UNIVERSITAT POMPEU FABRA

MASTER THESIS

Retrieval of Drum Samples by High-Level Descriptors

Author:
Javier ARREDONDO GARRIDO

Supervisor:
Dr. Frederic FONT CORBERA

A master thesis submitted in fulfillment of the requirements for the MSc of Sound and Music Computing

in the

Music Technology Group
Department of Information and Communication Technologies

September 3, 2017

Universitat Pompeu Fabra

Abstract

Universitat Pompeu Fabra

Department of Information and Communication Technologies

Sound and Music Computing

Retrieval of Drum Samples by High-Level Descriptors

by Javier ARREDONDO GARRIDO

Systems that manage audio databases for sound description and retrieval could be really useful within a context of music production. In this thesis, one of these systems has been created using classification techniques, by the application of learning algorithms and audio feature selections on two datasets. One of them is composed by commercial sounds and the other one is formed with sounds from an online repository of Creative Commons audio content. Due to the fact that drums are a fundamental element on most of the musical genres nowadays, it is the family of instruments chosen to train and test the presented classification models. Research is focused on finding generalist drum instrument class and category models, understanding category as the source nature of the sample (acoustic or digital). Generalization on these models lead us to be able to classify different drum sound datasets, achieving good model and prediction accuracies. Combining these models with an annotation of our datasets with specific values of audio high-level descriptors (Brightness, Hardness, Roughness and Depth), a drum samples retrieval tool could be created and would open new possibilities for database management within a music production framework.

Acknowledgements

First of all, I would like to thank my parents, who have supported me to study what I am passionate about, for the incredible education I have received from them. Secondly, to my two brothers for always guiding me along the hard way of life. Without their support I could not come to Barcelona.

I also want to express my gratitude to the entire group of people who form the Music Technology Group and most of the students of MSc Sound and Music Computing, since I have learned a lot during these two years thanks to all of them. Both researchers and classmates have opened my mind to a new world, technologically and musically. Especially thanks to my supervisor, Frederic Font, for helping me whenever I asked for it.

And last but not least, thanks to Daniela for listening to me every day talking about music technology and trying to understand what I was talking about. But above all, for always making me smile.

Contents

List of Figures

Chapter 1

Motivation

1.1 Introduction

This master thesis needs to be understood within the context of Sound and Music Computing. According to the definition provided by the Sound and Music Computing (SMC) research community[1], this is a "research field that studies sound and music communication chain by combining several methodologies, such as scientific, technological and artistic, through computational approaches". Although several disciplines can be named within this research field, those which are relevant to this thesis are mainly Music Information Retrieval (MIR), Digital Music Libraries and Music Production.

The aim of the project is to improve the integration of audio database management systems in standard audio production tools, which is one of the specific challenges proposed by Serra et al. (2013) related to the SMC field. Feature content-based searches can be done based on automatic models that have been created by the application of MIR techniques. Such models can be used to create useful tools for music producers that allow them to search audio content on the basis of their sonic qualities.

From 1990s onwards, the concept of music producer has considerably changed due to technological advances. Since then, it is feasible to make music directly from a computer. This fact made possible the concept of home-studio, which basically consist on an affordable imitation of a music studio inside a computer: sample libraries, virtual instruments, software recreations of famous hardware or even novel applications of signal processing. Nowadays, a Digital Audio Workstation (DAW) allows you to record, compose, edit, arrange, post-produce or whatever other process you need to build your tracks. People that make music using DAWs are commonly named as DAW-producers and they are precisely the intended users of the proposed system, which is outlined in Figure 1.1.

Sample organization and categorization is a common problem that often appears in an environment such as music production. As a producer's sample database may contains plenty of different musical or non-musical sounds, there is an obvious necessity of defining a concrete family of instruments to specify this research. *Drums* are the chosen option, due to previous successful researches about percussive instrument classification within the MIR field, and to the importance of this family in most current musical genres. In order to classify isolated drum sounds, a proper definition of a drum set is completely needed. It is important to clarify that drum

[1] Definition of SMC field: http://smcnetwork.org/roadmap/definition

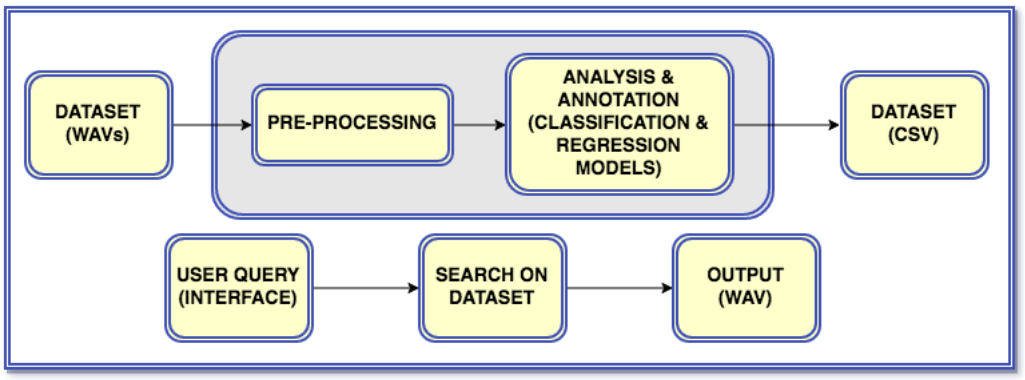

FIGURE 1.1: Global System Schema.

set and drum kit are different ways to call *Drums* depending on the origin of the speaker. From now on, in order to avoid misunderstandings, drum kit will be the term used along this document so as to distinguish this concept from the idea of *preset*, which can be found in many musical applications as a combination of sounds already setted by the manufacturer or the application developer.

As this thesis is intended to focus on a DAW-producer criteria, the drum kit concept has to change slightly from the typical drummer's point of view. From our perspective, it is not understood as a real physical instrument but as a combination of individual drum sounds inside a virtual instrument, commonly known as a drum machine. Here, drum patterns may be composed of isolated drum sounds that come from different sources or, even, that have been created synthetically. This fact has allowed producers to create new music genres by combining concrete elements of different styles.

1.2 Problem Statement

The problem we are facing is the creation of a system that automatically characterize isolated drum samples in terms of instrument class; instrument category, depending on the acoustic- or digital-nature of its source; and its semantic description, which depends on several high-level audio descriptors defined according to perceptual criteria. This characterization may be really useful within a sound retrieval framework, allowing users to speed up processes during music composition or production.

The importance of this research, besides the creation of a system that might help musicians or producers in their every day work-flow to quickly retrieval drum samples from a local or an on-line repository, is related to the improvement of sound description and database management. A breakthrough on the field of sound description was the proposal of a sound description standard (Peeters, McAdams, and Herrera, 2000) in the context of multimedia content description MPEG-7, which described some timbral features of sounds. However, there is still a lack of high-level information on databases and musical digital libraries due to the low level of specificity in any description standard.

MIR techniques have been lately applied to solve tasks such as automatic music transcription, music recommendation systems, tracks separation or even instrument recognition in polyphonic music. However, this research is focused on the particular case of monophonic files, which has its own specific challenges. Specifically, there is still room for progress in timbre-related studies. The originality of this work is based on the high level of specificity that the proposed system has for describing and characterizing isolated drum samples. These studies might be appropiate, not only for retrieving samples, but for improving audio content browsing or auto-tagging.

1.3 Goals

First of all, the idea of this thesis came up under the Audio Commons Initiative (ACI), whose purpose is to bring Creative Commons audio content into the creative industries (e.g. videogames, film and music industries). This initiative, as a supported research and innovation action, dedicates special attention on researching about intellectual property, audio ontologies and semantic description of audio content (Font and Serra, 2015).

In order to semantically describe audio content, state-of-the-art technologies for sound description are presented. Taking advantage of previous Audio Common's workpackages, the reliability of several high-level descriptors, defined to characterize isolated audio samples by their sonic qualities, is intended to be tested for the drum samples case. However, the main goal of this thesis in relation to sound description is the automatic classification of drum samples according to their instrument class and category, understanding category as the source nature of the sample: acoustic or digital.

Taxonomic classification of drum sounds is a topic that can be found in current literature, but still needs further revision. First of the two improvements presented along this thesis is the intention of model's generalization. One problem found in literature is that many models have been trained with an unique dataset, whose different classes tend to be composed by really similar sounds. Therefore, these models are not taking into account the huge diversity of drum samples that can be found either in commercial digital libraries or in online repositories of audio samples. Second improvement is the definition of two drum categories. In fact, drum category classification has been almost unnoticed in literature. In this thesis, one category model per each instrument is presented so as to study whether drum sounds can be automatically detected as acoustic or digital.

Apart from sound description, another goal related to the ACI is the creation of tools that make CC content easily available to incorporate it in the production workflows of the creative industries (in our specific case, music industry). Despite the amount of CC content available in online repositories, there are technical and practical issues that do not facilitate its usage. The unstructured nature of these platforms, where messy annotations tend to be the usual case, does not help for our purpose. Nevertheless, computational approaches, such as state-of-the-art classification or regression techniques, allow us to automatically characterize this CC content and make it quickly accessible for other applications or softwares.

The proposed system has been designed to be implemented into a Digital Audio Workstation further on. Ideally, as it can been seen in Figure 1.2, this system would

automatically classify a provided dataset, in terms of instrument and category, and annotate each sample with a numerical value of each of the defined high-level descriptors. After analyzing the provided dataset, it could be used to retrieve samples according to their category, instrumentation and sonic features.

1.4 Thesis Organization

This document is structured as follows. This first chapter is considered an introduction to the research done in the next chapters. In order to put the reader on context, Chapter 2 is the State-of-the-Art, where different approaches of the most relevant topics related to this thesis found in current literature are presented. In Chapter 3, those materials that have been used for dataset creation and annotation (in the case of high-level descriptors) are detailed. Chapters 4 and 5 could be considered as the methodology part of the master thesis, where different approaches have been taken into consideration and their corresponding results, in terms of model's accuracy, are exposed. Specifically, Chapter 4 is dedicated to the creation of a drum instrument model and Chapter 5, to the creation of category models per each instrument class. A holistic evaluation is made on Chapter 6, where a preliminary user evaluation has been done so as to test how good or bad the system performs, based on retrieved sounds. Finally, Chapter 7 is dedicated to conclusions and presentation of future work.

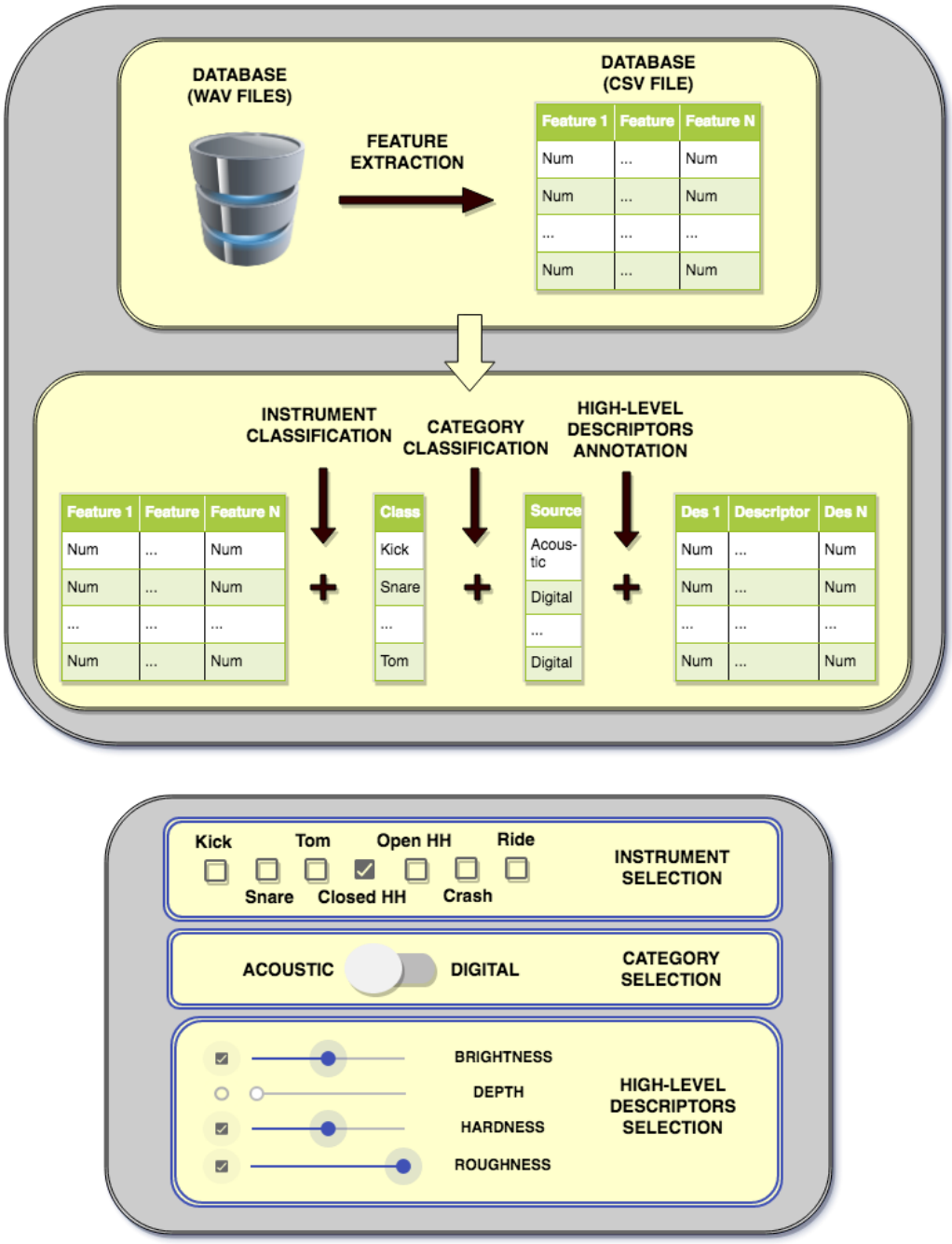

FIGURE 1.2: System: Analysis and Retrieval of Drum Samples.

Chapter 2

State of the Art

2.1 Introduction

Along this chapter, several research topics are analyzed: Sound Description and Retrieval, Automatic Classification of Percussive Instrument and Semantic Characterization of sounds. They will be pillars on which our methodology is based.

Instrument classification is an old but still challenging task within the MIR field. Most of the difficulty appears in those cases when the still poor definition of timbre is needed. In this thesis, after previous successful related studies made at MTG, the main part of our research will be focused on Percussive Instrument classification. A drum kit taxonomy definition accompanied by high-accuracy classification of individual isolated drum sounds from a particular kit (Stamellos, 2016) was already proposed and will serve as basement for our automatic classification approach.

On the other hand, Semantic Characterization has not been so common task on the field due to its complexity. Determining which descriptors are meaningful to explain a concrete sonic quality of a sound to some specific users, which may have different backgrounds, in a certain MIR application is not trivial at all. The well-known semantic gap of most MIR applications is the hole between the low-level features extracted from data and the high-level features that try to describe music or sounds from a human intelligence perspective; therefore, narrowing down this gap, by discovering meaningful relationships between low-level and high-level features, would be one of the most valuable contributions in MIR research. Related to the percussive instruments case, a Semantic Characterization of Drum Sounds was proposed in (Sá Pinto, 2015) and helped us to know the starting point of this project.

Sound description and retrieval is one of the most relevant applications of MIR since early days. In the following section, an overview of standards, EU-funded related-projects and different approaches to the topic are presented in order to understand the its importance and the context in which this thesis is undertaken.

2.2 Sound Description and Retrieval

As mentioned on the first chapter, this thesis' intention is to improve audio database management based on some MIR challenges proposed in (Serra et al., 2013), which include developing creative tools for commercial environments, producing descriptors based on musicological concepts and facilitating access to distributed data in digital libraries. In order to make these challenges possible and be able to boost the

audio database management, there is a fundamental initial topic to consider: Sound Description. Sound Retrieval would be a logical consequence of Sound Description. Once an homogeneous and meaningful description of sounds is made, a retrieval of specific audio samples, or even pieces of music, become a much easier and simpler task to be faced.

A world-wide standardization for multimedia content description, and probably the most relevant work in the sound description field, is MPEG-7. A complete overview can be seen in (Chang, Sikora, and Purl, 2001). This description standard use computational approaches that make possible a multimedia content characterization. Its purpose was precisely to allow fast and efficient multimedia retrieval. Focusing on the sound field, apart from format description (sampling rate or encoding format) and meta information (author name or copyright issues), MPEG-7 provide users the capacity to quickly access to low-level and some high-level features of an audio file. This low-level features are mathematical representations of temporal, spectral and timbral features of audio content that serve, among other applications, to extract semantic-related information of the audio file. Informations such as the type of musical event or the recorded instrument's name, which is also included as another descriptor on the meta-data information, are considered as high-level descriptor.

In MPEG-7's framework, directly linked with one of our main goals, a review of the proposal for instrument description was made in (Peeters, McAdams, and Herrera, 2000). In this well-known paper, which will be widely commented in further section, authors demonstrate that a good instrument classification model can be achieved for both sustained harmonic and non-sustained/percussive sounds families. High accuracy models were created by the usage of perceptual feature sets, which are actually included in MPEG-7 standard.

Following the spirit of the MPEG-7 standard, CUIDADO Project (Vinet, Herrera, and Pachet, 2002) aimed to develop new applications for accessing and manipulating audio content on music production and distribution industries, by systematically exploiting audio content descriptors. Both descriptors and descriptor schemes were designed for enabling content-based retrieval functions on large audio databases and for allowing users to manipulate audio content through high-level of specification.

Two applications were developed within this project: *Music Browser* and *Sound Palette*. The first one intended to be the first content-based music management tool for large music catalogs, implemented both descriptor-based and similarity-based searches by using rhythm, energy-based and timbre-based descriptors. The second one, Figure 2.1, offered audio sample management and feature editing based on sound content description. It combined management of a global sample repository; classification tools, which help the user to organize sounds by automatically providing an instrument class when introducing new sounds; and sample retrieval functions, combining classification criteria and query-by-example functions, based on similarity distances computed between each pair of samples across class boundaries.

After CUIDADO Project, another EU-funded project, called SIMAC, also shared same common goals with this thesis. SIMAC's main task was the development and usage of semantic descriptors for exploration, recommendation and retrieval purposes, by tagging audio and music with meaningful descriptors. Its objective was to find descriptors as similar as possible to the user's way of describing its content (Herrera et al., 2005). In relation to our work, one dimension of musical description

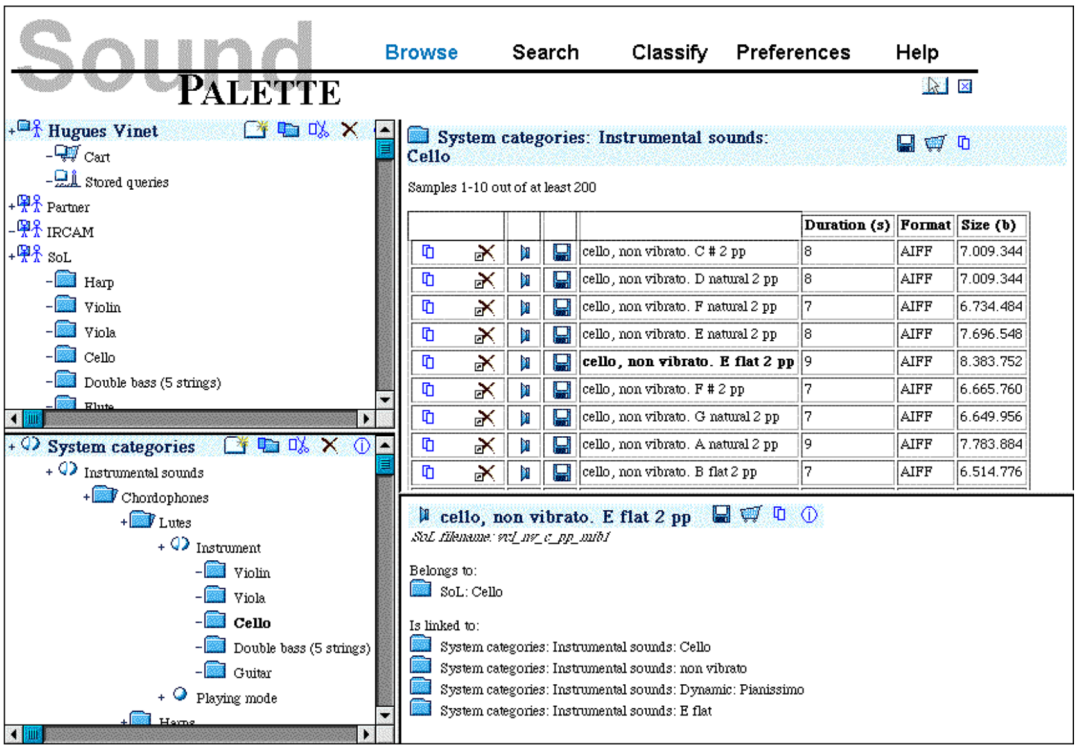

FIGURE 2.1: Screenshot of Sound Pallete from CUIDADO Project.

treated on this project was timbre, also known as instrumentation description. Although some percussion descriptors, which will be later explained, might be used for individual instrument description, most of the work was intended to retrieve information of whole pieces of music. Lead instrument recognition, solo detection or instrument profiling based on detection without performing any isolation or separation were the most relevant contributions.

Till that moment, huge advances were achieved on sound description for musical facets such as rhythm, melody, harmony, intensity/dynamics, segmentation, structure or even timbre, simplified as instrumentation. But a higher level of specificity on sound description, specially on timbre, is still needed. As it was explained in (Wiggins, 2009), the research made on sound and music information processing field for describing audio content lack semantically relevant information. Being able to describe sonic qualities of certain music or sounds, short audio samples in our case, can enable searching and browsing audio content in unprecedented and innovative ways for music production. In the framework of Audio Commons, as it is described in (Font et al., 2016), there is a large interest to improve the state-of-the-art in sound and music description and semantic representation technologies. Firstly, by researching on overlooked aspects in existing literature, such as the development of descriptors targeted to short music samples. And secondly, by developing reliable high-level semantic descriptors with the use of bigger and crowd-sourced datasets.

In the following subsections, an examination of different approaches to sound description and retrieval and an overview of sonic qualities described in actual literature, considered as high-level features, are detailed.

2.2.1 Perceptual Description and Taxonomic Classification

There are two main sound description and retrieval approaches, clearly explained in (Herrera-Boyer, Peeters, and Dubnov, 2003), on the context of automatic classification and characterization of musical instruments: perceptual description and taxonomic classification. Although they are computationally a bit different, they might be complementary in many cases. The former is used to develop perceptual similarity functions so as to use them for timbre clustering and its logical consequence, sound retrieval according to timbre similarity, which is also known as query-by-example. The latter is used to get indexes for labeling sounds after culture- or user-biased taxonomies. Perceptual description tries to find features that explain human perception of sounds, while taxonomic classification tries to assign to sounds some labels from a previously established taxonomy. Therefore, authors claimed that taxonomic classification approach might be considered deterministic, while perceptual description is derived from experimental results using human subjects.

Most of these perceptual experiments have applied Multidimensional Scaling (MDS) analysis to explore and visualize human judgments and typically have represented sounds in a two or three dimensional space, commonly named as "Timbre Space". Grey was the first to introduce the concept (Grey, 1977), which could be described as a representation of similarities, or dis-similarities, between sounds that are delimited by axes. These axes correspond to an acoustic or perceptual property that have an important role in the definition of a timbre sensation. After him, many authors have applied his timbre representation (McAdams et al., 1995; Terasawa, Slaney, and Berger, 2005); in the case of percussion studies, different authors (Lakatos, 2000; Herrera, Yeterian, and Gouyon, 2002; Brent, 2010) have also followed this representation. From their research, we extract that several features have been considered as a well representation of a perceptual space of percussive instruments: log-attack time, spectral centroid and temporal centroid.

These studies can be considered quantitative, because the description is based on those low-level features that best explain the dis-similarity judgment. Using multiple-regression techniques between feature values and sound positions in the timbre space, only the most correlated features are kept to determine the axes. This makes the perceptual description framework correspond to the best possible explanation of a sound, while the taxonomic classification framework can be seen as the best way to discriminate sounds between the considered classes. Another trend, which has not been too used lately, is the use of verbalizations to define axes of the timbre space, by numerically quantifying terms like "bright" or "attack". The main problem in such cases is that some of these terms might not be perceptually clear to users.

Following the perceptual description framework, some interesting papers have exploited the benefits of this approach for drum samples libraries organization and visualization. Pampalk, Hlavac, and Herrera (2004) presented a hierarchical user interface for efficient exploration and retrieval based on a computational model of similarity and self-organizing maps (SOM). In order to automatically create such an organization, authors adapted an auditory model and used clustering algorithms to create summaries of the collection and visualize the hierarchical structure. Although results seemed to be promising, artists that used the system proposed that a possible improvement would be to also classify samples according to semantic adjectives like "dark", "thick", "crispy"... Pampalk, Herrera, and Goto (2008) optimized and evaluated two computational model of similarity for sounds, specifically from the same

instrument classes: bass drums, snare drums and high and low toms. One model was based on auditory images and the other one, on the ISO/IEC MPEG-7 standard; resulting that auditory images perform clearly better than the MPEG-7 model. Turquois et al. (2016) proposed a design of an interface called *Drumspace*, as an alternative to the typical 1D scroll list for browsing drum samples (bass drums, snare drums and hi-hats), based on timbre similarity and arranged on a 2D plane. Proximity between each sample suggest their similarity and its position on the plane directly informs on some perceptual timbral qualities (not clearly defined). Although the short distance between the samples facilitates their playback, seemingly improving the user experience, it appears that the similarity-based arrangement was unexpectedly difficult to understand and to use efficiently.

2.2.2 Semantic Representation: High-level Descriptors

Besides previous approaches, there are several authors that have tried to represent sounds by the definition of semantic descriptors. Instead of trying to place a sound within a space, so as to make clusters based on timbre-related low-level features, or to describe a sound by an auditory model, they have tried to define high-level features of music and sound. As it is described in (Celma, Herrera, and Serra, 2006), in order to bridge the Music Semantic Gap and improve music retrieval effectiveness, the usage of automatically extracted semantic descriptors from audio files may represent an interesting and poorly explored alternative to the use of typical metadata. These semantic descriptors are generalizations that might be induced from the previously commented perceptual experiments, as derivations or combinations of lower-level descriptors, or from manually annotated databases, by means of the combination of signal processing, musical knowledge and the intensive application of machine learning.

Within the context of percussion instruments, four percussion-related semantic descriptors were defined (Herrera, Sandvold, and Gouyon, 2004) in the framework of SIMAC: *Percussion Index*, computed as the ratio between the number of percussive onset to the total number of onsets, which allow users to query for music with lot of percussion or not percussion at all; *Percussion Profile*, a refinement of the PI descriptor for describing the amount of percussion in certain classes; *Kick-Snare Crossings*, computed as the average number of changes from kick to snare or viceversa, which can tell us the complexity of a rhythm pattern; and *Percussivity*, the perceptual sensation of onset prominence. But as it was commented before, the usability of these descriptors are intended to work for music files, not for short audio samples, in order to later extract rhythmic or genre information. Only the *Percussivity* descriptor has correlation (Bell, 2015) with another semantic descriptor that have been used for short audio samples: *Hardness*.

Another interesting approach is the one taken by Bernardes, Davies, and Guedes (2015). In this paper, authors described a computational toolkit for real-time analysis of audio, not specifically for percussive instruments. This toolkit identifies sound objects and describes its attributes based on perceptual criteria of musical perception, which is grounded in sound-based theories. These musicological theories are Schaeffer's typo-morphology (Schaeffer, 1966), Smalley's spectromorphology (Smalley, 1997) and Thoresen's aural sonology (Thoresen and Hedman, 2007). Authors

claimed that the reason why they use a theoretical sound-based approach is the evident barrier for operating audio descriptors in creative music applications: the lack of meaningful labels adapted to application contexts and user preferences.

In the mentioned paper, its audio description schema is described by three perceptual sound criteria: mass, harmonic timbre and dynamic. Each of them is split into two categories: matter and form; matter is further divided into main and complementary. Those that are relevant to this thesis are: the main matter of the mass, called *Noisiness* and several main matters of the harmonic timbre, named as *Brightness*, *Width* and *Roughness (Sensory dissonance)*. *Noisiness* descriptor measures amount of noisy components in the signal as opposed to pitched components, ranged between zero and one, and it is computed by a combination of four low-level features: spectral flatness, tonalness, spectral kurtosis and spectral irregularity. *Brightness* is directly correlated with the spectral centroid and it is expressed in Hertz (Hz), limited to the audible range of human hearing (around 20 to 20000 Hz). *Width* characterizes the density, thickness or richness of the spectrum of a sound and it is measured also in Hz and computed, in a simplified version, using the low-level feature spectral spread. *Roughness* descriptor tries to perceptually regulate the pleasantness of a sound, it is described by the sensory dissonance and ranged between zero and one. Some of these semantic descriptors will be taken into account when designing our system.

In the framework of Audio Commons, Pearce, Brookes, and Mason (2016) have studied some timbral attributes that have potential value in relation to automatically-generated tags, which might be also relevant for this thesis purpose. First of all, a dictionary of timbral terms used to perceptually describe audio was compiled from academic literature and structured into a hierarchy. Secondly, a frequently used terms dictionary based on Freesound's searches was established to give an indication of their potential. As many attributes may refer to similar or opposite concepts, in order to reduce redundancy, the frequency-of-use for each term was summed into the hierarchical ordering[1]. Authors have claimed that these terms can be useful to retrieval samples by their perceptual sound qualities. Some of the most searched sound qualities are: *Hardness*, *Depth*, *Brightness* or *Metallic-nature*. According to these results, authors have implemented six perceptual models, using Python programming language, that can predict the four mentioned timbral attributes and also a couple more: *Roughness* and *Reverb* (Pearce, Brookes, and Mason, 2017). Although these studies have not dealt specifically with drum sounds, we will see further on that these descriptors may also worth attention for our purpose.

In strict relation with isolated drum sounds, the use of semantic adjectives has not been a common and standardized practice on MIR field, but current related literature will be reviewed in its particular section: Semantic Characterization of Drum Sounds.

2.2.3 Description and Retrieval on Commercial Applications

Apart from related academic publications, there are some commercial applications that deal with this concept of sound description in the context of music production.

[1] A plot of timbral attributes weighted by frequency of use and grouped by categories can be found here: audiocommons.org/assets/files/AC-WP5-SURREY-D5.1-Fig5.png

Mostly, audio database management systems are pre-built in some of the most famous DAWs, such as Logic Pro[2], Ableton Live[3] or Maschine[4]; or sometimes can be also found as audio plugins, which apart from sample browsing, sample editing is also possible. The former systems tend to allow users to browse directly through audio samples from commercial or user's libraries (Figures 2.2, 2.3 and 2.4) and the latter favor browsing through presets or kits. Drum kit concept was already presented in the first chapter of this thesis as a combination of individual drum sounds inside a virtual instrument (drum machine). The concept of drum preset refers to a pre-programmed setting on a musical instrument, which could be a synthesizer patch or a rhythm-pattern on a drum machine.

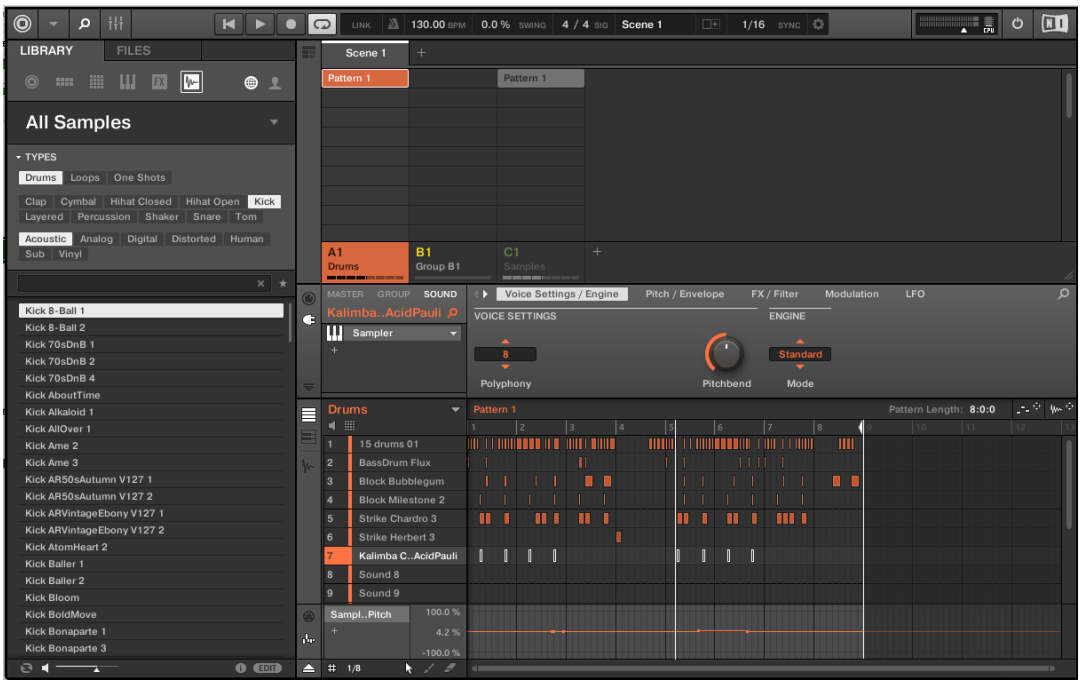

FIGURE 2.2: Maschine 2.0 Interface (Native Instruments).

In particular, Maschine software, Figure 2.2, has served as inspiration during the realization of this thesis due to their drum samples taxonomy and database. For the drum case, taxonomy is composed by clap, cymbal, hihat closed, hihat open, kick, percussion, shaker, snare, tom and layered samples. Each of these classes have several subclasses. In the cymbal case, options are acoustic, analog, china, crash, digital, human, ride, ride bell, splash and vinyl. Hihat closed case is composed by acoustic, analog, digital, human, pedal and vinyl. In the hihat open case, subclasses are acoustic, analog, digital, human and vinyl. In kick case, acoustic, analog, digital, distorted, human, sub and vinyl can be found. For snare sounds, options are acoustic, analog, brush, digital, human, rimshot, roll, side strick and vinyl. Finally, in the tom case, subclasses are acoustic, analog, digital, human and vinyl.

In Ableton Live software, Figure 2.3, drum samples management is similar but not so deeply structured, either there is no mention to different categories or subclasses. In the case of Apple's Logic Pro X, Figure 2.4, isolated drum samples are not considered

[2] apple.com/es/logic-pro/

[3] ableton.com/en/live/

[4] native-instruments.com/es/products/maschine/production-systems/maschine/

FIGURE 2.3: Live 9 Interface (Ableton).

but drum loops are well tagged by the usage of Apple's Loop format. In fact, a new tool called Drum Kit Designer[5] enables the user to select a whole drum kit that can be adjusted for each drum in terms of tuning, dampening and gain, apart from many other possibilities.

Some plugins that also manage drum samples databases, as well as facilitate the user tools for audio sample editing, are Kontakt and Battery from Native Instruments and Alchemy from Apple.

2.3 Automatic Classification of Percussive Instruments

Once sound description and retrieval topic has been analyzed, an overview of Automatic Classification of Percussive Instruments (mainly isolated drum sounds) needs to be presented along this section, in order to understand the methodology that needs to be followed for the creation of our system.

A task like this one always requires a well-defined taxonomy due, firstly, to the amount of different percussive instruments and, secondly, to the necessity of a balanced dataset to properly define and prepare experiments. Once the taxonomy is decided and the dataset prepared to be studied, audio feature extraction is the process of computing numerical representations that can be used to characterize a piece of audio.

Having extracted the desired audio features of the audio content from our dataset, feature selection is the next step on our model design so as to choose only the relevant features that enable us to solve the taxonomic classification task; taking into consideration that too many features may imply a high computational cost model. By the intensive application of machine learning algorithms over the dataset and the

[5] apple.com/logic-pro/plugins-and-sounds/

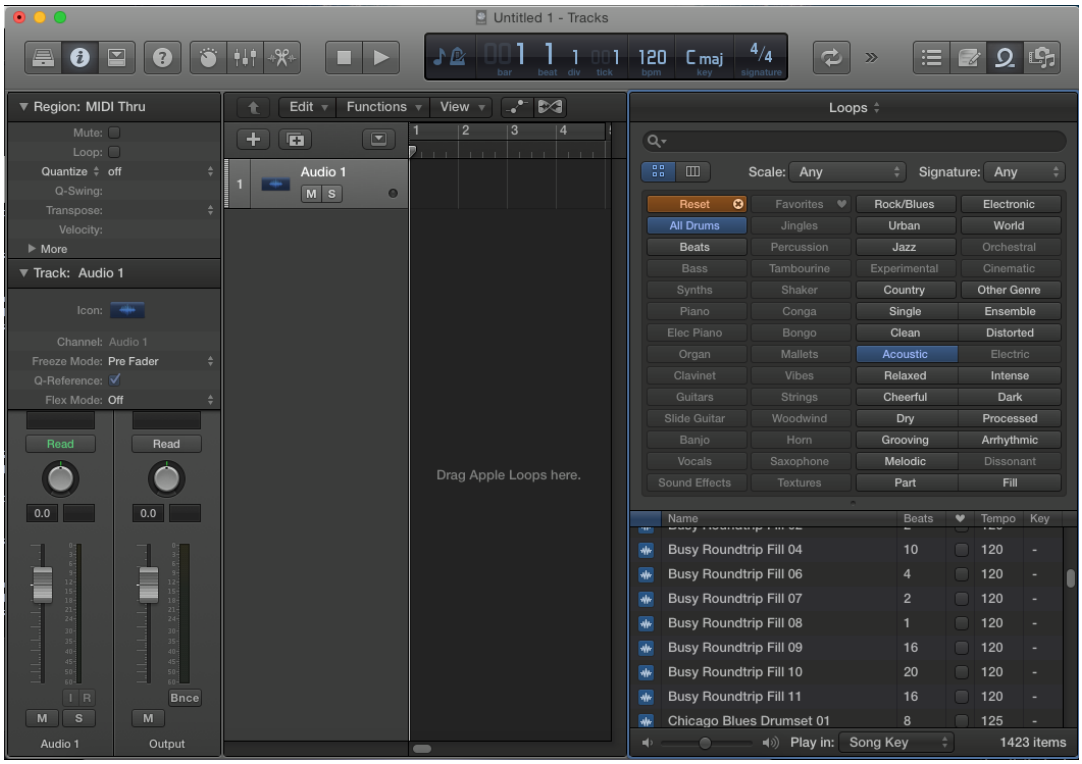

FIGURE 2.4: Logic Pro X Interface (Apple).

chosen features, a model that automatically assign labels to any drum sample input can be obtained.

2.3.1 Drum Kit Taxonomy

A broad taxonomy of individual drum sounds that defines a drum kit was introduced by Stamellos (2016), who proposed a drum kit standardization from a drummer point of view. Author made an exhaustive research on distinct taxonomies that appears on literature (Herrera, Yeterian, and Gouyon, 2002; Gouyon and Herrera, 2001; Shahar, 2010; Sillanpaa et al., 2000; Van Steelant et al., 2005), concluding that the one proposed by Herrera et al., which is composed of three level categories, was the most complete. Stamellos' approach, in fact, is based on Herrera's. In his case, a tree form taxonomy tries to standardize a modern drum kit.

Besides drum kits, other authors have approached the taxonomy definition problem by focusing in a certain family or group of percussive instruments or, even, different techniques of playing the same instrument. One example of the latter case is (Tindale, Kapur, and Fujinaga, 2004), where authors made a model that differentiate between snare's playing techniques, because the snare drum can be struck at different points along the radius of the batter head producing different timbres, due to the fact that different modes of the membrane are excited. Snares can also be struck with a wire brush or in such a way that the stick makes contact simultaneously with the snare head and the rim, producing a rimshot. A part from the snare case, the tabla case is also studied in (Tindale et al., 2005). In this paper, seven different stroke types

of snare, divided into three main categories: brush strokes, rimshots and standard strokes; and four tabla strokes are classified.

Most of the used datasets on literature are composed by kicks (or bass drum), snares (or snare drum) and some other instrument depending on the case, which are usually toms and hi-hats. There is an interesting and uncommon research made in (Souza, Batista, and Souza-Filho, 2015), where authors studied the cymbals case, achieving greats results when discriminating between china (edge an roll), crash (edge and roll), hi-hat (chick, closed and open), ride (bell, body and edge) and splash (choke and edge). Part of this approach is taken into consideration for the design of the model presented along this thesis.

2.3.2 Audio Features: Extraction and Selection

In case audio feature or audio descriptor is not a clear concept for the reader, see (Peeters, 2004; Peeters et al., 2011) for further information. There are several packages and libraries that allow us to extract audio features and that appear many times when studying current literature, such as LibROSA[6], Essentia[7] or MIR.EDU[8]. Essentia is the selected option for this master thesis. Developed at the MTG and utilized by the other two master thesis that firstly motivated this one, it is logical to maintain the same feature extraction library in order to compare results. Different audio analysis libraries do not usually use the same exact computation to obtain audio features and it results in slightly different numerical values.

Feature selection is the process of electing a subset of relevant features for model construction. It can be done manually, in those cases that groups of related features are compared between each other in order to understand which features are more useful for a certain task; or by applying feature selection techniques (De Silva and Leong, 2015), which are used to simplify models by the elimination of redundant and irrelevant features, denoting an easier interpretation, shorter training times and a reduced dimensionality.

A simple dimensionality reduction can be done by techniques such as principal component analysis (PCA), independent component analysis (ICA) or kernel principal component analysis (KPCA). But these techniques cannot handle the task we need to solve. When applying them, we reduce the number of dimensions by transforming our data into some abstract dimensions considered as the best representation of the given data. However, when applying feature selection techniques we can reduce dimensionality by throwing out irrelevant information and keeping real and relevant features. There are three main feature selection models: filter, wrapper and embedded. Filters are independent of the chosen learning algorithm and, therefore, purely data dependent. Wrappers optimize the feature set depending on the specific learning function, assessing the performance based on a cost function. Embedded models are built into the learning algorithm itself.

Filters are usually utilized to rank features, while wrapper and embedded models tend to be used for feature subset selection strategies. The two most used filter models are Correlation-based Feature Selection (CFS) and Information Gain. The former are also called similarity or dependency measures; in time-series prediction tasks on

[6] https://librosa.github.io/librosa/
[7] http://essentia.upf.edu/documentation/
[8] https://github.com/justinsalamon/miredu

stationary data, it can be used as a tool to help deciding if one time-series can be useful in predicting another. When a good individual feature is highly correlated with the target, the correlation measure can be used for ranking or subset selection. On the other hand, Information Gain evaluates the amount of information with respect to the classification target by measuring Shannon's information entropy. This metric may introduce unwanted bias in certain cases, so this has to be solved by introducing the intrinsic value of each attribute in the denominator, resulting in Information Gain Ratio measurement.

Directly related to automatic classification of drum sounds, (Herrera, Yeterian, and Gouyon, 2002) is a referent on the literature because of its comparison of feature selection methods and classification techniques. In the case of feature selection, one wrapper method (using Canonical Discriminant Analysis as wrapper for the selection) and two algorithm-independent methods were used for evaluating the relevance of the original feature set: CFS and ReliefF. The latter evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and for the nearest different class. CFS finally resulted as the best option, due to the higher value of the achieved accuracies. The original descriptor set was composed by attack-related descriptors, decay-related descriptors, relative energies for selected bands and Mel-Frequency Cepstral Coefficients (MFCC), means and variances.

Herrera et al. extended in (Herrera, Dehamel, and Gouyon, 2003) their previous research to unpitched percussion sounds, introducing several novel descriptors such as an extended version of the Bark scale critical bands and a logarithmic transformation of some features, in order to adapt them to a more Gaussian-like representation of the dataset. In this paper, authors studied global features instead of taking into account separately attack and decay parts of each instance. They also compared a selection of features made only by CFS against a combination between CFS and a wrapper, concluding that this combination of feature selection methods enables us to reduce the number of descriptors obtaining almost the same result.

Due to the necessity of reducing the number of features so as to be able to create a real-time system, in (Stamellos, 2016) the strategy was to apply Information Gain Ratio to rank the most a priori relevant features in each classification problem. Testing with different combination of high-ranked attributes and comparing performances was the approach to select a final feature set.

Another crucial consideration while feature selection and learning algorithm application is to avoid feature selection bias (De Silva and Leong, 2015). It is easy to over-fit a model by including too many degrees of freedom and it will lead to a poor generalization. If the same dataset is used for feature selection and the learning task, the model would be over-fitted and it won't work properly when new instances have to be classified. In order to manage effective feature selection and validation sets, 10-Cross validation is applied always when experimenting for model design.

2.3.3 Automatic Classification Techniques

During a learning task, many different Machine Learning algorithms can be used with distinct relevance depending on the task. Although many algorithms have been used for percussive instrument classification and drum sounds classification, achieving great results in terms of model's accuracy, it is hard to compare those

from most relevant related papers. Each one tries to classify different classes, to use various feature subsets or to apply distinct parameters of the algorithm or the feature selection method. So, inevitably, different learning algorithms should be studied for our drum classification tasks in order to find the proper one.

Within instrument classification framework in literature, Shahar (2010) as well as Tindale et al. (2004) compared K-Nearest Neighborhood (K-NN), Support Vector Machines (SVM) and Neural Networks (NN). Souza, Batista, and Souza-Filho (2015) achieved good results using SVM and Random Forest (RF). Herrera, Yeterian, and Gouyon (2002) analyzed K-NN, K*, C4.5 and Partial Decision Trees (PART). Stamellos (2016) decided to compare SVM, K-NN, RF, C4.5 and PART.

Only one paper has been found where the automatic classification of drum sounds was extended to more tasks than the purely instrument class labeling. Herrera, Dehamel, and Gouyon (2003) experimented, apart from the classification of unpitched percussion sounds, with the identification of drum manufacturers or models, which has served as a motivation for this thesis to study the acoustic- vs digital-nature of drum samples.

After reviewing literature and due to results on model's accuracies, K-NN and SVM seem to be the most reliable methods for our purpose. A brief explanation of them is presented to put the reader in context, but better explanations can be found in (Herrera-Boyer, Peeters, and Dubnov, 2003).

K-NN is a popular instance-based learning algorithm that stores the feature vectors of every training example and when it has to classify a new instance, finds a set of k nearest training examples in the feature space. Later, it assigns the new example to the class that has more examples in the set. This technique is not computing an abstract model, but storing every observation in memory and taking a decision based on the closer neighbor; because of that they are also known as lazy. The closeness is measured usually with Euclidean distance, but there are other versions where the algorithm measures the distance according to entropy (K*).

SVMs follow a statistical learning approach that tries to find the optimal linear hyperplane such that the expected classification error for unseen test samples has the lowest complexity, based on structural risk minimization inductive principle. Mapping from lineal map to high dimensional feature space is done by a kernel function and its procedure tend to be computationally intensive. As it was mentioned in (Herrera-Boyer, Peeters, and Dubnov, 2003), SVMs tend to perform better when classifying only two categories, one vs one classification approaches.

Finally, a conclusion for automatic classifications extracted by reviewing current literature is that each case needs its specific experimentation but good accuracies had always been obtained by an intensive application and comparison of the combination between different learning algorithms and the proper feature extraction and selection.

2.4 Semantic Characterization of Drum Sounds

The final section of our system is related to the semantic characterization of drum samples. As it was commented previously, the use of semantic adjectives to describe sonic qualities of percussive instruments has not been a standardized, neither

common, practice on MIR field. Published material related to this topic is quite insufficient.

Sá Pinto (2015) pointed directly to Bell as the main contributor due to his Percussive Audio Lexicon (PAL) (Bell, 2015), praising its potential for the promotion of an intelligent user-centered percussive classification system. Author of PAL assigns multiple descriptors to an isolated drum sound so as to provide a multifaceted overview of its features and qualities in a greater level of detail than any other attempt. Author claim that PAL can be used to learn about percussion, compare similarities and differences between sounds, and explore correlations between sound and source. Another application could be tagging percussion samples in sound libraries with a comprehensive set of semantic descriptors. Instrument morphology, excitation, spectral/temporal features, descriptive adjectives and onomatopoeias are different sections of the lexicon, which has a total of 2500 descriptors (encoded with alphanumeric IDs) organized hierarchically in a tree-branch interrelated structure that tries to put together Source, Sound and Subject into the same system.

PAL's intention was not to substitute current taxonomic classification models, as those presented on previous section, but to be complementary, either by integration or by operating in parallel with them. Descriptors like *Size* and *Material* from the Source tree, *Brightness* and *Pitch Strength* properties from Subject tree or *Reverberation*, *Spectral Width* and *Attack and Decay times* from Sound tree, could be some of the useful representations of drum samples' sonic qualities to be integrated within existing feature extraction algorithms for the purpose of automatically labeling sounds.

As it also happened in (Bernardes, Davies, and Guedes, 2015), Bell mentioned Schaeffer's work (1966) as an inspiration on the initial conception of his research, declaring it "the most significant formal classifications of sound in the 20th century". In fact, Bernardes, Davies, and Guedes (2015) used similar descriptors as some of those proposed in (Bell, 2015).

Another related research is the one done by Brent (2010), who tried to use verbalizations for timbre description within the percussive instrumentation domain; trying to fill the existent gap in the literature related to this sort of timbres. Brent's study confirmed the importance of spectral centroid and attack time duration as predictors of perceptual dimensions and defined two new exclusive dimensions for percussive timbres: *Dryness* and *Noisiness*, but without excessive success, specially in the latter case.

Sá Pinto (2015) summarized Bell's and Brent's work by choosing the most relevant and feasible concepts of their proposal. *Brightness*, *Hardness*, *Size*, *Ambiance* and *Tone* were the five semantic adjectives chosen for the semantic characterization. In order to find the possible relationships between the acoustic (low-level) descriptors and the semantic descriptors, author computed one lineal regression model per each of the mentioned descriptors, obtaining fairly good statistical results.

Although Pearce, Brookes, and Mason (2017) have not specifically studied timbre on the drum case, some of their proposed semantic descriptors are also in strict relation with some of the mentioned descriptors along this and previous sections. *Brightness* is the most common descriptor, not only on the percussive domain; *Hardness* is another usual percussive descriptor that has been related to the forcefulness of the attack; *Reverb* has been studied as *Dryness* or *Ambiance*; *Roughness* also as *Sensory*

Dissonance... Therefore, their approach will be explored and tested during the description of the presented system, taking advantage of its development within the Audio Commons framework.

2.5 Summary

Automatic classification of drum sounds is a topic that has been fairly studied, achieving good results on model's accuracies when discriminating between many classes. However, there is still room for progress on model's generalization. Accuracies tend to be excessively high and it may be caused by an overfitting on the model, due to the use of similar sound examples. In order to check this hypothesis, two different datasets are used for our methodology. Drum classification models found in literature also need a higher level of specificity, which is solved in this thesis by the study of the acoustic- or digital-nature of the drum samples.

More percussive instruments, more drum categories (like analog, composed by typical drum sounds from hardware drum machines) or drum manufacturers could be also taken into consideration when the goal is the creation of a system with a high level of specificity.

Developing and testing models that automatically detect specific sonic qualities of drum sounds has been also seen as a must for data management and retrieval application purposes. Due to the complexity of this topic, the are only a few papers where authors have studied this problem, hence there is no standardization at all. Pearce, Brookes, and Mason (2017) have presented some models based on linear regression that may be useful for our purpose and that need to be tested while developing the system.

Chapter 3

Materials

3.1 Introduction

Materials used for the shake of this thesis are: two different datasets, composed by audio drum samples; and several timbral models that try to describe numerically some sonic features, developed in the context of Audio Commons. One of these datasets is formed by samples found in commercial libraries and promotional sound packages or has been synthesized by my own. In the other one, sounds belong to an online repository of Creative Commons sounds.

While researching, both datasets have been studied and analyzed. Accuracy results related to classification models for both datasets are presented in following chapters. During preliminary user evaluation, the former dataset has been used as training dataset, while the later has been used as evaluation dataset. Ideally, the proposed system would be trained using the commercial dataset and the resulting classification model would be used to predict values from a drum samples dataset collected by the user.

During this chapter, these datasets and timbral models are presented; as well as the data preprocessing that make these sounds fit into our system. Another crucial fact commented along this chapter is the low-level features extraction. This processing is also made before the creation of our models, extracting audio features information from WAV files and expressing it as numerical values.

3.2 Drum kit Taxonomy

As found in literature and in commercial applications, a drum kit taxonomy is usually composed by kick, snare, tom, hi-hats and other cymbal sounds. Many others percussive instruments tend to be included in drum kits but they have been discarded due to the complexity of classifying too many classes.

3.2.1 Commercial Dataset

Four sources have been used to complete this dataset, as it is shown in Figure 3.1. The most relevant one is Maschine Native Instruments Library [1], contributing with 543 acoustic and 587 digital drum sounds. In order to analyze an equally sized

[1] native-instruments.com/en/products/maschine/maschine-expansions/

dataset, some classes needed to be completed. Due to this reason, 105 acoustic crash and ride sounds were downloaded from paiste.com/[2], an US cymbal store that enable costumers to download some free samples recorded with their products. 108 digital crash and ride sounds were synthesized using monophonic drum plug-ins[3] built for MASCHINE 2 software. Finally, the dataset was completed with 50 acoustic kick and snare sounds from (Stamellos, 2016) that were previously obtained from free online samples packages. Total dataset size is 1385 samples, around 200 per each of the 7 classes.

COMMERCIAL DATASET		Membrane			Plate				
		Kick	Snare	Tom	Closed HH	Open HH	Crash	Ride	TOTAL
NI Library	Acoustic	72	78	100	100	100	50	43	543
	Digital	100	100	90	100	98	65	34	587
Stamellos' Dataset	Acoustic	28	22	0	0	0	0	0	50
	Digital	0	0	0	0	0	0	0	0
Paiste.com	Acoustic	0	0	0	0	0	50	55	105
	Digital	0	0	0	0	0	0	0	0
Synthesis	Digital	0	0	10	0	0	35	63	108
	TOTAL	200	200	200	200	198	200	195	1393

FIGURE 3.1: Commercial Dataset.

As a big subset of the total dataset cannot be published due to copyright, a detailed description of the data is consider to be relevant. In Figure 3.1, the number of contributions per each Native Instrument Library (Expansion)[4] is presented. Anyway, drum samples names coming from these libraries, as well as their corresponding extracted low-level features, can be found in a CSV file at github.com/Javier-AG/SMC_thesis.

Number of samples per NI Library	Maschine Library	Queenbridge Lib	LoneForest Lib	VelvetLounge Lib	AstralFlutter Lib
	751	23	8	59	15
	MarbleRims Lib	LucidMission Lib	GoldenKingdom Lib	BlackArc Lib	NeonDrive Lib
	54	53	75	49	43

FIGURE 3.2: Number of samples per Native Instrument Libraries.

3.2.2 Free Dataset

This dataset was manually selected by querying instrument class names on freesound.org. Acoustic and digital sounds in this dataset are more easily differentiable due to the fact that they have been selected according to my own perceptual criteria and also taking into consideration related tags. In the previous case, drum samples were also manually selected, but they were already tagged as acoustic or digital according to NI taxonomy. As it is shown in Figure 3.3, total dataset size is 407 samples, around 60 per each of the 7 classes. Their freesound ID's can be found in the corresponding CSV file (name column) at github.com/Javier-AG/SMC_thesis.

[2] paiste.com/products/cymbals/sounds/discontinued/

[3] native-instruments.com/en/products/maschine/production-systems/maschine/sound-details/

[4] native-instruments.com/en/products/maschine/maschine-expansions/

FREE DATASET		Membrane			Plate				
		Kick	Snare	Tom	Closed HH	Open HH	Crash	Ride	TOTAL
Freesound.org	Acoustic	30	31	31	31	30	26	29	208
	Digital	30	31	29	30	26	28	25	199
	TOTAL	60	62	60	61	56	54	54	407

FIGURE 3.3: Free Dataset.

3.3 Preprocessing and Feature Extraction

In the first chapter of this thesis, a detailed schema of the proposed system was presented in Figure 1.2. First yellow block at the top of the diagram correspond to this section. Initially, audio files are slightly modified so as to adapt them for their features extraction.

In order to preserve the consistence of the experiments, every WAV file belonging to our dataset needs to be pre-processed before starting any process. Firstly, file's format has to be the same for each sample of our database. The selected format is the most common one, used in CDs: WAV file with a bit rate of 16 bits and a sampling rate of 44100 Hz. Samples from NI Library, as well as the synthesized ones, are already formatted that way, but samples from Stamellos' dataset had to be resampled from 22050 Hz to 44100 Hz. In the case of those samples downloaded from `paiste.com`, the original format was MP3. Secondly, as this thesis works directly with timbral features, the ideal situation for this study would be to have same pitch, loudness and duration in each sample. As we are working with unpitched percussive sounds, we only has to equalize loudness and duration of files. As the stretching of sounds tend to infer in timbral perception, the approach was to eliminate silences and try to select sounds of similar length inside each instrument class. Those sounds that needed to be shorten were some that came from NI Library and those downloaded from `paiste.com`.

Both processes are applied by SoX[5], which is a cross-platform command line utility that can transform formats of audio files and apply various effects to these sound files.

Once the data is already prepared, the feature extraction is made using an Essentia's out-of-box extractor, specifically the *streaming_extractor_freesound*[6], which is recommended for sound analysis and it is used by `freesound.org`, in order to provide sound analysis API and search by similar sounds functionality. It provides a wide selection of spectral, time-domain, rhythm and tonal features, as well as metadata information. As attributes computed over time are concerned for the timbre analysis, two distributions' statistical representations, arithmetic mean and variance of each attribute, are also included on the list of attributes.

The *streaming_extractor_freesound* mixes down all the audio files' channels to mono and applies an equal-loudness filter to the audio sample for the computation of those low-level features that needs to be loudness equalized in order to be computed properly. A subset of the list of features are discarded because they can be considered irrelevant for our purpose, as the rhythm and tonal features or the metadata information. Some features are also eliminated of the list when their value correspond to infinite or zero.

[5] `sox.sourceforge.net/`

[6] `essentia.upf.edu/documentation/extractors_out_of_box.html`

Finally, the extracted information per each audio sample is located at its corresponding JSON and YAML files. Then, information of every example of the different classes is ordered into a CSV file. During experimentation process of this thesis, two CSV files (corresponding to our two datasets) have been used on the application of learning algorithms, so as to create the classification models that are the basis of the proposed system.

3.4 High-Level Descriptors Annotation

Following the analysis schema of the system presented in Figure 1.2, current section correspond to the last summation element of the lower yellow block: High-Level Descriptors Annotation.

Last information added to our dataset is related to the semantic description of audio samples. No original research has been done during this section. On the contrary, the decision was to take advantage of a recent Audio Commons workpackage (Pearce, Brookes, and Mason, 2017), where authors have developed a Python implementation that perceptually describes an audio sample in terms of several timbral attributes. It consists on six perceptual models that try to predict hardness, depth, brightness, roughness, metallic-nature and reverb. The latter two were discarded due to the poor relation with our dataset, composed by drum sounds, and our purpose. These models are published in an online repository[7] and released under an open source license.

Each drum sample belonging to our dataset has been annotated by the application of the mentioned models. According to the documentation, their values should be between 0 and 100; but these models have been created using linear regression and sometimes a predicted value might go beyond the maximum value of 100 or even might not reach the minimum of 0. Due to this fact, post-processing is needed in order to normalize these values between 0 and 1.

However, some classes tend to have most of descriptors' values at the same region. Then, in order to avoid empty outputs when selecting a combination of values for high-level descriptors, these values are rearranged according to maximum and minimum ones found in each class of our dataset. Therefore, the proposed system is not representing exact measures of these timbral attributes. When selecting a position on the descriptor's slider, as it can be also seen in Figure 1.2, the intention is to let the user choose only among the available high-level descriptor values.

[7] github.com/AudioCommons/timbral_models

Chapter 4

Drum Instrument Classification

4.1 Introduction

Two different approaches have been undertaken to study this problem. In the first one, experimentation is made using the commercial dataset. The proposed system is thought to be trained with this dataset and to be used for instrument prediction of drum sounds coming from other datasets. For the shake of experimentation and system design, sounds from our free dataset are those that will be predicted.

After discovering that my model, and probably those models that can be found in current literature too, could be overfitted, another approach was undertaken. In this second case, both dataset are used for feature selection in order to check which features always appear as relevant, independently of the source from where we have obtained our drum sounds.

Once feature extraction has been applied, feature selection process is dedicated to select which of these features are going to be part of our models. Correlation-Based Feature Selection (CFS), Information Gain (Info Gain) and Information Gain Ratio (Ratio Gain) are the feature selection algorithms that have been utilized. Combining feature selection (both manually and computationally) and an intensive application of machine learning algorithms, we obtain some results that will motivate the final system's design.

WEKA[1] and Scikit-Learn[2] are the tools used to apply machine learning. WEKA is an open source software issued under the GNU General Public License, which contains tools for data pre-processing, classification, regression, clustering, association rules and visualization for data mining tasks. Scikit-Learn is a Python library which also contains simple and efficient tools for data analysis. The latter is used from 'Initial Model Testing' (subsection 4.2.4) to the end of the system description (Chapters 4 and 5).

In the following sections, the mentioned combinations of feature selection and learning algorithms and their corresponding performances are detailed and explained.

[1] http://www.cs.waikato.ac.nz/ml/weka/
[2] http://scikit-learn.org/

4.2 First Approach

4.2.1 Membranes vs Plates Experiment

Firstly, before any other experiment and as a first contact with the proposed problem, an initial task has been done. Due to the common categorization of drums between membranophones and idiophones, or membranes and plates in our dataset, creating a binary model to automatically label sounds based on these categories seems to be natural.

Surprisingly, using an unique low-level feature, selected by Info Gain and been applied a K-NN learning algorithm, a performance of 97.33% is obtained. Figure shows

This feature is the variance of **ERB Band 2**. In order to compute this descriptor, the algorithm applies a frequency domain filterbank to the signal using gammatone filters, so as to create an Equivalent Rectangular Bandwidth (ERB) scale (Moore and Glasberg, 1983), which gives an approximation to the bandwidths of the filters in human hearing. The magnitude of this descriptor in the second ERB band of the spectrum results to be fundamental to distinguish between these two classes. Kicks, snares and toms vary their values of the magnitude's variance of the second ERB band; while rides, crashes and hi-hats tend to have a zero value.

4.2.2 One vs All Experiments

As mentioned previously, the seven selected classes to make our model are: Kick, Snare, Tom, Open Hi-Hat, Closed Hi-Hat, Ride and Crash. A *one vs all* classification is studied per each class, resulting in a set of features that is composed by those features that achieve best performance on each experiment. An schema of this approach, as well as selected features are presented in Figure 4.1.

Importing the file with all the extracted features of each sample from our dataset into WEKA and applying machine learning algorithms, without making any previous selection of features, let us to a great model's accuracy of 91.19%, when applying SVM with polynomial kernel. However, using 278 features is not efficient at all and, even less, for a system that is intended to be used within a music production framework. Then, CFS was used to select a certain number of meaningful features. According to this feature selection algorithm, 34 features were relevant and an accuracy of 88.95% can be achieved, when applying SVM with puk kernel. Nevertheless, 34 features are still too many for a system like ours. Trying with Info Gain, instead of CFS, again provoke the use of too many features so as to achieve good performance results.

According to our referent (Stamellos, 2016), an efficient model to discriminate between drum instrument classes can be created by the use of just 9 features. That model was trained only with acoustic drum samples, so it may be natural if we finally need more features to distinguish between drums classes that are composed by acoustic and digital samples. The idea is to use the less features, the better; but we are still far away from an efficient result.

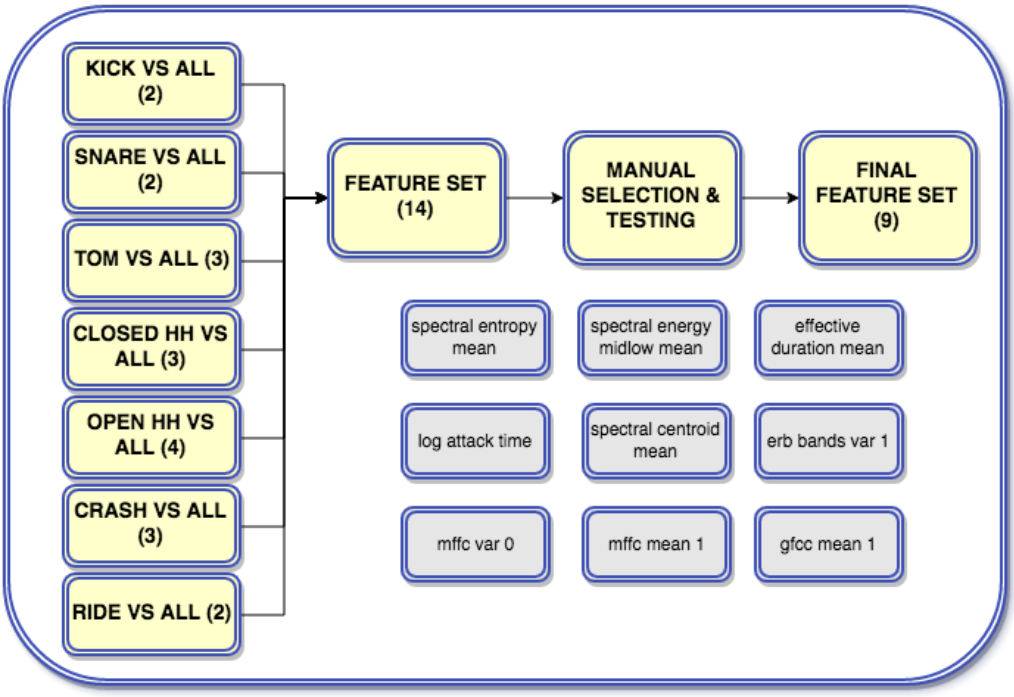

FIGURE 4.1: One vs All Schema.

Automatic classification problems tend to perform better when variables are simplified. So, next step is to make several *one vs all* binary-classification type of experiments. Taking this approach, each class was tested against an equilibrated combination of the other classes; i.e., *Crash vs All* experiment had 200 crash annotated instances vs 33 instances from each of the rest classes, annotated as non-crash. During these experiments, best results are obtained when applying K-NN algorithm instead of SVM.

Several features are shared on the different binary models, which may indicate that a low computational cost model could be achieved. An initial test with the 14 mentioned features was made, producing a model whose accuracy is around 83.54%, if it is trained with SVM and a puk kernel. This result is not bad but can be improved discarding some features that are not contributing on the performance. A final subset of 9 features performs quite good, achieving an accuracy of 84.11%, again by the application of SVM with puk kernel. These features are the following:

1. Mean of Spectral Entropy.
2. Mean of Spectral Energy of the Middle-Low Band.
3. Mean of Spectral Centroid.
4. Mean of Effective Duration.
5. Mean of Log-Attack Time.
6. Mean of GFCC Band 1.
7. Variance of ERB Band 1.

8. Variance of MFCC Band 0.
9. Mean of MFCC Band 1.

In order to evaluate how good this model performs, a new experiment was done. The result was a performance of 85.13% using Stamellos' model when applying SVM with puk kernel on our commercial dataset. Due to the fact that both models have same similar accuracies, another experiment was done taking advantage of the mentioned Stamellos' model so as to check if a higher accuracy can be achieved.

4.2.3 One vs One Experiments

Trying to improve the obtained results from the previous approach, the feature set used in (Stamellos, 2016) is studied for our specific dataset and boosted with new features that come from several *one vs one* classification experiments. This approach and the corresponding selected features are described in Figure 4.2.

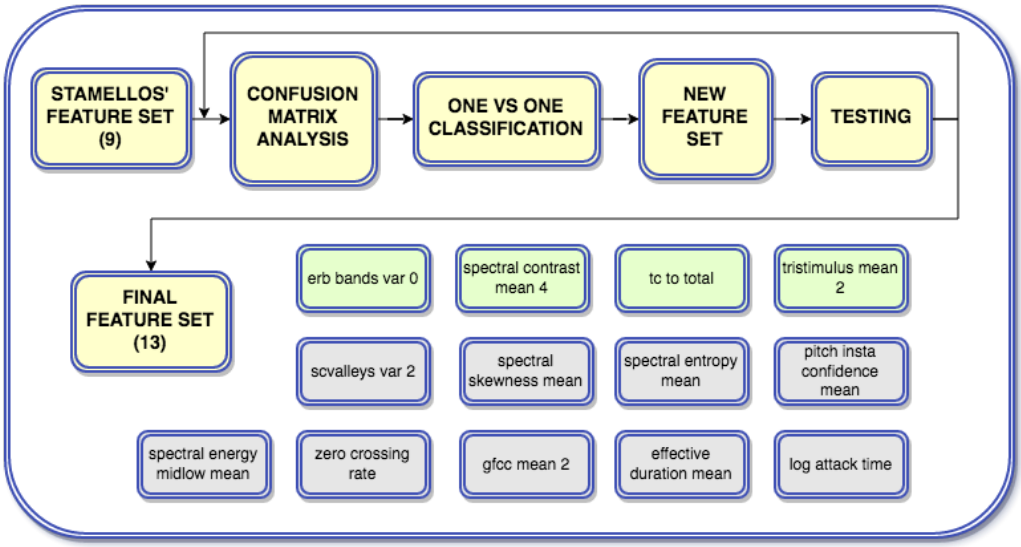

FIGURE 4.2: One vs One Schema.

Stamellos (2016) built a high accuracy model but, as commented before, the difference with our system resides in the fact that he only took into account acoustic and non-processed drum samples. Instead of re-inventing the wheel, taking advantage of his model could lead us to meaningful results. Firstly, the relevance of each feature was checked so as to know if the model's accuracy could improve without any of these features. As it was expected, the lack of any of them provokes a decrease on the performance. Therefore, better results could only be achieved by the increase of the amount of features.

Following this approach, we took a look at the confusion matrix obtained when applying Stamellos' selection of features and SVM, as learning algorithm, to our dataset. In a first confusion matrix evaluation, a necessity of improvement in the following cases are noticed, due to the amount of misclassification between classes: *Kick vs Tom*, *ClosedHH vs OpenHH* and *ClosedHH vs Snare*. As it happened on *one vs*

all binary classification, these *one vs one* binary classifications are better performed by the application of K-NN algorithm instead of SVM.

Being studied those cases that apparently caused more errors while classifying, four features were added to the previous set and tested on our dataset. Results are encouraging, obtaining 85.99% of correct classified instances. Moreover, discarding two of the four added features we still obtain the same accuracy. The two selected features that enhance the system are: **ERB Band 0 variance**, which improve the misclassification between kicks and toms, and **Spectral Contrast Band 4 mean**, which help us differentiate between closed and open hi-hats.

Despite this results, a better performance could still be achieved without adding too many features. There were other binary classifications such as *Crash vs OpenHH*, *Ride vs OpenHH*, *Crash vs Ride* or *Snare vs Tom* that could be revised in the search of improvements. These cases were studied, concluding in an addition of four features to the eleven-feature model proposed above.

Four features, from these that had been obtained through the *one vs one* approach, already belonged to our feature set and other two of them resulted in a deterioration of the performance when tested. Then, the four extra features chosen to join the previous set were: **Temporal Centroid to Total Length Ratio**, **Tristimulus 2 mean**, **Spectral Roll-Off mean** and **Bark Band 20 mean**. The obtained model's accuracy, when applying SVM with puk kernel and these fifteen features selection, is 85.70%.

Finally, the two latter were discarded. **Spectral Roll-Off mean**, due to its correlation with **Zero Crossing Rate mean**; and **Bark Band 20 mean**, because it was not really solving the misclassification problem. Therefore, a final set of 13 features is selected to from the model, achieving a result of 86.35% of correct classified instances. These features are the following:

1. Variance of Octave-based Spectral Contract of 3rd Valley.
2. Mean of Spectral Skewness.
3. Mean of Spectral Entropy.
4. Mean of Pitch Instantaneous Confidence.
5. Mean of Spectral Energy of the Middle-Low Band.
6. Mean of Zero Crossing Rate.
7. Mean of Effective Duration.
8. Mean of Log-Attack Time.
9. Mean of GFFC Band 2.
10. Variance of ERB Band 0.
11. Mean of Spectral Contrast 4.
12. Mean of Tristimulus 2.
13. Temporal Centroid to Total Length Ratio.

4.2.4 Initial Model Testing

In the previous two subsections, two models have been created to discriminate instrument classes among sounds of our commercial dataset. As it was mentioned in the summary of Chapter 2, it is necessary to test how these models perform when facing with a different dataset. Moreover, experiments are made using two machine learning softwares in order to check the stability of the models. Instead of applying SVM with puk kernel, due to the fact that Scickit-Learn does not include this type of kernel, SVM with linear kernel is used from now on. Therefore, in this subsection our two datasets are tested using Stamellos' model and our two created models from previous subsections by the application of SVM with linear kernel on WEKA and Scikit-Learn. In Figure 4.3, models and predictions accuracies are shown.

INSTRUMENT MODELS COMPARISON

MODEL		SVM linear kernel			
		All features (278)	Stamellos' Model (9)	Model 1 (9)	Model 2 (13)
NI Dataset	WEKA	91.33%	80.86%	79.20%	82.31%
	Scikit-learn	90.37%	79.62%	78.53%	81.42%
Freesound Dataset	WEKA	82.80%	72.72%	70.52%	72.97%
	Scikit-learn	79.62%	70.68%	66.30%	68.69%

PREDICTION		TRAIN	TEST	ACCURACY
SVM linear kernel	All features (278)	NI	Freesound	74.93%
		Freesound	NI	76.75%
	Stamellos' Model (9)	NI	Freesound	54.79%
		Freesound	NI	65.99%
	Model 1 (9)	NI	Freesound	69.77%
		Freesound	NI	69.45%
	Model 2 (13)	NI	Freesound	56.51%
		Freesound	NI	67.22%

FIGURE 4.3: Comparison of Models and Predictions Accuracies (Instrument).

Prediction process has been made using fit and predict functions of Scikit-Learn. According to these results, several facts need to be commented. First of all, accuracies on prediction process for Stamellos' model and the two created models seem to be unpromising when compared with the prediction accuracy of 'All features' model. These results are around 5-20% lower. This decrement means that these feature selections are not a real representation of the relevant data attributes. Apart from this fact, which could be considered as the main conclusion of this subsection, another important evidence is that the mentioned 'All features' model only achieve an accuracy of around 75%, which make us suppose that higher accuracies might be hard to be obtained.

It is important to mention that accuracies tend to be higher in the case of training with our free dataset and testing with our commercial dataset, only for Model 1 accuracies are similar. First reason could be the size of the studied dataset, bigger

on commercial dataset. When taking a look on model table in Figure 4.3, accuracies of our free dataset, in any feature selection case, are significantly lower. Models' accuracies decrease around 10% when using WEKA and bit more when using Scikit-Learn. A second reason could be that the used feature selections have overfitted the system.

Stamellos' model, whose accuracy when performing with his dataset is over 90%, performs around 80% and 70% of corrected classified instances for our commercial and our free dataset, respectively. A decrease of 10-20% could mean that this model was overfitted. An accuracy around 55% on the prediction process, when training with our commercial dataset and testing with our free dataset, and a prediction of 66%, when training and testing the other way around, express a poor generalization of the mentioned model.

In the case of Model 2, which is based on Stamellos' and composed by four more features that the other two studied models, accuracies appear to be quite similar to Stamellos' model. When creating the model, both Stamellos and Model 2 make us obtain better results than Model 1; but surprisingly these models perform a bit worse, when training with our commercial dataset and testing with our free dataset, and over 15% worse, when performing the other way around. Therefore, Model 2 should be also discarded as generalist model.

Finally, Model 1, created by the application of 'One vs All' classification approach, seems to be the more stable model of the studied ones. However, model and prediction accuracies corresponding to this feature selection are not super promising. Next section describes a new intention to create a model for discriminating drum instrument classes with a higher accuracy.

4.3 Definitive Model

Poor predictions accuracies obtained in previous section lead us to a new approach. In order to create a model as generalist as possible, our two datasets are taken into account when selecting which features are going to be part of our definitive model.

As it can bee seen in Figure 4.4, four feature selection have been studied. The two former have been selected while applying CFS and Ratio Gain algorithms in our free dataset. Feature Selection 2 is a reduced selection of Feature Selection 1, where Ratio Gain was applied after CFS. The two latter have been also selected while applying CFS, but on a mixed version of our two datasets. Feature Selection 4 is a reduced version of Feature Selection 3, which was manually selected by an intensive application of learning algorithms for different combination of features from Feature Selection 4. No exclusive feature selection based on our commercial dataset has been studied in this section because of its extensive study on the previous section.

In Figure 4.5, accuracies for each of the studied cases are shown. Worst performances are achieved again in the study of our free dataset, which has a significantly lower number instances. This fact confirms that the amount (and the variability) of sound examples in a dataset is a key factor when developing this kind of classification models.

Similar performances have been obtained during prediction process, as it can be seen in Figure 4.6. Only 'Feature Selection 3 model' shows a bit lower accuracy. Despite

INSTRUMENT MODEL

MODEL		SVM linear kernel					
		All features	FS 1 (44)	FS 2 (28)	FS 3 (37)	FS 4 (10)	*All features (278) vs Final (10) models*
Freesound Dataset	WEKA	82.80%	81.32%	83.29%	79.36%	75.67%	-7%
	Scikit-learn	79.62%	77.14%	77.83%	77.02%	73.48%	-6%
NI Dataset	WEKA	91.33%	89.24%	87%	88.01%	81.44%	-10%
	Scikit-learn	90.37%	87.78%	85.25%	87.13%	80.98%	-9%
Mixed Dataset	WEKA	89.84%	87.50%	84.65%	86.16%	80.80%	-9%
	Scikit-learn	88.20%	85.18%	82.39%	84.29%	79.42%	-9%

From freesound dataset	Feature Selection 1	CFS (44)
	Feature Selection 2	CFS + RatioGain (28)
From mixed dataset	Feature Selection 3	CFS (37)
	Feature Selection 4	CFS + Manual Selection (10)

FINAL FEATURES			
spectral contrast mean 1	spectral contrast mean 2	spectral contrast mean 3	
spectral contrast mean 4	spectral contrast mean 5	spectral contrast var 0	
spectral entropy mean	pitch instantaneous confidence	effective duration mean	log attack time mean

PREDICTION		TRAIN	TEST	ACCURACY
SVM linear kernel	All features	NI	Freesound	74.93%
		Freesound	NI	76.75%
	FS 1 (44)	NI	Freesound	72.48%
		Freesound	NI	77.25%
	FS 2 (28)	NI	Freesound	75.18%
		Freesound	NI	75.45%
	FS 3 (37)	NI	Freesound	65.60%
		Freesound	NI	73.35%
	FS 4 (10)	NI	Freesound	74.94%
		Freesound	NI	75.60%

All features (278) vs 10 features model	
NI -> FS	0.01%
FS -> NI	-1.15%

FIGURE 4.4: Definitive Instrument Model.

the fact that accuracy of 'Feature Selection 4' model (purple color in Figure 4.5) is always lower than the rest, its prediction results (Figure 4.6) are really promising. This model achieve same results than the 'All features' model, but selecting only 10 of 278 features, which is a significant feature reduction. This is the main reason why the final selected features are the ones belonging to this 'Feature Selection 4' model. Then, the proposed model for a drum instrument classification task is composed by SVM with linear kernal as learning algorithm and the following low-level features:

1. Mean of Spectral Entropy.
2. Mean of Pitch Instantaneous Confidence.
3. Mean of Effective Duration.
4. Mean of Log-Attack Time.
5. Mean of Spectral Contrast 1.
6. Mean of Spectral Contrast 2.
7. Mean of Spectral Contrast 3.
8. Mean of Spectral Contrast 4.
9. Mean of Spectral Contrast 5.

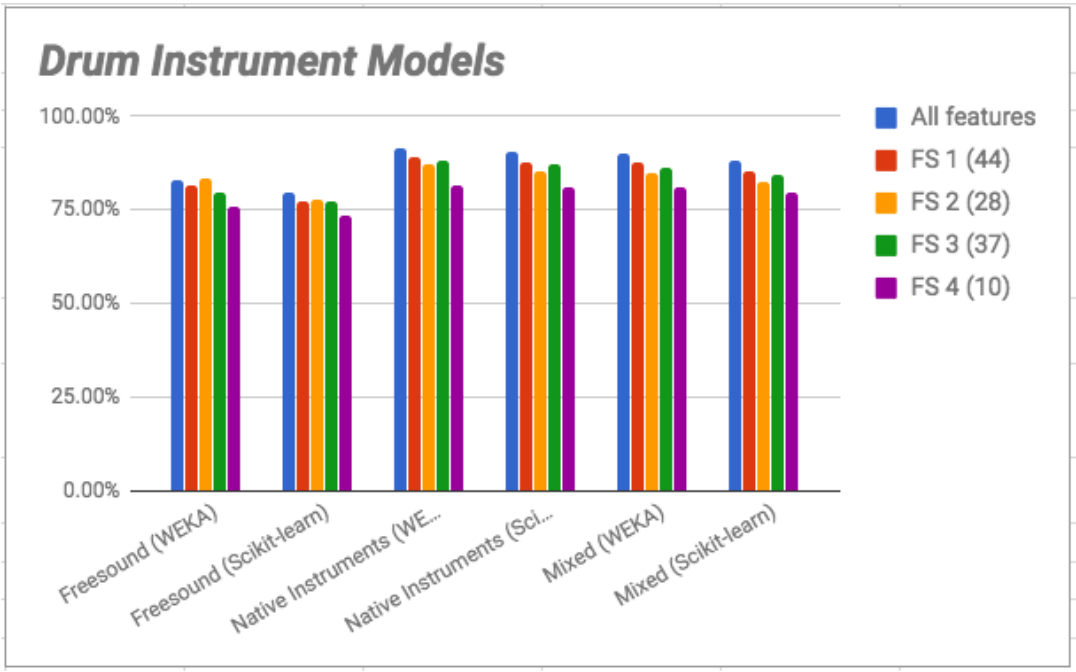

FIGURE 4.5: Accuracies for Different Feature Selections and Datasets.

10. Variance of Spectral Contrast 0.

An interesting finding is the appearance of each band of the **Spectral Contrast** descriptor (Akkermans, Serrà, and Herrera, 2009) as relevant features for a drum instrument classification task. Octave-based Spectral Contrast (OBSC) descriptor is defined as "the ratio between the magnitudes of the peaks and valleys within subbands of the frequency spectrum", measuring the relation of harmonic and non-harmonic frequency components of each sub-band (0, 200Hz, 400Hz, 800Hz, 1.6kHz, 3.2kHz, and 8kHz). It has been proven to be useful for genre classification.

Besides the relevance of OBSC descriptor, two more spectral low-level features has been found as important. **Spectral Entropy mean** is measured as the mean of Shannon entropy of the spectrum. It is usually used to quantify the peakiness of a distribution. **Pitch Instantaneous Confidence mean** is the confidence with which the pitch was detected, measured in a range [0,1]. If its values is zero, this means that the sound is unpitched. Although pitch is considered as a stable variable when studying timbre and apparently could be not taken into account, this feature is useful for discriminate plates from other drum sounds that could be considered as pseudo-pitched sounds or even pitched sounds in some cases.

TO DO: features distribution per each class.

Apart from the three mentioned spectral descriptors, there are two temporal descriptors that has been taken as fundamental for this task. **Log-Attack Time mean** and **Effective Duration mean** are great representations of the perceptual temporal evolution of drum sounds, as it has been proven several times. Log-Attack Time is calculated as the logarithm (base 10) of the attack time of a signal envelope; this attack time is defined as the time duration from when the sound becomes perceptually audible to when it reaches its maximum intensity. Effective Duration measures the time that the signal is perceptually meaningful, approximated by the time the

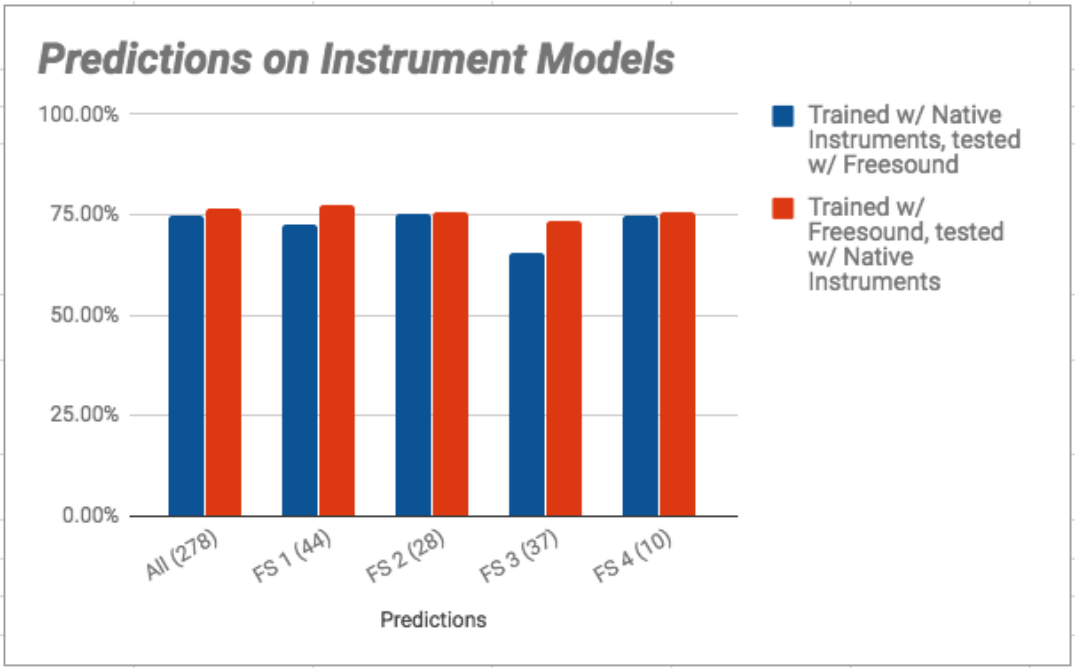

FIGURE 4.6: Predictions on Instrument Models.

envelope is above or equal to a given threshold and is above the -90db noise floor. Both features are sufficient to describe time envelope of our drum samples.

Comparing this final feature selection with our referent (Stamellos, 2016) feature selection lead us to another interesting finding. Spectral Entropy mean, Pitch Instantaneous Confidence mean, Log-Attack Time mean and Effective Duration mean are shared for both models. This fact confirms their relevance for drum instrument classification tasks. Moreover, Model 1 (*One vs One* classifications) from previous section also share Spectral Entropy mean, Log-Attack Time mean and Effective Duration mean as relevant features; which could be the reason why it was the model with better prediction accuracy of the previous section.

Comparing our final model with Model 1, the latter has a model accuracy and a prediction accuracy around 1-7% and 5-6% lower, respectively. Therefore, this final model is the ultimate and is the one implemented on the system.

Chapter 5

Drum Category Classification

Next step on our system description is dedicated to detect the acoustic- or digital-nature of a drum sample. As a first thought, one possible approach could be trying to classify 14 classes instead 7, splitting each class in two, resulting in both acoustic and digital classes of the same instrument. As it was commented before, experimenting with complex classification problems tend to produce low-accuracy models. In fact, results obtained when performing an experiment with 14 classes decrease the final model's accuracy to 71.62% for all features selection and SVM with polynomial kernal application, which means a reduction of 20% compared to the correct classified instances obtained when classifying 7 classes with the same dataset, same feature selection and same learning algorithm.

Discarding this complex classification problem, due to these unpromising results, two new binary classification approaches are studied. First of all, a category classification for the whole dataset in order to study if the source nature of drum samples can be identified without taken into consideration instrument information. And secondly, a category classification for each of the proposed classes, studying this source nature in each specific case.

5.1 First Approach

This first attempt tries to distinguish between acoustic and digital sounds in the whole dataset, without taking into account the corresponding instrument class. Following this approach, the selected features could be relevant to discriminate between acoustic- and digital-nature of any drum sample.

Three feature selections, as well as 'All features' selection (278 features), are used to create different models in order to find best performance. Feature Selection 3 (17) has been created by the application of CFS algorithm in a mixed version of our dataset. Same feature selection algorithm is used for Feature Selection 1 (10) and 2 (21) in our free and commercial datasets, respectively.

In Figure 5.1, a description of model and prediction accuracies can be seen. Prediction performances vary around 8%, reaching a maximum accuracy of 70.76% in the case of Feature Selection 2. This situation is not encouraging at all and it was not any surprise. A discrimination between acoustic and digital should be done using spectral features, those able to determine how the timbre of a concrete instrument is changing when an acoustic or digital sound is been studied. Considering different

CATEGORY GLOBAL MODEL

CATEGORY MODEL		SVM linear kernel			
		All features	FS 1 (10)	FS 2 (21)	FS 3 (17)
Freesound Dataset	WEKA	80.58%	74.93%	72.48%	68.79%
	Scikit-learn	72.87%	73.37%	70.88%	66.25%
NI Dataset	WEKA	76.24%	65.84%	73.57%	70.75%
	Scikit-learn	74.30%	64.33%	71.62%	69.52%
Mixed Dataset	WEKA	78.34%	67.57%	73.15%	70.14%

From freesound dataset	Feature Selection 1	CFS (10)
From NI dataset	Feature Selection 2	CFS (21)
From mixed dataset	Feature Selection 3	CFS (17)

PREDICTION	TRAIN	TEST	ACCURACY
All features	NI	Freesound	67.56%
	Freesound	NI	67.87%
FS 1	NI	Freesound	69.04%
	Freesound	NI	62.52%
FS 2	NI	Freesound	70.76%
	Freesound	NI	65.92%
FS 3	NI	Freesound	68.55%
	Freesound	NI	63.46%

FIGURE 5.1: Comparison of Models and Predictions Accuracies (Category).

instruments in a same dataset for a drum category task has been proved as a not valid approach.

5.2 Second Approach

Thinking about this problem from a perceptual point of view, a much more natural approach for humans seems to be the distinction between acoustic- or digital-nature of sounds from the same instrument class; i.e., when comparing an acoustic kick with a digital open hi-hat, humans can not usually contribute with much information about the source of the sample, but when comparing an acoustic snare with a digital one, we are commonly able to differentiate their nature.

During these experiments, feature selection has been made using Ratio Gain ranker algorithm and manual selection. There are two models, as well as the 'All features' model, for each class. Each of them corresponding to one of our two datasets.

5.2.1 Closed Hi-hat

In the closed hi-hat case, as it can be seen in Figure 5.2, the model that employs all features when performing on our free dataset achieve a high accuracy. On the contrary, when performing on our commercial dataset, performance is revealing that acoustic- or digital-nature of the studied sound is hard to predict. In Native Instrument taxonomy some of the sounds that are defined as acoustic or digital are perceptually hard to distinguish.

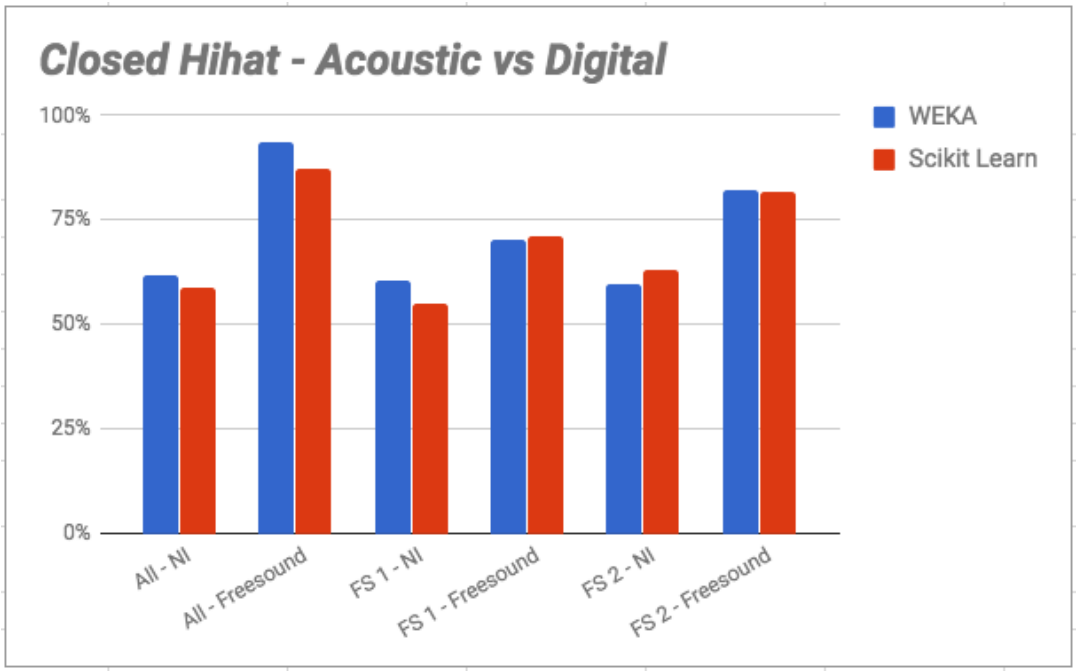

FIGURE 5.2: Feature Selection Comparison (Closed Hi-hat).

Predictions are shown in Figure 5.3. Models perform significantly better when testing with our free dataset, which confirms the previous assumption. Another interesting finding is that 'Feature Selection 2' model perform almost as well as 'All features' model. Because of this reason, the final feature selection is composed by the following low-level features:

1. Mean of Dissonance.
2. Mean of Barkbands Spread.
3. Mean of MFCC 1.
4. Variance of Tristimulus 0.

Dissonance mean is measured as the sensory dissonance of an audio signal given its spectral peaks, measuring perceptual roughness of the sound. **Barkbands Spread mean** is the variance of energy mean in Bark bands of all the spectrum. **MFCC 1 mean** is measured as the mean of Coefficient 1 of Mel-frequency Cepstrum Coefficients. **Tristimulus 0 var** is the variance of the first tristimulus, which measures the relative weight of the first harmonic. The former two descriptors could be considered as possible good representations of the 'acousticness' o 'digitalness' of a sound, at least for closed hi-hats.

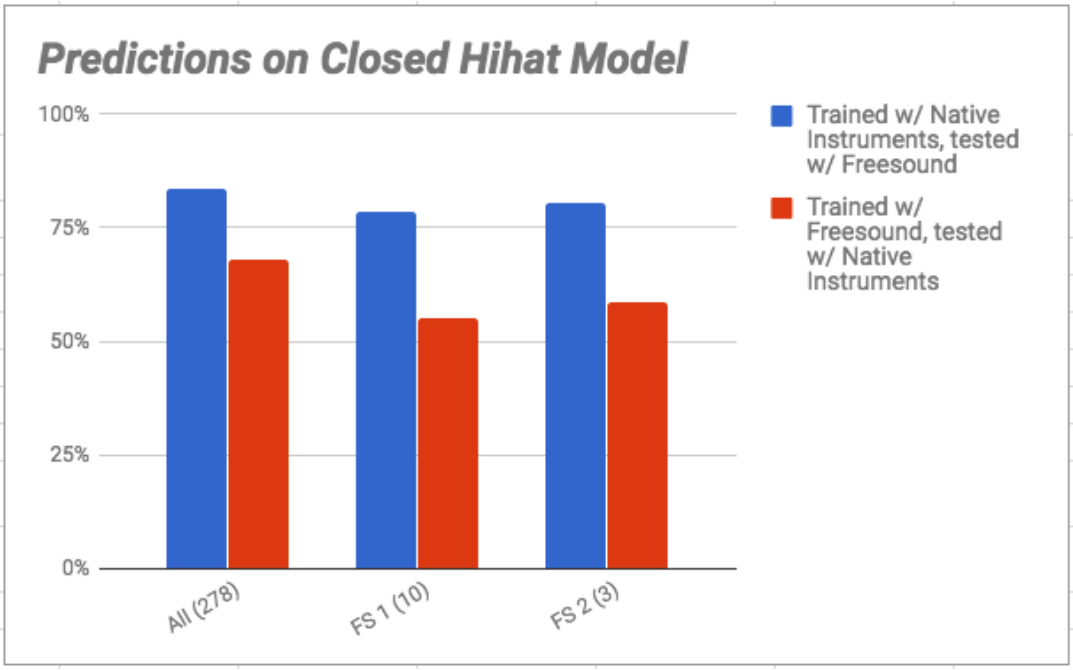

FIGURE 5.3: Predictions for Closed Hi-hat Models.

Closed Hi-hat model for category classification is described in Figure 5.16 at the end of this section.

5.2.2 Crash

In the crash case, as it can be seen in Figure 5.4, the model that employs all features when performing on our free dataset now achieve a poor accuracy. On the contrary, when performing on our commercial dataset, performance results promising. Predictions, Figure 5.5, show that using 'Feature Selection 1' accuracy reaches to 82% for our commercial dataset but decreases to 62% for our free dataset. This lack of generalization, observed in several experiments, is telling us that these created models might be overfitted to one of the two datasets. Nevertheless, FS1 is the final feature selection for crash model:

1. Mean of Spectral Entropy.
2. Mean of Spectral Spread.
3. Mean of Flatness.

Spectral Spread mean is similar to the previously commented Barkbands Spread, but taking into consideration the whole spectrum instead of splitting it into Bark bands. **Spectral Entropy mean** was explained in the previous Chapter as the mean of Shannon entropy of the spectrum that quantifies the peakiness of a distribution. **Flatness mean** is measured as the ratio between the geometric mean and the arithmetic mean of the spectrum.

Crash model for category classification is fully described in Figure 5.17 at the end of this section.

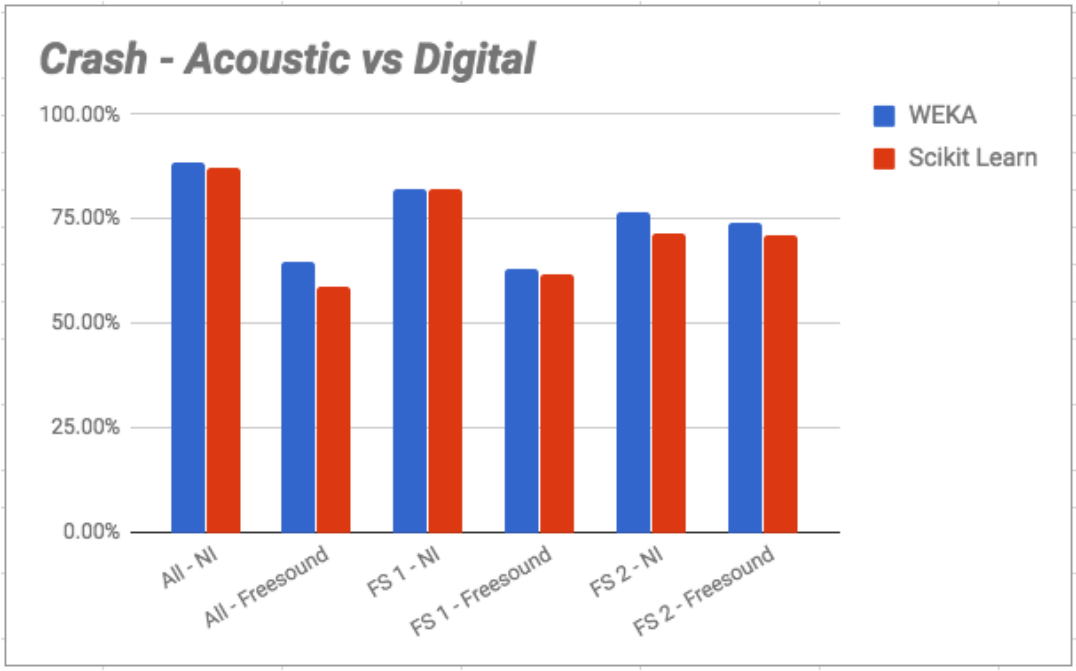

FIGURE 5.4: Feature Selection Comparison (Crash).

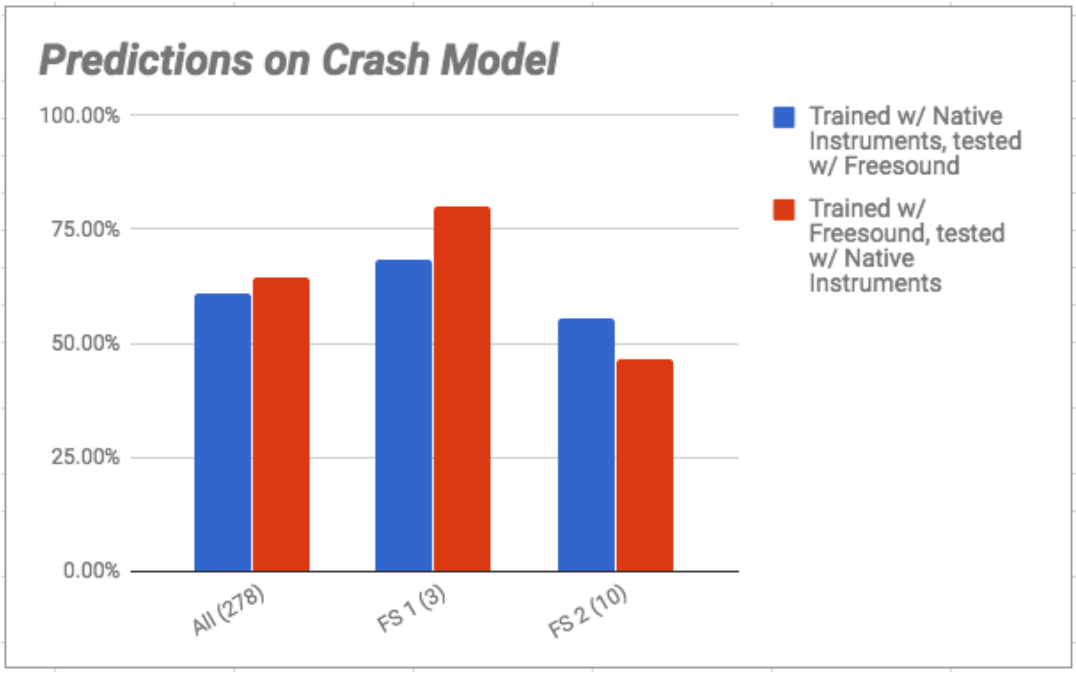

FIGURE 5.5: Predictions for Crash Models.

5.2.3 Kick

The kick case presents stable accuracies on the feature selection comparison, Figure 5.6. Best result is achieved for 'Feature Selection 1' model. Predictions, Figure 5.7, also shows promising results of 'Feature Selection 1' model, obtaining 75% and 79.5% of corrected predicted instances for training with commercial dataset and

testing with free dataset and vice versa. It is interesting to notice that models tend to perform better when testing with the same dataset as the one used for feature selection, which could induce into overffiting of the model. Nevertheless, if prediction accuracy when testing with the other dataset results in a similar value we can assume that this feature selection works fine for our purpose.

Selected features for Kick Model are the following:

1. Mean of Spectral Energy.
2. Variance of MFCC 4.
3. Variance of MFCC 7.

MFCC 4 and 7 are coefficients of Mel-frequency Cepstrum Coefficients. **Spectral Energy mean**, as its name indicates, is the mean of the computed energy of the whole spectrum.

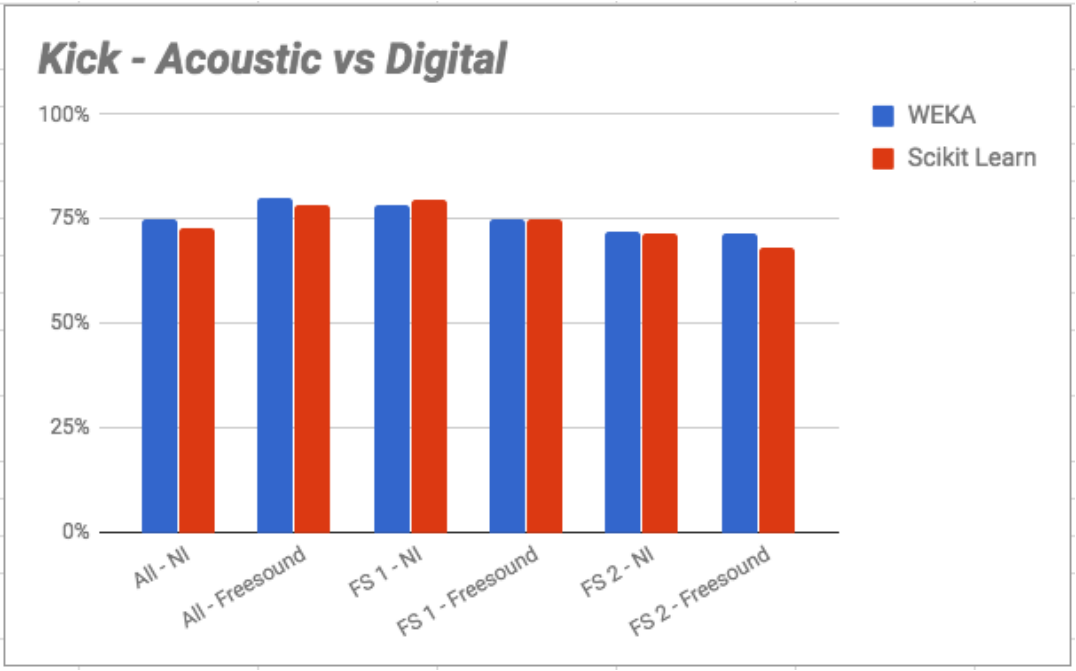

FIGURE 5.6: Feature Selection Comparison (Kick).

Kick model for category classification is fully described in Figure 5.18 at the end of this section.

5.2.4 Open Hi-hat

In the open hi-hat case, Figure 5.8, our free dataset seems to perform quite good for 'All features' model and even better when applying its feature selection, 'Feature Selection 2'. Our commercial dataset perform always over 70%, obtaining also promising results when modeling with the mentioned 'FS 2'. During the prediction process, Figure 5.9, testing with our free dataset achieve considerably greater results than doing it with our commercial dataset. Best results are achieved when using 'FS 2' model in both cases. Therefore, this is the final selection.

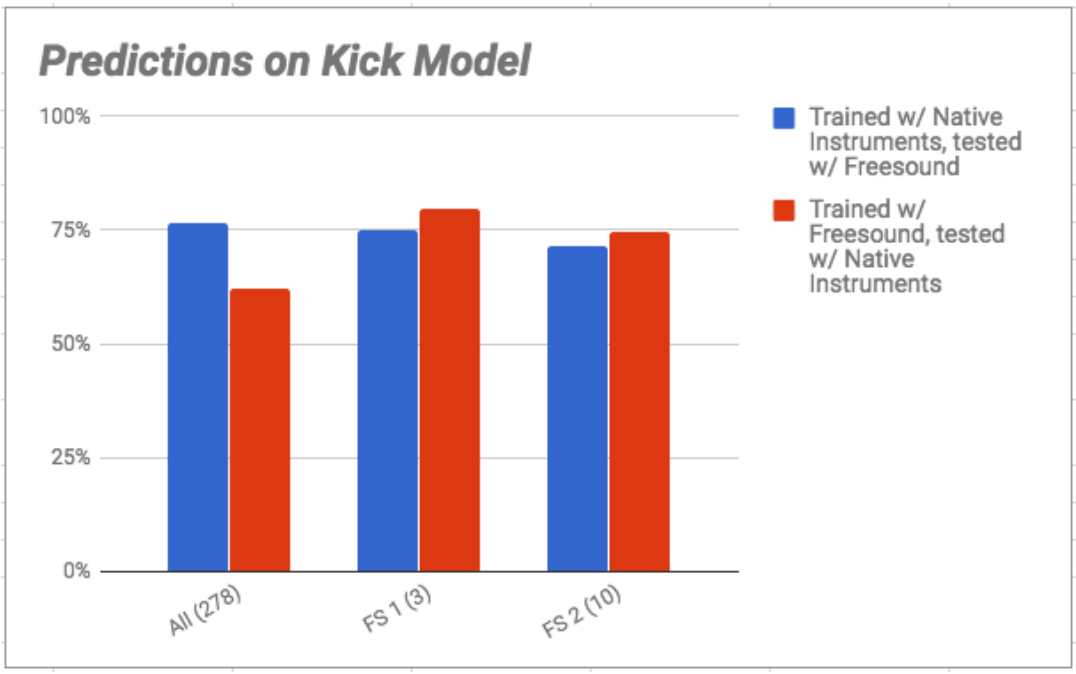

FIGURE 5.7: Predictions for Kick Models.

Therefore, selected features are the following:

1. Mean of Pitch.
2. Mean of Pitch Salience.
3. Mean of Spectral Strong Peak.
4. Mean of MFCC 1.
5. Mean of MFCC 7.
6. Variance of MFCC 9.
7. Mean of Tristimulus 2.

Pitch mean represents the mean of pitch over the signal. **Pitch Salience mean** is measured as the ratio of the highest auto correlation value of the spectrum to the non-shifted auto correlation value. It was designed as a quick measure of tone sensation. Unpitched sounds and pure tones have an average close to 0 whereas sounds containing several harmonics in the spectrum tend to have a higher value. **Spectral Strong Peak mean** is given by the ratio between the spectrum's maximum peak's magnitude and the "bandwidth" of the peak above a threshold (half its amplitude). This ratio reveals whether the spectrum presents a very "pronounced" maximum peak; which seems to be higher on the case of digital open hi-hat sounds. **MFCC 1 and 7 mean** and **MFCC 9 var** are coefficients of of Mel-frequency Cepstrum Coefficients. **Tristimulus 2 mean** is defined as the variance of the third tristimulus, which measures the relative weight of high harmonics (from the 5th harmonic).

Open Hi-hat model for category classification is described in Figure 5.19 at the end of this section.

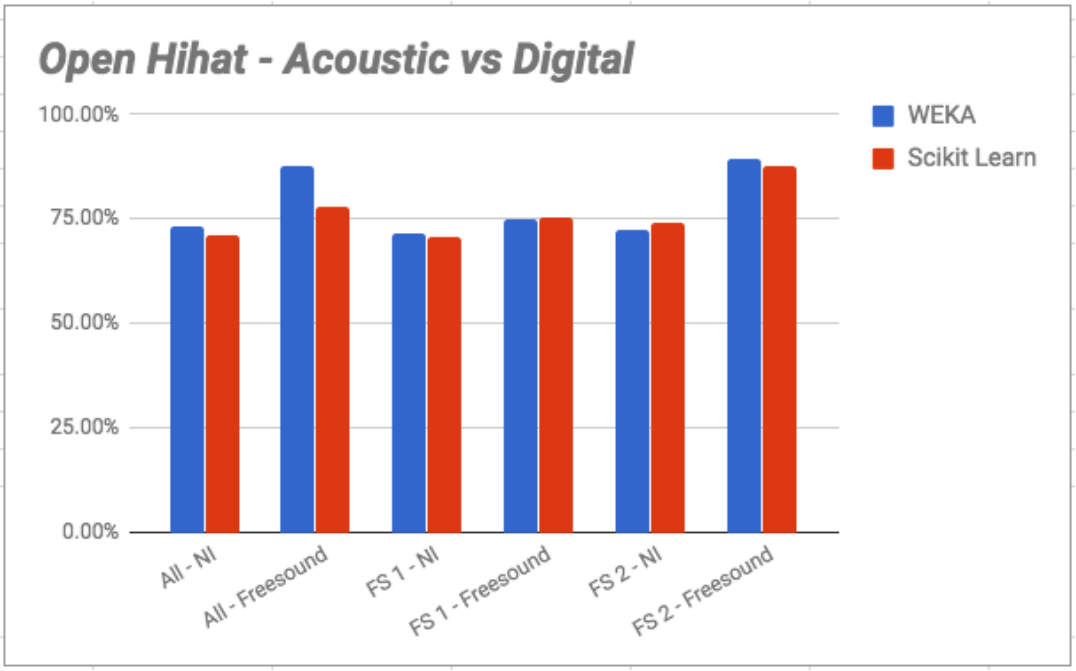

FIGURE 5.8: Feature Selection Comparison (Open Hi-hat).

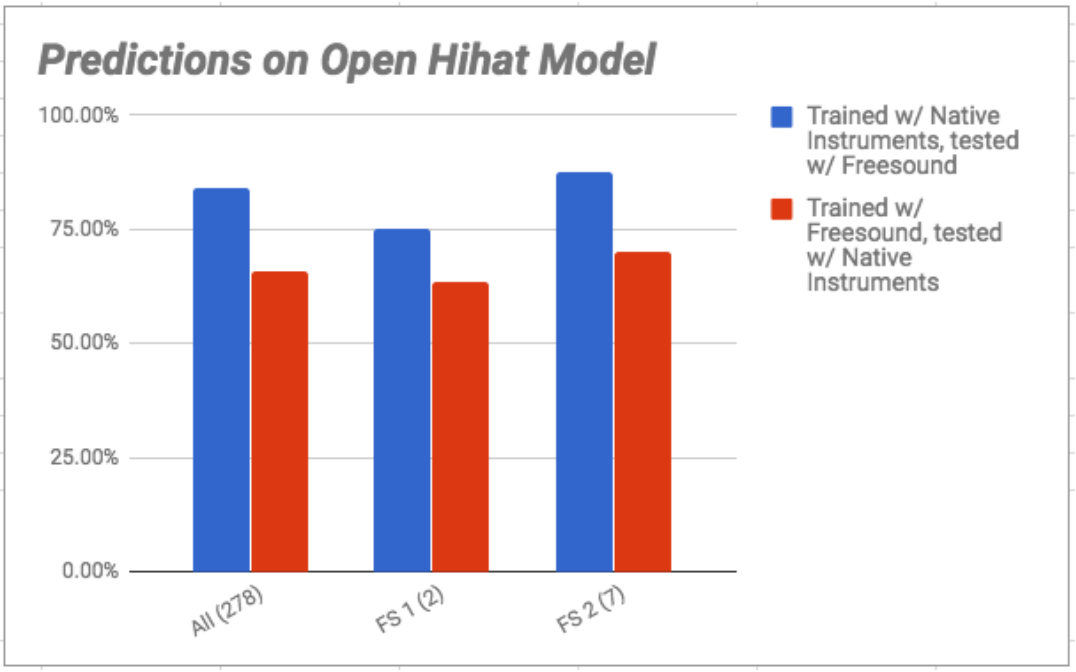

FIGURE 5.9: Predictions for Open Hi-hat Models.

5.2.5 Ride

There is a significant reduction on model accuracy when applying SVM with linear kernel for 'Feature Selection 2', Figure 5.10. As commented previously, it could lead us to an overfitted model. However, in Figure 5.11 predictions show that 'Feature Selection 1' does not perform better than 'FS 2'. Due to this fact and to the fact that

the system is supposed to be trained with our commercial dataset and tested with our free dataset or another one selected by the user, 'FS 2' is chosen as final selection.

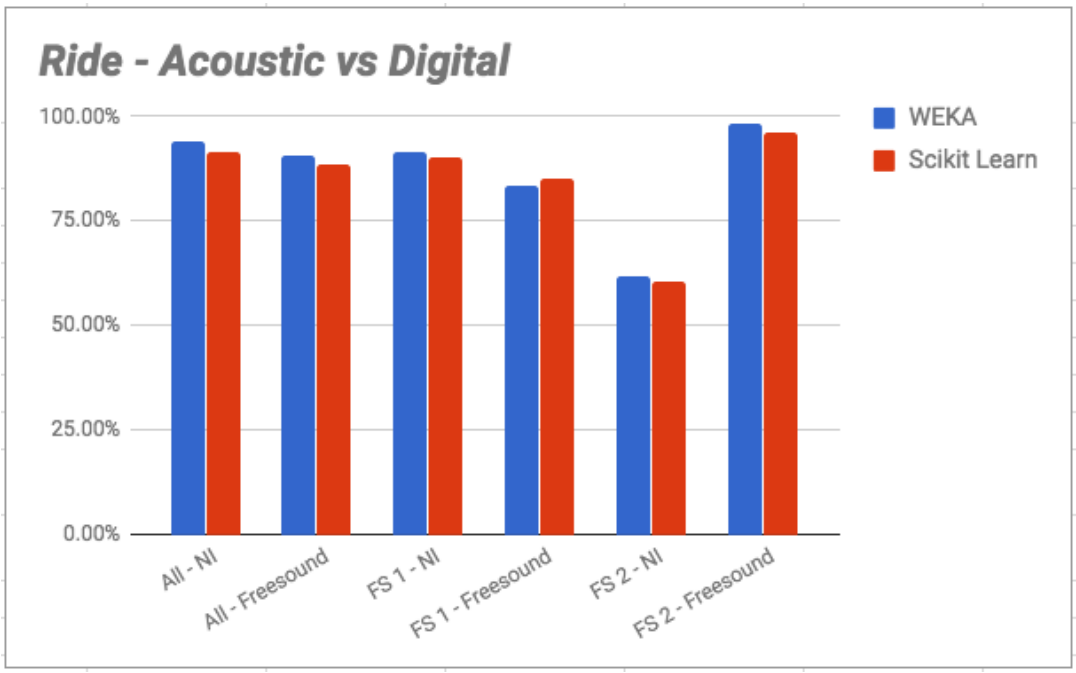

FIGURE 5.10: Feature Selection Comparison (Ride).

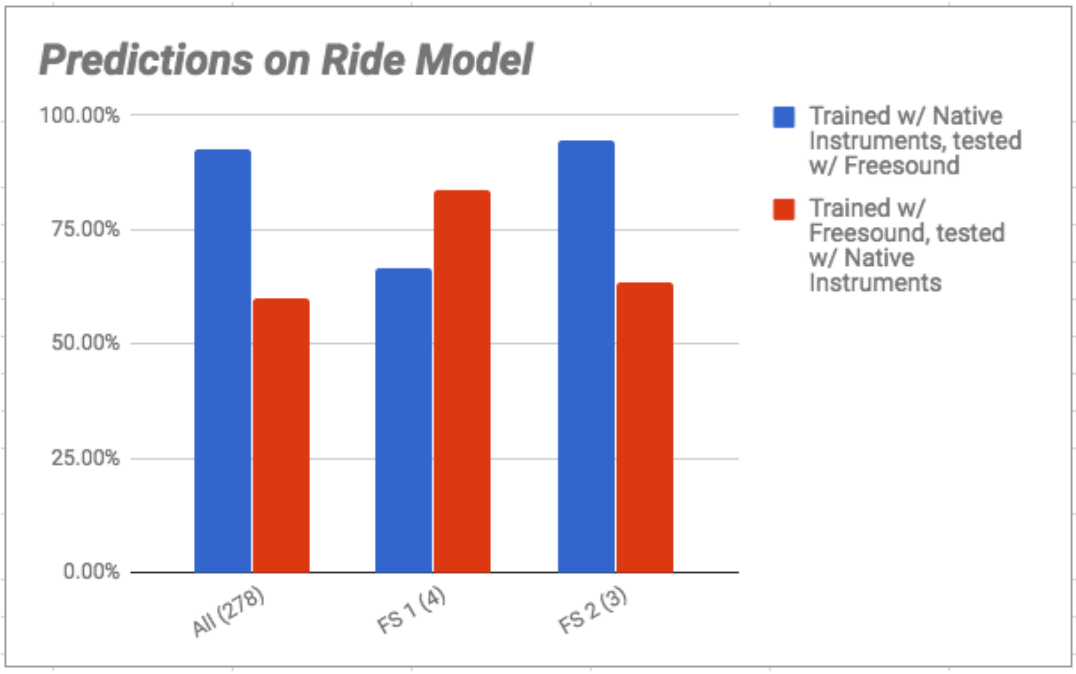

FIGURE 5.11: Predictions for Ride Models.

Final selected features to create Ride Model are:

1. Mean of Spectral Spread.
2. Mean of Spectral Entropy.

3. Mean of Silence Rate 60dB.

It is worth to mention that **Spectral Spread mean** and **Spectral Entropy mean** have appeared again as relevant features to measure drum category. **Silence Rate 60dB mean** estimates if a frame is silent, given a the threshold of 60dB. If a given frame's instant power is below this threshold, then each of the corresponding outputs will emit a 1. Although the latter is hard to understand why it could be relevant, it has been included on the model due to the obtained results.

Ride model for category classification is described in Figure 5.20 at the end of this section.

5.2.6 Snare

Apparently, snare case does not look so promising as previous classes, 5.12. Accuracies on 'All features' models are around 70-78% and, on most features selection models, below 70%; which may mean that snare sounds are not perceptually clear to differentiate. Best accuracy is found for 'Feature Selection 2' model when modeling our free dataset. Predictions, 5.13, confirm 'FS 2' as the most generalist model of the studied ones, reaching to predictions accuracies of 71.5-77.4%, for training with commercial dataset and testing with free dataset and vice versa, respectively; which is not as bad as it was supposed to be.

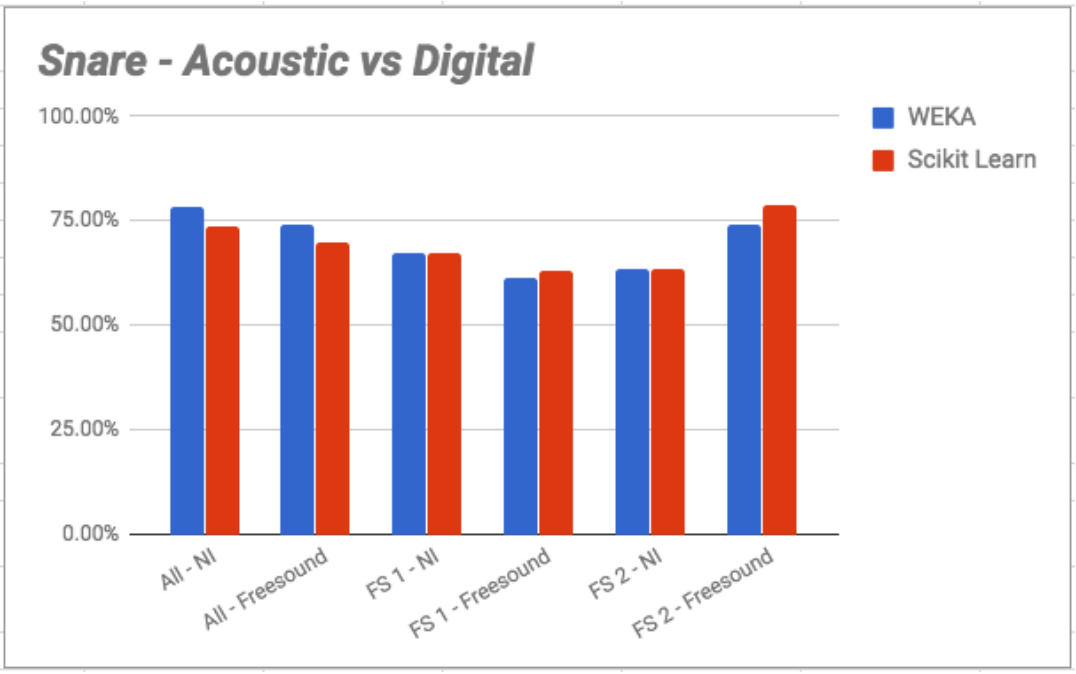

FIGURE 5.12: Feature Selection Comparison (Snare).

Final selected features to create Snare Model are:

1. Mean of Spectral Entropy.
2. Variance of ERB band 17.
3. Variance of Barkbands 22.

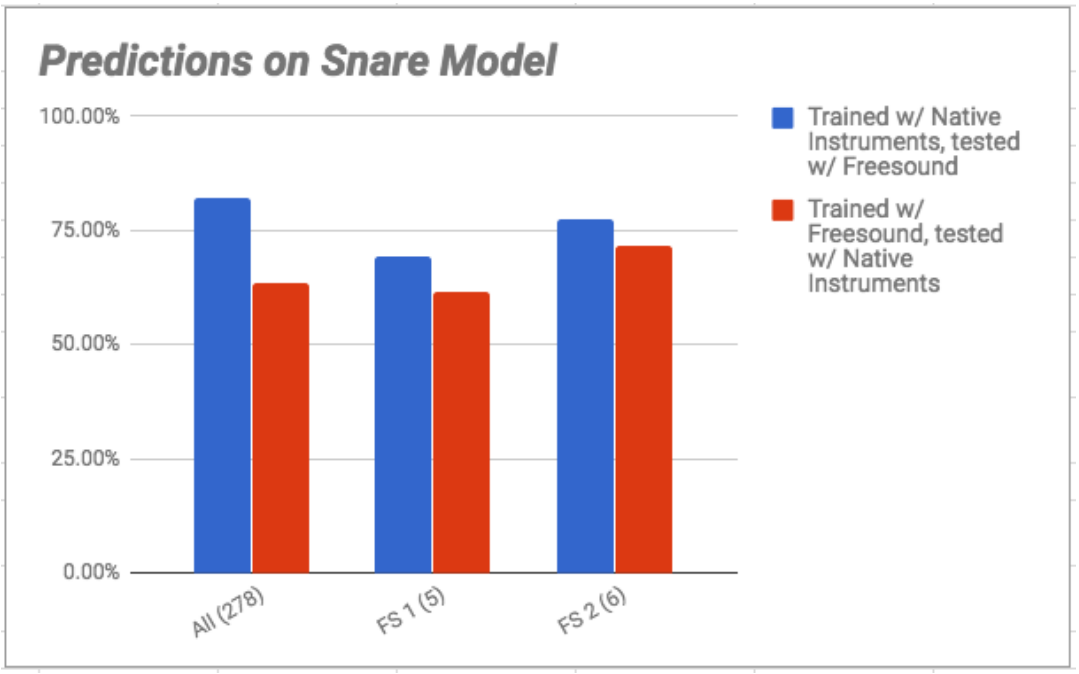

FIGURE 5.13: Predictions for Snare Models.

4. Variance of Barkbands 23.
5. Variance of Barkbands 24.
6. Mean of GFCC 1.

Despite of **Spectral Entropy mean**, mentioned before as relevant for other drum category models, the rest appear as difficult to relate with the 'acousticness' or 'digitalness' of a sound.

Snare model for category classification is fully described in Figure 5.21 at the end of this section.

5.2.7 Tom

Finally, tom case is probably the one that reached poorer results. In Figure 5.14, feature selection comparison shows that most model achieve a model accuracy lower than 70%, some of them even around 55%, which is present a non-promising situation. In Figure 5.15, results confirm the initial hypothesis. 'FS 2' model is again selected as the more generalist one and this is the reason why its features has been chosen as the selected ones. Although this feature selection is best of the studied ones, prediction accuracies reveal that it won't work as fine as the rest of drum category models created along this section.

Final selected features to create Tom Model are:

1. Mean of Barkbands 1.
2. Variance of GFCC 1.

Barkbands 1 mean is the mean of the computed energy in Bark band 1, which correspond to [50, 100] Hz. It could be explained by the prominence of energy of this low frequency band on digital sounds of our datasets. **GFCC 1 var** is the variance of the coefficient 1 of Gammatone-frequency cepstral coefficients.

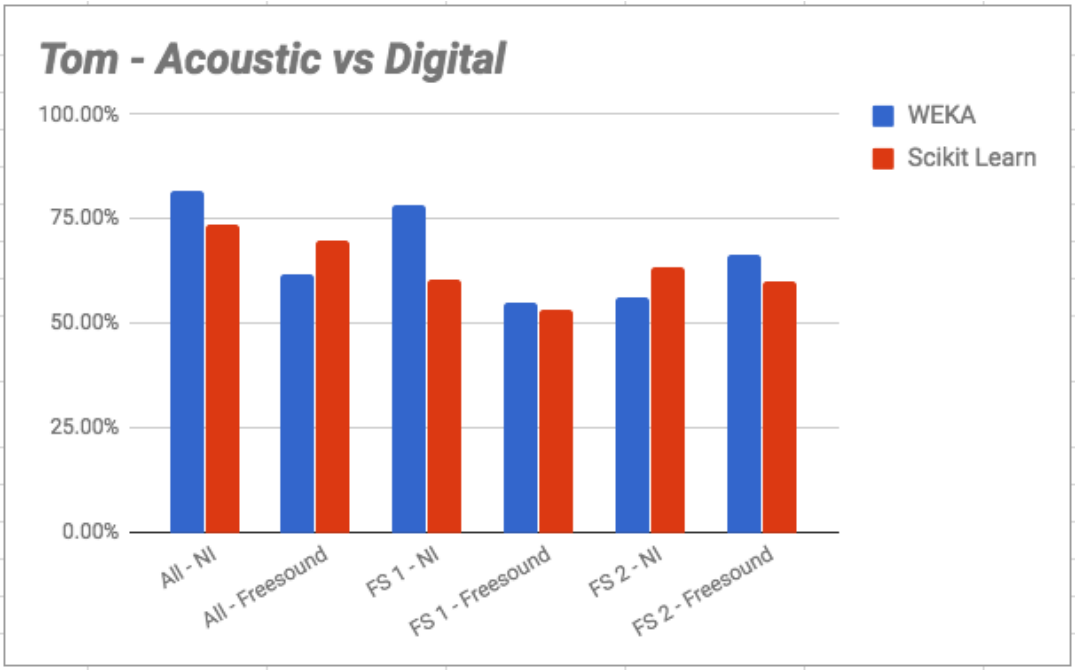

FIGURE 5.14: Feature Selection Comparison (Tom).

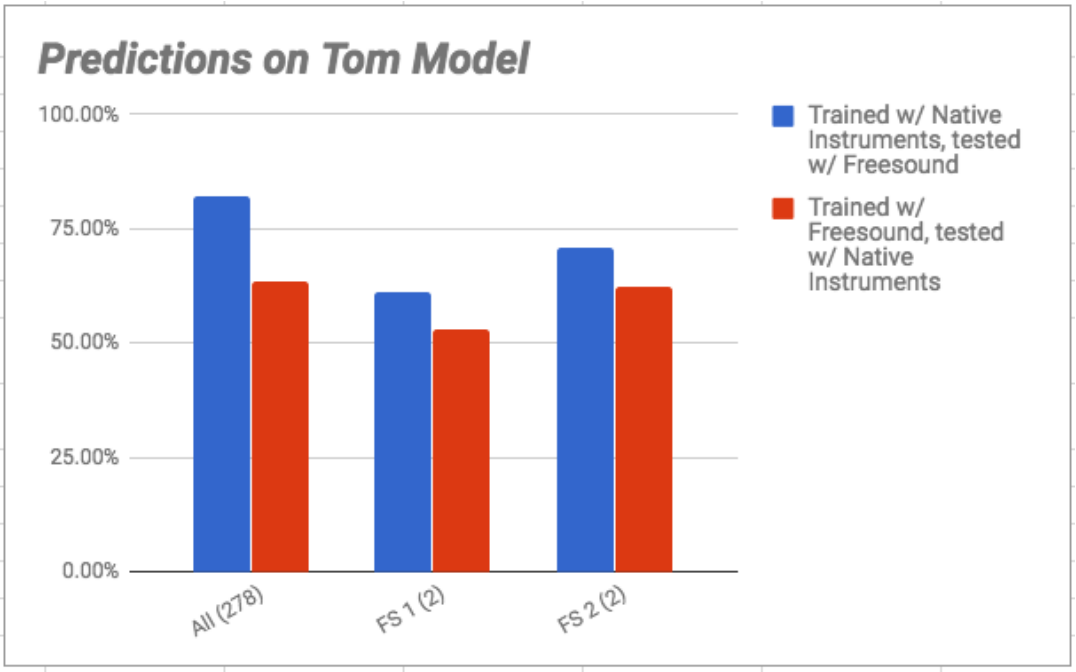

FIGURE 5.15: Predictions for Tom Models.

5.2.8 Discussion

Comparing the first approach with the second one, the latter tends to perform much better. In the former, best accuracy of correct predicted values when classifying drum category for the whole dataset achieve a value of only 70.76%. This accuracy corresponds to the model created by 'Feature Selection 2', the one selected by the application of CFS on our commercial dataset.

Second approach results in one model per each instrument class that determine the 'acousticness' or 'digitalness' of each drum sound. Closed Hi-hat sounds are predicted with an accuracy of 80.32% when modeling with four features. Crash sounds reach an accuracy of 68.5% (80% if the free dataset is used to train and the commercial one to test) when modeling with three features. Kick case shows an accuracy of 75% (79.5% if the free dataset is used for training and the commercial dataset for testing) when modeling with three features. Open Hi-hat case presents a model accuracy of 87.5% of correct predicted instances using only three features. In ride case, the highest accuracy of this approach is found. Sounds are predicted with an accuracy of 94.44% using a feature selection of three features. However, this result decreases around 30% when the model is trained and tested the other way around, meaning that it is not generalist as we would like. Snare case shows a more generalist model that achieves 77% of prediction accuracy. Finally, tom case presents the lowest prediction performance of this approach, 70.96% when using two features.

The mean of selected features in the second approach rounds the 4 features, which apparently let us to a low computational cost model, fulfilling our initial statement. Moreover, the mean of the obtained accuracies rounds 79.16% against 70.76% from the first approach, which represent an increment of almost 10% that reveals a meaningful improvement. Even though, not every class has appeared to be easily classified.

Some of the features considered as relevant for category classification tasks belong to some MFCCs, GFCCs, ERB Bands or Bark Bands. In these cases, it seems a bit difficult to explain why they can be considered as suitable for this type of task. An specific study should be done in order to experiment the influence of these descriptors and their corresponding bands and coefficients.

Nevertheless, some features have appeared as relevant for several models. **Dissonance**, **Pitch salience**, **Spectral Strong Peak**, **Spectral Energy** and **Flatness** come out only once in a final model, but they are some of the features with the highest frequency of appareance when applying feature selection algorithms. **Spectral Entropy** appears in three of the seven models and it could be considered as one of the most relevant features for category classification tasks. Considering **Spectral Spread** and **Barkbands Spread** as similar features, they also appear on three of the seven created models. **Tristimulus** coefficients appear twice in a final model. From the previously commented hard-to-relate descriptors, the most noticiable appareance are the GFCC and the MFCC both coefficients 1, which are included twice, for both cases, in a final model.

In the following pages category models per each instrument class are shown.

CLOSED HH MODEL

CLOSED HH	FEATURE SELECTION	TRAINNIG MODEL	WEKA	Scikit Learn
SVM linear kernel	All features	NI	62%	59%
		Freesound	93.44%	87.14%
	FS 1 (10)	NI	60.50%	55%
		Freesound	70.49%	70.95%
	FS 2 (3)	NI	59.50%	63%
		Freesound	81.96%	81.66%

FS 1 Ratio Gain	**NI dataset**	*10 features*			
FS 2 Ratio Gain	**Freesound dataset**	dissonance mean	barkbands spread mean	mffc mean 1	tristimulus 0 var

PREDICTION	TRAIN	TEST	ACCURACY
All features	NI	Freesound	83.60%
	Freesound	NI	68%
FS 1	NI	Freesound	78.68%
	Freesound	NI	55%
FS 2	NI	Freesound	80.32%
	Freesound	NI	58.50%

FIGURE 5.16: Closed Hi-hat Category Model.

CRASH MODEL				
CRASH	**FEATURE SELECTION**	**TRAINNIG MODEL**	**WEKA**	**Scikit Learn**
SVM linear kernel	All features	NI	88.50%	87.14%
		Freesound	64.81%	59%
	FS 1 (3)	NI	82%	82%
		Freesound	62.96%	62%
	FS 2 (10)	NI	76.50%	71.50%
		Freesound	74.07%	71%

FS 1 Ratio Gain	**NI dataset**	spectral spread mean	spectral entropy mean	flatness mean
FS 2 Ratio Gain	**Freesound dataset**	*10 features*		

PREDICTION	**TRAIN**	**TEST**	**ACCURACY**
All features	NI	Freesound	61.11%
	Freesound	NI	64.50%
FS 1	NI	Freesound	68.51%
	Freesound	NI	80%
FS 2	NI	Freesound	55.55%
	Freesound	NI	46.50%

FIGURE 5.17: Crash Category Model.

KICK MODEL				
KICK	FEATURE SELECTION	TRAINNIG MODEL	WEKA	Scikit Learn
SVM linear kernel	All features	NI	75%	73%
		Freesound	80%	78.33%
	FS 1 (3)	NI	78.50%	79.50%
		Freesound	75%	75%
	FS 2 (3)	NI	72%	71.50%
		Freesound	71.66%	68.33%

FS 1 Ratio Gain	NI dataset	mffc 7 var	mffc 4 var	spectral energy mean
FS 2 Ratio Gain	Freesound dataset	mffc 8 var	gffc 4 var	dissonance mean

PREDICTION	TRAIN	TEST	ACCURACY
All features	NI	Freesound	76.67%
	Freesound	NI	62%
FS 1	NI	Freesound	75%
	Freesound	NI	79.50%
FS 2	NI	Freesound	71.66%
	Freesound	NI	74.50%

FIGURE 5.18: Kick Category Model.

OPEN HH MODEL

OPEN HH	FEATURE SELECTION	TRAINNIG MODEL	WEKA	Scikit Learn
SVM linear kernel	All features	NI	73.23%	71.21%
		Freesound	87.50%	78%
	FS 1 (2)	NI	71.71%	70.60%
		Freesound	75%	75.33%
	FS 2 (7)	NI	72.22%	74.29%
		Freesound	89.28%	87.67%

FS 1 Ratio Gain	NI dataset	mfcc 1 mean	flatness mean		
FS 2 Ratio Gain	Freesound dataset	pitch mean	pitch salience mean	spectral strong peak mean	
		mfcc 1 mean	mfcc 7 mean	mfcc 9 var	tristimulus 2 mean

PREDICTION	TRAIN	TEST	ACCURACY
All features	NI	Freesound	83.93%
	Freesound	NI	65.65%
FS 1	NI	Freesound	75%
	Freesound	NI	63.63%
FS 2	NI	Freesound	87.50%
	Freesound	NI	70.20%

FIGURE 5.19: Open Hi-hat Category Model.

RIDE MODEL

RIDE	FEATURE SELECTION	TRAINNIG MODEL	WEKA	Scikit Learn
SVM linear kernel	All features	NI	93.91%	91.65%
		Freesound	90.74%	88.33%
	FS 1 (4)	NI	91.37%	90.13%
		Freesound	83.33%	85%
	FS 2 (3)	NI	61.92%	60.73%
		Freesound	98.15%	96%

FS 1 Ratio Gain	**NI dataset**	dissonance mean	hfc mean	gfcc 0 mean	gfcc 1 mean
FS 2 Ratio Gain	**Freesound dataset**	spectral spread mean	silence rate 60dB mean	spectral entropy mean	

PREDICTION	TRAIN	TEST	ACCURACY
All features	NI	Freesound	92.59%
	Freesound	NI	59.89%
FS 1	NI	Freesound	66.67%
	Freesound	NI	83.75%
FS 2	NI	Freesound	94.44%
	Freesound	NI	63.45%

FIGURE 5.20: Ride Category Model.

SNARE MODEL

SNARE	FEATURE SELECTION	TRAINNIG MODEL	WEKA	Scikit Learn
SVM linear kernel	All features	NI	78.50%	73.50%
		Freesound	74.19%	70%
	FS 1 (5)	NI	67.50%	67.50%
		Freesound	61.29%	62.92%
	FS 2 (6)	NI	63.50%	63.50%
		Freesound	74.19%	78.75%

FS 1 Ratio Gain	**NI dataset**	erb bands 13 mean	erb bands 14 var	barkbands 18 mean
		barkbands 19 mean	barkbands 25 mean	
FS 2 Ratio Gain	**Freesound dataset**	erb bands 17 var	spectral entropy mean	barkbands 22 var
		barkbands 23 var	barkbands 24 var	gffc 1 mean

PREDICTION	TRAIN	TEST	ACCURACY
All features	NI	Freesound	82.25%
	Freesound	NI	63.50%
FS 1	NI	Freesound	69.35%
	Freesound	NI	61.50%
FS 2	NI	Freesound	77.41%
	Freesound	NI	71.50%

FIGURE 5.21: Snare Category Model.

TOM MODEL

TOM	FEATURE SELECTION	TRAINNIG MODEL	WEKA	Scikit Learn
SVM linear kernel	All features	NI	81.57%	73.50%
		Freesound	61.66%	70%
	FS 1 (2)	NI	78.42%	60.50%
		Freesound	55%	53.33%
	FS 2 (2)	NI	56.31%	63.50%
		Freesound	66.67%	60%

FS 1 Ratio Gain	**NI dataset**	silence rate 30db mean	barkbands spectral spread
FS 2 Ratio Gain	**Freesound dataset**	barkbands 1 mean	gffc 1 var

PREDICTION	TRAIN	TEST	ACCURACY
All features	NI	Freesound	82.25%
	Freesound	NI	63.50%
FS 1	NI	Freesound	61.29%
	Freesound	NI	53%
FS 2	NI	Freesound	70.96%
	Freesound	NI	62.50%

FIGURE 5.22: Tom Category Model.

Chapter 6

Holistic Evaluation

As it has been described in the two previous chapters, the taxonomic classification part of the system is composed by two steps. In the first one, the system is intended to automatically distinguish between seven drum instruments: kick, snare and tom (membranes) and ride, crash, open hi-hat and closed hi-hat (plates). In the second step, the system is designed to be able to differentiate between the acoustic- o digital-nature of these drum samples. Both models are arranged in series in order to make them as simple, in terms of amount of computed features, and as efficient, in terms of model's accuracy, as possible.

In order to evaluate how the designed system actually performs, a preliminary user evaluation is proposed. A Python library called Ipywidgets [1] that provides some interactive HTML widgets for Jupyter Notebook has been used. A prototype of the retrieval system, which can be seen in Figure 6.1, has been created as similar as possible to the second table of Figure 1.2. Users have evaluated if the retrieved sounds belong to the selected instrument class, if its category is perceptually clear and if high-level descriptor selections show results that make sense.

The following three questions were asked to users:

1. Does the retrieved sample really correspond to the expected drum instrument class?
2. Does the retrieved sample really correspond to the expected drum category class?
3. Select how you think the system has interpreted your selection based on High-Level Descriptors.

In the two former questions possible answers are 'NO', 'NOT SURE' and 'YES'. While in the last question, possibilities are 'BAD', 'NOT SURE' and 'GOOD'. The aim of this preliminary evaluation is not a full evaluation about the usefulness of a possible commercial application based on the designed system, but only to get user first impression about the prediction accuracy of the system.

This preliminary prototype and evaluation, as well as user results, can be found at `github.com/Javier-AG/SMC_thesis`.

Each subject was asked to complete 10 evaluations, no matter which combination of instrument, category or high-level descriptors. This decision gives the user freedom to test the system according to his/her preferences. As it has been designed within a context of music production, music producers are the intended users of the system

[1] `https://github.com/jupyter-widgets/ipywidgets`

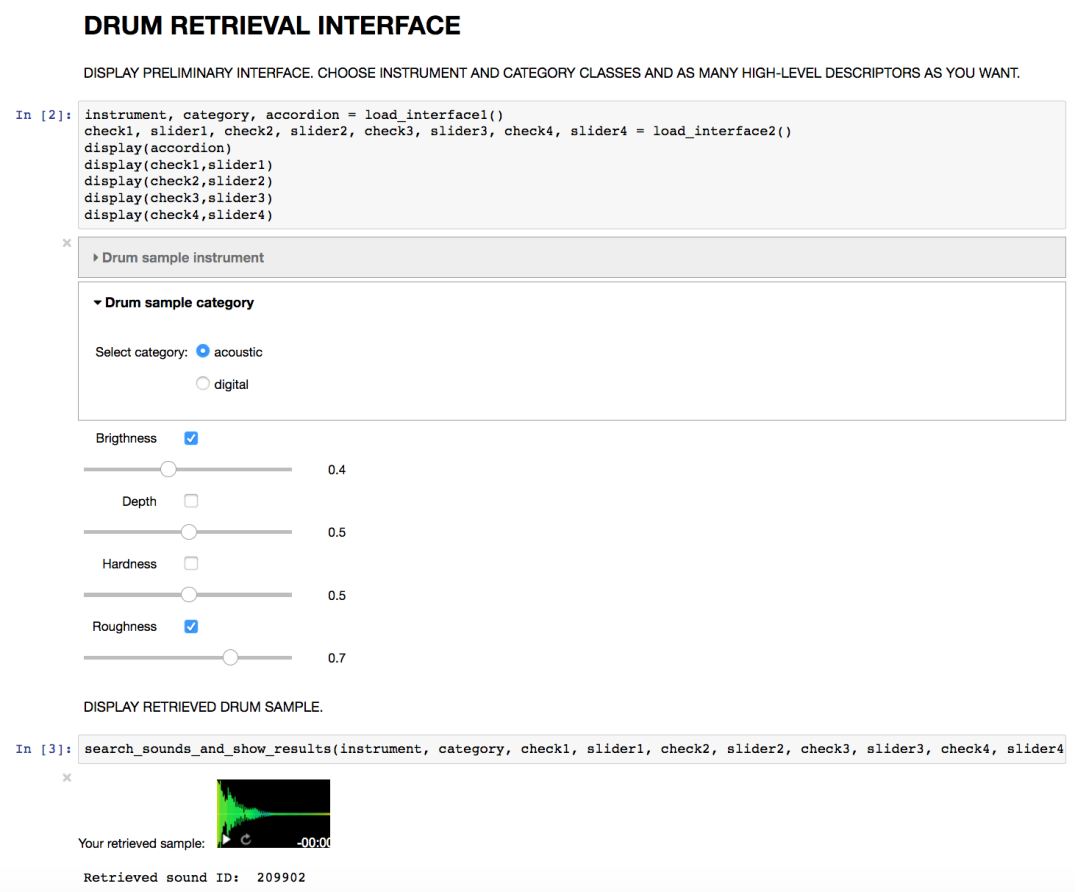

FIGURE 6.1: Screenshot of Preliminary Prototype.

but drummer's opinion is also interesting to consider. Finally, 7 subjects completed the questionnaire; 6 of them are music producers, while the other one is a drummer. Everyone has a professional experience of more than 5 years in the field. Their opinion serve us to test the proposed system from a professional point of view, due to the fact that their sound perception is highly trained.

Results, in Figure 6.2, show an apparently positive reaction to the proposal. Answers to the first question determine an accuracy of 77.63% for the instrument model. While 21.05% of subjects say that the retrieved drum sample does not correspond to the selected instrument class, only 1.31% are not sure about this fact. This last result confirms that instrument class is perceptually clear to distinguish for music professionals.

Number of searches are not equally distributed for each instrument class. Kick (21), snare (17), tom (15) and closed hi-hat (12) are the most wanted classes, probably because they tend to be the most used drum instruments in modern music genres.

In the case of category models, a global result of 71.05% of favorable answers implies that, in general, these models provide fairly good results. When analyzing each model separately, not every one results to be so promising. From the mentioned four most wanted instruments, the kick case results in 85.71% of good category predictions; followed by closed hi-hat and snare cases, 75.0% and 70.58%, respectively; but the tom case provide an unfavorable result of 40% on proper predictions, according to user's ratings.

PRELIMINARY EVALUATION RESULTS

MODELS ACCURACIES BY USERS		NO / BAD	NOT SURE	YES / GOOD
Instrument		21.05%	1.31%	77.63%
Category	**Global**	27.63%	1.31%	71.05%
	Closed HH	25.00%	0.00%	75.00%
	Open HH	22.22%	0.00%	77.77%
	Crash	0.00%	0.00%	100% *
	Ride	0.00%	0.00%	100% *
	Kick	14.28%	0.00%	85.71%
	Snare	17.64%	5.88%	70.58%
	Tom	60.00%	0.00%	40.00%
High-level Descriptors		19.73%	18.42%	61.84%

REPEATED SELECTION BY USERS	INSTRUMENT	CATEGORY	COUNT
	Closed HH	Acoustic	10
		Digital	2
	Open HH	Acoustic	6
		Digital	3
	Crash	Acoustic	0 *
		Digital	1
	Ride	Acoustic	0 *
		Digital	1
	Kick	Acoustic	10
		Digital	11
	Snare	Acoustic	12
		Digital	5
	Tom	Acoustic	8
		Digital	7

FIGURE 6.2: Preliminary User Evaluation Results.

Although no research was done during the high-level description section, the intention was to check if these high-level descriptors are useful in the task of describing drum sounds sonic features. Evaluations have not shown exceptional results, but somehow can be considered quite promising. Users have found that their retrieved sounds, according to their descriptors selection, have been acceptable on 61.84% of the times.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

During this thesis, a system design for drum samples description and retrieval has been presented. Both models' accuracies and user evaluation results have certified its usefulness for drum samples management in a music production environment. Drum instrument classification is a MIR task, as commented in State-of-the-Art chapter, that has been studied earlier. The aim of this project was trying to improve current classification models. Nevertheless, drum category classification is practically unnoticed in literature. In this case, the goal was to introduce a possible new MIR task and to preset a first attempt.

7.1.1 Instrument Model

The initial hypothesis of overfitted models found in literature for drum instrument classification tasks has been confirmed. Learning algorithms and low-level features can do a great job, but bigger and more diverse datasets are completely necessary if our intention is to create a generalist model that works fine for different type of drum sounds, such as more or less processed sounds, with or without room reverberation sounds, different types of strokes...

In relation to the contributions done in this topic, a fairly generalist instrument classification model for drum sounds has been exhibited. It is worth to mention that Spectral Contrast Descriptor, for each of its bands, has appeared as a significant descriptor for this kind of task. Assuming its six bands values as an unique descriptor, a drum instrument model can be created by the application of SVM with linear kernel and the following feature selection: **Spectral Contrast**, **Spectral Entropy**, **Pitch Instantaneous Confidence**, **Effective Duration** and **Log-Attack Time**. The two latter describe time envelopes and the other three characterize sound spectrum.

Ideally, the proposed system should query directly to `freesound.org` so as to get drum sounds. However, there is a technical problem. The created instrument classification models only discriminate between drum sounds, which means that drum samples in `freesound.org` should be already tagged as drums to be able to discriminate between the different classes that form this family of sounds. Due to this limitation, searches made by users on our prototype are querying local datasets.

During the preliminary user evaluation presented on the previous chapter, subjects confirmed that this model tend to perform properly. Several mistakes were found mainly in the tom case and, secondly, on the snare case. The former was confused

with kicks and the latter with open or closed hi-hats. During personal interviews, subjects have ratified this model as a valuable tool for automatic sound recognition and database management on a possible future commercial application.

7.1.2 Category Model

Discriminating between acoustic and digital sounds has been proved to make more sense when analyzed sounds belong to the same instrument class. Category prediction of any drum sample tend to be easier and faster when classification models already know the instrument name. Due to this reason, particular category models should be designed per each class. Another contribution related to this topic is the discovery of several low-level features that have proven to be relevant on the task of distinguishing between the acoustic- and digital-nature of a specific sound.

Spectral Entropy, which describes the peakiness of a distribution, seems to be a really relevant feature for this task, as well as **Spectral Spread** or **Barkbands Spread**. Several features have not been so fundamental but are also worth to mention due to its appearance on different category models: **Dissonance**, **Pitch salience**, **Spectral Strong Peak**, **Spectral Energy** and **Flatness**. The study of some descriptors and their corresponding coefficients, such as **Tristimulus**, **MFCC** and **GFCC**, should be also taken into account for future works related to this topic.

The fact that particular category models need to be created for each instrument provoke a limitation on queries to `freesound.org`. Drum samples should be already tagged with the corresponding instrument class to enable the system to determine its source nature. Therefore, same as in instrument model, searches made by users on our preliminary prototype are only querying local datasets.

During preliminary evaluation, users have shown positive reactions for the category model, although some of them have expressed on personal interviews the subjectivity of this measure.

7.1.3 High-Level Descriptors

Timbral models used on the high-level description section have return enoughly good results according to user evaluations. However, they have not been confirmed as meaningful high-level descriptors for drum sounds. During interviews, some users have proposed to recreate these timbral models. In this recreation, instead of having one model per high-level descriptor, the proposal would be creating a specific model per instrument and descriptor; which is the opinion that I personally agree. This would imply, for instance, the creation of a different brightness (or any other descriptor) model for the kick case than for the hi-hat case.

7.2 Future Work

Regarding research done during this thesis, several aspects could be improved in a future work. In relation to the instrument model, the final set of selected low-level descriptors would need to be verified on more drum sounds datasets. Experiments

with bigger and more diverse datasets could confirm if these features are as relevant as they seem to be according to our methodology.

In the case of category models, a new task has been exposed and need to be further studied. In order to avoid subjectivity, a possible solution could be the usage of datasets that have been previously annotated by experts. This fact would allow us to create models that distinguish between perceptually clear sound categories. Including more categories is also another possible future work. Following Native Instruments taxonomy, there are several categories that usually appear for many drum instrument classes, such as 'digital', 'human' or 'vinyl'.

Concerning system design, only a preliminary prototype has been presented. A real prototype should have a fancier interface and should work within a real-time music production environment. A Max for Live object[1] could be created so as to integrate the proposed system into a professional DAW, like Ableton Live.

[1] ableton.com/en/live/max-for-live/

Bibliography

Akkermans, Vincent, Joan Serrà, and Perfecto Herrera (2009). "Shape-based spectral contrast descriptor". In: *Proc. of the Sound and Music Computing Conf.(SMC)*, pp. 143–148.

Bell, Robert (2015). "PAL: The Percussive Audio Lexicon". PhD thesis. Swinburne University of Technology, Melbourne, Australia.

Bernardes, Gilberto, Matthew EP Davies, and Carlos Guedes (2015). "A Pure Data Spectro-Morphological Analysis Toolkit for Sound-Based Composition". In: *Proc. of the Electroacoustic Winds, Aveiro, PT*, pp. 31–38.

Brent, William (2010). "Physical and perceptual aspects of percussive timbre". In:

Celma, Oscar, Perfecto Herrera, and Xavier Serra (2006). "Bridging the music semantic gap". In:

Chang, Shih-Fu, Thomas Sikora, and Atul Purl (2001). "Overview of the MPEG-7 standard". In: *IEEE Transactions on circuits and systems for video technology* 11.6, pp. 688–695.

De Silva, Anthony Mihirana and Philip HW Leong (2015). *Grammar-based feature generation for time-series prediction*. Springer.

Font, Frederic and Xavier Serra (2015). "The Audio Commons Initiative". In: *International Society for Music Information Retrieval Conference (ISMIR)*.

Font, Frederic et al. (2016). "Audio Commons: bringing Creative Commons audio content to the creative industries". In: *Audio Engineering Society Conference: 61st International Conference: Audio for Games*. Audio Engineering Society.

Gouyon, Fabien and Perfecto Herrera (2001). "Exploration of techniques for automatic labeling of audio drum tracks instruments". In: *Proceedings of MOSART: Workshop on Current Directions in Computer Music*.

Grey, John M (1977). "Multidimensional perceptual scaling of musical timbres". In: *the Journal of the Acoustical Society of America* 61.5, pp. 1270–1277.

Herrera, Perfecto, Amaury Dehamel, and Fabien Gouyon (2003). "Automatic labeling of unpitched percussion sounds". In: *Audio Engineering Society Convention 114*. Audio Engineering Society.

Herrera, Perfecto, Vegard Sandvold, and Fabien Gouyon (2004). "Percussion-related semantic descriptors of music audio files". In: *Proc. of 25th International AES Conference*. Citeseer.

Herrera, Perfecto, Alexandre Yeterian, and Fabien Gouyon (2002). "Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques". In: *Music and Artificial Intelligence*. Springer, pp. 69–80.

Herrera, Perfecto et al. (2005). "Simac: Semantic interaction with music audio contents". In: *Integration of Knowledge, Semantics and Digital Media Technology, 2005. EWIMT 2005. The 2nd European Workshop on the (Ref. No. 2005/11099)*. IET, pp. 399–406.

Herrera-Boyer, Perfecto, Geoffroy Peeters, and Shlomo Dubnov (2003). "Automatic classification of musical instrument sounds". In: *Journal of New Music Research* 32.1, pp. 3–21.

Lakatos, Stephen (2000). "A common perceptual space for harmonic and percussive timbres". In: *Perception & psychophysics* 62.7, pp. 1426–1439.

McAdams, Stephen et al. (1995). "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes". In: *Psychological research* 58.3, pp. 177–192.

Moore, Brian CJ and Brian R Glasberg (1983). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns". In: *The Journal of the Acoustical Society of America* 74.3, pp. 750–753.

Pampalk, Elias, Perfecto Herrera, and Masataka Goto (2008). "Computational models of similarity for drum samples". In: *IEEE transactions on audio, speech, and language processing* 16.2, pp. 408–423.

Pampalk, Elias, Peter Hlavac, and Perfecto Herrera (2004). "Hierarchical organization and visualization of drum sample libraries". In: *Proc. Int. Conf. Digital Audio Effects (DAFx)*, pp. 378–383.

Pearce, Andy, Tim Brookes, and Russell Mason (2016). *D5.1 Hierarchical Ontology of Timbral Semantic Descriptors*. Deliverable Work-package. Audio Commons.

– (2017). *D5.2 First Prototype of Timbral Characterisaiton Tools for Semantically Annotating non-musical Content*. Deliverable Work-package. Audio Commons.

Peeters, Geoffroy (2004). "A large set of audio features for sound description (similarity and classification) in the CUIDADO project". In:

Peeters, Geoffroy, Stephen McAdams, and Perfecto Herrera (2000). "Instrument sound description in the context of MPEG-7". In: *ICMC: International Computer Music Conference*, pp. 166–169.

Peeters, Geoffroy et al. (2011). "The timbre toolbox: Extracting audio descriptors from musical signals". In: *The Journal of the Acoustical Society of America* 130.5, pp. 2902–2916.

Sá Pinto, António (2015). "Automatic Semantic Characterization of Drum Sounds". MSc thesis. Universitat Pompeu Fabra.

Schaeffer, Pierre (1966). *Traité des objets musicaux*. Le Seuil.

Serra, Xavier et al. (2013). *Roadmap for Music Information ReSearch*, p. 88. ISBN: 978-2-9540351-1-6.

Shahar, E. (2010). *Automatic Drum Samples Classification*. URL: http://alumni.media.mit.edu/~persones/pattern_rec/patternrec.html (visited on 03/05/2017).

Sillanpaa, Jukka et al. (2000). "Recognition of acoustic noise mixtures by combined bottom-up and top-down processing". In: *Signal Processing Conference, 2000 10th European*. IEEE, pp. 1–4.

Smalley, Denis (1997). "Spectromorphology: explaining sound-shapes". In: *Organised sound* 2.02, pp. 107–126.

Souza, Vinícius MA, Gustavo EAPA Batista, and Nilson E Souza-Filho (2015). "Automatic classification of drum sounds with indefinite pitch". In: *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, pp. 1–8.

Stamellos, Lefteris (2016). "Freesound drum sets using unconventional sounds". MSc thesis. Universitat Pompeu Fabra.

Terasawa, Hiroko, Malcolm Slaney, and Jonathan Berger (2005). "The thirteen colors of timbre". In: *Applications of Signal Processing to Audio and Acoustics, 2005. IEEE Workshop on*. IEEE, pp. 323–326.

Thoresen, Lasse and Andreas Hedman (2007). "Spectromorphological analysis of sound objects: an adaptation of Pierre Schaeffer's typomorphology". In: *Organised Sound* 12.02, pp. 129–141.

Tindale, Adam R, Ajay Kapur, and Ichiro Fujinaga (2004). "Towards Timbre Recognition of Percussive Sounds." In: *ICMC*.

Tindale, Adam R et al. (2004). "Retrieval of percussion gestures using timbre classification techniques." In: *ISMIR*.

Tindale, Adam R et al. (2005). "Indirect acquisition of percussion gestures using timbre recognition". In: *Proc. Conf. on Interdisciplinary Musicology (CIM)*.

Turquois, Chloé et al. (2016). "Exploring the Benefits of 2D Visualizations for Drum Samples Retrieval". In: *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. ACM, pp. 329–332.

Van Steelant, Dirk et al. (2005). "Support vector machines for bass and snare drum recognition". In: *Classification—the Ubiquitous Challenge*. Springer, pp. 616–623.

Vinet, Hugues, Perfecto Herrera, and François Pachet (2002). "The cuidado project". In: *International Conference on Music Information Retrieval*, pp. 197–203.

Wiggins, Geraint A (2009). "Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music". In: *Multimedia, 2009. ISM'09. 11th IEEE International Symposium on*. IEEE, pp. 477–482.