. A bio-inspired feature extraction for robust speech recognition. *SpringerPlus*, 3:651, 2014.