Automatic Classification of Musical Instrument Samples

Author: Daniele Scarano

MASTER THESIS UPF / 2016
Master in Sound and Music Computing

Master thesis supervisors:
Dmitry Bogdanov, Frederic Font
Department of Information and Communication Technology
Universitat Pompeu Fabra, Barcellona

This work is dedicated to Raffaella that is incredibly patient, helpful and never stopped encouraging me during the course of this year.

Acknowledgment First of all I would like to thank Professor Xavier Serra that gave me the possibility to spend this year attending the Sound and Music Computing master programme at the Universitat Pompeu Fabra in Barcelona. My two incredible supervisors Frederic Font and Dmitry Bogdanov for their support during my thesis work, their invaluable help and advices and their patience. All my professors and all the people in the Music Technology Group that shared with me an incredible amount of knowledge, informations, documentation and much more: Emilia Gomez, Perfecto Herrera, Rafael Ramirez, Agustín Martorell, Marti Sanchez Fibla, Davinia Hernandez, Sergi Jordà, Alastair Porter, Jordi Janer, Panos Papiotis and all the other.

My special gratitude goes to my family in Italy that supported me in many ways, my mother, my father my Children Riccardo and Valerio. I cannot forget my fellow classmates and the time we spent together studying, discussing, programming, drinking, eating and enjoying our stay in Barcelona. A special thanks to Kosmas for his lovely presence, Felipe for his enthusiasm, Roman for the sharing spirit, Pere for the geek attitude, Angel for the interesting talks, Blai for his reserved perfectionism, Pedro for the spanish recipies, Albin for his kindness, Javi and Javi respectively for the family hamburgers and the hospitality... and my gratitude goes also to all the ones that shared their time, their food, their house and other amenities with me and are not mentioned here.

The last thanks, but not the least, goes to all the people that supported the Chromahelix project, Andrès Lewin, Sonia Espì and all the Phonos foundation.

This year at the MTG is one of the most valuable experiences I've ever made and I have to thank many people in this small page space, if you are not mentioned here you are for sure in my thoughts, I will not forget this period and all of you.

Abstract

Automatic classification of musical instrument is an old research topic in music information retrieval. In this work we address the problem of the classification using musical instrument single note samples from Freesound and we put the accent on the content analysis of the sound and how those content information are connected to the physical characteristics of each instrument. We build a taxonomy based on instrument families and mode of excitation. The musical instruments play a central role in this work and the studies on timbre are used as a methodological base to apply feature selection to our complete set of descriptors, the aim of this is to find which descriptors are relevant to describe a specific instrument. The machine learning then is used as an instrument to evaluate our choices, to identify weakness and problem in the current implementation of audio descriptors.

Table of Contents

List of Figures

List of Tables

Introduction

0.1 Foreword

Musical instrument sample classification is still an open research topic, the first attempts to address this problem where in the '70 [Wessel, 1979; Grey, 1977], during that decade the timbre was registered as standard by the American National Standard Institute (ANSI) and a lot of experiments to better understand this multidimensional characteristic of the sound were conducted.

The increasing computational power of today's computers combined with the size of available sound collections offer the possibility to work on instrument sample classification with considerably larger number than in the past years. Freesound [Font et al., 2013] is a large crowd sourced database of sound that contains more than 300.000 samples. In this huge collection we can find many instrument samples in form of single notes or musical phrases. This is the source for our dataset and we will see in chapter 2 the detailed information on how we created it.

The Music Information Retrieval field has evolved and nowadays we have many libraries that offer the possibility to extract low-level descriptors from an audio file. The number of MIR algorithms is increasing every day and this gives us the possibility to describe sounds in a very accurate way, we can group the audio features in 4 categories [Schedl et al., 2014]:

- Time domain representation
- Spectral domain representation
- Pitch content description
- Rhythm

In this thesis we will extract the content information using the Essentia library [Bogdanov et al., 2013] developed by the Music Technology Group (MTG) at the Universitat Pompeu Fabra. This library is used in Freesound to analyze the sounds and store the audio features in the database and those descriptors are used to search for similar sounds. In chapter 1 we describe the similarity approach used in many

studies on classification [Grey, 1977; Wessel, 1979; Krumhansl, 1989; McAdams et al., 1995] to understand how the low level descriptors are correlated with the human perception and timbre.

Machine learning is receiving a lot of attention from the Industry and the Academia and this produces improvements to the algorithms quality and the availability of tools that are fundamental for classification. Many techniques as regression trees, Naive Bayes, Key Nearest Neighbour (K-NN), Neural Networks and Support Vector Machine (SVM) are used to classify musical instruments. Some of them, as decision trees, give us the possibility to understand better the audio descriptors and their relation with audio. Others like SVM and neural networks produce interesting results in term of accuracy but the correlation of the data with the audio is not clear. We will use the gaia library[1] to accomplish this task with the SVM algorithm for classification. At the same time we will address the problem of the feature selection from a physical and perceptual perspective to learn as much as possible on the object of our study, the sound of musical instruments.

0.2 Motivation

The huge size of the sound collections used in the industry and the availability of open access content as Freesound are the two main reasons why we are trying to address the musical instrument sound samples classification. The organisation and visualisation of this content in a 'human readable' way can be achieved improving the tools we use to describe and classify sounds and those tools are relevant for two main reasons:

1. They can be used to speed up the work of sound designers in the TV/Movie and Videogame industry

2. The content organisation and accessibility will increase the attention of the industry on open access databases as Freesound

While the first reason is more practical and is relevant for the scope of the development of new audio technology the second is a virtuous circle in which Freesound offers an important service to the research community in terms of quality and quantity of available sound material, the attention of the industry to this content increases the number of contributors as long as the quality of the contributions. In addition to this the creative industry, the musicians and all the sound hackers around the world will benefit from this kind of research because it will increase usefulness and usability of Freesound content. Moreover the classification of musical instrument samples is a complete journey in the music technology, we will

[1] https://github.com/MTG/gaia

deal with signal processing, physics of sounds, MIR and other concepts and tools that are relevant in this moment in the study of sound in general, we consider this as an important outcome of this project.

0.3 Goals

We will focus on one classification techniques and we will try to improve the accuracy of the algorithm using a content based description approach. We present here a list of goals:

- Collect musical instrument samples and organise them in different datasets
- Use the current state of the art descriptors and classifier to create a baseline
- Find a correlation between instrument physical models and audio descriptors
- Classify the dataset using small sets of selected descriptors
- Evaluate the results and identify weakness and possible improvements

Chapter 1

STATE OF THE ART

Music is part of human society since ancient times and the musical instruments used to compose and perform music have been object of attention in almost every culture. The ethnomusicologists have studied in deep the classification schemas and their difference between culture and in her complete dissertation on this topic Kartomi [Kartomi, 2001] shows that the hierarchical categorisation of instruments and their families is often subject to social and cultural biases derived from the society in which it is generated. The complete presentation of defined classification schemas is outside the scope of this work, but if we want to categorise musical instrument we have to know a little bit of this story to choose the schema that best fits our goals.

The first classification schema suitable for use worldwide was created by Victor Charles Mahillon in 1880 [Kartomi, 2001]. This study is based on an ancient indian classification schema, but it extends the scope of the original work. The well known Hornbostel and Sachs [1] classification schema is similar to the fourth classes Mahillon schema and is partially based on it. The studies in organology of the nineteenth century are mostly based on the Hornboster model that offer a general and schematic approach the instrument categorisation easy to use in different contexts. There have been also some improvements and expansions to the original proposal, the introduction of the electrophones group, made to include the electroacoustics instruments in the schema, is one example of contribution made to the model and was proposed for the first time by Galpin (Galpin, 1937). We should note that the scope of each classification schema depends on the individual researcher goal and responds to a scientific view, a museological organisation or other approaches. The classification of museum instrument collections and the study of social and cultural implication are present in many musicological studies between the publication of the Hornbostel and Sachs model in 1914 and the

[1] https://en.wikipedia.org/wiki/Hornbostel-Sachs

Author	Year	Dataset	Descriptors	Feature Selection	Classifier
John M. grey	1977	16 Synthesized Musical instrument tones (perceptually equalized for loudness, pitch and duration(?))	Spectral energy distribution, synchronicity in the transient of higher harmonics with the spectral fluctuation, low-amplitude high-frequency energy in the initial attack segment	MDS INDSCAL and Hierarchical cluster analysis	2 Listening experiments with human subjects.
Lakatos	2000	34 sounds from the McGill University Master Samples(MUMS): 17 pitched instruments and 18 percussive instruments	Spectral Centroid and rise time alone for acoustically dive timbres. Tmbral richness' adds to the other two as percussive instrument third dimension.	MDS CLASCAL (multi Dimensional Scaling) used to evaluate not to select. EXTREE to evaluate non spacial representable characteristics	Listening experiments with human subjects
Deng et al.	2008	UIOWA MIS collection. Solo recording extracted from the University of Otago Library CD collection	MFCC(26). MPEG-7(7). IPEM(11)	MDS techniques: PCA and Isomap	K-NN; naive Bayes; multilayer perceptron(MLP) and radial basis functions(RBF)
Liu, Jing Xie, Lingyun	2010	2177 musical clip of 30 seconds length from solo recording	Spectral centroid, spectral roloff, spectral flux, time domain zero crossing, MFCC, spectral crest factor(SCF), spectral flatness measure(SFM). 16 attributes of STFT, 52 of MFCC, 96 of SCF and also 96 of SFM.		SVM and SMO (Sequential Minimal Optimization) to train the model
Bhalke et al.	2015	UIOWA MIS collection: 20 musical instruments. Piano samples from Virtual Studio Technology (VST) instruments: four different kinds of single note of pianos.	MRP (multi resolution plots), spectrogram images		Deep convolutional neural network

Figure 1.1: Table of Citations

present. From our perspective an exhaustive classification of musical instrument is based on the sound they produce is related to the sound they produce and we will search for it in the literature of musical instrument classification in MIR.

The digitalisation of our world and the broad use of computers in almost every human activity leads naturally towards the creation of automatic systems for musical instrument classification, to accomplish this task a deep knowledge of the sound characteristic of each instrument is necessary. In this work we aim at understand better the correlation between low-level descriptors and the sound and more in detail to learn what are those features that best represent each musical instrument. In this chapter we present the theoretical background behind this work and a short literature review on musical instrument classification, the figure 1.1 present a schematic representation of the studies we used as a basis for this work.

1.1 Content-based Approach

The content-based approach is completely based on informations that can be extracted from the audio. There have been many attempts to describe timbre and measure timbre similarity between instruments. Krumhansl [Krumhansl, 1989] conducted some psychoacoustical experiments using human subjects to measure perceptive similarity in musical instrument sounds. The timbre space described by McAdams [McAdams et al., 1995] is an attempt to identify the low-level descriptors that are relevant to measure perceptual similarity and can be correlated with the higher-level concepts as instrument family. In his works he identified Spectral Centroid, Spectral Flux and Log Attack Time as the three more relevant features. The identification of the characteristics of the sound that are relevant

for describing timbres is made more complex by the fact that the importance of a feature is context dependent[2]. Lakatos [Lakatos, 2000] identifies the spectral center of gravity and the rise time as the two main characteristics of an instrument timbre, those two characteristics are relevant both in harmonic sounds and percussive sound perception (that reinforces the idea that our basic timbre perception is bidimensional) while the third dimension seems to be context dependent, where the context is the sounds used in the dataset. In the last mentioned experiment conducted by Lakatos [Lakatos, 2000] in fact the Multi Dimensional Scaling (MDS) identifies two different timbre spaces, one bidimensional related to the harmonic sounds and another three dimensional for the percussive sounds. The earlier experiment of Grey [Grey, 1977] yielded a three dimensional timbre space, also obtained using MDS, where the three dimensions are assigned to the spectral energy, the synchronicity in the transient of higher harmonics with the spectral fluctuation and low-amplitude high-frequency energy in the initial attack segment. Krumhansl [Krumhansl, 1989] using a different dataset obtained results similar to Grey's, but the third dimension was assigned to the spectral flux. The experiment made by McAdams [McAdams et al., 1995] used the same dataset as Krumhansl but a different MDS technique that assigned the 3 dimensions to the spectral centroid, the log attack time and spectral flux. We can argue that the features are dependent on the dataset used, the results of McAdams [McAdams et al., 1995] and Krumhansl [Krumhansl, 1989] agree on the three dimensional space even if they use different MDS techniques, but even in the same dataset as we've seen in Lakatos [Lakatos, 2000] the features are related to the timbre of the sounds.

1.2 Musical Instrument Physical Models

Musical instruments are traditionally divided into families for practical reasons, in the western orchestra there is a standard way of grouping the instruments and this can be used as taxonomy for instrument classification. Nevertheless Howard e Angus [Howard and Angus, 2009] propose a categorisation based on four big families represented by string, wind, percussion instruments and speaking/singing voice as a fourth big family. This categorisation is based on the fact that each of those families can be explained with a distinct physical model. Each family can then be divided into subcategories using the mode of excitation of the instrument, this physical aspect describes how the mass that produces the perturbation in the air is indeed excited. The mode of excitation is important because it determines the timbral characteristics of the sound produced by the instrument.

[2]Different instrument classes can be described by distinct sets of features

To explain better this concept we will use the violin as an example, the two main techniques to play it are the bow and the pizzicato. We can generalise the pizzicato technique as a string pluck; in this case when the string is plucked at his center, that is the ideal condition, the odd harmonics are maximally excited while the even are not excited at all. We will not observe this perfect behaviour in a sound produced by a real musical instrument but the presence of the odd harmonics is certainly proportional to the position in which the string is plucked and possibly higher than the even ones.

When the bow is used the hairs make the string to move at a constant speed until it is no more gripped and returns rapidly to his normal position, this cycle creates a particular timbre that is very similar to a sawtooth wave. The spectral characteristic of this waveshape is the presence of all the harmonics with an amplitude proportional to the number of the harmonic

$$A = 1/n \tag{1.1}$$

where *A* is the amplitude of each harmonic and *n* is the number of the harmonic.

In the case of of the piano where the string is excited by a struck we observe a shift in the harmonics whose pitch is higher than the harmonic series built on the fundamental, this slightly different timbre is due also to the stiffness of the piano strings that act like a bar. The produced characteristic pitch shifting of the timbre can be defined inharmonicity and can be measured using the implementation available in the Essentia library named *Inharmonicity*[3] used for computing the sound descriptors in Freesound. In 1998 Scalcon, Rocchesso and Borin [Scalcon et al., 1998] proved the inharmonicity perceptual relevance in piano notes, they conducted an experiment in which some pianists had to decide if they were listening to synthesized piano sounds or real ones. The result was that the inharmonicity plays an important role in the decay portion of the sound; they also discovered that there is a frequency threshold above which the inharmonicity become more and more perceptual irrelevant and this threshold can be computed as a function of the *f0*.

The physical characteristics of the instruments and their modes of excitation offer a basis to create a taxonomy for classification because as the examples of the violin and the piano show we found a correlation between some low-level descriptors and the physical models. We will separate the instrument samples into three big families derived from the study of the physical models:

- Plucked string
- Bowed string

[3] http://essentia.upf.edu/documentation/reference/std_Inharmonicity.html

- Wind instruments

Liu and Xie in their research citepLiu2010 used a similar taxonomy with 4 classes, plucked strings, bowed strings, wind instrument and percussive instruments, they used this schema to classify a dataset of western and chinese instruments, they argue that this categorization where the playing techniques has an central role can generalise better than a categorization based on the classical western orchestra instrument families. In our case we will focus only on harmonic sounds and avoid the percussive instruments family.

1.3 The Dataset

In automatic musical instrument classification different kind of audio samples are used depending on how the problem is addressed in each specific research. The latest tendency is to work with polyphonic signals where studio or live recording of musical pieces are used. The traditional approach instead attempts to classify the instruments using musical phrases, single notes or synthesised instrument samples in solo recordings[Herrera-Boyer et al., 2003]. We will review some works related to the second and older approach to identify which kind of sounds have been used and why.

I the early studies on timbre similarity and also in the recent research in which psychological experiments are used to evaluate the audio material and create the baseline, small sets of sounds are used, Grey [Grey, 1977] used a set of 16 synthetised musical instruments and Lakatos [Lakatos, 2000] used 34 from the McGill University Master Samples(MUMS). In the case of automatic instruments classification instead larger dataset are used, Liu and Xie [Liu and Xie, 2010] used 2177 sound excerpts from solo recordings in their research on automatic musical instrument classification. The size of the dataset is also related to the number of features used for classification, common practice is to have a dataset with a number of samples considerably larger than the number of features.

Another important aspect is the choice of multiple datasets that can be useful to generalise the results and to perform cross-collection validation. Livshin and Rodet [Livshin and Rodet, 2003] made an experiment to identify which evaluation method performs better for classification and the *Minus-1 DB* was demonstrated being the best. This approach is characterised by the use of multiple datasets and each of them is classified with a model trained using the other N datasets. Deng [Deng et al., 2008] in his research used two datasets, the UIOWA MIS collection that contains 761 instrument notes recorded in an anechoic chamber and to generalise the results he conducted the same experiment using a dataset of solo recording extracted from the University of Otago Library CD collection.

In this research we will create one dataset based on Freesound content. Roma in his PhD dissertation attempted a large scale classification of the Freesound content [Roma, 2015] and other works are based on Freesound content as the master thesis of Carlos Vaquero [Vaquero, 2012]. Our goal is to identify the sounds that are available for the instrument classification task and share the dataset with the MIR community in order to encourage future use of it. Another dataset will be synthesised using the soundfont technology[4].

1.4 Low-level Descriptors

Nowadays there are many low-level descriptors schemes used in research with content based approach, but which is the best for instrument classification is still not known. The studies on timbre similarity offer a good basis to learn the most relevant low-level features, but as we?ve seen earlier in this chapter they don't agree on the 3 dimensions of the timbre space. In automatic instrument classification more complex and larger schemas are used and the problem of identifying the more accurate is far from being solved. Deng in his research on feature analysis [Deng et al., 2008] identified the Mel Frequency Cepstral Coefficients (MFCC), 5 mpeg-7 descriptors, harmonic deviation, harmonic spread, harmonic variation, spectral centroid, and temporal centroid, and 4 perceptual based descriptors. Another scheme is used by Liu and Xing [Liu and Xie, 2010] that comprehend Spectral centroid, spectral rolloff, spectral flux, time domain zero crossing, MFCC, spectral crest factor, spectral flatness measure. A list of features sets used in the literature has been collected and stored in the features table presented in figure 1.2. In the context of Freesound a features schema based on Essentia library is used to search for similar sounds. The feature set is made of 89 audio descriptors divided into four categories: tonal, low-level, sfx and rhythm, and for some of them statistic measures are computed. A complete list of descriptors and the list of statistics are available in the Freesound API documentation[5]. In this thesis we will start using this schema to identify possible improvement to the existing descriptors.

1.5 Feature Selection

When we describe the timbre of a sound we use terms like bright or dark and usually terms related to the affect/emotion, a detailed description of the words we use for this reason deserve a specific research and is outside of the scope of

[4] http://www.synthfont.com/sfspec24.pdf

[5] http://freesound.org/docs/api/analysis_index.html

Author	Year	Descriptors
John M. grey	1977	Spectral energy distribution synchronicity in the transient of higher harmonics with the spectral fluctuation low-amplitude high-frequency energy in the initial attack segment Average pitch Δ ratio
Martin, Keith D. Kim, Youngmoo E.	1998	Pitch variance Pitch variance Δ ratio Average spectral centroid (Hz) Spectral centroid Δ ratio Variance of spectral centroid Spectral centroid variance Δ ratio Average normalized spectral centroid Normalized spectral centroid Δ ratio Variance of normalized spectral centroid Normalized spectral centroid variance Δ ratio Maximum slope of onset (dB/msec) Onset duration (msec) Vibrato frequency (Hz) Vibrato amplitude Vibrato heuristic strength Tremolo frequency Tremolo strength Tremolo heuristic strength Spectral centroid modulation Spectral centroid modulation strength Spectral centroid modulation heuristic strength Normalized spectral centroid modulation Normalized spectral centroid modulation strength Normalized spectral centroid modulation heuristic strength Slope of onset harmonic skew Intercept of onset harmonic skew Variance of onset harmonic skew Post-onset slope of amplitude decay Odd/even harmonic ratio
Lakatos	2000	- Spectral Centroid - Rise time - Tmbral 'richness'
Deng et al.	2008	MFCC(26). MPEG-7(7). IPEM(11)
Liu, Jing Xie, Lingyun	2010	16 attributes of STFT: Spectral centroid Spectral roloff Spectral flux Time domain zero crossing MFCC(52) Spectral crest factor(SCF, 96) Spectral flatness measure(SFM, 96)
Bhalke et al.	2015	Multi resolution plots(MRP) spectrogram images

Figure 1.2: Features set used in the referenced papers

this work. Anyway we argue that this higher-level description is related to our perception and our hearing system and can be recognised as caused by changes in the spectral and temporal characteristics of the sound. When we compute the low-level features that contains most of the temporal and spectral characteristics of a sound we do not have a direct relationship between them and our perceptual description. The attempt to find a correlation between those two worlds is one of the main challenges of MIR research and the distance between them is known as the semantic gap. The categorisation we are using here is strictly related to the timbre and the identification of the relevant perceptual descriptor is part of our goals.

The investigation on timbre opened the problem of identifying which content informations are relevant for our perception and can describe our sounds. The descriptor used as the three dimensions of the timbre spaces defined by Grey [Grey, 1977] and McAdams [McAdams et al., 1995], to cite two examples, are the outcome of this investigation. The results obtained through listening experiments are later analysed with algorithms that can extract the features that best represent the categorisation of the stimuli, various techniques are used to remove noise and understand better the correlation with timbre. The MDS is used in several studies, Grey [Grey, 1977] used the INDSCAL algorithm and Lakatos [Lakatos, 2000] use a new and improved MDS algorithm named CLASCAL. Another approach to accomplish this task is the use of Rough Sets considered the first non statistical analysis of data, the theory of rough sets has been presented by Pawlak the first time in 1982 and proposed again in Rough Sets Theory and its Application [Pawlak, 1998]. This is a simplistic description of this theory: a vague concept can be represented by two level of approximation, the lower approximation that contains all the characteristics that for sure are part of the concept and the upper approximation that contains some characteristics that can be part or not of the concept. This brief explanation seems to fit perfectly with the relation between content analysis, that represents the characteristics we have to choose, and the audio that in this scenario is the *vague concept*. In musical instrument classification rough sets have been used by Wieczorkowska [Wieczorkowska and Czyżewski, 2003] to perform feature selection and reduce her set of descriptors to a reasonable number of perceptual relevant characteristics.

The scope of this task cannot be restricted to reduce noise in the obtained data or reduce the number of dimensions used in the classification problem, but must offer an improvement to our knowledge on the relation between low-level descriptors and musical instrument sounds. In this work we will select a reduced set of features based on the acoustic models presented earlier in this chapter, only descriptors that match a specific characteristic of the model will be used.

1.6 Machine Learning

A well known method that gives good results in classification tasks characterised by multiple dimensions is SVM and in the literature we find many examples of its use for instrument sample classification. In 2010 Liu and Xie [Liu and Xie, 2010] used SVM to perform classification of a large dataset made of western and Chinese instruments. They used the SVM with different features set to classify into 4 instrument classes their dataset. The accuracy of the SVM combined with a specific set of features permitted to evaluate which was the feature set that best represented the instrument taxonomy. Nowadays other popular techniques exist for example in a recent attempts to address the musical instrument classification problem Bhalke [Bhalke et al., 2015] to exploit the capability of deep convolutional neural networks in image classification used spectrogram and multi resolution plot (MRP) images to classify his dataset of instrument samples. Also other methods are used for this task like k-NN, decision trees, naive Bayes and others.

The SVM is very popular in MIR research and at MTG it is well integrated in the common tool used. The Essentia library and Gaia are used together to perform classification of genre and mood where a high dimensional space is used in both tasks [Bogdanov et al., 2013]. We will use the SVM algorithm to take advantage of those well established techniques and concentrate our attention on the evaluation of the low-level features. The main use of this technique is for genre and mood classification and we want to prove the benefits of this approach for different tasks.

Chapter 2

METHODOLOGY

In this chapter we present all the tasks involved in the project, presented in chronological order to make the process as clear as possible and discuss on every aspect of the work. We present how we create the taxonomy, how we collect the audio samples to create the datasets then we present the tools used to analyse the sounds and how we worked on the feature selection and finally we describe the classification task ,which tools we used and some results. In this chapter we follow a logical path that resemble many studies on automatic classification of musical instruments and we try to clarify any of the personal decisions we made during the process.

2.1 Taxonomies

The selection of the set of musical instruments is related to the scope of the task as we have pointed out in the state of the art, our scope is to find a subdivision optimal in terms of how it can represent our acoustic models and how clearly each individual instrument fits into one model. Because of the parallel of the models with some western instruments like violin and organ the individual instruments are based on a subset of the western orchestra instruments. The root of the taxonomy is composed by three families grouped by combining the acoustic models and the modes of excitation, the taxonomy is represented in figure 2.1.

We generated this taxonomy for this study based on the work of Liu and Xie [2010] and on the acoustic models presented by Howard and Angus [2009] in their psychoacoustic book. In this model there is a direct connection between the individual instruments and the corresponding higher level class, that is representative of an acoustic model. We decided to keep this three family configuration that is similar to the one used by Liu and Xie in 2010 and then for the individuals we based on the western orchestra instruments a subdivision method used in many

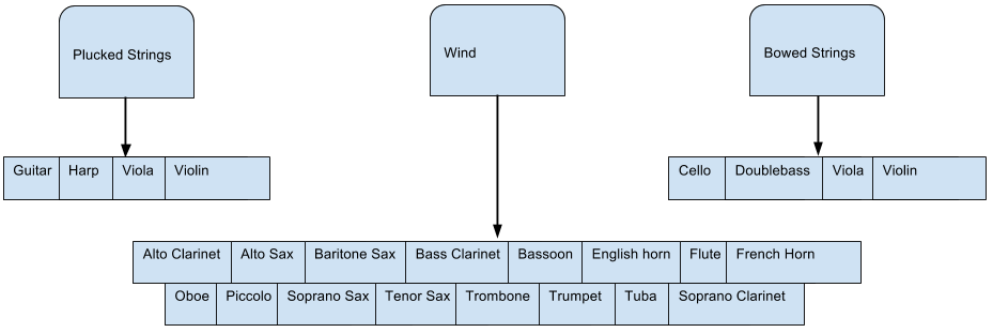

Figure 2.1: Musical Instrument Taxonomy

other studies. We also had in mind that the material present in Freesound could fit this taxonomy and we will see some consideration on this point when we will speak about the dataset creation.

2.2 Dataset Creation

For this project we decided to create two datasets, one based on Freesound content and the other based on the SoundFonts technology.

We want to use both real and synthesised sounds for our classification task because in the literature we found datasets composed by synthesised sound, used especially, but not only, in the early studies on timbre [McAdams et al., 1995] and datasets composed by recordings of real instruments like the IOWA MIS Dataset recorded in a controlled situation into an anechoic chamber. All this variability is useful to generalise our classifier and test different type of sounds that are present in our musical ecosystem.

We started with the assumption that Freesound contains mostly recordings of real instruments and we selected the SoundFonts technology to generate a dataset of synthesised sounds. In the next paragraphs we will review the details on how we built our datasets.

2.2.1 Freesound Dataset

One dataset used in this thesis is created using Freesound content. Freesound contains more than 300,000 sounds to be freely used according to their creative commons license. We extracted samples from a set of western orchestra instruments from the huge collection available according to our taxonomy. Freesound offers a

complete API to access all the relevant information present in the database and we created python scripts that helped us to identify and download the sounds for our dataset. The scripts are based on the python library developed at the MTG[1] that implement most of the services available in the Freesound API. Our approach to this task was very simple, we designed a python script that performs a text search on Freesound, we used only the instrument name in order to get every sound that contains that specific word in the name, in the description or in the tags and we downloaded all the retrieved sounds locally and we can see in table 2.1 the set of classes and the number of samples per class.

Instrument	Samples N°
Basoon	86
Cello	595
Clarinet	830
Doublebass	169
English Horn	1
Flute	582
Harp	79
Horn	62
Oboe	8
Trombone	5
Trumpet	405
Tuba	44
Viola	85
Violin	876

Table 2.1: Freesound Dataset Instrument Classes

Our goal was to create a ground-truth dataset of musical instruments single notes. Once we downloaded all the sounds we verified that the audio files were coherent with our necessity, to achieve this we listened the downloaded sounds one by one to eliminate the sounds that were not musical instrument and to separate the single notes from the musical phrases. After this work of manual annotation we can consider this dataset of musical instruments single note samples as our ground-truth.

During this process we identified some problems: we had some difficulties to match our taxonomy with the sounds available and we had to discard some classes due to the lack of sounds; the resulting dataset was not balanced, so we had to find a balance despite the decrease in size. The final dataset contains 14 classes and 3824 audio samples, most of them are real recordings with just a few synthesised

[1] https://github.com/MTG/freesound-python

Instrument	Samples N°	Instrument	Samples N°
Alto Clarinet	37	Harp	80
Alto Sax	30	Oboe	32
Baritone Sax	37	Piccolo	32
Bass Clarinet	37	Soprano Clarinet	35
Bassoon	40	Soprano Sax	30
Cello	49	Tenor Sax	30
Doublebass	35	Trombone	33
English Horn	26	Trumpet	42
Flute	36	Tuba	37
French Horn	43	Viola	36
Guitar	37	Violin	43

Table 2.2: SoundFonts Dataset Instrument Classes

sounds. We used the mp3 high quality previews in this work and we want to note that Freesound offer the possibility to download also the original audio files.

2.2.2 SoundFonts Dataset

The SoundFonts technology was created to play realistic sounds using MIDI, it is based on a wavetable synthesis to generate notes in the pitch range 1-127. We used a set of musical instrument samples from the MuseScore software[2]. For each instrument we generated a chromatic scale from C1 to D7 and then extracted only the notes in the specific range as we can see approximately in figure 2.2

The resulting dataset contains 849 samples divided into 22 classes visible in the table 2.3 *SoundFonts Dataset Instrument Classes* 2.2.2, that we used to train our models for classification. This is a set of single notes in wav format per each instrument.

2.3 Complete Audio Descriptors Set

We used part of the descriptors computed using the Freesound extractor (essentia_streaming_extractor_freesound), an extractor is a program that analyse an audio file and extract a set of audio descriptors available as algorithms in the Essentia library, there are many kind of extractors tailored for specific use and we select this one because it's already used in the context of Freesound. First the project aim to improve the usability of Freesound database so we want to use and test

[2]https://musescore.org/

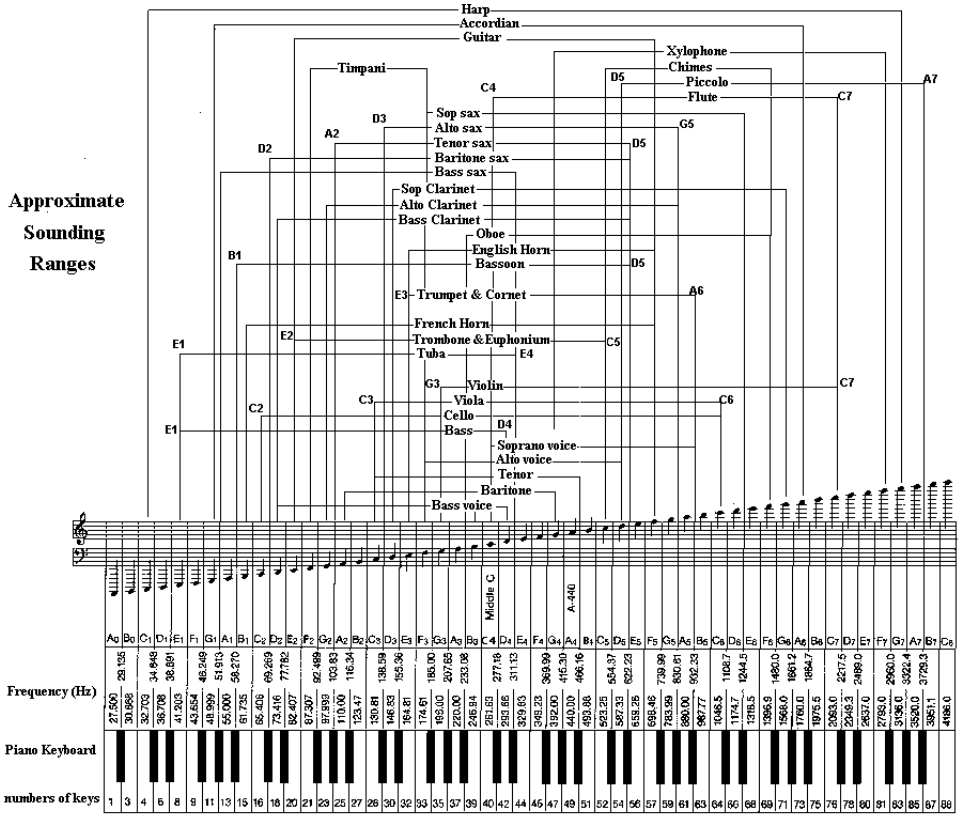

Figure 2.2: Musical Instruments Notes Range

the current tools used to extract the audio descriptors; second we are focused on musical instrument samples and this extractor has been designed to analyse this kind of audio material. Freesound contains many other kind of samples together with musical instrument samples so we had to select only some descriptors that are the ones contained in this two major groups:

- low-level descriptors (lowlevel)
- sfx descriptors (sfx)

We discarded tonal descriptors because we work with single notes and we are not interested in melody and chords; we also discarded rhythmic descriptors because we work with single notes and we think that those descriptor are not relevant to recognise different instruments. The complete list of descriptor can be found in the Freesound documentation[3].

[3]https://www.freesound.org/docs/api/analysis_docs.html

2.4 Audio Features and Physical Models

An acoustic model is composed of two parts, the sound source and the sound modifiers, when applied to a musical instrument those are referred to the vibrating body, the string or the air column and the resonator, e.g. the body of the violin, that enhances or attenuate certain frequencies. Both those components determine the spectral characteristics of the musical instrument sound. When we deal with real sounds the space in which the sound is produced play an important role in the modification of the spectral characteristics of the final sound, this effect is known as room acoustic. The specific reverberation of a space modifies the sound produced by the instrument in terms of its spectral components; the room acoustic can be considered as a second sound modifier similar to the body of the instrument defined earlier. The utilisation of perceptual scales as Bark bands, gammatone filter banks are some of the examples of MIR algorithms that try to combine our hearing system characteristics with the acoustic of the sound sources, our hearing system can be considered as another level of sound modification.

The string instruments is a big family and includes violin, viola, cello, double bass that can be played either plucking or bowing; while guitar, lute and harpsichord are plucked only instruments, at least for traditional playing techniques; we include also the piano in this family, but it is slightly different because the strings are excited by a struck. In musical instruments usually the length of a string is fixed and his tuning is adjusted changing the string tension. In the case of a string fixed at both sides we can calculate the frequency components of the sound using the position of the pluck in respect of the length of the string and the closer end. In this way we compute the nodes and antinodes produced by the stationary waves that traverse the string according to the pluck position using this formula:

$$ModenotExcided = m[L/d] \tag{2.1}$$

where:

- *m* are the modes (e.g. 1,2,3, ecc..)
- *L* is the length of the string
- *d* is the distance of the pluck from the closest end

This characteristic of the timbre can be expressed using a specific sound descriptors from our set of features: odd to even harmonic energy ratio[4]

This behaviour applies to all the plucked string instruments vibrating mass, but for the case of the piano we can add some other consideration. As we already

[4]http://essentia.upf.edu/documentation/reference/std_OddToEvenHarmonicEnergyRatio.html

said the piano strings are struck and not plucked as the ones of the harpsichord for example. When the string is struck it reacts as a bar because of its stiffness, and the physical effect of this characteristic is a slightly increasing pitch in the harmonics that follow the fundamental. We can describe this timbral characteristic as inharmonicity and we have a descriptor that computes this from the audio, the inharmonicity from the Essentia library computes exactly the signal spectral peaks deviation from the harmonic serie computed using the estimated fundamental frequency.

The bow creates another specific behaviour of the string, the hairs grip the string and move it from his rest position until the grip stops and the string return fast in its rest position. The movement of the string can be described as circular movement around the rest point and the periodicity of the movement determines the fundamental frequency of the produced note. The wave shape formed by this mode of excitation is similar to a sawtooth in which the amplitude of the harmonics is proportional to the number of the harmonic.

Another important characteristic of the sound is the spectral shape, as we've seen the amplitude of the harmonics is determined by the mode of excitation; when we move our attention to the sound modifiers defined in the acoustic model we see that they act as an acoustic resonant filter. In the string instruments the sound modifier is the resonating body; top and bottom plates vibrate and resonate to specific frequencies enhancing those upon the others. The air produces other resonances according to the size and the shape of the body in which it is contained. For those reasons it's reasonable to use some descriptors that capture the spectral shape of the sound like MFCC. Those descriptors are computed using the MEL scale for the computation of Cepstrum. The values clustered near the origin contain informations related to the modifiers while the values far from the origin contain information on the mode of excitation [Bhalke et al., 2015].

For a look into the wind instruments we start talking about the organ flue pipes that offer the possibility to explain the general mechanism that create sounds into a flue. The source of sound in this case is the flue that is the small obstacle that the air has to pass in order to flow into the pipe that act as a resonator. When the air is injected from the bottom it encounters the mouth of the pipe where the upper lip lets pass some air on one side and some on the other side creating a turbulent movement that with the increase of the air pressure become stable and produce the air flow that enter the pipe and resonates. The turbulence of the air produces noise, and this noise is audible and has to be considered as part of the sound so in this case analysing the sound and looking at the descriptors that try to capture the noisiness of a sound we find another match between the model and the descriptors the zero crossing rate, this algorithm used for the pitch analysis in the time domain can also measure the noisiness of a signal and can be an indicator of this behaviour of the flue pipes.

Audio Descriptor	*Acoustic Property*
Odd to even harmonic energy ratio	String pluck position
Inharmonicity	Stiffness of a string(e.g. Piano)
Waveshape (e.g envelope)	Mode of excitation (e.g bow = sawtooth)
Spectral shape (MFCC, GFCC)	Body and plates resonances
Noisiness	Air turbulence in flues

Table 2.3: Acoustic Models and Audio Descriptors Correlation

2.4.1 Other Audio Features

The GFCC is a synthetic representation of the spectral shape by means of few coefficients, it is computed as the MFCC, but instead of the MEL scale the ERB Gammatone Filterbank (improvement of the Bark bands defined by Zwicker) is used [Roma, 2015].

- The spectral centroid: captures the central frequency of the spectrum distribution and it is an important factor to discriminate sounds of instruments which have different tessitura.

- The spectral spread: measure the bandwidth of the spectrum and due to the characteristic of each instrument and the specific harmonics produced this descriptor is considered relevant.

- The log attack time: chosen for two reasons, first because it is determined by the mode of excitation and second because in the timbre space built by McAdams it is one of the three dimensions.

2.5 Classification

The automatic classification is entirely based on machine learning techniques that permit us to aggregate large amount of data and analyse them and extract some intelligence from those data. A general classifier is modelled using a specific flow of information shown in figure 2.3 and this flow is not based on the algorithm that we choose but is a common practice. Our goal in this project is to find a correlation between the audio descriptors and the acoustic models so we selected the Gaia library to build our classifier and we choose the SVM algorithm to train our model. SVM has been used in automatic classification of musical instruments with good results and it is one of the best models to support multidimensional spaces with many dimensions that is exactly the case of audio. As in many studies

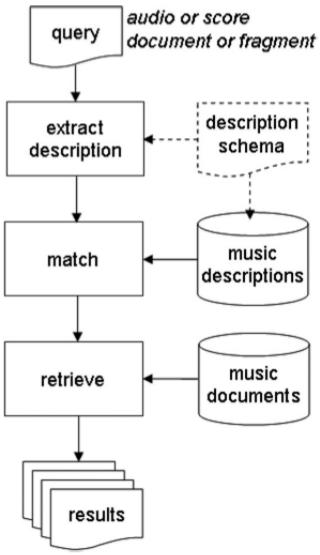

Figure 2.3: Classification Algorithm Model [Casey et al., 2008]

on classification our goal is to find the space with the smallest number of dimensions to represent the musical instruments and classify them and we try to achieve this goal selecting those descriptors that are relevant in the acoustic models. To evaluate our selection we started with a baseline composed by all the audio descriptors available and compared the accuracy of this classification with the one obtained using smaller sets of descriptors.

In this project we are using for the first time Gaia to classify musical instruments and we are combining the Freesound extractor with the SVM extractor (essentia_streaming_extractor_svm). The SVM extractor is used to classify new instances based on a previously trained SVM model; we encountered some problems to achieve this goal due to some inconsistencies in the audio descriptors extracted by the Freesound extractor.

We created a set of scripts that permit us to train Gaia SVM models with the data extracted using the Freesound extractor. The process we followed is very simple and can be described in a few steps:

- Analyse the audio files and create a folder structure to contain the extracted data
- Preprocess the data files and eliminate the unwanted descriptors as explained in section 2.3
- Convert the statistics files into the format readable from Gaia (YAML to .sig)

- Balance the dataset selecting the same number of statistics files per class (balanced only)
- Train the SVM model using the *sig* files and get the results

This process is applied to all the datasets we are using so we can compare the results of each of them because they follow the same training process.

2.5.1 Baseline

Our baseline is computed using all the *sfx* and *lowlevel* descriptors and give us the base accuracy that we want to reach using the small sets of descriptors that we designed according to the acoustic models.

Dataset	*Accuracy*
Freesound (FS)	94.33%
FS Balanced	80.05%
FS Families	97.22%
FS Fam. Balanced	86.25%
SoundFonts (SF)	85.55%
SF Balanced	99.22%
SF Families	96.49%
SF Fam. Balanced	97.26%

Table 2.4: Baseline Accuracy

2.5.2 SVM Model Training

The computational process to train a model is based on continuous reiteration on the dataset elements. The parameters of the SVM model are changed on every iteration and once a representative set of parameters has been used we select the model that gives the best accuracy and record it. We used a 10-fold cross validation to train the models changing the type of kernel and adjusting the gamma value per each kernel. This process is heuristic, but gives optimal results and at the moment we don't have any other way to select which parameters fit best with a specific set of data. In this study we used a C-SVC type of SVM algorithm with two types of kernels

1. Polynominal
2. RBF

For every kernel we used a set of C values and Gamma values in every possible combination

1. C: [-5, -3, -1, 1, 3, 5, 7, 9, 11]
2. gamma: [3, 1, -1, -3, -5, -7, -9, -11]

Every iteration is computed with a 10-cross fold validation as mentioned before.

2.5.3 Features Sets

To close this chapter we present a list of feature sets used in this study, some are related to our approach using the acoustic models, other are tested because of their relevance in the literature(e.g. McAdams Timbre Space) and in previous research at MTG.

FEATURES SETS:

```
Baseline: 61 (number of descriptors)
  lowlevel: 40 descriptors
            Total number of lowlevel features: 285
  sfx: 21 descriptors
            Total number of sfx features: 39

Acoustic Model 1: 10
  Spectral centroid
  Spectral Spread
  Inharmonicity
  Oddtoevenharmonicenergyratio
  tristimulus

Acoustic Model 2: 36
  Spectral centroid
  Spectral Spread
  Inharmonicity
  Oddtoevenharmonicenergyratio
  Tristimulus
  MFCC

Acoustic Model 3: 16
  Spectral centroid
  Spectral Spread
  Inharmonicity
```

```
  Oddtoevenharmonicenergyratio
  Tristimulus
  logattacktime

MFCC: 26
  MFCC Mean
  MFCC Variance

GFCC: 26
  GFCC Mean
  GFCC Variance

MFCC + GFCC: 52
  MFCC Mean
  MFCC Variance
  GFCC Mean
  GFCC Variance

McAdams Timbre Space: 6
  spectral_centroid
  logattacktime
  spectral_flux
```

Chapter 3

CONCLUSIONS AND FUTURE WORK

The results of the classifier are interesting and show that the small sets of descriptors give good accuracy in the classification of musical instruments. If we look at the table we observe that the accuracy obtained identifying the instrument family is really high for any set of descriptors used so our first conclusion is that we can build a hierarchical classifier that first separate the dataset into families and then each family is classified using a different set of descriptors that is relevant for its specificity. This hierarchical classification process is strongly encouraged after the results obtained, the very high accuracy obtained in the identification of the instrument family is a relevant result not only in this study but even in similar research projects[Liu and Xie, 2010].

Dataset	Baseline	A. Model 1	A. Model 2	MFCC	GFCC	MFCC+GFCC
Freesound(FS)	94.33%	95.04%	95.23%	93.89%	94.82%	95.98%
FS Balanced	80.05%	**63.56%**	82.44%	80.05%	81.64%	82.44%
FS Families	97.22%	97.60%	97.47%	97.47%	96.97%	97.88%
FS Fam. Bal.	86.25%	86.77%	93.38%	93.38%	92.21%	92.99%
SoundFonts(SF)	85.55%	**65.41%**	81.33%	83.05%	79.86%	83.61%
SF Balanced	99.22%	**64.54%**	82.06%	82.06%	79.68%	83.61%
SF Families	96.49%	92.08%	95.99%	96.11%	95.40%	97.64%
FS Fam. Bal.	97.26%	88.01%	93.15%	92.46%	92.12%	91.43%

Table 3.1: Classification Accuracy

The first column represent our reference to measure how our classifier works, the high accuracy obtained is probably due to overfitting because we are using a large set of descriptors compared to the number of samples in the datasets. Starting from the second column we used reduced sets of descriptors to avoid overfitting and try to identify the relevant features to classify musical instrument

samples. The set *Acoustic model 1* shows nice performance in most of the cases, but the *Freesound Balanced* and the *SoundFonts balanced and unbalanced*. The *Freesound Balanced* dataset is built random selecting samples from the complete dataset, we observed that the results where changing according to the sample selection and the variance in the results was higher than the other cases. If we look at the models we see that the algorithm parameters selected in each run where different, the selected kernel is always RBF but *gamma* and *C* values are different. In this study we used a method to discover the best performing algorithm, on every training of the classifier we where using a set of parameters to fine tune our SVM model, but the risk in which we occurred is that our model is biased to our data. From this observation we argue that we need to improve the quality of the data present in the dataset and increase the number of samples for the smallest classes, in fact while the Freesound dataset complete contains more than 3800 samples the balanced one is really smaller and is set to 440 samples in total.

The performance of the classifier with the SoundFonts dataset and the Acoustic model 1 feature set that are in the second column, *65.41%* on the unbalanced dataset and *64.54%* on the balanced one, can be explained analysing the data. The confusion matrix in figure 3.1 show that the errors occur mostly on Saxophone and clarinet instruments.

Accuracy: 65.555555556

Predicted (%)

	alto_clarinet	alto_sax	baritone_sax	bass_clarinet	bassoon	cello	doublebass	flute	french_horn	guitar	harp	oboe	piccolo	soprano_clarinet	soprano_sax	tenor_sax	trombone	trumpet	tuba	viola	violin		Proportion
alto_clarinet	10.00	0.00	0.00	66.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	16.67	6.67	0.00	0.00	0.00	0.00	0.00	0.00	alto_clarinet	4.76 %
alto_sax	0.00	50.00	36.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33	6.67	0.00	3.33	0.00	0.00	0.00	alto_sax	4.76 %
baritone_sax	0.00	29.33	53.33	0.00	3.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	16.67	0.00	3.33	0.00	0.00	0.00	baritone_sax	4.76 %
bass_clarinet	56.67	0.00	0.00	20.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	10.00	6.67	0.00	3.33	0.00	0.00	3.33	0.00	bass_clarinet	4.76 %
bassoon	3.33	0.00	0.00	0.00	70.00	0.00	0.00	0.00	0.00	0.00	10.00	0.00	0.00	0.00	0.00	0.00	13.33	0.00	3.33	0.00	0.00	bassoon	4.76 %
cello	0.00	0.00	0.00	0.00	0.00	70.00	30.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	cello	4.76 %
doublebass	0.00	0.00	0.00	0.00	0.00	20.00	80.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	doublebass	4.76 %
flute	0.00	0.00	0.00	0.00	0.00	0.00	0.00	90.00	0.00	0.00	6.67	0.00	13.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	flute	4.76 %
french_horn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	french_horn	4.76 %
guitar	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33	0.00	96.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	guitar	4.76 %
harp	3.33	0.00	0.00	0.00	0.00	0.00	0.00	13.33	0.00	0.00	66.67	0.00	16.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	harp	4.76 %
oboe	0.00	0.00	3.33	3.33	3.33	0.00	0.00	0.00	0.00	0.00	0.00	83.33	0.00	3.33	0.00	0.00	3.33	6.67	0.00	13.33	0.00	oboe	4.76 %
piccolo	0.00	0.00	0.00	0.00	0.00	0.00	0.00	16.67	0.00	0.00	10.00	0.00	70.00	3.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	piccolo	4.76 %
soprano_clarinet	26.67	0.00	0.00	90.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33	3.33	3.33	20.00	3.33	0.00	0.00	6.67	0.00	3.33	0.00	soprano_clarinet	4.76 %
soprano_sax	13.33	0.00	0.00	3.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33	6.67	56.67	0.00	3.33	10.00	0.00	3.33	0.00	soprano_sax	4.76 %
tenor_sax	0.00	13.33	13.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.67	0.00	0.00	0.00	50.67	0.00	3.33	0.00	3.33	3.33	tenor_sax	4.76 %
trombone	6.67	0.00	0.00	0.00	3.33	0.00	0.00	0.00	0.00	0.00	0.00	3.33	0.00	0.00	0.00	3.33	83.33	0.00	0.00	0.00	0.00	trombone	4.76 %
trumpet	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33	6.67	6.67	0.00	0.00	83.33	0.00	0.00	0.00	trumpet	4.76 %
tuba	3.33	0.00	0.00	0.00	6.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.33	0.00	0.00	86.67	0.00	0.00	tuba	4.76 %
viola	3.33	0.00	0.00	13.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	13.33	0.00	0.00	0.00	0.00	0.00	3.33	0.00	66.67	0.00	viola	4.76 %
violin	0.00	0.00	6.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	93.33	violin	4.76 %

Actual (%)

Figure 3.1: Confusion Matrix of Acoustic model 1 applied to the SoundFonts balanced dataset

The SoundFonts technology uses a wavetable oscillator on a real instrument sample and changes the frequency rate to obtain different pitches, in the case of

those two instrument the sample of the original sound used to generate the notes is the same and this explains why the algorithm miss the correct class for the samples that overlap in the pitch range.

In the column 4 and 5 we put the accuracy of a model trained using MFCC and GFCC, we decided to compare those two descriptors because they are similar in terms of computation, the same algorithm is applied on a MEL scale in the computation of MFCC and a Bark scale in the case of GFCC and from a conceptual perspective in fact both use a perceptual scale to preprocess the spectral information. What we observed is that their accuracy is almost the same, MFCC perform better than GFCC but the difference is far from being relevant. The sum of the two descriptor does not give better accuracy and the increment in the number of the descriptor is much more relevant.

3.1 Content Descriptor Evaluation

The content description approach gives good results in terms of accuracy and permit to discover if a correlation exists between a specific descriptor and a characteristic of a sound. In our approach we went a step further and tried to match a sound characteristic with a specific physical property of a musical instrument. The results are interesting in terms of accuracy and we think that this kind of correlation can be further investigated in order to minimise the error in the classification performance based on content description. The content evaluation will not solve completely the task and combining this approach with the context information already available in Freesound in the form of tags, descriptions and user comments will increase the accuracy of the classifiers.

3.1.1 Odd to Even Harmonic Energy Ratio

During the work we identified some descriptors that describe the behaviour of the acoustic model, one of those descriptors is the one that measures the ratio between the odd and even harmonics. Once we computed this descriptor for our datasets we discovered some inconsistent values in the statistics because many samples where discarded because of the presence of infinite (*Inf*) on null (*NaN*) values in some statistical measures. The preprocessing task of the Gaia library checks if the values in the dataset are consistent and usable in the training of the model, NaN and Inf values are discarded and this behaviour claimed our attention on this specific descriptor and we investigated to understand the reason for those values to be present in the statistics. We identified some implementation inconsistency and fixed it with a workaround.

The solution implemented is designed to avoid division by zero, the ratio between the even harmonics energy and the odd harmonics energy was clipped to the maximum float value allowed in the C++ language when the even energy was equal to zero. This approach make sense to avoid the error in the computation but results in an inconsistent value in the computation of some statistics, e.g. the mean; adding together the maximum float allowed gives a *Inf* value. To avoid the *Inf* value we clipped the value to 1000 that is a reasonable energy ratio that represent the fact that the even energy is irrelevant compared to the odd energy. This workaround solved the problem of the discarded samples and permitted to generate balanced dataset with the maximum number of samples available. The changes to the algorithm are visible in the branch oddtoevenharmonicenergyratio_fix[1] of the Essetia github repository.

One of the assumption of this work was that this methodology can be used to identify weakness in the algorithms used to compute the audio descriptors and what we described earlier in this paragraph is a first prove that this methodology serves this scope.

3.2 Multiple Datasets and Generalisation

The risk in the automatic classification using a specific dataset can be identified in two main aspects:

1. The overfitting
2. The impossibility to generalise

The overfitting problem compares when we have a complex model where the number of samples is too small compared to the number of parameters. This can held to very high accuracy in the results and we can prevent this error keeping the number of parameters small compared to the size of the dataset. In practice we can reduce the number of variables we are using to classify our samples until the moment we see the performance decreasing. If our model is overfitting it hardly generalise to other data so another way to understand if we are in this situation is to classify another set of data using the trained model and look at the results. The two problems are strictly connected to each other and we tried a to work on the classification task using different datasets to validate the trained models. Unfortunately we couldn't perform cross-validation using our datasets because we should first fix some problem in the descriptors algorithms, so a good continuation of this work is to analyse all the extracted data, identify other problems as the one in the oddtoevenharmonicenergyratio and fix them to make possible to use the SVM

[1] https://github.com/hellska/essentia/tree/oddtoevenharmonicenergyratio_fix

extractor with the Freesound extractor data. Another line of investigation is the association of every descriptor in the Freesound extractor with a specific characteristic of the acoustic models and maybe identify which characteristics are not covered by our algorithms.

Bibliography

Bhalke, D. G., Rao, C. B. R., and Bormane, D. S. (2015). Automatic musical instrument classification using fractional fourier transform based- MFCC features and counter propagation neural network. *Journal of Intelligent Information Systems*.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J. R., and Serra, X. (2013). ESSENTIA: an Audio Analysis Library for Music Information Retrieval. *International Society for Music Information Retrieval Conference (ISMIR'13)*.

Casey, M., Veltkamp, R., and Goto, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the ...*, 96(4):668–696.

Deng, J., Simmermacher, C., and Cranefield, S. (2008). A Study on Feature Analysis for Musical Instrument Classification. *Systems, Man, and ...*, 38(2):429–438.

Font, F., Roma, G., and Serra, X. (2013). Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia - MM '13*, pages 411–412, New York, New York, USA. ACM Press.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *The Journal of the Acoustical Society of America*, 61(5):1270.

Herrera-Boyer, P., Peeters, G., and Dubnov, S. (2003). Automatic Classification of Musical Instrument Sounds. *Journal of New Music Research*, 32(1):3–21.

Howard, D. and Angus, J. (2009). Acoustic Model for Musical Instruments. In *Acoustics and Psychoacoustics*, chapter 4, pages 167–230. Elsevier Ltd. All, fourth edi edition.

Kartomi, M. (2001). The Classification of Musical Instruments: Changing Trends in Research from the Late Nineteenth Century, with Special Reference to the 1990s. *Ethnomusicology*, 45(2):283–314.

Krumhansl, C. L. (1989). Why is musical timbre so hard to understand? In Nielzen, S. and Olsson, O., editors, *Structure and perception of electroacoustic sound and music*, volume 9, pages 43–53. Elsevier, Amsterdam.

Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception & psychophysics*, 62(7):1426–1439.

Liu, J. and Xie, L. (2010). SVM-based automatic classification of musical instruments. *2010 International Conference on Intelligent Computation Technology and Automation, ICICTA 2010*, 3:669–673.

Livshin, A. and Rodet, X. (2003). The Importance of Cross Database Evaluation in Sound Classification. *Society*, pages 1–2.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58(3):177–192.

Pawlak, Z. (1998). Rough set theory and its applications. *Journal of Telecommunications and Information Technology*, 29(7):7–10.

Roma, G. (2015). Algorithms and representations for supporting online music creation with large-scale audio databases.

Scalcon, F., Rocchesso, D., and Borin, G. (1998). Subjective Evaluation of the Inharmonicity of Synthetic Piano Tones. *International Computer Music Conference*, (1):1–4.

Schedl, M., Gomez, E., and Urbano, J. (2014). Music Information Retrieval: Recent Developments and Applications. In *Foundations and Trends in Information Retrieval*, volume 8, pages 127–261.

Vaquero, C. P. (2012). *Improving the description of instrumental sounds by using ontologies and automatic content analysis*. PhD thesis.

Wessel, D. (1979). Timbre Space as a Musical Control Structure. *Computer Music Journal*, 3(2):45–52.

Wieczorkowska, A. and Czyżewski, A. (2003). Rough Set Based Automatic Classification of Musical Instrument Sounds. *Electronic Notes in Theoretical Computer Science*, 82(4):298–309.

This document is powered by LaTeX