

Improving the description of instrumental sounds by using ontologies and automatic content analysis

Carlos Vaquero Patricio

MASTER THESIS – UPF 2012

Master in Sound and Music Computing

August 26th, 2012

Master Thesis Supervisor:
Dr. Xavier Serra

Master Thesis Co-supervisor:
Msc. Frederic Font

Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona

to Luna and Teo.

Copyright: © 2012 Carlos Vaquero Patricio. This is an open-access document distributed under the terms of the Creative Commons Attribution License 3.0 Unported, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Acknowledgements

I would, first of all, like to thank my thesis supervisor, Prof. Xavier Serra, for giving me the opportunity to follow the SMC Master's studies and join the MTG. Also for all the things I learned from him and for showing an extraordinary exercise of patience and kindness in all the meetings we had.

I must thank very much my thesis Co-supervisor, Frederic Font, and Gerard Roma, who have helped me since the very first day of this thesis in any of the technical or conceptual problems that were encountered through.

I should also thank Mohamed Sordo, for his availability to help whenever I would need him, Hendrik Purwins and Perfecto Herrera, for clarifying so many concepts in the design or in the understanding of the techniques being applied, and any of those at the MTG and Phonos that have contributed to make my studies successful, either by teaching a good lesson, having an inspiring conversation, helping with homework problems or giving me a good smile.

There are no words to express my gratitude to my parents, Pilar and Carlos, sisters, Cristina and Elena, my extended family and my partner, Sonja Gruijs. Your love and support has been very much needed during this time and you couldn't have given me more of it. Any success I should make is to be shared with all of you.

I would also like to thank Fundació "La Caixa" for the financial support received during the last two years, giving me the opportunity to grow so much in a professional and personal level in such a wonderful city as Barcelona.

Abstract

Browsing sound collections in a social database is a complex task when no uniformity in the classification of the sounds and tags is applied; the relation between tag concepts and sounds can vary extremely from one user to another, as well as the types of sounds associated to them.

Social databases are environments that allow users to have many different approaches and classification systems when tagging their contents, since a sound can be described and tagged according to different characteristics and with different perspectives and purposes. Browsing through such databases is often complex and inaccurate, returning results are often very distant in either their sound characteristics or the tags being used.

Collaborative sound databases are a perfect environment to study the problems derived from an inaccurate or a multidimensional description. Using an ontology as the basis of the classification tags applied to sounds, may not only ease the browsing of sounds through the collection, but also help to define common definitions within the community of its users.

This thesis defines a methodology to build a sound collection by using a proposed ontology of tags and the content analysis of its sounds. A corpus of 700 samples has therefore been recorded, classified according to a designed ontology, integrated in the database and analyzed. In addition, similarity measures between content based descriptions and semantic descriptions of this sounds is defined by extracting six different models, providing the possibility of automatically describing eventual new sounds to be integrated within our collection. Finally, the proposed models are evaluated within three different experiments and a preliminary survey of expert users acceptance.

Contents

Acknowledgements.....	3
Abstract.....	4
Contents.....	5
1 Introduction.....	7
1.1. Motivation and Context of the Thesis.....	7
1.2. Scope of the thesis	9
1.3. Contributions	9
2 Literature review	10
2.1. Organology.....	10
2.2. Audio Signal Processing.....	13
2.3. Machine Learning	20
2.3.1. K-Nearest-Neighbor	22
2.3.2. Support Vector Machines.....	23
2.3.3. Decision Tree	24
2.3.4. Other classification techniques	25
2.4. Feature Selection and Projection	25
2.4.1. Resampling of decompensated classes	25
2.4.2. Manual Feature Selection	26
2.4.3. Feature projection – PCA	26
2.4.4. Standardization of the features.....	27
2.5. Evaluation measures	27
2.6. Browsing Sound Collections.....	30
2.6.1. Freesound.....	30
2.6.2. Free and Commercial Instrumental Sound Collections Alternatives.....	31
2.7. Differences between Ontology and Taxonomy	32
3 Methodology	34
3.1. Ontology Design within the framework of Freesound	34
3.2. Corpus Building	39
3.2.1. Samples recordings	40
3.2.2. Samples selection	41
3.2.3. Natural Language Description.....	43
4 Experiments.....	44
4.1. Experiment 1 – Self built Data set	44
4.2. Experiment 2 – Testing in Freesound	48
4.3. Experiment 3 - Ontology Models Precision	50

4.4. User Survey and Proof of Concept	52
5 Conclusions and Future work	54
5.1. Conclusions	54
5.2. Future Work	55
References	57
Appendix A: A review of Audio Signal Processing basic concepts	62
Appendix B: Ontology.....	65
Appendix C: The Music Ontology.....	70
Appendix D: Decision Trees.....	72
Appendix E: Ontology nomenclature	76

CHAPTER 1

Introduction

1.1. Motivation and Context of the Thesis

In 1936, Varèse decides to call his music “organized sounds” as a reaction to the traditional differentiation between noise and music (Varèse & Wen-chung, 1966). Extending this definition, we can say that music happens whenever transmitted or perceived sounds are organized either at the sender or receiver perspective of a communication process.

Organizing sounds is also needed in the process of manipulating sounds, whatever the purpose of the collection of them will be, and it implies the challenge of how to create associations among them.

In the history of humanity, several classification systems of musical instruments have been used according to aspects such as the cultural and social roles of the sounds, the relations established between player and the instrument, the perceptual characteristics of the sounds, or the physical aspects or resonating properties of the musical instruments.

Freesound¹, a collaborative database of sounds hosted and maintained by the MTG-UPF, is the perfect environment to explore some of the problems classification of sounds. Having at this moment more than 2 million registered users, 150.000 sound recordings and 5000 users that have uploaded sound recordings, we can observe several different approaches in tagging and browsing sounds, as well as many of the cultural differences in their description.

From this variety of sounds and tags, browsing sounds becomes inaccurate; the uniformity among the tags in sounds is as diverse as the amount of uploaders tagging them. It is not rare to have situations in which existing sounds in the database are hardly ever found because the tags and description being used are inaccessible with the words with which these are being browsed. We are therefore confronting a rather challenging database system.

The question that arises at this point is whether users and social networks communities could eventually benefit from proposing an automatic instrument classification system (eventually

¹ <http://www.freesound.org/>

extended to tag propagation²) by using the available technologies today, and by means of an effective methodology to design and implement them within a determined context.

The CLOSED project (Houix et al., 2007) combined sound perception and classification studies trying to resume some of the most general characteristics. Some of the conclusions of this project can be resumed in:

- A classification done on a free basis, without any hint about the origin of the sound, implies many different strategies, sometimes according to individual schemes.
- Verbalization and semantic classification occurs more when the sounds have fewer possible causes, whereas association of sounds with high values in acoustic similarities is very effective.
- Expertise plays a role in the easiness with which sounds are classified. Naive participants are effective classifying sounds when they can identify them whereas experts use their acoustic “education” to classify sounds.
- Mirroring physical production qualities of the sound is effective: participants of the experiments are able to classify sounds according to the classes solid, liquid, gas and electric sound

Other studies (Srinivasan, 2002) show that human perception in instrument classification might be problematic since accuracy in timbre association to musical instruments can vary from 28% to 90% depending on individual training and surrounding-environmental circumstances. This accuracy will obviously become worse as we approach instrumental sounds with non-conventional performance techniques.

As we can observe from the above references, the description variety of sounds is mainly a cultural and perceptual issue.

Also, the approach in which sounds are categorize implies many issues; from a creative and research perspective the name given to the technique with which an instrument is played might eventually be as relevant as the instrument with which is being played, for instance a pizzicato might be easier to identify as a name than the instrument with which is being played. An ontology of concepts related to the content analysis of sounds might, therefore, solve many of these relations and let us observe different tendencies in social databases tagging.

² Tag propagation is the collateral effect of having a social database in which tag recommendation is combined with collaborative tagging establishing relations of new tags proposed by users. (Sordo, 2012)

1.2. Scope of the thesis

The goal of this work is proposing a *proof of concept* to improve browsing in Freesound by using an automatic classification methodology based on a proposed ontology and machine learning techniques combined with analysis.

After developing an ontology based on a corpus of instrumental sounds, and comparing its results to data extracted by using signal processing, different models for automatically tagging sounds are being derived.

Much of the historical background in organology and the state-of-the-art in automatic classification of monophonic instrumental sounds, linked to extended practices in signal processing and machine learning techniques, will be reviewed along this work, offering an overview of the areas that sound databases using automatic classification could benefit from.

1.3. Contributions

- 700 new isolated instrumental samples properly recorded, segmented and classified available at Freesound under the attribution non-commercial creative commons license.
- A OWL ontology designed in Protégé to be extended and included in any other ontologies being developed or available.
- A study of three different classification techniques within the proposed ontology.
- Six different classification models using J-48 decision trees for further implementation
- A methodology to record and classify according to the proposed ontology instrumental sounds.
- A dataset available at MTG/UPF for research purposes combining samples from different sound libraries.
- A study of the presence of tags related to the proposed ontology in Freesound.
- A comprehensive state-of-the-art review of monophonic isolated instrumental samples and a possible use of ontologies related to them.

CHAPTER 2

Literature review

In this chapter a study of the different traditions in instrument classification as well as the state of the art techniques that can be used for this purpose will be presented.

An overview in Organology, Signal Processing, Semantic Web, Machine Learning Techniques as well as Browsing systems is offered as the basis for the proposed methodology to be developed afterwards.

2.1. Organology

Organology is the science that studies musical instruments in terms of their history and social function, design, construction and relation to performance.

The access to knowledge about Organology in history within different cultures has influenced considerably the evolution of the aesthetics of music, conditioning the different classification techniques of music instruments.

Even though the influences among different cultures might be evident, and the classification of music instruments among them might share many properties, the attempt to design a *categorical* ontology, valid for every culture and type of instrument, will impose a perspective in the way the classification is conceived.

Finding relations among different classifications will, on the contrary, incline the user to recognize and discover essential characteristics of its different cultures and properties that might enrich the identity of each of the taxonomies encompassed. This might be a necessary complementary perspective when studying relevant musical aspects such as rhythm, melody or structure (Liu, 2009).

Through history we can find several precedents in instrument classification:

Chinese and Hindu Cultures:

Around 2233 BC, in China, a classification based in the material construction of different instruments was used. These categories are metal (*chin*), stone (*shih*), silk (*ssu*), bamboo

(*chu*), gourd (*p'ao*), clay (*t'u*), leather (*ko*), and wood (*mu*), and they are linked to the 8 seasons and 8 winds of Chinese culture (Kartomi, 2001).

In Hindu culture, chapter 28 of the Sanskrit treatise “*Natyasastra*” (200 BC) we can observe the following categories: “stretched”, “covered”, “hollow” and “solid”. Besides their organological properties these categories correspond as well to different hierarchical status, in society or in music, and to their different soloist (major limbs) or accompaniment (minor limbs) roles, determining in a great way essential aspects of Carnatic music.

In both Chinese and Hindustani antecedents, we can see that the instrument classification developed obeys to characteristics that reach contextual and cultural aspects of other relevance rather than the exclusive acoustical properties of their instruments.

Praetorius:

Michael Praetorius (1571 - 1621) wrote between 1614 and 1620 one of the most extensive surviving treatises of the XVII century, the *Syntagma Musicum*. On it many of the “timbre” characteristics to organize western and non-western instruments is already employed. His work is considered a milestone in instrument classification and a very relevant resource for musicologists and performers of early music from late XVI century.

Mahillon:

In 1888, Victor-Charles Mahillon, curator of the “Conservatoire museum” in Brussels, published a catalogue of the instruments collection that was extended in 1890 to develop much of the classifications of Idiophones.

The division that Mahillon made consists of four broad categories according to the sound production material: air column, string, membrane, and body of the instrument. As we can observe, these categories resemble in several aspects those ones of the *Natyasastra* but it is also considered as the basis of the one developed by Sachs and Hornbostel.

Hornbostel and Sachs:

Erich Moritz von Hornbostel joined Curt Sachs in 1909 to work together for the “Berlin Phonogram-Archiv”. Five years later, in 1914, they published in the “*Zeitschrift für Ethnologie*” (Hornbostel-Sachs, 1914) a classification system proposal that has been used and extended till nowadays, being the predominant system in most of the music instrument museums around the world.

The original Hornbostel and Sachs system divided the instruments in four main categories depending on the physiology or the sound production:

- Aerophones (air column)
- Chordophones (string)

- Idiophones (own bodies)
- Membranophones (elastic membrane)

In 1961 this classification had a major contribution made by Anthony Baines and Klaus P. Wachsmann with the addition of a fifth category (Pegg, 2012):

- Electrophones (electrical means)

The main reason why this system was embraced by so many users is that, with a numerical method, it allows to find easily the category as well as to classify it without getting too restrictive within cultural or linguistic frontiers (Kartomi, 2001) This easiness made of it the predominant classification system in most of the music instrument museums around the world through the XXth century.

In the following lines an example of how this classification would work for a Transverse Flute (HS 421.12) it is demonstrated:

4 – Aerophones

Instruments where the sound is primarily produced by vibrating air.

42 - Non-Free Aerophones

Instruments where the vibrating air is contained within the instrument. Most of the instruments “wind instruments”, such as flutes and bugles, are included in this group, but as well “odd” instruments such as conch shells.

421 - Edge-Blown Aerophones

The player's breath is directed either by the player or by the instruments against a “splitting edge” that causes the air to vibrate.

421.1 - Flutes without a duct

The player's lips direct a stream of air to the splitting edge.

421.12 - Side-blown flutes

The player blows against the sharp rim of a hole in the side of the tube.

The Hornbostel and Sachs system has been criticized mainly because of the different uses it makes of non-western idiophones. Some instruments cannot follow the tree structure as they belong to two or more levels of different categories (*e.g.* Kalimba). As a solution, Elschek proposed a system combining Hornbostel and Sachs with a bottom-up approach based on instrument attributes (Elschek, 1969). Similar limitations can be observed when trying to classify Javanese musical instruments.

After Hornbostel and Sachs:

André Schaeffner, in 1932, proposed a system trying to cover “all real and conceivable instruments (Kartomi, 2001)”. Schaeffner classified instruments depending on their vibrating capabilities, solving efficiently some of the problems encountered when trying to classify some percussion instruments within the Hornbostel and Sachs system.

With the appearance of computers and the developments in audio signal processing, the classification of instruments immediately gained more possibilities. Yet, the main challenge nowadays relies in the approach in classifications; giving a bigger relevance to cultural or acoustical properties and, ultimately, finding whether the design of the classification comes from a bottom-up or top-down perspective. The advances in technology and development of social networking databases might bring together different dimensions in any of these approaches from which similarities among them could be derived and being able to find some interesting relations and commonalities.

2.2. Audio Signal Processing

Computers can be used to distinguish among the different layers contained in a sound that, once summed, have significance to us ³. This can be done by means of the features that describe a sound. There are different steps to consider when extracting and analyzing features.

Figure 2.1 illustrates one of the most complete frameworks explained in (Peeters, 2004) including aspects such as modeling, or feature computation that, will be treated in following chapters.

According to Schedl, features can be categorized between two types (Schedl, 2008):

- **Context-based:** ID3-tags (genre, artist, album, title), user defined tags, information from Web pages (“cultural features”).
- **Content-based** (audio signal-based): energy, pitch, beat, rhythm, harmony, timbre, and melody. Extracted after doing signal processing and analysis capturing the essential characteristics of an audio signal.

Literature also usually distinguishes between **Low Level** (content based) and **High Level** features (context based) (Casey et al.2008); often a **Mid Level** category, in which the relations among Low and High cannot be separated, is also used.

³ In Appendix-A a review on the basic concepts to understand audio signal processing is presented. Being acknowledged with these concepts is essential to understand the rest of this subchapter.

In our particular case of study, Mid-level description is done individually with each of the tags proposed by our ontology or the users and the analyzed content linked to it.

Low-level features can be expressed in as many ways as signal processing allows. The analysis of low-level features and their comparison in different dimensions (instantaneous, segment or global modeling) is what makes feasible bringing the audio signal extracted into a higher or mid-level description of the musical domain, closer to the semantic representation of it, or its symbolic meaning.

Some of the features that have been most extensively used in this work and are considered most relevant are named below. For a more complete description of their use or design, the reader can refer to (Peeters, 2004; Herrera, 2006; Tzanetakis & Cook, 2002; Klapuri, 2004; Fuhrmann, 2012):

Information over frame boundaries from a sound can be extracted from the time-domain signal. This type of features can be classified as Temporal Features:

- Root Mean Square - related to the energy of the signal.
- Log Attack Time - how fast is the attack of a sound.
- Zero Crossing Rate - harshness of sound, measured after rating the sign-changes along a signal (Tzanetakis & Cook, 2002); helps to distinguish between periodic sounds, low ZCR, and noisy sounds, high ZCR.
- Temporal Centroid - normalize by decay of the sample, good to differentiate long from short decay sounds (Peeters, 2004).

Information over the spectral content of the signal is obtained by doing different transformation such as the DFT (using the FFT when possible) or STFT (can also be multiresolution). Some of them are:

- Spectral Centroid - where the “center of mass” of the spectrum is; good for finding out how bright the sound is (Tzanetakis & Cook, 2002).
- Spectral Kurtosis - flatness of a distribution around its mean value (Peeters, 2004).
- Spectral Flux - Measures how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame (Couvreur et al., 2008).
- Spectral Spread - variance of a distribution around its mean value; timbre characterization.
- Spectral skewness - measure of the asymmetry of a distribution around its mean value; timbre characterization.

- Spectral roll-off frequency under which some percentage of the total energy of the spectrum is contained; to distinguish between harmonic and noisy sounds (Tzanetakis & Cook, 2002).
- Barkbands - algorithm extracting the 28 Bark band values of a Spectrum and doing different aggregations. Among them all the spectral representations mentioned. It used to recognize perceptual description of sounds, since the scale ranges from 1 to 24 and corresponds to the first 24 critical bands of hearing (Zwicker, 1961).
- Mel Frequency Cepstrum Coefficients – MFCC are a compact representation of the spectral envelope, originally speech oriented procedure very useful to detect timbre of instruments (Klapuri and Davy, 2006). It allows the deconvolution of source and filter taking the logarithm of the amplitude spectrum, scaling it to Mel-scale bands and finally doing a Discrete Cosine Transform of it. MFCC's were originally proposed by (Logan, 2000) and they have been extensively used since then overall and in the models that we will propose later, since they have proven to be very efficient as “all-purpose” descriptors within our classification. The first MFCC is normally discarded in literature due to its high correlation with the signal power. High MFCC are correlated with pitch and when the focus of the classification would be timbre they could be discarded.
- Tristimuli - are 3 different types of energy ratio: the first value corresponds to the relative weight of the first harmonic, the second to that of the 2nd, 3rd, and 4th harmonics, and the third to the weight of the rest. Is an equivalent concept for timbre to the three primary colors of vision (Pollard et al., 1982).

Analysis parameters such as window size (related to the framing) or type, overlap factor between frames from different signals or size of the FFT (when used) must be the same when comparing features from different signals since they will determine the trade off in frequency vs. time resolution.

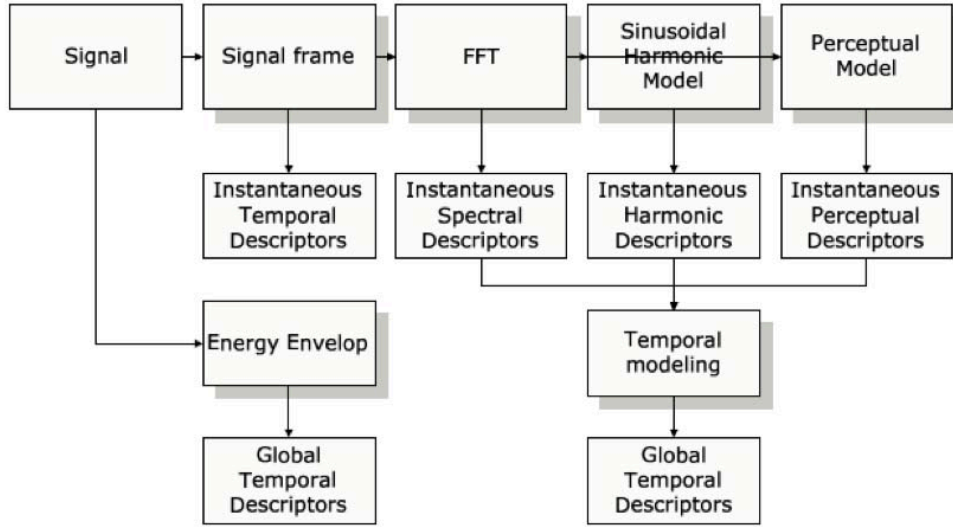


Figure 2.1: CUIDADO project features extraction framework (Peeters, 2004)

To illustrate the use of Low Level Features lets see a practical example:

A good way to separate a violin producing a pizzicato sound from one producing a tenuto sound would be to differentiate them by measuring the value distances between the Log Attack Time of them in combination to the spectral mean centroid.

The Log Attack Time (LAT) is one of the most important perceptual descriptors. It is as well a clear example to observe how different windows and hop sizes influence in the results of the values. The log attack time can be obtained by:

$$lat = \log_{10}(stop_{attack} - start_{attack})$$

The spectral centroid (SC) indicates where de center of the mass is. In it, the spectrum frequencies are related to the distribution's values and its magnitudes to the observation probabilities. The centroid describes the distribution's barycenter:

$$centroid = \frac{\sum_{i=1}^N X_i f_i}{\sum_{i=1}^N X_i},$$

where X_i is the value of the feature and f_i is the center frequency value of FFT bin i .

In Figure 2.2 violin sounds having a tenuto attack (sustained and soft) or a pizzicato attack (short and sharp) are classified according to two features Log Attack Time (LAT) and Spectral Mean Centroid (SMC).

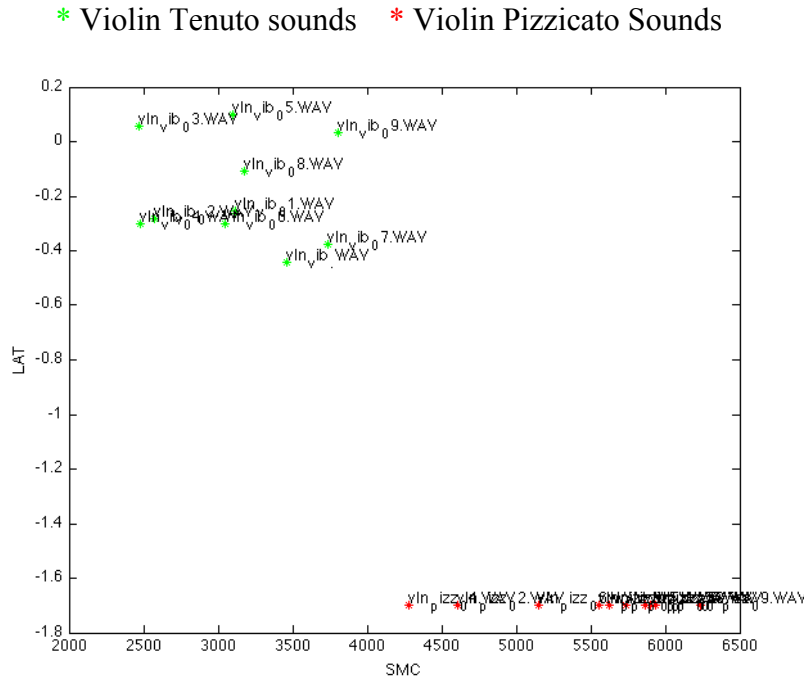


Figure 2.2: Automatic Classification using SMC and LAT descriptors

As we can see in Figure 2.2, the classification works very effectively. Sounds that are pizzicato have a high spectral centroid mean and a low log attack time, whereas sounds with a tenuto attack have the opposite. The fact that the value of the spectral centroid in the pizzicato samples is higher is partly due to the high frequency components of the (fast) attack of the plucked string.

The difference in the x-axis location of the sounds is due to the fact that the pitches are different. The overlapped values in the pizzicato part are repeated pitches (notes).

Applying transformations to the descriptors (such as logarithm or normalization of their values) can help to obtain a better separation among them as well as standardize results.

Many other combinations of descriptors are possible in order to classify sounds correctly. While a selection of the most well known behavior of descriptors for certain sounds is possible and used, in the next chapter, three techniques to find the best “combinations” of them is explained.

According to Wessel, timbre is, after pitch and loudness, one of the most important properties to differentiate sounds. Timbre is often referred to as the “color” or quality of sounds (Wessel, 1979). An interesting approach related has been done by McAdams, who proposed classification of timbre in spaces, to help to define the differences in their perception. To define them similarity judgments between pairs of sounds is combined with multidimensional scaling (responding to perceptual space) and audio analysis of the descriptors (McAdams, 1995).

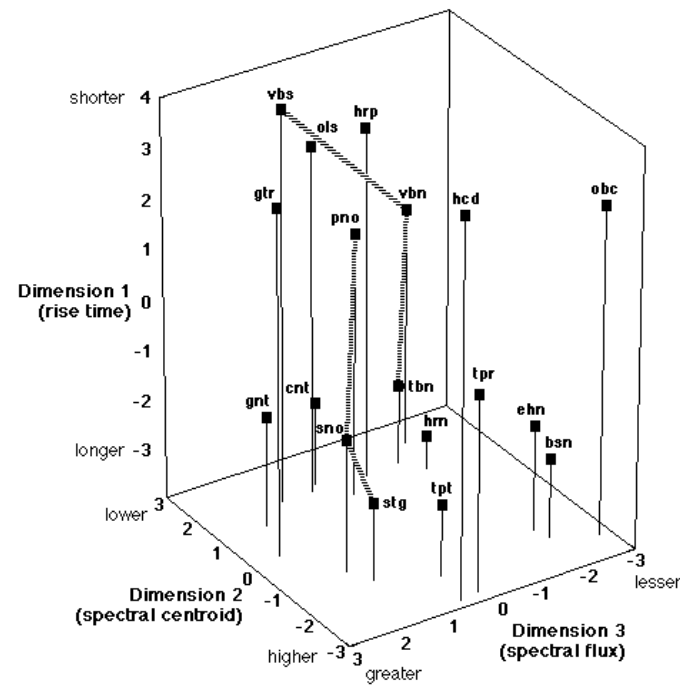


Figure 2.3: Timbre spaced obtained by (McAdams, 1995)

Essentia:

Essentia is the core library that analyzes each sound being uploaded to Freesound’s database allowing, within different formats, finding characteristics related to the sounds as well as similarities among them ⁴.

Essentia provides a reusable collection of algorithms and descriptors mainly used to extract features from audio files. By using these algorithms or the library tools, different descriptions

⁴ <http://www.freesound.org/docs/api/resources.html?highlight=essentia>

related to features classified in the categories Metadata, High-level, Low-level, Rhythm, SFX or Tonal can be obtained.

The content analysis of Freesound is therefore determined by the efficiency of Essentia as well as the proper use of the implemented descriptors in it.

The Essentia descriptors that will be used for this thesis are those contained within the SFX and Low-level descriptors. In total they are 517 different attributes (including transformations of the descriptors) that have been used for this thesis work.

The results that each of the descriptors in Essentia returns are averaged over the length of the whole sample and is represented as a histogram of values. This means that analyzing the temporal evolution and envelope of the signal is impossible. Some of these descriptors, yet, can give a clue of the envelope within time for short signals and when knowing the content by looking at their histogram (e.g. the skewness).

Analysis parameter selection in Freesound:

The analysis parameters implemented in the Extractor function of Essentia are not adaptable to sounds depending on their content. This is set like that as users are expected to upload either deterministic or stochastic sounds of very different qualities.

Window type-size, hop-size cannot be manipulated *a posteriori* and, since we are working in the environment of Freesound, the analysis to be carried as well as the descriptors to be used should be coherent with the preset parameters of Essentia's extractor.

For this reason, comparisons in the content-based perspective of the sounds has been made using the same extractor parameters.

Two common frameworks in content description:

We shouldn't finish this subchapter without briefly mentioning two relevant and extended frameworks in Audio and Music Analysis and, specifically, in Low Level content descriptors.

The **Cuidado project** (2003) suggested a large set of audio features implemented including those related to the temporal shape, temporal features, energy features, spectral shape features, harmonic features and perceptual features and an extraction framework (Peeters, 2004).

The **MPEG-7** contains a set of low-level tools designed to provide the basis for the construction of higher-level audio applications. Approved at the end of 2001 it has the intention of providing complementary functionality to the previous MPEG standards, representing information about the content, not the content itself ("the bits about the bits").

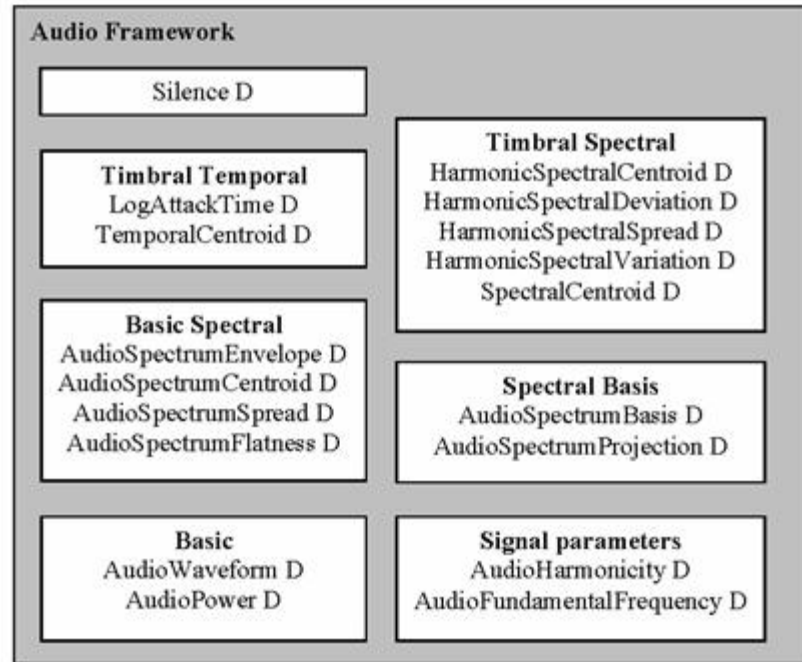


Figure 2.1: Low Level Descriptors used in the MPEG-7 Audio Framework (MPEG7)

2.3. Machine Learning

Computers can be programmed to make associations between tags and sounds by using classification techniques in which sound processing and machine learning techniques are combined. In multimedia, the interest on automatic classification of sounds arises in the 90's decade, proposing a solution to the already mentioned humans difficulty to discriminate instrumental timbres.

Automatic Tagging refers to the task of automatically classifying sound samples with respect to high-level concepts from diverse music facets such as Emotion, Musical instruments, Genre or Usage (Marques et al, 2011).

Automatic Tagging is made using Machine learning, the artificial intelligence discipline by which computers can infer and evolve behaviors based on empirical data. The most common techniques for instrument classification based on content description are explained in detail in (Herrera, 2006) and (Fuhrman, 2012).

Other studies show that “content-based retrieval systems can not classify, identify and retrieve as well as humans can (Sordo & Celma, 2007)”. Consequently, they propose a system for tag recommendation⁵ having an improvement of 38% over a 40% annotated collection.

Automatic classification methods are the basis of tag recommendation. A further use of the models suggested by this work could be the adoption of tag recommendation systems in Freesound.

Supervised Learning:

Supervised Learning is an approach to machine-learning by which a system is able to infer a function from labeled training data. Having a ground truth of labeled data, a model can be obtained after training it according to a defined technique or algorithm. Supervised learning is expressed by:

$$y = g(x | \theta)$$

$g()$ is the Supervised learning model being used.

θ are the parameters to be solved

x is the input

y is the output

The value of the output y can be a discrete value, if we are considering classification, or a continuous function of real-valued elements, in the case of regression.

In automatic tagging, x is defined as the acoustic features extracted from the audio and the output, y , will be the tag defined by the vocabulary.

Our supervised learning examples will normally contain an input object (our audio descriptors vector) and a desired output value (the tags belonging to our ontology).

When extracting features from the training set and deriving similarity measures according to machine learning techniques, a rule can be derived from which we can classify a new observation into the existing classes.

⁵ Tag recommendation in music refers to the task in which after the content of some sound excerpt has been analyzed the system will suggest a number of tags based on content similarity to an existing database.

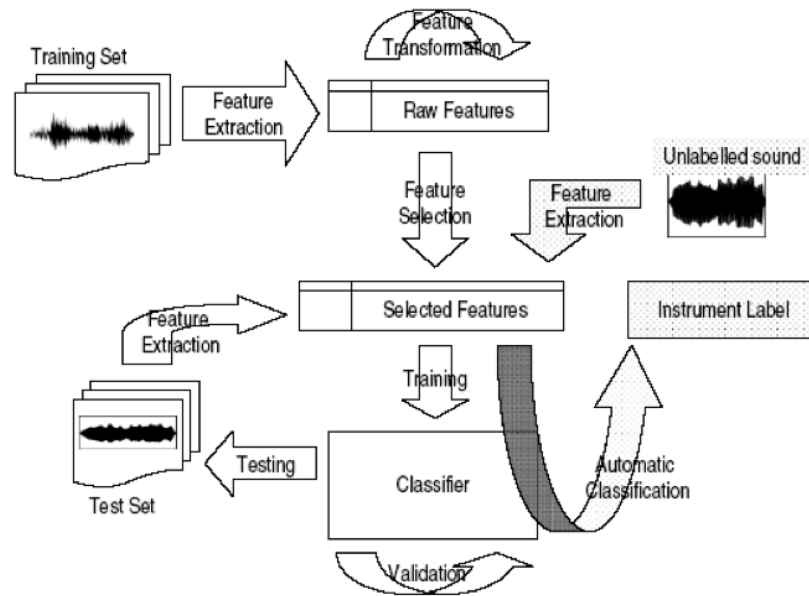


Figure 2.6: Operations involved in setting up an automatic classification system for musical instrument sounds (Herrera, 2006)

Weka⁶ is a standard software tool and library that allows trying a great number of algorithms for machine learning and visualize their behavior. To test in a fast and confident way different solutions Weka has been used within this thesis, being able to decide which classification technique would be most suitable for our purpose.

The techniques that have been tested for our system are:

2.3.1. K-Nearest-Neighbor

K-Nearest Neighbor is one of the simplest methods for instance-based learning. Being a lazy learning algorithm is only approximated locally and all computation is postponed till the classification task is completed. Each instance is classified by a majority vote of k nearest neighbors, previously trained, around the feature space. Therefore the more examples of one class will conform it. This can be a problem as irrelevant features might become relevant when dominating the distance metrics.

Cano et al. used k-NN with results of around 85% (955 audio files) in 6 classes of harmonic instruments using 89 features and a taxonomy based on Wordnet (Cano et al, 2004).

Wordnet⁷ is a lexical database that can be used as a lexical ontology due to the relations

⁶ <http://www.cs.waikato.ac.nz/ml/WEKA/>

⁷ <http://wordnet.princeton.edu/>

between conceptual categories derived from the type of relations among the noun synsets (one or more synonyms).

As Cano points it out, one of the best advantages of using NN classifiers is that they do not need to be retrained when a new class of sounds is added to the system (Cano et al, 2004). The main problem of this technique is that in a social database as Freesound in which sounds are being uploaded constantly the classifiers might not be consistent for possible new values description.

Another recent study using k-NN in monophonic instrument recognition has been carried by Yu & Slotine. The difference in their approach is that they used instead of acoustic features, sample blocks according to different scales and then are convolved with the test spectrogram in each time- frequency points. The minimum difference is then stored at the equivalent feature vector position. The accuracy obtained with this method was of 85.5% for a seven instruments classification (Yu & Slotine, 2009).

2.3.2. Support Vector Machines

The support vector machines (SVM) are supervised learning models with associated learning algorithms that can be used for classification and regression. The SVM find the maximum margin hyperplane separating two classes of data.

If the data are linearly separable, the best hyper-plane is the one that separates them in such a way that the distance from it to the closest points is maximum.

If the data are not linearly separable in the feature space they can be projected into a higher dimensional space by means of a kernel, $K(x_i, x_j)$. Often used different kernel functions are Linear, Polynomial, Radial Basis Function or Sigmoid.

Only the inner products of the data points in this higher dimensional space are necessary, so the projection can be implicit if such an inner product can be computed directly. The space of possible classifier functions consists of weighted linear combinations of key training instances in this kernel space (Cristianini & Shawe-Taylor, 2000).

The SVM training algorithm chooses these “support vectors” and weights to optimize the margin between classifier boundary and training examples (Mandel & Ellis, 2005).

“Essid et al. (2006b) evaluated SVM classifiers together with several feature selection algorithms and methods, plus a set of proposed low-level audio features. The system was applied to a large corpus of monophonic recordings from 10 classical instruments and evaluated against a baseline approach using GMM prototype models. Classification decisions were derived by performing a voting among the classifiers’ predictions along a given decision length. Results showed that SVMs outperformed the baseline system for all tested parameter variants and that both pair-wise feature selection and pair-wise classification strategies were

beneficial for the recognition accuracy. Moreover, longer decision length always improved recognition performance, indicating the importance of the musical context (i.e. the integration of information along several instances in time) for recognition, as similarly observed in perceptual studies (Fuhrmann, 2012).”

2.3.3. Decision Tree

A *decision tree* is a hierarchical model for supervised learning in which the local region is identified in a sequence of recursive splits in a smaller number of steps. A decision tree is composed of internal decision nodes and terminal leaves (Alpaydn, 2010).

The output from the decision tree algorithm is a table of decisions that is obtained after the algorithm has verified which attribute is the most representative in each of the leaves and which values should each node have. These decisions are made by testing all possibilities and finding a probability distribution of all possible classifications.

To be able to have an efficient model that will be valid for other datasets and avoid over fitting, many attributes have to be discarded. With this goal decision trees are combined with **pruning**, a technique that removes those sections of the nodes that provide poor values to classify instances.

It might happen though that by setting the pruning too low some important attributes could be discarded and overall accuracy classification will be worse.

Decision trees are one of the most simple and machine learning techniques but, in comparison with other more sophisticated methods such as SVM or HMM, they provide explicit data structure very intuitive for human understanding (Breiman, 1984). For this reason, and the little difference in results previously explained, they have been chosen as the main technique to be used in our proposed models. The type of binary tree used on this work is J-48 (open source implementation of the C4.5).

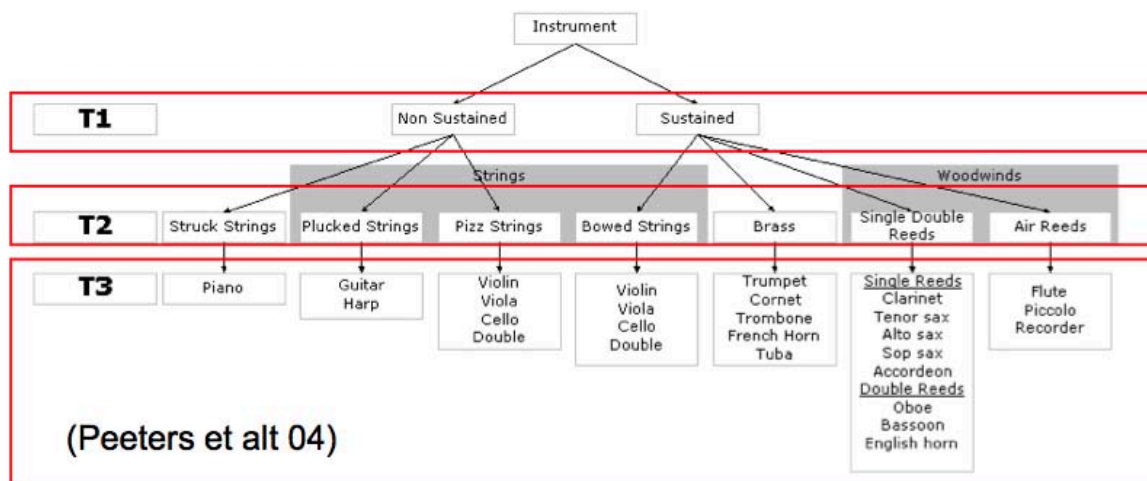


Figure 2.2: Decision Tree of Musical Instruments Classification (Peeters et al, 2004)

2.3.4. Other classification techniques

Linear Discriminant Analysis, Hidden Markov Models, Artificial Neural Networks or Gaussian Mixture Models are other techniques commonly used in MIR that have not been tested on the presented work.

Table 2.2 presents a relation of the State of the Art experiments up to 2006. Conclusions are difficult to be drawn as the number of instances, classes and techniques vary from one study to the other. Yet it can be observed that the less classes being trained, the more precise results are obtained (similar difficulty to human perception). Obviously, the dataset used will also have an incidence in the modeling of the classification system.

2.4. Feature Selection and Projection

Freesound's API allows the possibility to access the analysis results of the feature extraction done by Essentia. This analysis attends to Temporal features, Spectral features and Harmonic features. High Level, Tonal and Rhythm Descriptors have not been considered for analysis, as the sounds we are aiming to classify are isolated.

517 descriptors, corresponding to the Low Level and SFX categories of Essentia's library have been extracted.

One of the goals of the work here presented is finding out the minimum number of features possible to classify, by content, sounds belonging to our ontology in the studied context. Gaining efficiency and simplicity in the different nodes models is a direct derivation of the outcome of it.

2.4.1. Resampling of decompensated classes

The first step to have a model that will work efficiently is having an equal number of sound samples per class. A different number of instances per class, when doing a classification among the classes, will vary enormously the results and create a biased model.

To avoid this problem the re-sampling function of Weka has been applied when the instances would differ by more than the 5% between the different classes.

Resampling produces a random subsample of a dataset using either sampling with replacement or without replacement. In Weka, the bias to Uniform Class method can be adjusted; in the cases in which we have needed to use resampling, this parameter has been set to 1.0.

2.4.2. Manual Feature Selection

Features have been, in most of the cases, manually selected to obtain an optimal number of them. This selection has been done attending to references to be found in the State of the Art literature, Freesound documentation and Advice from different experts at the MTG.

As each of the Ontology nodes attends to different relevant sound characteristics, a study of each of the 517 attributes has been done.

Avoiding over-fitting in the attribute selection has been also taken into account by using the ten to one rule (>10 instances /1 attribute rule) in the predicted performance of the model to be developed, and to gain more generality when using it in other test sets (Rubens, 2011).

Another well-known technique for Feature selection is “Attribution selection”, by which the features that are most correlated are chosen after choosing different evaluator and searching methods.

Feature transformations such as Square root, Log, Inverse or Arcsine-root are often used to improve the Gaussianity implicit in many classification techniques.

Essentia has different statistics available in each of the features to be able to obtain the most relevant aspects that differentiate between sounds are the arithmetic mean, maximum value, minimum value, variance, mean of the derivative, mean of the second derivative, variance of the derivative, variance of the second derivative, covariance and inverse covariance.⁸

2.4.3. Feature projection – PCA

Dimensionality reduction is needed as manual selection becomes less intuitive. The difficulty in classifying increases proportionally with the number of classes treated, it is, therefore, to be expected that we will encounter a bigger difficulty to classify certain Ontology nodes in the selection of features that could differentiate a certain class - type of sounds.

This has been observed in the nodes (Extended) Techniques and String Instruments, in which manual selection has not provided an accurate performance and a dimensionality reduction of the feature space was needed.

Principal Component Analysis (PCA) is a common technique used in dimensionality reduction that linearly decorrelates a number of correlated variables by using an orthogonal transformation. This transformation is done by maximizing the variance of the projected data, combining features that reduce the dimensionality of data, resulting each combination in a different component.

⁸ A view of the models and Features chosen per node can be found on Appendix B.

No limit in the number of features has been defined when creating the different components; we therefore have set the argument of maximum Attributes in Weka to -1.

2.4.4. Standardization of the features

Standardization of the features takes the statistic mean of the training set values of each descriptor and divides it by its deviation.

In comparison to normalization it doesn't apply a Gaussian curve over the values so outliers might cause a problem since they will set the mean in a non-desired point.

The instances being used for this experiment are all supervised and the recordings count with rather homogenous characteristics in each of the three different sound libraries used, so values are supposed to not count with a great deal of outliers.

A problem with using Standardization or Normalization of the training set is that, to keep coherence in training and testing results, the values obtained for the mean and deviation from the training set must be taken into account when using the testing set or in the implementation of the model.

2.5. Evaluation measures

To evaluate the performance of our proposed classifiers we need to know how accurately the sounds being uploaded are recognized and correctly tagged by our system.

This performance can be obtained by looking at the accuracy of the classification, by measuring it with Precision, Recall and F-measure (Olson et al., 2008) as well as the relevance and ratio of the tagged sounds by using the truthfulness/falseness of the null hypothesis in hypothesis testing (Lehmann & Romano, 2005).

Table 2.1: Contingency table

	Null hypothesis (H0) is true	Null hypothesis (H0) is false
Reject null hypothesis	Type I error False Positive	Correct outcome True Positive
Fail to reject null hypothesis	Correct outcome True Negative	Type II error False Negative

Precision, measures the ratio of predicted classes that are relevant. The equation is given by:

$$P = \frac{TP}{TP + FP}$$

Recall, measures the ratio of relevant classes that were predicted. The equation is:

$$R = \frac{TP}{TP + FN}$$

Accuracy,

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

Ten-fold cross-validation, allows the use of the same data set for training and for testing an algorithm. The original sample is randomly partitioned into 10 subsamples. Of these, a single subsample is retained as the validation data for testing, and the remaining 10 – 1 subsamples are used as training data. The cross-validation is then repeated 10 times (the *folds*), with each of the k subsamples used exactly once as the validation data. The 10 results from the folds then can be averaged (or otherwise combined) to produce a single estimation.

The strongest point of the cross-validation method, over repeated random sub-sampling, is that all observations are used for both training and validation, and each observation is used for validation exactly once (Witten & Frank, 2005).

Table 2.2: State of the art of classification
of isolated instrumental sounds in 2006 (Herrera, 2006)

Author, year [ref]	Total instances ^a	NC	Acoustic features ^b	Classifi- cation algo- rithm	Instru- ment perform- ance (%)	Family perform- ance (%)
Eronen, 2001 [174]	5286 (MUMS, Iowa, SOL, RolandXP30, own recordings)	29	MFCC (attack-steady), F0, ATT, onset features, SC, Crest Factor, AM	<i>k</i> -NN	35 (H: 30)	77 (H: 75)
Livshin et al., 2003 [413]	4381 (SOL, Iowa, MUMS, Prosonus, Vitous)	16	SC, ATT, temporal decrease, TRI, HD, SKW, KUR, SV, SS, MFCC, noisiness	LDA & <i>k</i> -NN	47-69	62-92
Peeters, 2003 [511]	4163 (SOL, Iowa, MUMS, Microsoft MI, Prosonus, Vitous)	23	same as above	LDA & GMM (hierar- chical)	54 (H: 64)	81 (H: 85)
Eronen, 2003 [175]	5895 (MUMS, Iowa, SOL, Martin, own recordings)	7	MFCC, delta-MFCC + ICA	HMM	68	n/a
Kitahara et al., 2003 [345]	6247 (RWC)	19	SC, OER, F0 relative energy, KUR, SKW, FM, amplitude envelope slope, onset energy	Bayes (<i>k</i> -NN after PCA & LDA)	80	91
Kostek et al., 2004 [362]	n/a (CMIS, MUMS)	12	Wavelet-based energy bands, MPEG-7 features	MLP	71	n/a
Szczuko et al., 2004 [615]	2517 (CMIS, MUMS)	16	MPEG-7 features, OER, F0	MLP (2-stage hierarchi- cal MLP)	86 (H: 89)	n/a
Park et al., 2005 [496]	829 (several commercial instrument-sample CDs)	12	SS, SC, harmonic slope, LPC noise, harmonic expansion/contraction, spectral jitter and shimmer, spectral flux, TC, ZCR	MLP with el- liptical/ radial basis functions	71	88
Ch��try et al., 2005 [88]	4415 (Iowa, RWC, voice)	11	Line spectrum frequencies	K-means derived codebook	95	n/a

2.6. Browsing Sound Collections

2.6.1. Freesound

Freesound is a collaborative database of Creative Commons Licensed sounds in which people from different disciplines and with different purposes can share sample recordings and synthesized sounds.

The last data collected shows that in Freesound are more than two million registered users that have access to more than 150.000 samples and that 5000 users are active uploaders (around 80% of the users have uploaded less than 20 samples).

One of the main challenges of data classification in any social database is improving the efficiency in the searching process. As each of the users provides very different classification techniques or uses distinct perceptual tags, the vocabulary of words is extensive and the aggrupation of terminologies as well.

According to Mika's studies in Social Networks Semantics, "Ontology creation necessitates a social presence as it requires an actor to reliably predict how other members of the community would interpret the symbols of an ontology based on their limited description. (Mika, 2007)". An ontology of tags, related among them by associative and parent-child relationships, might help to define concrete areas of knowledge and extend it to other users.

In multimedia, if the content analysis is linked, the use of an ontology may improve significantly the process of browsing and have many creative or research applications after having a correct coherence in the similarities derived from the audio content (Roma et al., 2010).

In Freesound the mentioned challenge in searching is still very present; it is not rare to type in the searcher an instrument of a specific timbre (*i.e.* tag: "guitar") and retrieve another one of which it's content has very little to do with the one being searched (*i.e.* piano). The reasons for this are various:

- The architecture of Freesound is composed by different technologies and resources that determine the efficiency of it (**Figure 2.3**). Amongst the most relevant related to search are GAIA and SOLR⁹.
- GAIA is a statistic tool used by Freesound that has the function of finding similarities between the content of the sounds. In Freesound the similarities derived when browsing sounds are divided from an average analysis of all the descriptors between different sounds, a relevance match of the tags, title and description. The order of the results obtained when doing a query is ordered in

⁹ <http://lucene.apache.org/solr/>

weight descending order depending on the number of downloads. This tag similarities preset punishes in a great manner the positioning of new uploaded samples and it is an accumulative problem that could be improved by implementing a vector model (Baeza, 2011).

- The fact of dealing with creative commons licenses determines in a great way the use of it as well as the average quality of the samples. This quality homogeneity influences very much in the results obtained when trying to derive similarities among samples. Many of them are grouped under packs that have very different descriptions.

In (Font, 2012) we can observe that amongst the most used tags in Freesound there are a great number that describe environmental recordings and *concrete* sounds. There is also an increasing number of music recordings and spoken documentation.

Freesound has also recently incorporated the possibility of using geotags querying which is a very interesting way of browsing that could be of many research applications in cultural domains by combining MIR tools within an ontology of sounds tagged by places.

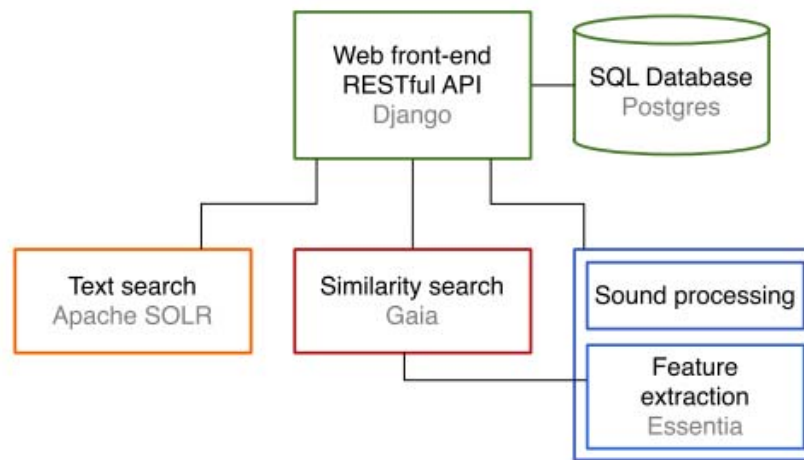


Figure 2.3: Freesound architecture framework (Akkermans et al, 2011)

2.6.2. Free and Commercial Instrumental Sound Collections Alternatives

Other previous works and references in automatic sound classification systems can be found in the studio-online, musclefish, audioclas or soundfisher. The only online resource nowadays is the soundfisher, but it is designed as a search engine production environment and,

therefore, it does not provides a collection of samples that would be of use to train our system.

Other examples of information management systems than Freesound are available as research or commercial sound collections and community databases. Some of these have been used in our experiments as part of the train sets.

Non-commercial sound libraries:

The London Philharmonia Orchestra Library and the Instruments samples collection of the University of IOWA have an attribution non-commercial alike license extensive collection of symphonic instruments samples that cover some of the nodes presented by our ontology. For this reason this material can be of a big interest when defining the machine learning models that will be used for machine tagging.

These libraries have, however, some restrictions in the use to be made of their samples. Among them the impossibility to upload them to Freesound even when maintaining their license and attribution. The classification terminology is similar to the one proposed by our ontology in the dynamics, octave division, notes, sustain and some of the attacks.

Commercial sound libraries:

The Vienna Symphonic Library is probably the most extensively used sound sample library in industrial and media productions due to its very complete and high quality samples. The classification system used to tag their sounds has many parallelisms with the one being proposed by our ontology.

2.7. Differences between Ontology and Taxonomy

The boundaries between using one and another term have been the cause of misunderstanding in the different communities in which they are commonly used. To understand them within this work it is necessary to establish clear definitions and differences between them.

A **Taxonomy** is a collection of controlled vocabulary terms organized into a hierarchical structure. Each term in a taxonomy is in one or more parent-child relationships to other terms in the taxonomy (Pidcock, 2003). A Taxonomy is normally defined by a “tree” structure, in which the hierarchical relations between the nodes are explained by the “is-a” subset definition.

According to Gruber "an **ontology** defines a set of representational primitives with which to model a domain of knowledge or discourse (...) about individuals, their attributes, and their relationships to other individuals (Gruber, 2009)".

A formal **Ontology** is a controlled vocabulary expressed in an ontology representation language. This language has a grammar for using vocabulary terms to express something meaningful within a specified domain of interest (Pidcock, 2003)". Their relationships can be associative, in addition to the parent-child relationships referred by a Taxonomy.

Ontology deals with questions concerning what entities exist or can be said to exist, and how such entities can be grouped, related within a hierarchy, and subdivided according to similarities and differences. The definition of Ontology implies the possibility of encompassing several Taxonomies; if a Taxonomy could be seen as a tree structure, an ontology could be understood as a forest.

Ontologies in philosophy have been traditionally approached in two ways (Marcela Sanchez, 2002) both can be associated to a cognitive approach:

- 1- obeys to the intrinsic reason to identify objects as "being" or not. For that the object being observed (studied) must be reflected and compared with all the other things in the universe and try to assign words to it. Our cognition is ruled by a bottom-up approach
- 2- looks for the essence, this will be on the top of all the beings, the "thing" in common ontologies *jargon*, and it will allow giving a hierarchy to all the beings. By dividing beings in high and low levels, a distinction between general properties and common characteristics can be established. Our cognition is ruled by a top-down approach.

An ontology is normally broader in scope than a taxonomy and it is also usually referred when the aim is representing knowledge "in a way that computers could derive meaning by traversing the various relationships among its concepts. Therefore, the application scenarios vary in how the ontology itself will be used" (Uschold, 2009).

The ultimate goal of search engines in social databases is narrowing the semantic gap by finding relations between high and low level descriptions of its content (whatever the content they deal would be).

For practical reasons in the methodology, an strictly definition of ontology has not being developed, but rather a taxonomy that might serve as a basis for further development of an ontology when other users would add different tags that will be interrelated with the tags already proposed.

For the above reason, using the Ontology terminology seems therefore as the most appropriate within the domain of study here presented.

CHAPTER 3

Methodology

The goal of our study is improving the browsing in Freesound; in this chapter we resume the methodology to set up the base to conduct three different experiments that will be presented in Chapter 5 and categorize and upload them within Freesound. With this purpose, the steps to be followed include an ontology design, a corpus building and an integration of both within Freesound's database.

The ontology design is a very important part of this work since it is going to be the ground truth of tags being applied.

A study of the presence of the terms proposed for the ontology and the relevance of them in Freesound has also been done to check what impact and feasibility this ontology might have within the context of Freesound.

To be able to derive distances and designing automatic classification models over the content, data had to be gathered.

Recordings have been realized using some standard techniques, and an approach to automatically describe the sounds using tags as key concepts has been applied.

3.1. Ontology Design within the framework of Freesound

In this subchapter an ontology designed for a community of users with some previous formal knowledge, or professional experience, in acoustic musical instruments, is presented. An study of the feasibility of the integration of the proposed terms (based on comparisons with other sound libraries) as well as the vocabulary presence in Freesound of those terms, has also been carried.

The utility of the proposed Ontology is the description of sounds afterwards in the context of Freesound. These terms can be used as tags and keywords when doing automatic tagging and description as an approach of natural language description.

Proposed Ontology:

The ontology proposed for this work is not intended to cover every possible concept treated in performance or sound classification but rather build a ground over which the experiments of

automatic classification could be carried. Ideally, this ontology should be further developed, and propagated to other sounds, either by Freesound users or other researchers.

The selection of the terms (tags) and corresponding sounds of our ontology is not arbitrary. Much of the extensive literature has been studied trying to get from it the most useful terms and discarding those definitions that would not be of much use in the context we are dealing with. The classes have also been chosen within a feasible framework during the recording or collecting process.

The (Extended) Technique node has been included as there are hardly any studies or resources within this type of sounds in automatic classification, and it has been considered a valuable contribution for both this work or future approaches. A good explanation of the significance of each of the tags being used can be found in (Michels, 1977) or in Appendix D.

Another node we should explain is the link of the node A between note and diapason, being usually the reference note to tune instruments and determining if the tuning of those is characteristic of an specific musical period in history.

Many of the most common sound libraries classifications and ontologies of sounds are based on the Instruments Classification system developed by Hornbostel and Sachs. Probably the most complete in instrumental isolated sounds are the IOWA, Vienna Symphony Library and the Philharmonia Sample Library. A specialized group of users with a certain cultural background in music theory should be familiarized with them. Among these users we should include any music related professional or student with a music theory background in western “classical”/academic music¹⁰.

The terms not to be found in any of the commercial or non-commercial sound libraries or in the traditional classification systems, such as the ones of extended techniques, obey to a tradition in the academic nomenclature that is common among instrumentalists, composers or professionals of audio/music that would work within this context. These instrumental techniques are all of standard use in contemporary instrumental practice. Other Figures illustrating the result of our Ontology in Protégé can be found in Appendix-B.¹¹

¹⁰ Appendix-B contains several screenshots of some of the terms being used by these libraries.

¹¹ An online JavaScript version of the ontology and Protégé OWLs can be found at:
<http://www.carlosvaquero.com/Sites/Msc.html>

In Chapter 5 a link to the results of user's feedback on the method suggested are shown and analyzed.



Figure 3.1: Proposed ontology made with Protégé (OWL)

Vocabulary presence:

According to Stanoevska-Slabeva, “a community can be characterized by the vocabulary that is shared among its members (Stanoevska-Slabeva, 2002)”.

The users that will make use of our ontology will, ideally, have some familiarity with the terminology proposed or will acquire it after using it for some time. Still, an study of the presence of common vocabulary before designing the Ontology should be done in order to identify if there are any tags already applied within this Social Database.

In (Font et al, 2012) an equation to evaluate the importance of the tags being used is exposed. On it the number of occurrences of a tag is divided by the longevity of it in the database.

A complementary measure to define a potential community could be done taking into account the number of different users that have used those tags and how many of the tags belonging to the designed Ontology are shared among this potential subgroup of users

In Table 3.1 we can observe which are the most common tags being used in Freesound. According to Font “interaction among users (...) tend to form clearly identifiable sub groups, which suggests that further growth could be achieved by supporting sub communities (Font et al, 2012)”.

Table 3.1: - 20 Most Frequent Tags in Freesound (Font et al, 2012)

#	Tag	Occ.	#	Tag	Occ.
1	field-recording	12556	11	percussion	4726
2	noise	8781	12	drone	4597
3	loop	7993	13	processed	4468
4	drum	7314	14	soundscape	4075
5	voice	7129	15	metal	4028
6	ambient	6719	16	pad	3829
7	electronic	6264	17	water	3698
8	synth	6170	18	vocal	3590
9	multisample	5472	19	glitch	3506
10	bass	4886	20	ambience	3498

On Table 3.2 can be observed a relation of the already used proposed tags that have been used previously in Freesound. On it, the number of sounds tagged corresponding to our ontology that are already uploaded into Freesound, as well as the amount of different users that have used it. Far from defining a user community with this amount of data, we can still see how many users are interested in exchanging tags that could be used within an ontology of western “classical” instrumental sounds.

As it will be noticed, many of the tags of the proposed ontology have never been used, but other tags have already been proposed by some users. This fact make us believe that a potential community of users could find themselves identified with this descriptions and eventually extend the ontology with a tag recommendation algorithm such as the proposed by (Marques et al., 2011) or (Font at al, 2012).

Table 3.2: A “potential” Freesound sub-community already using the ontology tags proposed
n.b. Data is extracted before uploading the new dataset

Tag used	Tagged sounds	Different users
“violin”	643	77
“glissando”	177	25
“violoncello” or “cello”	237	46
“viola”	42	8
“double-bass”	15	7
“flute” or “transverse-flute”	427	96
“bassoon”	24	3
“recorder”	64	29
“classical-guitar”	5	1
“vibrato”	190	27
“non-vibrato	41	1
“pizzicato	161	17
“aerophone	16	2
“ricochet	41	10
“sul-ponticello	1	1
“white-noise	811	72
“guitar”	4065	455

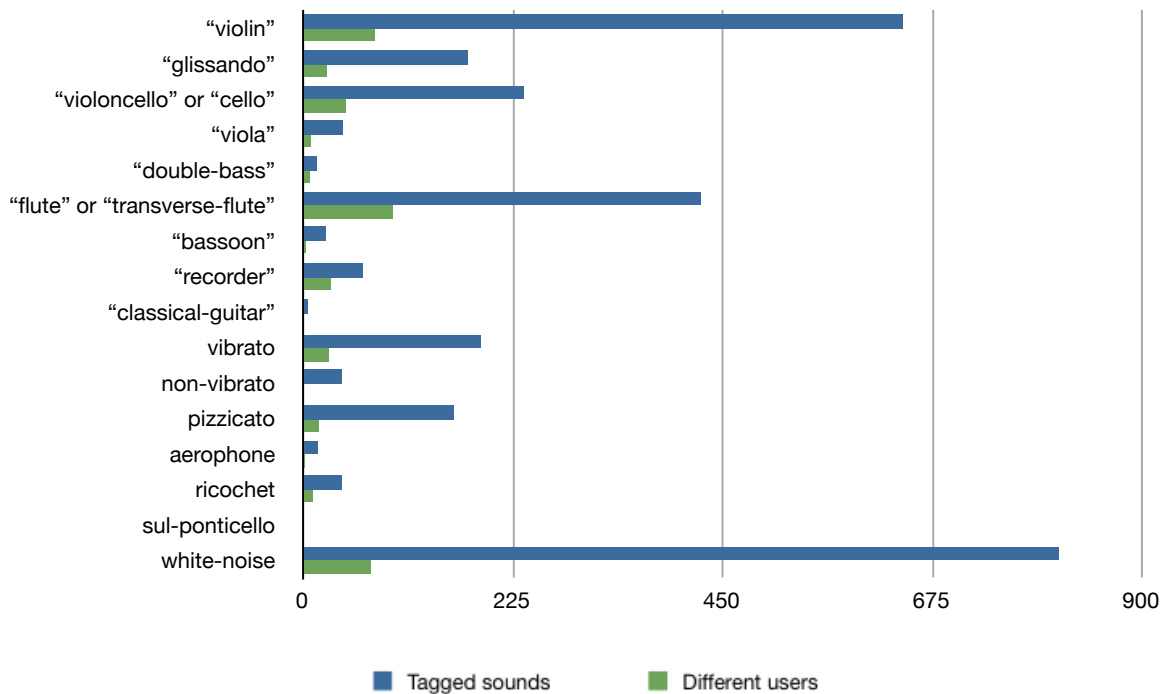


Figure 3.2: Tags of our proposed ontology already being used in Freesound

3.2. Corpus Building

Following the ontology of terms previously chosen, a corpus of audio samples has been collected in order to be able to test the use of the automatic classification and other eventual uses of the ontology within Freesound.

After studying the availability of sounds related to our ontology in Freesound (either by tags or by manual similarity selection) a great lack of training/testing material was encountered. Not having a free alternative to upload sounds into Freesound, recordings and data had to be gathered manually.

For testing purposes, the Philharmonia Orchestra Sound Library and IOWA Sounds Library have been used in addition to the samples recorded. These two sound libraries are a complete resource similar to the uploaded material but due to restrictions in their terms of use, none of them could be uploaded to Freesound.

3.2.1. Samples recordings

Recordings have been made in different studios in Den Haag, Barcelona and Madrid. A number of volunteers have participated in these recordings:

- | | |
|-------------------------------|---------------------------------|
| - María Berengué | (violin) |
| - Alex Benedicto | (viola) |
| - Roser Ávila | (cello) |
| - Jorge Soria García | (double-bass) |
| - Nahia Gastearoro Etxebarria | (bassoon) |
| - Sonja Gruys | (alto and soprano recorders) |
| - Carlos Vaquero | (transverse flute and psaltery) |

Except the bassoon samples (recorded with Schoeps MK-4 in an XY microphone setup) all of the recordings were made using a Zoom H2n handheld recorder device (XY as well). Sample Frequency used during the recording has been 48 KHz with a bitrate of 24.

The reason for choosing this microphone setup and equipment has been obtaining good quality digital samples while maintaining a simple setup for later audio processing and analysis.

Distance between the microphone capsules and the instruments has varied between instrument recordings between 50 cm and 75 cm, trying to capture as few indirect reflexions of the room as possible and giving priority to the sources “direct sound”. The room frequency response conditions in each of the family instruments differ, however the reverberation of the studios where the recordings were made is rather dry, between 300 ms and 700 ms.

Attention to the microphone orientation has also been paid to avoid stationary waves between capsules and source or common derived vibrations in the resonant source of the instrument body or the microphone itself.

The different rooms in which the recordings have been made were semi-isolated, and external noises have been tried to be reduced as much as possible due to the derived influences when doing analysis and processing of the sounds. The background noise level has therefore oscillated between 20 dB in the best situations and 40 dB in the worst.

The number of samples recorded per instrumentalist varied depending on their availability. We therefore count with a great number of samples in some instruments and very few in others but most of them were asked to play within a perceptual dynamic intention of mezzoforte to try to keep uniformity in the gathering process. For the same reason, all of them were asked though to play, at least, the following samples:

- 1 Chromatic scale: pizzicato / staccato
- 1 Chromatic scale: tenuto - vibrato
- 1 Chromatic scale: tenuto - non-vibrato
- Some random (extended) techniques samples

Most of the samples recorded coincide in pitch at least in the same Octave range (4th - 5th), which is very useful to study timbre differences and accuracy in timbre descriptors automatic classification.

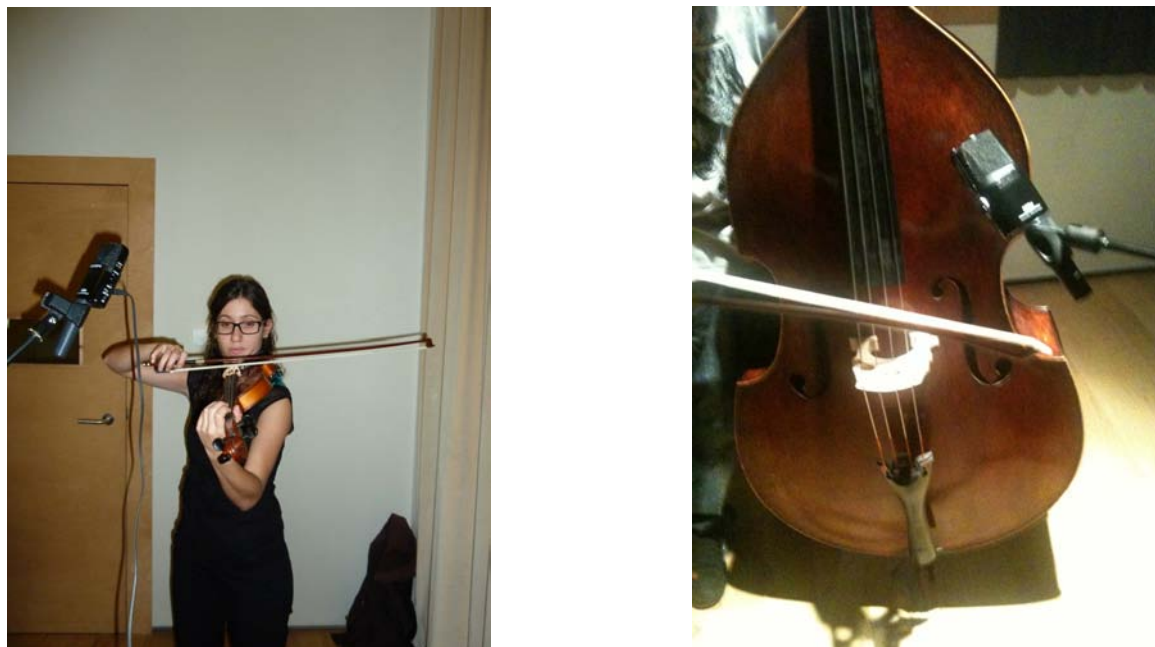


Figure 3.3: Recording sessions at ESMUC, Barcelona, 2012

3.2.2. Samples selection

820 samples have been segmented and chosen from the recording sessions, of which 700 samples were appropriately tagged and uploaded to Freesound in 24 different packs of sounds following the previously presented Ontology¹².

The editing of the samples has been done preserving all the qualities of attack and decay of the sound, facilitating the future user the possibility to adapt it according to the desired threshold. To setup the thresholds in the editing, users can refer to an extensive number of scripts / commercial applications available for this purpose.

Having a different amount of silence from some samples to other ones will have a direct influence when extracting certain descriptor values such as mean, however the difference of

¹² The resulting data set can be downloaded at:
http://www.freesound.org/people/Carlos_Vaquero/packs/

this amounts of silence is very small and we believe it might respond much more to the reality of samples from different users uploaded to Freesound.

freesound

Carlos_Vaquero Messages 0 Settings Log Out Upload

Home Sounds Forums People Help

search sounds

Packs by Carlos_Vaquero

Sort by

name last update number of sounds number of downloads

previous next 1 2 3 | 40 packs

Alto Recorder: 415 Staccato F4-F6 (25 sounds)
by Carlos_Vaquero. Downloaded 3 times.
Tags of sounds inside the pack: [aerophone](#) [alto-recorder](#) [mezzoforte](#) [multisample](#) [non-vibrato](#) [staccato](#) [woodwind](#) [zoom-h2n](#) -5 a a-4 a-sharp-4 a-sharp-5 b-4 b-5 c-5 c-6 c-sharp-5 c-sharp-6 d-5 ...

Alto Recorder: 440 Staccato F4-F6 (25 sounds)
by Carlos_Vaquero. Downloaded 6 times.
Tags of sounds inside the pack: [aerophone](#) [alto-recorder](#) [mezzoforte](#) [multisample](#) [non-vibrato](#) [staccato](#) [woodwind](#) [zoom-h2n](#) -5 a a-4 a-sharp-4 a-sharp-5 b-4 b-5 c-5 c-6 c-sharp-5 c-sharp-6 d-5 ...

Bassoon: Fortissimo and Pianissimo samples G3 (4 sounds)
by Carlos_Vaquero. Downloaded 7 times.
Tags of sounds inside the pack: [aerophone](#) [bassoon](#) [double-reed](#) [mezzoforte](#) [multisample](#) [non-vibrato](#) [schoeps-mk4](#) [tenuto](#) [woodwind](#)

Bassoon: Staccato Non Vibrato C2-C4 (25 sounds)
by Carlos_Vaquero. Downloaded 22 times.
Tags of sounds inside the pack: [aerophone](#) [bassoon](#) [double-reed](#) [mezzoforte](#) [multisample](#) [non-vibrato](#) [schoeps-mk4](#) [staccato](#) [woodwind](#) a-2 a-3 a-sharp-2 a-sharp-3 b-2 b-3 c-2 c-3 c-4 c-sharp-2 c-sharp-3 ...

Bassoon: Tenuto Non Vibrato C2-C4 (25 sounds)
by Carlos_Vaquero. Downloaded 15 times.
Tags of sounds inside the pack: [aerophone](#) [bassoon](#) [double-reed](#) [mezzoforte](#) [multisample](#) [non-vibrato](#) [schoeps-mk4](#) [tenuto](#) [woodwind](#) a-2 a-3 a-sharp-2 a-sharp-3 b-2 b-3 c-2 c-3 c-4 c-sharp-2 c-sharp-3 ...

Some random samples from the pack:

Some random samples from the pack:

Some random samples from the pack:

Some random samples from the pack:

Some random samples from the pack:

Figure 3.3: Screenshot of the uploaded packs with and preview of the tags and waveforms

3.2.3. Natural Language Description

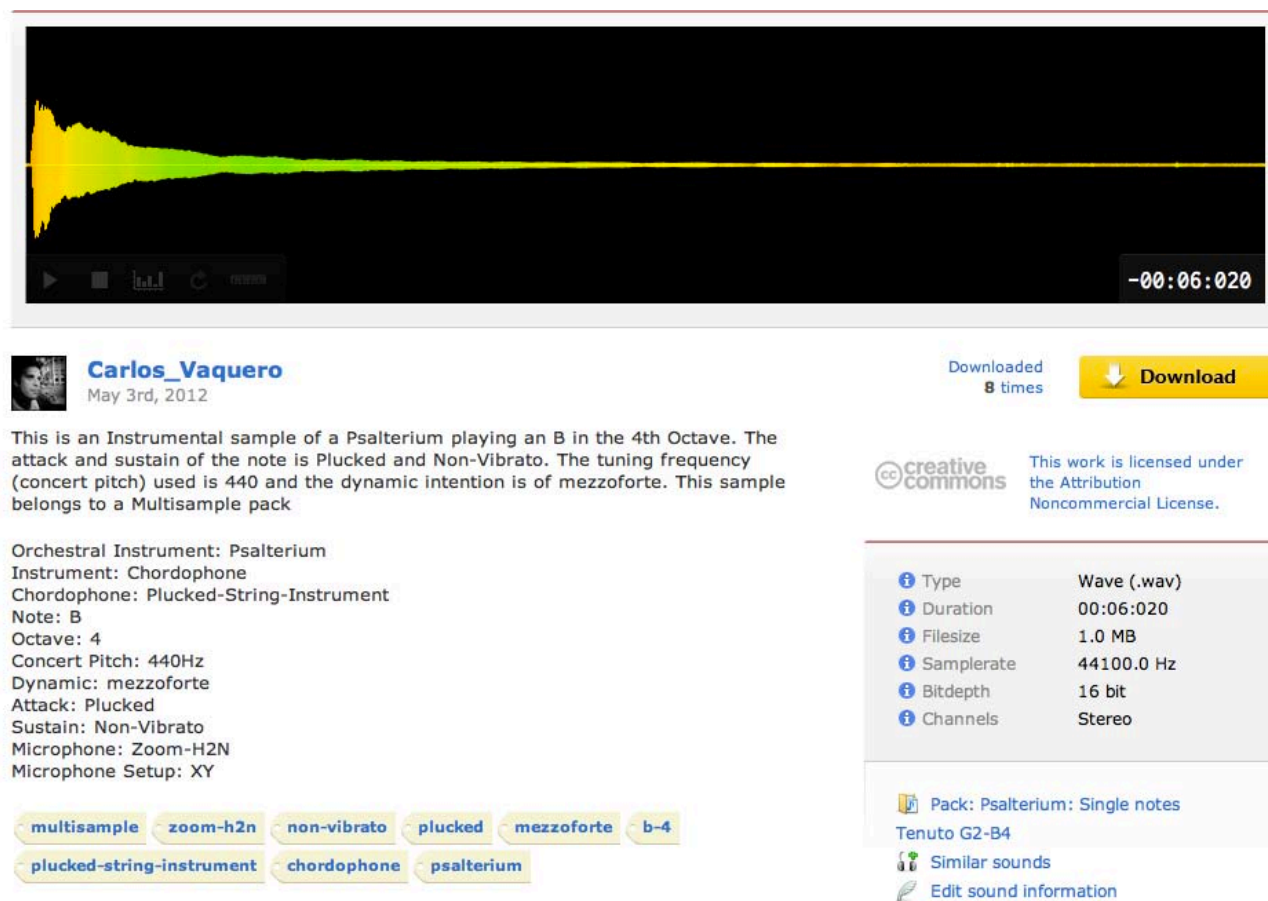
Automatic description of the samples has been used with two purposes:

- Approaching Natural language description by using the developed ontology and replacing several keys with the appropriate nodes nouns. The purpose of it is making it user friendly but also to allow possible further experiments similar to those previously done in the context of a studio framework such as (Wilmering, 2011).

- Giving the possibility of integrating this sounds using this triples into a possible OWL based browser to be developed in Freesound by using AI parsers (Bechhofer, 2004)

For this purpose a method in which certain variables are being replaced by the keywords belonging to the ontology has been used.

On Figure 3.4 we can see an example of the resulting description of the uploaded sound in Freesound.



The screenshot shows a Freesound upload page for a sample by Carlos_Vaquero, uploaded on May 3rd, 2012. The sample is a 6-second instrumental of a Psalterium. The waveform at the top shows a sharp attack followed by a sustained note. The description text states: "This is an Instrumental sample of a Psalterium playing an B in the 4th Octave. The attack and sustain of the note is Plucked and Non-Vibrato. The tuning frequency (concert pitch) used is 440 and the dynamic intention is of mezzoforte. This sample belongs to a Multisample pack". The Creative Commons license is Attribution-NonCommercial. A table of technical details is provided, including Type (Wave .wav), Duration (00:06:020), Filesize (1.0 MB), Samplerate (44100.0 Hz), Bitdepth (16 bit), and Channels (Stereo). A list of tags is shown at the bottom, including multisample, zoom-h2n, non-vibrato, plucked, mezzoforte, b-4, plucked-string-instrument, chordophone, and psalterium. A pack name 'Psalterium: Single notes' is also listed.

Carlos_Vaquero
May 3rd, 2012

Downloaded 8 times [Download](#)

This work is licensed under the Attribution Noncommercial License.

Type	Wave (.wav)
Duration	00:06:020
Filesize	1.0 MB
Samplerate	44100.0 Hz
Bitdepth	16 bit
Channels	Stereo

Pack: Psalterium: Single notes
Tenuto G2-B4

[Similar sounds](#)
[Edit sound information](#)

Tags: multisample, zoom-h2n, non-vibrato, plucked, mezzoforte, b-4, plucked-string-instrument, chordophone, psalterium

Figure 3.4: Screenshot of an uploaded sample tagged according to the proposed ontology in Freesound.

CHAPTER 4

Experiments

To be able to evaluate the efficiency of our ontology and find out which models fit better the automatic classification, three experiments have been carried. On the experiments the goal has been to propose different case scenarios in which the models could be derived and used.

It must be clarified that all the steps in feature selection explained in 3.3 have been followed for the realization of each of the experiments here presented.

Different nodes of the ontology have been chosen to select the different models. This nodes selection has been done in accordance to the relevance of the contribution to be made by this work and the classification techniques. Most of the state of the art presented so far covers widely the study of much of the instrumental sounds classification, however literature in automatic classification hardly covers certain type of attacks and instrument techniques covered by the ontology here presented.

For this same reason, nodes related to pitch classification: notes, tuning frequency and octave, and respective children, have not been tested.

Features selection:

The procedure to choose features has already been explained in 3.3. Using a trial and error method combined with PCA, in some nodes, features have been selected according to the information of them trying to keep the smallest number possible. A relation of the final selected features can be found in the models proposed on Appendix D.

4.1. Experiment 1 – Self built Data set

In (Livshing & Rodet, 2003) and (Herrera *et al*, 2003) we can observe that, to be able to generalize a system, it is needed to train it with multiple samples from different scenarios.

In the realization of the first experiment the goal has been to test the behavior of the descriptors available, trying three different possible classification techniques with 10 fold cross validation, and by using a context in which the recording of the samples has been rather homogeneous along each of the instances of the classes even though the material sources come from three different libraries.

A train set consisting in all the dataset proposed in 3.2 of 632 recorded samples mixed with another 100 samples from the IOWA library and the Philharmonia Samples library has been used to define the different models according to the nodes of the ontology selected. The aim has been obtaining objective results with different recording environments.

The same dataset has been tested afterwards by using Ten Fold Cross Validation.

Due to the rather consistent properties of the data being classified, the margin of error is not very big when using appropriate machine learning techniques. Most of the samples belonging to any of the three sounds collections being used (IOWA, Philharmonia and the one uploaded Freesound) are rather noise free and possible external noises coming from either sources or recording conditions are not very present.

Model selection:

As it has been mentioned in 2.4, extensive literature covers different machine learning techniques thoroughly used for instrumental sound classification having among the most common k-Nearest Neighbors, Support Vector Machines and Decision trees (Herrera *et al.*, 2006). After experimenting in Weka these machine learning techniques it has been observed that the cases with the greatest difference in performance results, either using one technique or another, was of a 8 % (see Table 4.1).

Overall, the technique that has provided better results is Support Vector Machines. Within SVM, different kernel functions have been tried obtaining overall a difference of 4 %.

From the Decision trees classifiers, pruned trees have been obtained, being these the resulting models implemented within our proposed automatic tagger algorithm.

Table 4.1: Classification accuracy using three different M.L. techniques

Ontology nodes	No. Classes	No. Instances	% Correctly Classified Instances Accuracy		
			k- NN	SVM	J-48
Aerophones – Chordophones	2	732	91.6341	92.7948	92.4774
Non-vibrato – Vibrato	2	647	88.9078	89.7801	84.2548
Staccato – Tenuto	2	673	91.7834	93.6324	88.4286
Wind - Instruments (Extended) Techniques	3	357	88.1489	91.8945	87.4286
Strings	8	46	85.9649	87.6571	85.1429
	6	375	82.9809	84.3089	78.4034

Observations:

As it has been previously mentioned, the number of classes determines the accuracy in the classification. The acoustic properties of the sounds to be classified are, obviously, of great relevance. Classifying between sounds of different classes with very similar acoustic characteristics among them it is more complex than differentiating within distant timbres (*e.g.* different string instruments playing in the same octave range with same sort of attack).

The cases in which binary classification is applied and the distance among the spectral characteristics of the sounds is large, the accuracy in classification is much higher.

The classifier **Aerophones – Chordophones** works with the highest accuracy. This is a very important classifier since it is the first filter to be used when doing automatic classification within the whole ontology.

The model obtained from the **Staccato – Tenuto** classification is very accurate. The name of this model should not be misunderstood since, within it, there are also sounds belonging to them among which this nomenclature does not apply, formally speaking. An arbitrary equivalence has been made between pizzicato (in this case just string instruments) and staccato (in this case just wind instruments).

The difference in this nomenclature is complicate. String instruments such as violins also play with a staccato attack but this is done with the bow technique, instead of by plucking with the fingers. Since in the wind instruments the use of staccato is related to producing a short sound by using a tongue articulation combined with a blowing technique, an arbitrary association has been made to have a broader dataset that would differentiate short sounds with a fast attack from tenuto ones.

The classifier **Non-Vibrato – Vibrato**, has a poorer performance. Many of the sounds contained in this dataset have a short attack, a very short sustain and a fast decay. All of the sounds that were wrongly classified (False Negatives) were those with a pizzicato attack (confused as vibrato).

Other sounds that have been wrongly classified are those in which the variation over time was not steady enough due to technical issues in the performance or to the spectral distance (complexity) of the extended technique sample. Being able to use descriptors with different values of the temporal evolution of these sounds (instead of an average) would help to classify better between these two families.

In the node **Wind instruments**, the classification is very accurate. This accuracy could be surprising since the properties in timbre qualities and envelope of the flute and the recorder are not very dissimilar. The fact that the recording conditions along each of the classes are so uniform might be helping to differentiate between each of these three classes.

When adding the tristimulus descriptors to the selected features, we get a two percent decrease in accuracy on J48 and SVM. Even when these would be supposed to work very

effectively in the total of our classification, have shown a negative influence in the studied classifications.

A few errors have been encountered when differentiating between the high register of the bassoon and the low register of the transverse flute. This could be due to the smaller presence of harmonics in the notes of the bassoon and, also, to a less precise attack in both instruments.

In the case of “**(Extended) Techniques**” the differences in spectral qualities of each of the child’s belonging to this node is rather noticeable, therefore, even though the number of classes is 8, the accuracy in results is rather high, we can also suppose that the reduced amount of samples within this class (48 instances).

It is remarkable that Extended techniques for “White noise” (psalterium and guitar) and “Sul-Ponticello” (viola and double-bass) are not entirely well classified, probably due to the short distance occurring in the non-deterministic components (noise part) of these types of sounds.

Double pizzicato has also been a problematic class when differentiating it with “Ricochet”, as some of the double pizzicato samples are not totally accurate in the attack. A good way to differentiate these double pizzicato sounds would be by using being able to identify the envelope of the sounds.

Possibly the best “transformation” available to differentiate these descriptors are those of the spectral skewness.

The “ricochet” samples that have been wrongly classified as “glissando” are due to a possible instability in the intonation of them.

Essentia’s library counts with a number of descriptors such as the “pitch instantaneous confidence” (Brossier, 2007) that can work for monophonic pitch identification or instrument transcription. Some of these are also very effective when using them in other nodes in which pitch relation to the energy envelope of the signal is varying, *i.e.* *glissandi* sounds.

We must recall that the selection of the descriptors of the Extended Technique node has been done combining manual selection with PCA technique. Among them the most remarkable in the pruned tree resulting are: spectral kurtosis (timbre characterization) (Peeters, 2004), contrast (noisy signals) (Akkermans et al., 2009), roll off (harmonic vs. noisy sounds), inharmonicity, odd to even harmonic energy ratio (Martin et al., 1998).

When adding more sounds and classes to the “**(Extended) Techniques**” node it is to be expected that the range of accuracy in other nodes in the same level of hierarchy, would give worse results. It must be reckoned that some of the extended techniques in contemporary practice take the timbral possibilities of each of the instruments to extremes. This can create problems in recognizing them among certain family nodes from both a perceptual and a signal processing-analysis perspective.

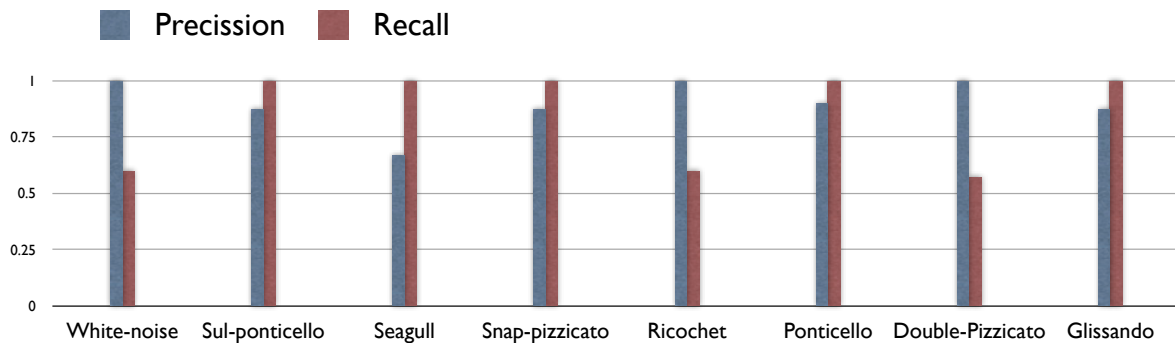


Figure 4.1: Precision and recall in the classification of
(Extended) Technique instrumental sounds

With Experiment 1 we have been able to analyze the results of three different techniques, seeing that the content material has been more or less consisting within the mistakes. Having the need to understand the behavior of the classifiers, and facilitating further implementation of them within Freesound or web clients, J48 Decision Trees have been chosen as models for the proposed automatic classification system.

In Appendix D a relation of all the pruned trees obtained can be observed and, within them, the descriptors that have proven to be most representative for the studied ontology nodes.

4.2. Experiment 2 – Testing in Freesound

With the Experiment 2 the aim has been contrasting the feasibility of the models obtained in Experiment 1 within the context of Freesound.

With this purpose, all the sounds that would have previously been tagged according to the ontology tags have been retrieved massively from Freesound (*e.g.* all the sounds tagged under violin).

To avoid having many long melody lines, and trying to retrieve isolated samples, the length of these samples has been restricted to 6 seconds, which is estimated as a reasonable length for the samples of the studied musical instruments in a very reverberant environment.

The main problem of this approach is that many of the sounds obtained under these tags are polyphonic, so the content analysis of them can have a predominance of other sources / instruments than the ones our model is looking for.

To better understand the consequences of this problem let's see an example. Since our system does not cover all the tags that can be added to a sound and will only be looking for those of our ontology, sounds with a bigger content of Aerophones (*e.g.* guitar with accordion on top of it) might be classified as such. Since our ontology lacks other tags (accordion), it will

categorize such as belonging to the Chordophones class, and therefore, the model would give a false-negative result.

In (Fuhrmann, 2012) a number of techniques and an algorithm is proposed to be able to discriminate instruments in polyphonic sources. It would be very useful to incorporate this algorithm to our ontology system in a future work to filter all the monophonic sounds before applying any of the models suggested.

In the realization of this experiment some adaptations had to be done to be able to test as many models and solutions possible as in Experiment 1 and to have a bigger test set in the context of Freesound:

- In the node Aerophones – Chordophones around 15% of the instances tagging was manually done according to other uploaded sounds tagged as “wind-instrument”.
- In the node of Wind Instruments, the “recorder” tagged sounds had to be ignored since the majority of the sounds tagged as such in Freesound obey to the recording machine and not to the musical instrument. The resampling in this node has been also rather extreme due to the small amount of “bassoon” tagged samples.
- No samples have been found in which equivalence to the sounds of the “(extended) technique” node could be done so we haven’t been able to try this model in the context of Freesound.
- None of the retrieved sounds for testing belonged to the ones previously uploaded for the design/training of the model. All these sounds have been filtered out to avoid having bias in the results of this experiment.

Table 4.2: Accuracy of sounds belonging to Freesound classified with the proposed models

Freesound Samples	No. Classes	No. Instances	J-48 - Accuracy
Vibrato-Non vibrato	2	195	82%
Staccato-Tenuto	2	84	78%
Aerophones-Chordophones	2	1881	74%
Wind Instruments	3	189	70%
String Instruments	6	7050	73%

The results shown in Table 4.2, therefore, are not completely representative of the behavior of the proposed model in a context in which only isolated sounds would be retrieved.

Despite this problem, the accuracy in recognition of the sounds associated to the retrieved tags is rather high if we reckon the diversity of material available at Freesound under the tags. This reinforces the idea that the models previously suggested are strong enough to be used and adapted into any context in which tag recommendation would be implemented.

4.3. Experiment 3 - Ontology Models Precision

An evaluation of the performance of the models for automatic tagging in the whole training dataset (self collected combined with around a 10 % of IOWA and Philharmonia samples) has been carried.

On the ontology evaluation, the Precision in the accuracy of the tags has been considered to be more relevant in an automatic tagging system.

The Precision will give us a measure of the behavior of the models in the whole dataset and how well are classified the sounds according to these, rather than just the percentage of relevance of the tags chosen in the whole context.

The Recall measure can also be confusing since the amount of tags per sound varies in each node extension and the models suggested do not cover all of them. In addition, in a real case scenario, each user applies a different complementary amount of tags. This problem should eventually be solved by developing the ontology, however the recall measure and models to evaluate the ontology should be changing each time a new tag would be added to it.

In the design of the evaluation script the fact of doing tree discrimination through the existing nodes might penalize the results of it. A system in which tags would be proposed wouldn't need to have such a discrimination; as a consequence, the dataset would be filtered while advancing through the ontology nodes, and results would therefore be improved. This however has not been a problem to discriminate between aerophones and chordophones sounds.

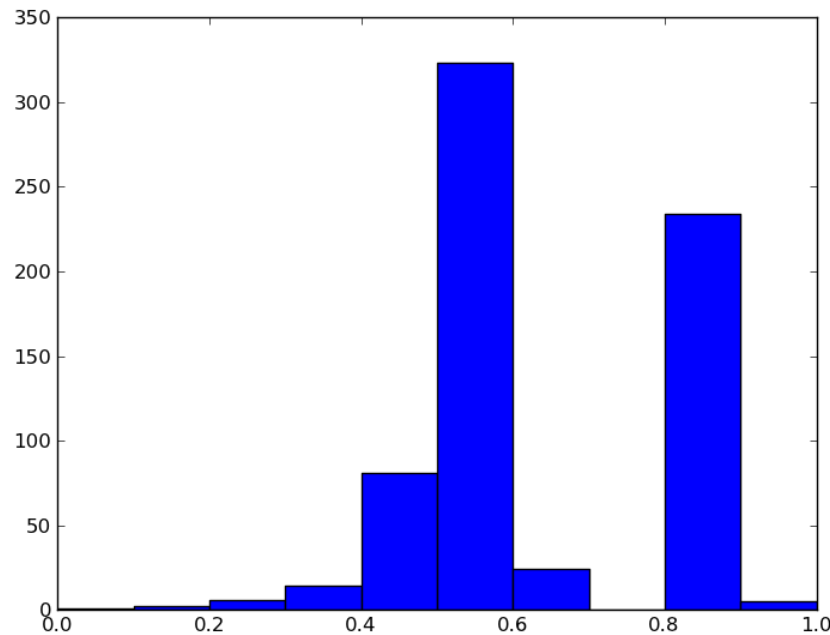


Figure 4.2: Tag Precision Histogram (Experiment 3): 0.626956521739

Since the **Aerophones - Chordophones** classifiers have proven to work efficiently, a pre-filtering of the sounds according to this node has been implemented having an error of around 8% in the general classification. We can therefore deduce that penalization from this node in the tree is not a big problem.

The **Non-vibrato** tagged sounds have been worse classified than the ones with vibrato (around 12 % worse). This result is rather surprising since the amount of vibrato varieties in performance is much bigger. However, producing a non-vibrato sound (one in which an oscillation would occur above the 20 Hz) does not mean that is a completely steady sound production. This is a relevant issue and a good example in which model generalization can fail.

The instruments in which non-vibrato was weaker in classification have been all of those in which there was a higher amount of testing samples than in training set (20% or more). Also, those sounds with a questionable non-vibrato steadiness, such as flute, double-bass in which certain fluctuations in the energy of the sound could be interpreted as vibrato.

Those instruments that contained a high number of extended technique samples have been also been problematic when being identified as Non-vibrato. Still, for a tag recommendation function, this classifier has around an 82 % of accuracy.

The **Extended- technique** model has worked very well but these results are not very representative due to the few samples available. Some problems have been caused by the double-pizzicato effect, due to the fact that the low level descriptors used are not designed to find within the envelope two or more separated sounds.

Within the **wind instruments**, the bassoon as well as the recorder classification has been absolutely precise (100% True Positive), but the flute classification has proven to be quite weak still. The number of samples available within the transverse flute classification are broader than other wind instrument samples and the model itself might escape to certain deviations in which the timbre of if it is varied within different registers depending on the amount of air (noise) produced in each note.

Despite the guitar, viola, and specially the double bass, in which the number of samples has been also significantly bigger and more diverse than in other chordophones, the **string instruments** have worked with a rather good accuracy for the other three classes.

Most of the problems derived in the classification come from the fact that these high False Negative classified instruments contain a bigger number of extended technique samples. Differentiating these samples from an instrument to another requires definitely a bigger number of samples training the models.

The automatic classification of the difficult nodes and the performance of the models could be improved when training them with a bigger amount of samples.

After analyzing the precision presented together with the False Positives and False Negatives results (and pertinent samples), it has been observed that the models work worse are:

- String Instruments with a broad amount of pizzicato (or plucked) samples that are similar in their envelope as in their spectral characteristics (*e.g.* pizzicato samples from viola *vs.* pizzicato samples from violin).
- String Instruments when samples categorized as such have acoustic properties that are difficult to be recognized and separated from each other (*e.g.* glissando, snap-pizzicato, Sul-Ponticello).
- Extended techniques in which there is a bigger variation in pitch or the timbre of the sound might change due to the characteristics of the technique.
- Non-vibrato samples in which keeping an steady pitch during the production of the note depends on the performers ability (*e.g.* wind or bowed instruments in comparison to guitar or psaltery)

4.4. User Survey and Proof of Concept

Building a Proof of Concept as the one that has been presented in this Thesis implies a final implementation in which users feedback is taken into account. Ideally, the implementation would count with a JavaScript application making used of the *search by content* function of Freesound's API. For this, HCI methodologies will have a relevant role when developing a possible interface, and results might vary from users experience.

Unfortunately, the descriptors that Freesound's uses in Essentia are not totally implemented yet and most of the models presented to do the automatic classification count with this type of descriptors (MFCC vectors). For this reason, the final implementation of it will require some technical progresses in the back-end level of Freesound's architecture before.

A preliminary **user survey** regarding the ontology, and the possibility of browsing sounds through it, has been done among expert users of Freesound. Expert users for this survey are considered those ones that have uploaded more than 300 sounds and use Freesound on a weekly basis.

90% of the survey participants the acceptance had a positive feedback to the use of an ontology to tag or browse through Freesound. Some of the most valuable information obtained after the survey is:

- Expert users in Freesound do not necessarily use a classification system when tagging their sounds (only 60% of the survey participants do it on a regular basis and without a unified system for different types of sounds).
- Most of the participants (80%) prioritize over certain aspects in the classification when tagging their sounds. The most important aspect is the sound it produces followed by the instrument that is being played with or the place in which it has been recorded.
- Users prefer to have the more tags possible to describe sounds, when tags are related between them and refer to relevant characteristics to describe the uploaded sound.
- The taxonomy proposed has been found very useful and 80% of the survey participants would like to use a similar taxonomy to tag these types of sounds.
- Some users would find very interesting to be able to browse through such an ontology.

CHAPTER 5

Conclusions and Future work

5.1. Conclusions

In this thesis, a proof of concept for improving instrumental sound description and browsing of it using automatic content analysis has been explained. The main conclusions that can be extracted from this work are:

- A proper methodology for tagging sounds based on an ontology can help in the process of classifying sounds and when deriving machine learning models with this purpose.
- Automatic tagging by using an ontology can be an efficient tool, but the design of the automatic classification algorithm for a general ontology is a delicate task that can influence a possible accumulation of error when filtering “child’s” nodes according to a tree structure.
- The study of “extended” technique instrumental sounds is a great basis to research the possibilities of an ontology since the quality of the sound is easy to classify, but the identification of its family of instruments not. Using appropriate descriptors and models based on the physical resonances of the instrument might help to make those associations if wanted.
- Automatic tagging should not be the only tool to improve classifications but it can give a certain degree of confidence in the system to recommend tags to users. It is therefore an appropriate solution to be extended by using tag recommendation.
- Freesound could eventually benefit from these models and tag ontology if a pre-filtering to discriminate polyphonic from monophonic sources would also be implemented. Also a bigger amount of instrumental isolated samples (either polyphonic or monophonic is need) is needed in Freesound to reinforce an eventual community of users that could be interested in these, for either creative or research purposes.
- Tag Ontologies do not represent collaborative tagging but a recommendation based on a tag ontology and content based automatic tagging is a good ground basis for tag propagation, as it might inspire tag *quorum*, within a potential community of users, and therefore help to make the community stronger in the sharing of concepts.

- Having available an ontology representation when tagging sounds can also be a good resource for tagging and extending a basis of concepts knowledge within a users community as most expert users would like to be able to navigate through such and incorporate a similar classification system for their sounds.

5.2. Future Work

The most immediate application of the work here presented would be its implementation in Freesound. Making the necessary changes to be able to “search by content” using all the descriptors available in Essentia, and treated in the models proposed, would make feasible the alternative browsing system suggested within this work or using the ontology available at the mentioned JavaScript prototype.

Another important approach to be carried from Freesound could be done by combining the owl taxonomy proposed and integrating it with the Music Ontology (Raimond, 2008).

Other machine learning techniques common in literature such as Latent Semantic Analysis or Neural Networks, for automatic classification should be tested, especially for the worse working nodes such as the strings one. Also a testing in Freesound of the SVM models could be very interesting since the performance has proven to be slightly better in Experiment 1.

Having a broader number of instances to be classified could improve the proposed models. A direct approach for this proposal would be using reinforcement learning, a machine learning procedure in which a system is constantly being trained adapting each of the models to a new number of instances.

A pseudo code algorithm is presented beneath. The system would recommend a tag based on the models proposed, if the user would not recognize the sound belonging to that concept an alternative concept/tag could be typed. If this users typed tag would belong to the existing ontology of tags, the model belonging to it would reinforce itself by incorporating to the test set and deriving (re-training) a new model (classification tree in our case).

The reinforce learning proposed system - algorithm could be :

```
> upload sound
> analyze
> recommend tag
> if tag is approved by user:
    > classify sound in the ontology
    > reinforce model adapting it to new content
> else:
    > user must input “user tag”
    > if “user tag” is in ontology:
        > classify sound in the ontology
> reinforce model adapting it to new content
```

This reinforce model has a direct advantage of making the system much more robust, however, if it is wrongly applied (tags are not correctly identified), it runs the risk of deforming the model and the scope of the tag to which it belongs. For instance when users would classify synthesized sounds that would have a very different content analysis than the ones already existing at the ontology. (e.g. synthesized cello added into the cello models).

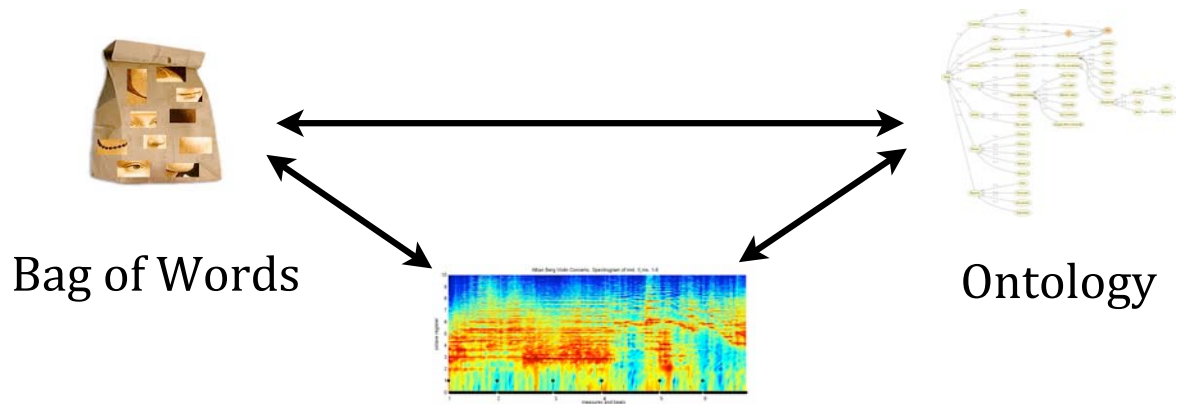


Figure 5.1: Proposed reinforce learning schema to be included in Future Work

References

- Akkermans, V., Serrà, J., & Herrera, P. (2009). Octave-based Spectral Contrast Shape-based Spectral Contrast. *Genre*, (July), 23–25.
- Alpaydin, E. (2004). *Introduction to Machine Learning*. (T. Dietterich, C. Bishop, D. Heckerman, M. Jordan, & M. Kearns, Eds.) *Machine Learning* (Vol. 15, pp. 1–153). MIT Press
- Antoniou, G., & Van Harmelen, F. (2004). *A Semantic Web Primer*. MIT Press (p. xx + 238). The MIT Press.
- Aur, M., Langlois, T., Gouyon, F., Lisboa, D., & Porto, I. (2011). *Gonc. Information Retrieval*, (Ismir), 795–800.
- Baeza-Yates, R. and B. Ribeiro-Neto (1999). *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. (K. Aberer & K.-S. Choi, Eds.) *Scientific American*, 284(5), 34–43
- Brossier, P. M. (2006). *Automatic Annotation of Musical Audio for Interactive Applications*. Doctor, Diploma of (August), 215
- Cano, P., & Koppenberger, M. Automatic sound annotation., (2004). *Proceedings of the 2004 14th IEEE Signal Processing Society Workshop Machine Learning for Signal Processing 2004* 391–400.
- Casey, M. A., Velkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: current directions and future challenges. *Proceedings of the IEEE*, 96(4), 668.
- Concepts, B., Model, T. B., Weighting, T., Model, T. V., & Model, P. (n.d.). *Modern Information Retrieval*.
- Couvreur, L., Bettens, F., Drugman, T., Dubuisson, T., Dupong, S., Frisson, C., Mancas, M. (2008). *Audio thumbnailing (QPSR Vol. I No. 2)*.
- Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. (C. U. Press, Ed.) *booksgooglecom* (Vol. 1, p. 204). Cambridge University Press.
- Eschenbach, C., & Grüninger, M. (2008). *Formal Ontology in Information Systems*. FOIS.

- Font, F., Roma, G., Herrera, P., & Serra, X. (2012). Characterization of the Freesound Online Community. Third International Workshop on Cognitive Information Processing. Baiona, Spain.
- Font, F., Serra, X., Gorup, M. T., & Fabra, U. P. (2012). Folksonomy-based tag recommendation for online audio clip sharing.
- Fuhrmann, F. (2012). Automatic musical instrument recognition from polyphonic music audio signals.
- Gruber, T. R. (2009) Encyclopedia of Database Systems, chapter Ontology. Springer- Verlag, 2009.
- Herrera-Boyer, P., Klapuri, A., & Davy, M. (2006). Automatic Classification of Pitched Musical Instrument Sounds. In A. Klapuri & M. Davy (Eds.), (pp. 163–200). Springer US.
- Herrera-Boyer, P., Peeters, G., Dubnov, S., & Fabra, U. P. (2003). Automatic Classification of Musical Instrument Sounds. *Journal of New Music Research*, 32(1), 3–21.
- Hornbostel, E. M. v., Sachs, C. (1914) “Systematik der musikinstrumente: Einversuch. Zeitschrift für Ethnologie”, Translated by A. Baines and K. Wachsmann as A Classification of Musical Instruments. *Galpin Society Journal*
- Houix, O., & Lemaitre, G. (2006). Closing the Loop of Sound Evaluation and Design (CLOSED) Everyday sound classification : Sound perception, interaction and synthesis Part I - State of the art. Design.
- Houix, O., Lemaitre, G., Misdariis, N., & Ircam, P. S. (2007). Closing the Loop of Sound Evaluation and Design (CLOSED) Everyday sound classification Part 2 Experimental classification of everyday sounds. *Design*, 2(4).
- 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.
- Klapuri, A., & Davy, M. (2006). Signal Processing Methods for Music Transcription. (Anssi Klapuri & M. Davy, Eds.) *Signal Processing* (Vol. 1, p. 440). Springer.
- Kolozali, S, Barthet, M., Fazekas, G., & Sandler, M. (2010). Towards the Automatic Generation of a Semantic Web Ontology for Musical Instruments. in *Proc 5th International Conference on Semantic and Digital Media Technologies SAMT 2010 Saarbrücken Germany*, 186–187.
- Kolozali, Sefki, Barthet, M., & Sandler, M. (2011). Knowledge Representation Issues in Musical Instrument Ontology design. *Information Retrieval*, (Ismir), 465–470.
- Liu, Y., Xiang, Q., Wang, Y., & Cai, L. (2009). Cultural style based music classification of

- audio signals. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 978-1-4244, 1–4.
- Livshin, A., Peeters, G., & Rodet, X. (2003). Studies and Improvements in Automatic Classification of Musical Sound Samples. *Proceedings ICMC 2003* (pp. 171–178).
- Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *Proceedings of 1st International Conference on Music Information Retrieval*. Plymouth, MA.
- Mandel, M., & Ellis, D. (2005). Song-level features and support vector machines for music classification. In J. D. Reiss & G. A. Wiggins (Eds.), *Proc ISMIR* (Vol. 5, pp. 594–599).
- Martin, K. D., Scheirer, E. D., & Vercoe, B. L. (1998). Music content analysis through models of audition. *Proc ACM Multimedia Workshop on Content Processing of Music for Multimedia Applications*.
- McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psychological research*, 58(3), 177–92.
- Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. (Y. Gil, E. Motta, V. R. Benjamins, & M. A. Musen, Eds.) *Web Semantics Science Services and Agents on the World Wide Web*, 5(1), 5–15.
- Michels, U. (1977) G Vogel Deutscher Taschenburg Verlag GmbH & Co. KG., München
- Moore. (1985). *Digital audio signal processing: an anthology*. (J. Strawn, Ed.). William Kaufmann, INC.
- Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *CUIDADO IST Project Report*, 54(version 1.0), 1–25.
- Pegg, C. (2012) "Ethnomusicology." Retrieved from Grove Music Online. Oxford Music Online. 29 Mar. 2012
- Pidcock, Woody (2003) "What are the differences between a vocabulary, a taxonomy, a thesaurus, an ontology, and a meta-model?"
<http://infogrid.org/trac/wiki/Reference/PidcockArticle>
- Raimond, Y., Abdallah, S., Sandler, M., & Giasson, F. (2007). The Music Ontology. *ISMIR* (pp. 1–6).
- Raimond, Y. (2008) *A Distributed Music Information System*. University of London. 2008.
- Roma, G., Janer, J., Kersten, S., Schirosa, M., Herrera, P., & Serra, X. (2010). Ecological Acoustics Perspective for Content-Based Retrieval of Environmental Sounds.

- EURASIP Journal on Audio Speech and Music Processing, 2010, 1–11.
- Sánchez, D. M., Cavero, J. M. and Martínez, E.M. (2007) in "Ontologies - A Handbook of Principles, Concepts and Applications in Information Systems"; Rajiv Kishore, Ram Ramesh; Springer; 2007
- Schedl, M. (2008). Automatically Extracting, Analyzing, and Visualizing Information on Music Artists from the World Wide Web. Analysis. Johannes Kepler University Linz.
- Serra, X. (1997) "Musical Signal Processing", chapter Musical Sound Modeling with Sinusoids plus Noise, pp. 91–122, Swets & Zeitlinger, Lisse, the Netherlands.
- Serra, X. (1989). A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition. (Doctoral dissertation, Stanford University).
- Smith, J. O. (2012). Spectral audio signal processing. Retrieved from <https://ccrma.stanford.edu/~jos/sasp/>. California Technical Publishing.
- Sordo, M., Celma, O., Blech, M., & Gaus, E. (2008). The Quest for Musical Genres: Do the Experts and the Wisdom of Crowds 9th International Conference on Music Information Retrieval, 255–260.
- Srinivasan, D. Sullivan, and I.Fujinaga. (2002) Recognition of isolated instrument tones by conservatory students. In 7th Int. Conf. on Music Perception and Cognition, pp. 720–723
- Tempelaars, S. (1996). Signal processing, speech and music. In M. Leman & P. Berg (Eds.). (Chap. 2 and 4). Swets & Zeitlinger.
- Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing, 10 (5), 293–302.
- Stanoevska-slabeva, K. (2002). Toward a Community-Oriented Design of Internet Platforms. International Journal of Electronic Commerce, 6(3), 71–95.
- Varèse, E., & Wen-chung, C. (1966). The Liberation of Sound. (E. Schwartz & B. Childs, Eds.) *Perspectives of New Music*, 5(1), 11–19.
- Wessel, D. L. (1979). Timbre Space as a Musical Control Structure. (C. Roads & J. Strawn, Eds.) *Computer Music Journal*, 3(2), 45–52.
- Wilmering, T., & Sandler, M. (2010). RDFx: Audio Effects Utilising Musical Metadata. 2010 IEEE Fourth International Conference on Semantic Computing, 452–453.
- Witten, I., & Frank, E. (2005). Data mining: practical machine learning tools and techniques. Elsevier.

- Yu, G. Y. G., & Slotine, J. J. (2008). Audio classification from time-frequency texture. 2009 IEEE International Conference on Acoustics Speech and Signal Processing, 1677–1680.
- Zwicker, E. & Fastl, H. (1999) Psychoacoustics, Facts and Models, Berlin: Springer Verlag, 1990.

Appendix A: A review of Audio Signal Processing basic concepts

Frequency and Pitch:

Frequency and Pitch are two descriptions of the regular period of sounds. Humans are able to hear sounds if their period is between $\frac{1}{20}$ and $\frac{1}{20000}$ of a second. (Moore, 1985) That is, between 20 and 20000 Hertz (Hz), being Hz the unit of frequency.

$$T = \frac{1}{f}$$

where T is the period (period duration)
 f is the frequency

Frequency is a physical quantity without a reference to our perception whereas Pitch is human's subjective evaluation of the frequency as a logarithmic scaling of it.

Amplitude:

It measures the strength of the pressure deviation with respect to the mean of the atmospheric pressure. The waveform representation can be measured as a function of time (Moore, 1985).

Phase:

The initial phase angle refers to the relative starting position of the oscillation. (Tempelaars, 1996):

$$y(t) = A \times \sin(360^\circ \frac{t}{T} + \emptyset)$$

where \emptyset is the initial phase angle and $T = \frac{1}{\text{frequency}}$
 A is the amplitude
 T is the period (period duration)

Fourier's Theorem:

Jean Baptiste Joseph Fourier (1768 - 1830) studied how periodic functions can be explained as a summation of a possibly infinite number of trigonometric functions, each with a particular amplitude and phase (Moore, 1985).

Fourier's theorem associates sinusoidal vibrations and non-sinusoidal (or arbitrary) vibrations. If we imagine a tuning fork as a harmonic oscillator producing a sinusoid, this

would vibrate forever without damping. In theory, the sound of a violin could be decomposed into a possibly infinite number of harmonic oscillators (Tempelaars, 1996). The Fourier transform (and Fourier analysis) is applied in digital processing units for the calculation of the signal spectrum in audio and image processing. The Fourier series is a possibly infinite number of sinusoidal components of the periodic signal. These components (harmonics) include frequencies that are multiples of the fundamental F0 (Tempelaars, 1996).

Spectrum representation of a signal:

It is the representation of the frequency and phase values of the sinusoidal components of the signal. The spectrum describes the amplitude over frequency bins for a given instant or time lapse and it might resemble the perception of pitch in the human auditory system (Serra, 1989). This filtering process resembles the cochlea behavior when detecting different frequencies from sounds.

It is possible that a similar filtering process in the cochlea detects frequencies from incoming sounds. Depending on the harmonics of the fundamental frequency we will hear a sound with a color or another.

Fourier transform:

Splitting an arbitrary signal into its components is referred to as a conversion from the time domain of the waveform to the frequency domain or signal spectrum.

The Fourier transform can be defined as (Serra, 1989):

$$X_{(\omega)} \triangleq \int_{-\infty}^{+\infty} x(t)e^{-j\omega t} dt$$

where \triangleq is equal by definition,
 $x(t)$ is the signal in the time domain,
 $X_{(\omega)}$ is the signal in the frequency domain
 ω is radians per second

The frequency spectrum is a complex function that can be decomposed into two real functions, the spectrum amplitude and the spectrum phase. The continuous frequency indexes are complex numbers that specify the frequency and the phase of each sinusoidal component. They are usually represented as $X(\omega)$ in the whole frequency spectrum (Serra, 1989).

By doing the inverse of the Fourier transform it is possible to reconstruct an arbitrary function from the sinusoidal components. This process is called the Fourier synthesis.

Short-Time Fourier Transform

With the Short-Time Fourier Transform (STFT) is possible to analyze signals varying over time and it is obtained by dividing the signal into successive audio frames and performing a succession of Fourier transform calculations.

The STFT representation of the amplitude as a function of frequency in a set of spectra is known as an spectrogram.

Discrete Fourier Transform

Is basically a Fourier transform for periodic (non varying over time) and discrete waveforms (Smith, 2012). The DFT is a heavy process and, as a solution to implement it, Fast Fourier Transform (FFT) algorithm is normally used.

The **FFT** calculates the frequency spectrum of a discrete time-domain signal of finite duration and it is computed in each audio frame of a signal.

The **DFT** of a signal X can be defined as:

$$X(\omega_{\kappa}) \triangleq \sum_{n=0}^{N-1} x(t_n) e^{-j\omega_{\kappa}t_n}, \kappa = 0, 1, 2, \dots, N-1$$

where,

$x(t_n)$ input signal amplitude (real or complex) at time t_n (in seconds)

$t_n \triangleq nT = \text{nth sampling instant (seconds), } n \text{ is an integer } \geq 0$

$X(\omega_{\kappa})$ spectrum of x (complex valued), at frequency ω_{κ}

$\omega_{\kappa} \triangleq \kappa\Omega = \kappa\text{th frequency sample (in radians/second)}$

$N = \text{number of time samples} = \text{number of frequency samples (integer number)}.$

Discrete Cosine Transform

Even though in theory is not the same, for real signals, just using the real part as an input of the DFT is a kind of DCT (Smith, 2010).

The DCT is used in the MFCC descriptors to obtain the logarithmic square of the mel-spectrum.

Appendix B: Ontology

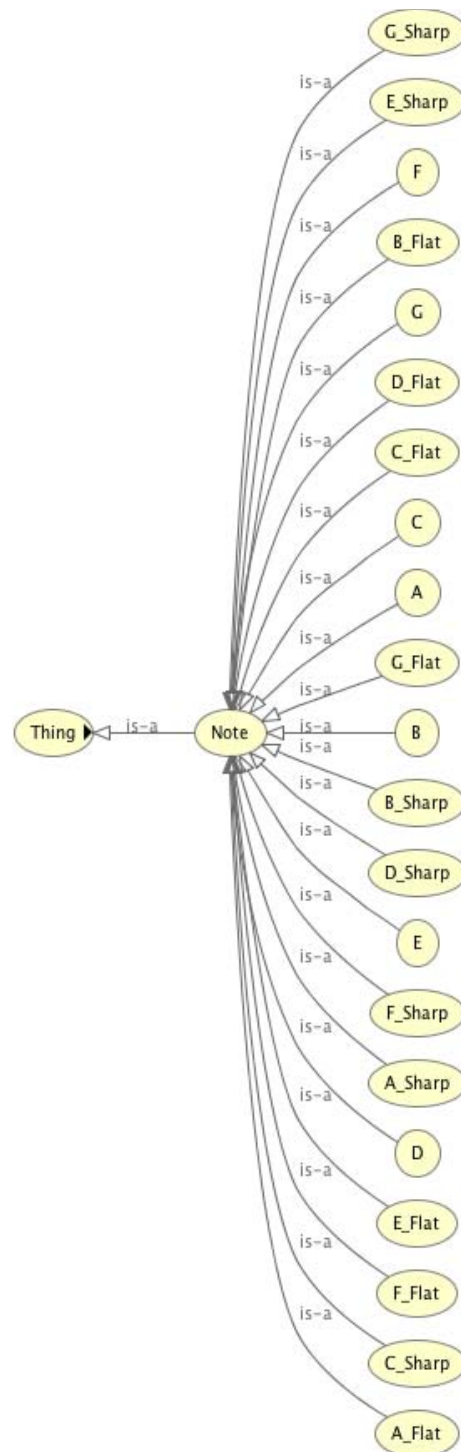


Figure 1: “Notes” node in the proposed OWL made with Protégé

The screenshot shows the Vienna Symphonic Library (VSL) website. The header features the VSL logo and navigation links: COMPANY, PRODUCTS, DEMO ZONE, USER AREA, and COMMUNITY. Below the header, a breadcrumb trail reads: VIENNA ACADEMY > Woodwinds > Bassoons > Bassoon > PLAYING TECHNIQUES. The left sidebar contains a list of categories: BRIEF DESCRIPTION, HISTORY, NOTATION, RANGE, SOUND PRODUCTION, > PLAYING TECHNIQUES (selected), SOUND CHARACTERISTICS, SOUND COMBINATIONS, and REPERTOIRE. A 'German' button is at the bottom of the sidebar. The main content area is titled 'PLAYING TECHNIQUES' and includes a 'General' section with the text: 'For an instrument of its size the bassoon is amazingly agile. Rapid passages – except with the very lowest notes – are possible as are large leaps.' A table of techniques is also present:

GENERAL	GLISSANDO
SINGLE TONGUING	MUTE
VIBRATO	LEGATO
SFORZANDO	RUNS
MULTIPLE TONGUING	MULTIPHONICS
FLUTTER TONGUING	KEY-SLAP
TRILL	SLAP TONGUE
TREMOLO	MODERN

Figure 2: Screenshot of VSL in which different attacks of bassoon are presented

The screenshot shows the Vienna Symphonic Library (VSL) website. The header features the VSL logo and navigation links: COMPANY, PRODUCTS, DEMO ZONE, > USER AREA, and COMMUNITY. Below the header, a breadcrumb trail reads: VIENNA ACADEMY > Strings > Violin > PLAYING TECHNIQUES (LEFT HAND). The left sidebar contains a list of categories: BRIEF DESCRIPTION, CONSTRUCTION, THE BOW, HISTORY, NOTATION, RANGE, SOUND PRODUCTION, > PLAYING TECHNIQUES (LEFT HAND) (selected), PLAYING TECHNIQUES (RIGHT HAND), SOUND CHARACTERISTICS, SOUND COMBINATIONS, and REPERTOIRE. A 'German' button is at the bottom of the sidebar. The main content area is titled 'PLAYING TECHNIQUES (LEFT H)' and includes a 'Double stops' section with the text: 'Two-part fingering on adjacent strings. The easiest double stops are those with an open string (all intervals are possible, including very wide ones). If both strings are fingered the most common intervals range from unison to a tenth. Intervals from a third to an octave are relatively easy if playing remains within one position. The fingering of larger intervals such as fifths and sixths is easier than that of thirds and fourths.' A table of techniques is also present:

DOUBLE STOPS	VIBRATO
TRIPLE STOPS	PORTAMENTO
QUADRUPLE STOPS	GLISSANDO
TRILLS	FINGER PIZZICATO
FINGER TREMOLO	CON SORDINO
NATURAL HARMONIC	SCORDATURA
ARTIFICIAL HARMONIC	

Figure 3: Screenshot of VSL in which different attacks of violin are presented

philharmonia orchestra the sound

Group Start Options x

Group Starts: **on key** If key is between **C4** and

Group Starts: **always**

Mapping Editor: **Edit** [Auto] [Solo] [List View] [Select Zone via Midi] [Root] [Sample: 037_7]

K.Range: C#1 - C#1 | Velocity: 1 - 127 | Root: C#1 | Volume:

Pitch Mod: +2 oct

Sample Library

Thousands of free, downloadable sound samples specially recorded by Philharmonia Orchestra players.

These samples are suitable for creating any kind of music, no matter what you're into. From piccolo to double bass there are single notes, phrases and whole orchestra samples to choose from.

License: You are free to use these samples as you wish, including releasing them as part of a commercial work. The only restriction is they must not be sold or made available 'as is' (i.e. as sampler or as a sampler instrument).

Sample Library

Please choose from the menus below to find the sample you are looking for:

bassoon mezzo forte --Articulation-- Clear

download all bassoon samples

Instrument	Pitch	Duration	Attack	Articulation
bassoon	A3	1"	mezzo forte	normal
bassoon	A#2	phrase	mezzo forte	tongued slur
bassoon	F3	phrase	mezzo forte	nonlegato
bassoon	F#2	0.5"	mezzo forte	tenuto
bassoon	C3	phrase	mezzo forte	staccato
bassoon	G3	phrase	mezzo forte	staccatissimo
bassoon	G3	1"	mezzo forte	legato
bassoon	A#2	phrase	mezzo forte	normal
bassoon	C4	phrase	mezzo forte	staccato
bassoon	F3	phrase	mezzo forte	staccatissimo
bassoon	G3	phrase	mezzo forte	staccato
bassoon	A#2	1"	mezzo forte	tongued slur
bassoon	A#1	phrase	mezzo forte	normal
bassoon	D#2	phrase	mezzo forte	tenuto
bassoon	F2	phrase	mezzo forte	staccato
bassoon	G3	phrase	mezzo forte	tongued slur
bassoon	G3	phrase	mezzo forte	staccato

Home Backstage The Orchestra Make Music Live Projects

Sample Library

Back to Samples Home Page

- Tuba Samples

Tuba Samples

Artists

Figure 4: Screenshot of London Philharmonia Orchestra web resource with different types of bassoon notes attacks

The screenshot displays the London Philharmonia Orchestra software interface. At the top, there is a control bar with buttons for 'List View', 'Select Zone via Midi', and 'Root'. Below this, a MIDI piano roll shows a sequence of notes. The interface is divided into three main sections: 'Home' (blue), 'Backstage' (purple), and 'The Orchestra' (orange). The 'Sample Library' section is active, showing a search bar with 'violin' entered and a dropdown menu for 'Articulation' open. The dropdown menu lists various violin articulations, including 'arco normal', 'molto vibrato', 'arco glissando', 'natural harmonic', 'artificial harmonic', 'arco sul tasto', 'arco staccato', 'arco col legno tratto', 'arco tremolo', 'arco legato', 'arco sul ponticello', 'arco col legno battuto', 'con sord', 'non vibrato', 'arco au talon', 'arco punta d'arco', 'arco major trill', 'snap pizz', 'arco spiccato', 'pizz normal', 'arco martele', 'harmonic glissando', 'arco tenuto', 'arco detache', 'pizz quasi guitar', 'pizz tremolo', and 'pizz glissando'. Below the search bar, a table lists various violin samples with columns for Instrument, Pitch, and Duration.

Instrument	Pitch	Duration
violin	G#3	0.5"
violin	A#5	long
violin	A#5	phrase
violin	D6	1"
violin	G7	0.5"
violin	B3	0.25"
violin	A6	1"
violin	F6	0.5"
violin	F4	1.5"
violin	G3	1"
violin	F4	0.25"
violin	G4	1"
violin	E6	phrase
violin	D6	0.5"
violin	C#4	1"
violin	D#7	1.5"
violin	D4	very long
violin	A#3	0.5"
violin	F7	0.25"
violin	E5	0.25"
violin	C7	1"
violin	B4	1"
violin	D6	0.5"

Figure 5: Screenshot of London Philharmonia Orchestra with different types of violin note attacks



Please consider [making a tax-deductible donation](#) to fund the next phase of this project. Beginning in 2012, instruments are being recorded at 24/96 with three mics for mono and stereo files that are archived into Zip files. For online listening, these notes are also organized into 16/44.1 chromatic scale files. Each instrument in the collection will be re-recorded with a variety of articulations, legato, glissandi, multiphonics, extended techniques, and in combination with other instruments. New instruments, such as the recently added guitar, will also be recorded. These freely available recordings have been used by countless musicians and in over 270 research papers. When [making a donation](#), please write "Electronic Music Studios" in the comments field.

		Arco	Pizzicato
Instrument	Violin	Violin.arco.pp.sulG.G3B3.aiff (4.1mb)	Violin.pizz.pp.sulG.G3B3.aiff (2.0mb)
Model	NA		
Performer	Josue	Violin.arco.pp.sulG.C4B4.aiff (8.9mb)	Violin.pizz.pp.sulG.C4B4.aiff (5.7mb)
	Jean-Francois	Violin.arco.pp.sulG.C5Bb5.aiff (7.7mb)	Violin.pizz.pp.sulG.C5Ab5.aiff (3.8mb)
Date		Violin.arco.pp.sulD.D4B4.aiff (8.2mb)	Violin.pizz.pp.sulD.D4B4.aiff (3.8mb)
Location	Anechoic Chamber	Violin.arco.pp.sulD.C5B5.aiff (9.5mb)	Violin.pizz.pp.sulD.C5B5.aiff (3.4mb)
Technician	Jean-Paul Perrotte	Violin.arco.pp.sulD.C6Eb6.aiff (2.8mb)	Violin.pizz.pp.sulA.A4B4.aiff (0.9mb)
Distance	5 feet	Violin.arco.pp.sulA.A4B4.aiff (2.7mb)	Violin.pizz.pp.sulA.C5B5.aiff (3.6mb)
Microphone	Neumann KM 84	Violin.arco.pp.sulA.C5B5.aiff (10.2mb)	Violin.pizz.pp.sulA.C6Ab6.aiff (2.3mb)
Mixer	Mackie 1402-VLZ	Violin.arco.pp.sulA.C6Bb6.aiff (9.8mb)	Violin.pizz.pp.sulE.E5B5.aiff (3.1mb)
Recorder	Panasonic SV-3800 DAT	Violin.arco.pp.sulE.E5B5.aiff (7.3mb)	Violin.pizz.pp.sulE.C6B6.aiff (4.5mb)
Format	16-bit, 44.1 kHz, mono	Violin.arco.pp.sulE.C6B6.aiff (10.9mb) Violin.arco.pp.sulE.C7B7.aiff (10.4mb)	Violin.pizz.pp.sulE.C7G7.aiff (2.7mb)

Figure 6: Screenshot of IOWA Sound Library with different types of violin note attacks

Appendix C: The Music Ontology

The Music Ontology Specification is a formal framework to provide a vocabulary for dealing with music-related information on the Semantic Web, including editorial, cultural and acoustic information (Music Ontology).

The semantic web is a very interesting step towards narrowing the semantic gap. In music, this gap is quite evident but using modern technologies many relations could eventually be found within different cultural backgrounds and classification systems. Even though the scope of this thesis has not a direct link with the music ontology it has been reckoned deeply

The semantic web is "a web of data that can be processed directly and indirectly by machines (Berners-Lee, 2001)". It aims at: finding, sharing and combining information more easily. For that, makes use of specific languages and technologies designed to handle data. An extensive description of these technologies can be found at (Antoniou, 2008).

Ontologies in music can benefit very much from the technologies linked to the Semantic Web specification. The most relevant work done so far is that within the Music Ontology.

Many of the tags/Concepts being used in Ontologies such as the Music Ontology (Raimond, 2008) miss a great number of the musical preliminary terms considered to cover with our ontology.

“The Music Ontology does not cover every music related concept but it provides extension points where a domain specific ontology, such as a musical instrument, a performance terminology or a genre ontology may be integrated (Kolozali, 2011).”

In Freesound, the context in which this work is developed, the Music Ontology cannot be integrated yet due to technical limitations. For this reason, within the developed ontology for this work, the OWL specification has been used and could eventually be merged into the music ontology description.

Recent additions to the Music Ontology framework include the Audio Features Ontology, in which concepts for the representation of features of audio signals are defined (Raimond, 2006, 2007). Another framework to approach metadata from the audio fx has been proposed to incorporate it automatically to the original audio metadata once audio has been transformed. The goal of this framework is translating semantic feature descriptors to control data for the DSP algorithms (Wilmering, 2011).

One of the approaches our Ontology could be linked and benefit more from the Music Ontology is the classification of instruments. In 2007 Ivan Herman added the Hornbostel and Sachs musical instruments taxonomy into the Music Ontology.

The characteristics of the music ontology framework could be extended in the future using many of the features proposed by our ontology and link them to a theoretical (solfeggio) musical approach instead of a production one.

There have been other small attempts to adapt existing taxonomies or vocabulary frameworks into ontologies within music and audio mostly for research purposes but despite the “Music Ontology” there is not a solid framework of research alternative at the moment.

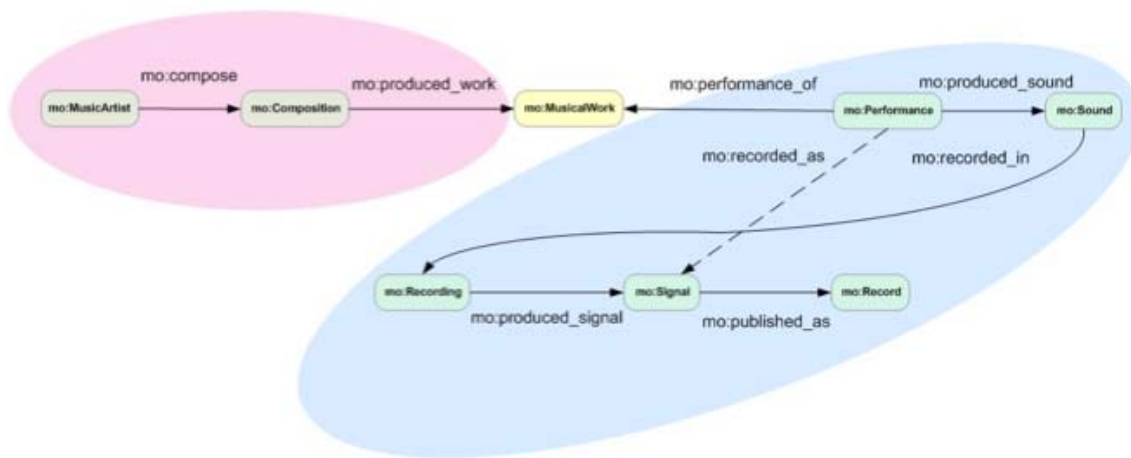


Figure 7: Music Ontology framework

Appendix D: Decision Trees

The pruned decision trees here presented are the resulting models after testing within different setups. On them we can observe which are the descriptors that best fit the node of classification within the resulting pruned tree.

Aerophone – Chordophone (722 instances)

```

4mfcc.mean <= 6.181388
| 1mfcc.mean <= -904.557678: aerophone (169.0/3.0)
| 1mfcc.mean > -904.557678
| | 1mfcc.mean <= -884.356262: aerophone (4.0)
| | 1mfcc.mean > -884.356262: chordophone (5.0)
4mfcc.mean > 6.181388
| spectral_rms.dmean2 <= 0.000364: chordophone (202.0/1.0)
| spectral_rms.dmean2 > 0.000364
| | spectral_complexity.max <= 8: aerophone (54.0/1.0)
| | spectral_complexity.max > 8: chordophone (19.0)

```

Staccato – Tenuto (647)

```

spectral_complexity.mean <= 2.785714
| spectral_strongpeak.mean <= 1.207702
| | spectral_flatness_db.dvar <= 0.000389
| | | spectral_centroid.dmean <= 30.26156: staccato (25.0/1.0)
| | | spectral_centroid.dmean > 30.26156: tenuto (3.0)
| | spectral_flatness_db.dvar > 0.000389: staccato (165.0)
| spectral_strongpeak.mean > 1.207702
| | spectral_flatness_db.dmean <= 0.036603: tenuto (43.0)
| | spectral_flatness_db.dmean > 0.036603: staccato (31.0)
spectral_complexity.mean > 2.785714
| spectral_flatness_db.dvar <= 0.000944
| | pitch.dmean <= 33.102779: tenuto (199.0)
| | pitch.dmean > 33.102779
| | | spectral_flatness_db.dmean <= 0.024105: tenuto (19.0/1.0)
| | | spectral_flatness_db.dmean > 0.024105: staccato (3.0)
| spectral_flatness_db.dvar > 0.000944
| | spectral_complexity.mean <= 5.230769: staccato (14.0)
| | spectral_complexity.mean > 5.230769: tenuto (7.0)

```


Vibrato – Non-vibrato (722)

```

spectral_flux.max <= 0.070686
| spectral_flux.mean <= 0.005855
| | spectral_flux.mean <= 0.003876: Non_vibrato (42.0)
| | spectral_flux.mean > 0.003876
| | | zerocrossingrate.max <= 0.046875
| | | | spectral_flux.min <= 0.000465: Non_vibrato (3.0)
| | | | spectral_flux.min > 0.000465
| | | | .pitch_after_max_to_before_max_energy_ratio <= 12.745884: Vibrato (16.0)
| | | | .pitch_after_max_to_before_max_energy_ratio > 12.745884
| | | | | zerocrossingrate.dmean <= 0.000384: Vibrato (3.0)
| | | | | zerocrossingrate.dmean > 0.000384: Non_vibrato (4.0)
| | | zerocrossingrate.max > 0.046875
| | | | .pitch_centroid <= 78.045197: Non_vibrato (41.0/2.0)
| | | | .pitch_centroid > 78.045197
| | | | .pitch_centroid <= 83.683365: Vibrato (5.0)
| | | | .pitch_centroid > 83.683365: Non_vibrato (2.0)
| spectral_flux.mean > 0.005855
| | .pitch_centroid <= 69.44706
| | | zerocrossingrate.min <= 0.022949
| | | | .pitch_centroid <= 36.121502: Non_vibrato (18.0)
| | | | .pitch_centroid > 36.121502
| | | | | spectral_flux.dmean <= 0.003583
| | | | | | zerocrossingrate.dvar <= 0.000034
| | | | | | | .pitch_min_to_total <= 0.02521
| | | | | | | | spectral_flux.min <= 0.0005: Vibrato (16.0/1.0)
| | | | | | | | spectral_flux.min > 0.0005
| | | | | | | | | .pitch_centroid <= 56.163342: Vibrato (15.0/1.0)
| | | | | | | | | .pitch_centroid > 56.163342: Non_vibrato (14.0)
| | | | | | | | | .pitch_min_to_total > 0.02521: Vibrato (15.0/1.0)
| | | | | | | | | zerocrossingrate.dvar > 0.000034: Non_vibrato (5.0)
| | | | | | | | | spectral_flux.dmean > 0.003583: Vibrato (34.0)
| | | | | | | | | zerocrossingrate.min > 0.022949: Non_vibrato (33.0)
| | | | | .pitch_centroid > 69.44706
| | | | | | .pitch_min_to_total <= 0.963483: Vibrato (185.0/1.0)
| | | | | | .pitch_min_to_total > 0.963483
| | | | | | | .pitch_centroid <= 80.991486: Non_vibrato (3.0)
| | | | | | | .pitch_centroid > 80.991486: Vibrato (7.0)
spectral_flux.max > 0.070686: Non_vibrato (157.0/1.0)

```

(Extended) Technique (46)

```

zerocrossingrate.var <= 0.000127
| 2barkbands.dmean2 <= 0.000005
| | 17barkbands.mean <= 0.000012: double-pizzicato (7.0)
| | 17barkbands.mean > 0.000012: snap-pizzicato (7.0)
| 2barkbands.dmean2 > 0.000005: sul-ponticello (14.0)
zerocrossingrate.var > 0.000127
| 1tristimulus.min <= 0.001145
| | spectral_complexity.dmean2 <= 1.071429
| | | spectral_energyband_high.dmean <= 0.000019: seagull (6.0)
| | | spectral_energyband_high.dmean > 0.000019: glissando (12.0)
| | spectral_complexity.dmean2 > 1.071429
| | | .average_loudness <= 0.982963: ponticello (9.0)
| | | .average_loudness > 0.982963: ricochet (10.0)
| 1tristimulus.min > 0.001145
| | 19barkbands.mean <= 0.000186: unknown (13.0)
| | 19barkbands.mean > 0.000186: white-noise (5.0)

```

Wind Instruments (307)

```

2mfcc.mean <= 196.394058
| 4mfcc.mean <= -4.096696
| | 4mfcc.mean <= -8.937961: Flute (61.0)
| | 4mfcc.mean > -8.937961
| | | 3mfcc.mean <= -129.253861: Flute (7.0)
| | | 3mfcc.mean > -129.253861: Recorder (7.0/1.0)
| 4mfcc.mean > -4.096696
| | 2mfcc.mean <= -12.105719: Flute (3.0/1.0)
| | 2mfcc.mean > -12.105719: Recorder (67.0)
2mfcc.mean > 196.394058: Bassoon (79.0)

```

String Instruments (415)

```

4barkbands.max <= 0.000084
| 2mfcc.mean <= 226.051941
| | .average_loudness <= 0.082286: psalterium (31.0)
| | .average_loudness > 0.082286
| | | 24barkbands.max <= 0.000064
| | | | silence_rate_60dB.mean <= 0.388889
| | | | 3tristimulus.min <= 0.000433: double-bass (18.0)
| | | | 3tristimulus.min > 0.000433: violin (2.0)
| | | | silence_rate_60dB.mean > 0.388889
| | | | pitch_salience.dmean <= 0.054337
| | | | 18barkbands.mean <= 0.000466

```

```

| | | | | 5spectral_contrast.var <= 0.000844
| | | | | | pitch.min <= 262.5: violin (3.0)
| | | | | | pitch.min > 262.5: viola (17.0)
| | | | | 5spectral_contrast.var > 0.000844
| | | | | | pitch_instantaneous_confidence.mean <= 0.802125: viola (5.0)
| | | | | | pitch_instantaneous_confidence.mean > 0.802125
| | | | | | | pitch_salience.min <= 0.33438: violin (72.0/1.0)
| | | | | | | pitch_salience.min > 0.33438: viola (2.0)
| | | | | 18barkbands.mean > 0.000466: viola (20.0)
| | | | | pitch_salience.dmean > 0.054337
| | | | | pitch.min <= 390.265472: viola (35.0)
| | | | | pitch.min > 390.265472: double-bass (2.0)
| | | 24barkbands.max > 0.000064: psalterium (4.0)
| 2mfcc.mean > 226.051941: violoncello (8.0)
4barkbands.max > 0.000084
| .pitch_centroid <= 114.145409
| | pitch_instantaneous_confidence.dmean <= 0.007878
| | | 2barkbands.dmean2 <= 0.000016
| | | | 3spectral_contrast.var <= 0.000613: classical-guitar (65.0)
| | | | 3spectral_contrast.var > 0.000613: psalterium (2.0)
| | | | 2barkbands.dmean2 > 0.000016: double-bass (2.0)
| | | pitch_instantaneous_confidence.dmean > 0.007878
| | | pitch_instantaneous_confidence.mean <= 0.902482
| | | | spectral_spread.mean <= 4330540.5
| | | | 7mfcc.mean <= 3.45603
| | | | | inharmonicity.dvar2 <= 0.000048: classical-guitar (2.0)
| | | | | inharmonicity.dvar2 > 0.000048
| | | | | 5mfcc.mean <= -32.48238: double-bass (14.0)
| | | | | 5mfcc.mean > -32.48238
| | | | | 1tristimulus.dvar <= 0.084742: violoncello (68.0/1.0)
| | | | | 1tristimulus.dvar > 0.084742: double-bass (7.0)
| | | | 7mfcc.mean > 3.45603: double-bass (35.0)
| | | | spectral_spread.mean > 4330540.5: classical-guitar (5.0)
| | | | pitch_instantaneous_confidence.mean > 0.902482: classical-guitar (14.0/1.0)
| .pitch_centroid > 114.145409: psalterium

```

Appendix E: Ontology nomenclature

In this Appendix some nomenclatures used for the proposed ontology are explained. The reader could also refer to http://www.freesound.org/people/Carlos_Vaquero/packs/ to get an impression of the sounding characteristics of them. For a complete description of them the reader could refer to (Michels, 1977) , Vienna Symphonic Library Academy¹³ and Grove Dictionary of Music¹⁴.

Col-legno:

(String instruments)

Can be translated as “with the wood” (referring to the hit) It is produced by striking the string with the stick of the bow, rather than by drawing the hair of the bow across the strings. This results in a quiet percussive sound.

“Bartók” pizzicato (“snap” pizzicato):

(String instruments)

The string is lifted with two fingers of the right hand so that it snaps back onto the fingerboard when let go. This produces a very percussive and resounding sound. It is named after the composer Béla Bartók due to the extensive and innovative use he made of it.

Pizzicato:

(String instruments)

Plucking of the strings with the right hand. This technique was originated on the lute in the XVIth century and it is applied to chords or thirds (double or triple pizzicato) as well as single notes.

It can also be done with the left hand. The left hand touching plucks a string usually with one finger. “Normal” pizzicato is performed by the right hand.

Ricochet:

(String instruments)

Played both with the upstroke and the downstroke it is produced by doing several *saltato* leaps on one stroke. “The bow does not fall on the string with the force of its own weight but is thrown onto it so that the leaping effect continues in the same direction. Following its first impact on the strings the bow performs a precise number of leaps, usually three or

¹³ <http://www.vsl.co.at/en/70/149/150/46.vsl>

¹⁴ <http://www.oxfordmusiconline.com/>

four. Groups of three or four notes can therefore be played before the bow has to be thrown again. As far as *p* and *mf* (Vienna Academy, 2012)".

Staccato:

(Italian for *detached*) A note of shortened duration. It is used in both string and wind instruments nomenclature. In the baroque period was used to specify that the note had it was of certain relevance within the context of the line being played.

"Seagull" effect:

(String instruments - Cello technique)

Most famously used by George Crumb's work, "*Vox Balaenae*".

It is a stopped harmonic beginning in a very high position with an octave span between the fingers. Glissando down the length of the fingerboard (without adjusting the space between the fingers) As a result of the finger spacing, higher partials of low fundamentals of the cello are activated while the left-hand motion in the scroll (neck), this makes repeated starts of the glissando.

Sul-ponticello:

(String instruments)

The bow makes contact with the string near the bridge.

If the bow is very close to the bridge the volume gets louder. The timbre becomes brighter but at the same time glassy, shrill, eerie, pale and thin. The number of partials increases.

The technique is also used as a tremolo.

Tenuto:

(String and wind instruments)

Sustaining of a note given its full value.

Vibrato:

(String instruments)

Vibrations of the left hand are transferred onto the string. The result is a fluctuating pitch and loudness. The vibrato depends on the extent of these fluctuations and the speed.

Vibrato can be produced by movement of the finger, hand or arm or a combination of these. Exactly how it develops depends on the position and playing technique.

(Wind instruments)

Microtonal periodic fluctuations in pitch and/or volume that are produced by movements of the diaphragm, larynx and/or lips, depending on the sound desired or the physical properties of the instrument.