UNIVERSITAT
POMPEU FABRA

Audio content processing for automatic music genre classification: descriptors, databases, and classifiers.

Enric Guaus i Termens

A dissertation submitted to the Department of Information and Communication Technologies at the Universitat Pompeu Fabra for the program in Computer Science and Digital Communications in partial fulfilment of the requirements for the degree of

—

Doctor per la Universitat Pompeu Fabra

Doctoral dissertation direction:

Doctor Xavier Serra
Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona

Barcelona, 2009

This research was performed at the Music Technology Group of the Universitat Pompeu Fabra in Barcelona, Spain. This research was partially funded by the HARMOS eContent project EDC-11189 and by the SIMAC project IST-FP6-507142.

A mon pare...

Acknowledgements

I would like to thank all the people of the Music Technology Group and specially Xavier Serra, Perfecto Herrera, Emilia Gómez, Jordi Bonada, Cyril Laurier, Joan Serrà, Alfonso Pérez, Esteban Maestre, Merlijn Blaauw. I would also thank the former members Sebastian Streich, Bee Suan Ong, Jaume Masip, Vadim Tarasov, José Pedro and Eloi Batlle.

It is also important to recognize the unconditional support from the people at the ESMUC, specially Enric Giné, Ferran Conangla, Josep Maria Comajuncosas, Emilia Gómez (again), Perfecto Herrera (again) and Roser Galí.

I would like to mention here the people who introduced me in the research at the Universitat Ramon Llull: Josep Martí and Robert Barti.

In addition, i would like to thank all my friends and specially those who I meet every thursday, those who I meet with thousand of childs running under our legs and also that who makes brings me part of the responsability of Aitana. And Paola, who supported my in everything, everytime and everywhere.

Finally, I would like to thank all my family who guided me in the difficult moments, and specially thanks to that who can't call me *Doctor Guaus*.

Abstract

This dissertation presents, discusses, and sheds some light on the problems that appear when computers try to automatically classify musical genres from audio signals. In particular, a method is proposed for the automatic music genre classification by using a computational approach that is inspired in music cognition and musicology in addition to Music Information Retrieval techniques. In this context, we design a set of experiments by combining the different elements that may affect the accuracy in the classification (audio descriptors, machine learning algorithms, etc.). We evaluate, compare and analyze the obtained results in order to explain the existing glass-ceiling in genre classification, and propose new strategies to overcome it. Moreover, starting from the polyphonic audio content processing we include musical and cultural aspects of musical genre that have usually been neglected in the current state of the art approaches.

This work studies different families of audio descriptors related to timbre, rhythm, tonality and other facets of music, which have not been frequently addressed in the literature. Some of these descriptors are proposed by the author and others come from previous existing studies. We also compare machine learning techniques commonly used for classification and analyze how they can deal with the genre classification problem. We also present a discussion on their ability to represent the different classification models proposed in cognitive science. Moreover, the classification results using the machine learning techniques are contrasted with the results of some listening experiments proposed. This comparison drive us to think of a specific architecture of classifiers that will be justified and described in detail. It is also one of the objectives of this dissertation to compare results under different data configurations, that is, using different datasets, mixing them and reproducing some real scenarios in which genre classifiers could be used (huge datasets). As a conclusion, we discuss how the classification architecture here proposed can break the existing glass-ceiling effect in automatic genre classification.

To sum up, this dissertation contributes to the field of automatic genre classification: a) It provides a multidisciplinary review of musical genres and its classification; b) It provides a qualitative and quantitative evaluation of families of audio descriptors used for automatic classification; c) It evaluates different machine learning techniques and their pros and cons in the context of genre classification; d) It proposes a new architecture of classifiers after analyzing music genre classification from different disciplines; e) It analyzes the behavior of this proposed architecture in different environments consisting of huge or mixed datasets.

Resum

Aquesta tesi versa sobre la classificació automàtica de gèneres musicals, basada en l'anàlisi del contingut del senyal d'àudio, plantejant-ne els problemes i proposant solucions.

Es proposa un estudi de la classificació de gèneres musicals des del punt de vista computacional però inspirat en teories dels camps de la musicologia i de la percepció. D'aquesta manera, els experiments presentats combinen diferents elements que influeixen en l'encert o fracàs de la classificació, com ara els descriptors d'àudio, les tècniques d'aprenentatge, etc. L'objectiu és avaluar i comparar els resultats obtinguts d'aquests experiments per tal d'explicar els límits d'encert dels algorismes actuals, i proposar noves estratègies per tal de superar-los. A més a més, partint del processat de la informació d'àudio, s'inclouen aspectes musicals i culturals referents al gènere que tradicionalment no han estat tinguts en compte en els estudis existents.

En aquest contexte, es proposa l'estudi de diferents famílies de descriptors d'àudio referents al timbre, ritme, tonalitat o altres aspectes de la música. Alguns d'aquests descriptors són proposats pel propi autor mentre que d'altres ja són perfectament coneguts. D'altra banda, també es comparen les tècniques d'aprenentatge artificial que s'usen tradicionalment en aquest camp i s'analitza el seu comportament davant el nostre problema de classificació. També es presenta una discussió sobre la seva capacitat per representar els diferents models de classificació proposats en el camp de la percepció. Els resultats de la classificació es comparen amb un seguit de tests i enquestes realitzades sobre un conjunt d'individus. Com a resultat d'aquesta comparativa es proposa una arquitectura específica de classificadors que també està raonada i explicada en detall. Finalment, es fa un especial èmfasi en comparar resultats dels classificadors automàtics en diferents escenaris que pressuposen la barreja de bases de dades, la comparació entre bases de dades grans i petites, etc. A títol de conclusió, es mostra com l'arquitectura de classificació proposada, justificada pels resultats dels diferents anàlisis, pot trencar el límit actual en tasques de classificació automàtica de gèneres musicals.

De manera condensada, podem dir que aquesta tesi contribueix al camp de la classificació de gèneres musicals en els següents aspectes: a) Proporciona una revisió multidisciplinar dels gèneres musicals i la seva classificació; b) Presenta una avaluació qualitativa i quantitativa de les famílies de descriptors d'àudio davant el problema de la classificació de gèneres; c) Avalua els pros i contres de les diferents tècniques d'aprenentatge artificial davant el gènere; d) Proposa una arquitectura nova de classificador d'acord amb una visió interdisciplinar dels gèneres musicals; e) Analitza el comportament de l'arquitectura proposada davant d'entorns molt diversos en el que es podria implementar el classificador.

Resumen

Esta tesis estudia la clasificación automática de géneros musicales, basada en el análisis del contenido de la señal de audio, planteando sus problemas y proponiendo soluciones.

Se propone un estudio de la clasificación de los géneros musicales desde el punto de vista computacional, pero inspirado en teorías de los campos de la musicología y la percepción, De este modo, los experimentos persentados combinan distintos elementos que influyen en el acierto o fracaso de la clasificación, como por ejemplo los descriptores de audio, las técnicas de aprondizaje, etc. El objetivo es comparar y evaluar los resultados obtenidos de estos experimentos para explicar los límites de las tasas de acierto de los algorismos actuales, y proponer nuevas estrategias para superarlos. Además, partiendo del procesado de la información de Audio, se han incluido aspectos musicales y culturales al género que tradicionalmente no han sido tomados en cuenta en los estudios existentes.

En este contexto, se propone el estudio de distintas famílias de descriptores de audio referentes al timbre, al ritmo, a la tonalidad o a otros aspectos de la música. Algunos de los descriptores son propuestos por el mismo autor, mientras que otros son perfectamente conocidos. Por otra parte, también se comparan las técnicas de aprendizaje artificial que se usan tradicionalmente, y analizamos su comportamiento en frente de nuestro problema de clasificación. Tambien planteamos una discusión sobre su capacidad para representar los diferentes modelos de clasificación propuestos en el campo de la percepción. Estos resultados de la clasificación se comparan con los resultados de unos tests y encuestas realizados sobre un conjunto de individuos. Como resultado de esta comparativa se propone una arquitectura específica de clasificadores que tambien está razonada y detallada en el cuerpo de la tesis. Finalmente, se hace un émfasis especial en comparar los resultados de los clasificadores automáticos en distintos escenarios que assumen la mezcla de bases de datos, algunas muy grandes y otras muy pequeñas, etc. Como conclusión, mostraremos como la arquitectura de clasificación propuesta permite romper el límite actual en el ámbito de la classificación automática de géneros musicales.

De forma condensada, podemos decir que esta tesis contribuye en el campo de la clasificación de los géneros musicales el los siguientes aspectos: a) Proporciona una revisión multidisciplinar de los géneros musicales y su clasificación; b) Presenta una evaluación cualitativa y cuantitativa de las famílias de descriptores de audio para la clasificación de géneros musicales; c) Evalua los pros y contras de las distintas técnicas de aprendizaje artificial delante del género; d) Propone una arquitectura nueva del clasificador de acuerdo con una visión interdisciplinar de los géneros musicales; e) Analiza el comportamiento de la arquitectura propuesta delante de entornos muy diversos en los que se podria implementar el clasificador.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Motivation

Wide-band connection to the Internet has become quite a common resource in our lifestyle. Among others, it allows users to store and share thousands of audiovisual content in their hard disk, portable media player or cellular phone. On-line distributors like iTunes[1], Yahoo! Music[2] or Amazon[3] take benefit of this situation and contribute to the metamorphosis that music industry is living. The physical CD is becoming obsolete as a commercial product in benefit of MP3, AAC, WMA or other file formats in which the content can be easily shared by users. On the other hand, the pervasive Peer2Peer networks also contribute to this change, but some legal issues are still unclear.

During the last thirty years, music has been traditionally sold in a physical format (vinyl, CD, etc.) organized according to a rigid structure based on a set of 10 or 15 songs, usually from the same artist, grouped in an album. There exist thousands of exceptions to this organization (compilations, double CDs, etc.) but all of them are deviations from that basic structure. Nowadays, digital databases and stores allow the user to download individual songs from different artists, create their own compilations and decide how to exchange musical experiences with the rest of the community. Portals like mySpace[4] allow unknown and new bands to grow using different ways than those traditionally established by the music industry.

Under these conditions, the organization of huge databases becomes a real problem for music enthusiasts and professionals. New methodologies to discover, recommend and classify music must emerge from the computer music industry and research groups.

The computer music community is a relative small group in the big field of computer science. Most of the people in this community is greatly enthusiastic about music. The problem arises when computers meet music. Sometimes, in

[1] http://www.itunes.com

[2] http://music.yahoo.com

[3] http://www.amazon.com

[4] http://www.myspace.com

this world of numbers, probabilities and sinusoids, everything about music can be forgotten and the research becomes quite far from the final user expectations. Our research, focused into the Music Information Retrieval field (MIR), tries to join these two worlds but, sometimes, it is a difficult task. From our point of view, the research in MIR should take into account different aspects of music such as (1) the objective description of music: basic musical concepts like BPM, melody, timbre, etc., (2) musicological description of music: formal studies can provide our community the theoretical background we have to deal with using computers, and (3) psychological aspects of music: it is important to know how different musical stimulus affects the human behavior.

Music can be classified according to its genre, which is probably one of the most often used descriptors for music. Heittola (2003) explores how to manage huge databases that can be stored in a personal computer in terms of musical genre classification. Classifying music into a particular musical genre is a useful way of describing qualities that it shares with other music from the same genre, and separating it from other music. Generally, music within the same musical genre has certain similar characteristics, for example, similar instruments, rhythmic patterns, or harmonic/melodic features. In this thesis, we will discuss about the use of different techniques to extract these properties from the audio, and we will establish different relationships between the files stored in our hard disk in terms of musical genres defined by a specific taxonomy.

There are many disciplines involved in this issue, such as information retrieval, signal processing, statistics, musicology and cognition. We will focus on the methods and techniques proposed by the music content processing and music information retrieval fields but we will not completely forget about the rest.

1.2 Music Content Processing

Let us imagine that we are in a CD store. Our decision to buy a specific CD will depend on many different aspects like genre, danceability, instrumentation, etc. Basically, the information we have in the store is limited to the genre, artist and album, but sometimes this information is not enough to take the right decision. Then, it would be useful to retrieve music according to different aspects on its content but, what is the *content*?

The word *content* is defined as "*the ideas that are contained in a piece of writing, a speech or a film*" (Cambridge International Dictionary, 2008). This concept applied to a piece of music can be seen as the implicit information that is related to this piece and that is represented in the piece itself. Aspects to be included in this concept are, for instance, structural, rhythmic, instrumental, or melodic properties of the piece. Polotti & Rocchesso (2008) remark that the Society of Motion Picture and Television Engineers (SMPTE) and the European Broadcasting Union (EBU) propose a more practical definition of the term *content*, as the combination of two concepts: *metadata* and *essence*. Essence is "*the raw program material itself, the data that directly encodes pictures, sounds, text, video, etc.*". On the other hand, Metadata is "*used to describe the essence and its different manifestations*", that can be splitted in different categories:

- Essential: Meta information that is necessary to reproduce the essence

- Access: Provides legal access and control to the essence
- Parametric: Defines the parameters of the capturing methods
- Relational: Allows synchronization to different content components
- Descriptive: Gives information to the essence. It facilities the search, retrieval, cataloging, etc.

According to Merriam-Webster Online (2008), the concept of *content-analysis* is defined as the "*analysis of the manifest and latent content of a body of communicated material (as a book or film) through a classification, tabulation, and evaluation of its key symbols and themes in order to ascertain its meaning and probable effect*". Music content analysis and processing is a topic of research that has become very relevant in the last few years. As explained in Section 1.1, the main reason for this is the high amount of audio material that has been made accessible for personal purposes through networks and available storage supports. This makes it necessary to develop tools intended to interact with this audio material in an easy and meaningful way. Several techniques are included under the concept of *"Music-content analysis"* such as techniques for automatic transcription, rhythm and melodic characterization, instrument recognition and genre classification. The aim of all of them is to describe any aspect related to the content of music.

1.3 Music Information Retrieval

It is difficult to establish the starting point or the key paper in the field of Music Information Retrieval (MIR). As shown by Polotti & Rocchesso (2008), the pioneers in MIR discipline were Kassler (1966) and Lincoln (1967). According to them, Music Information Retrieval can be defined as "*the task of extracting, from a large quantity of data, the portions that data with respect to which some particular musicological statement is true*". They also show three ideas that should be the goals of MIR: (1) the elimination of manual transcription, (2) the creation of an effective input language for music and (3) economic way for printing music.

MIR is an interdisciplinary science. Fingerhut & Donin (2006) propose a map with all the disciplines related to MIR. Figure 1.1 shows a simplified version of this map. In the left, we observe the kind of information we have (in the musician's brain, digitally stored or just metadata information). On the right, we observe the disciplines that are related to data for each level of abstraction. For automatic genre classification of music we need information from digital, symbolic and semantic levels. Ideally, we should also include information from the cognition level, but the state of the art is quite far to provide such information.

The relevance of MIR in the music industry is summarized by Downie (2003). Every year, 10000 new albums are released and 100,000 works registered for copyright (Uitdenbogerd & Zobel, 1999). In 2000, the US industry shipped 1.08 billion units of recorded music (CDs, cassettes, etc.) valued at 14.3 billion dollars (Recording Industry Association of America, 2001). Vivendi Universal (the parent company of Universal Music Group) bought MP3.com for 372 million dollars. Although this quantity was still so far from the overall

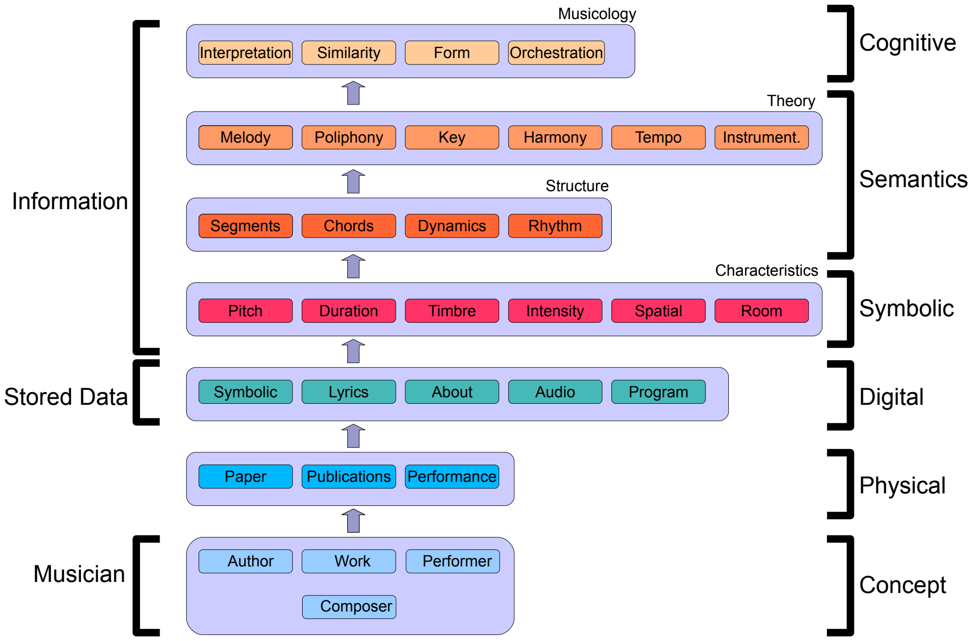


Figure 1.1: Simplified version of the MIR map proposed by Fingerhut & Donin (2006)

music business, it is obvious that it is not negligible. In 2001, Wordspot (an Internet consulting company that tracks queries submitted to Internet search engines) reported that MP3 format queries displaced the search for sex-related materials. In 2007, the search and download of MP3 files is, at least, as relevant as other traditional sources for consuming music. In front of this scenario, it is obvious that MIR has to provide solutions to the problems presented by this new way of listening to the music.

1.4 Application contexts

Some of the application contexts in which automatic genre classification becomes relevant are listed here:

- Organization of personal collections: Downloading music from the Internet has become a common task for most of music enthusiasts. Automatic classifiers provide a good starting point for the organization of big databases. Classifiers should allow users to organize music according to their own taxonomy and should allow them to include new manually labelled data defined by other users.

- Multiple relationships in music stores and library databases: In many cases, music can not be completely fitted into a unique musical genre. A distance measure for a specific audio file belonging to many given categories can help users in their music search. Catalogs can be defined using these distances in such a way that the same audio file can be found in different groups.

- Automatic labeling in radiostations: Audio fingerprinting and monitoring is a crucial task for author societies. The whole system can be spread into different specialized identification nodes according to musical genres. For instance, classical music radio stations do not modify their database as often as commercial pop music radio stations. In this context, the configuration of fingerprint systems is different for each musical genre. Automatic genre classification provides their initial filter.

- Music Recommendation: Due to the availability of high amounts of audio files in the Internet, music recommenders systems emerge from research labs. The most common systems are based on collaborative filtering or other techniques based on the feedback provided by users (p.ex. the *Customers who bough this item also bought* recommendation at Amazon.com). Other systems use content based and Music Information Retrieval techniques in which the genre classification plays an important role.

- Playlist generation: Many times, we would like to listen to a specific kind of music we have stored in our portable media player. According to some given examples, the system should be able to propose a list of similar songs in a playlist. Here again, a distance measure from the songs in our database to a given list of musical genres is a very valuable information.

In the applications here presented, the genre classifier is not the unique technique to be used. It needs to be complemented with the other active topics in MIR such as rhythm detection, chord estimation or cover identification, among others.

1.5 Goals

We present here the goals of this PhD dissertation, which are related to the hypothesis that we want to verify:

- Review current efforts in audio genre classification. This multidisciplinary study comprises the current literature related to genre classification, music categorization theories, taxonomy definition, signal processing methods used for the feature extraction and machine learning algorithms commonly used.

- Justify of the role of automatic genre classification algorithms in the MIR and industry communities.

- Study the influence of taxonomies, audio descriptors and machine learning algorithms in the overall performance of automatic genre classifiers.

- Propose alternative methods to make automatic classification a more comprehensive and musical task, according to the current state of the art and musicological knowledge.

- Provide a flexible classification technique that is capable to deal with different environments and applications.
- Provide a quantitative evaluation of the proposed approaches by using different collections.

A part from the technical content, this dissertation also presents the question about why humans need to classify music into genres and how they do that. The usefulness of classification is discussed in different environments where sometimes they follow some logical rules, but sometimes not. Maybe genre classification can not be performed in the same way than animals, figures or colors classification. And maybe traditional genre classification is not the best way to classify music.

1.6 Summary of the PhD work

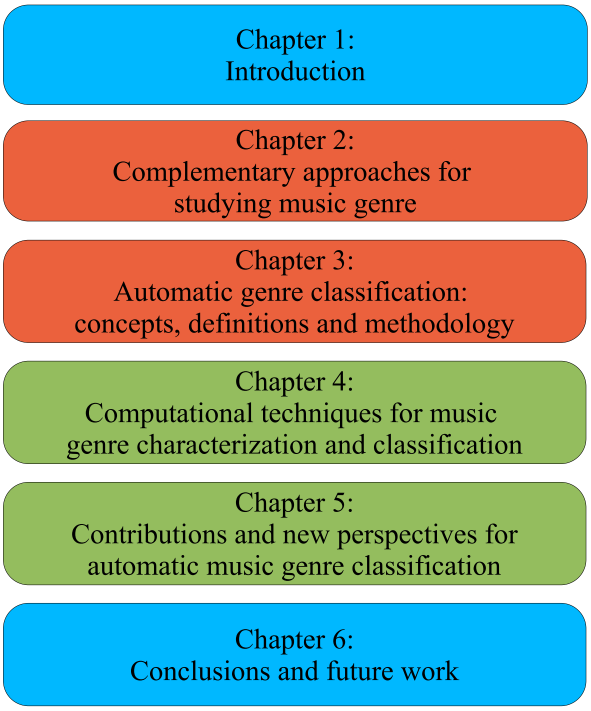
The aim of this work is to provide significant contributions to the state of the art in automatic music genre classification. Genre classification is not a trivial task. There exist a large amount of disciplines involved in it, from the objective analysis of musical parameters to the marketing strategies of music retailers. In this context, specialized classifiers with high performance rates can become complete useless if we change the dataset or we need to include a new musical genre.

We study the behavior of the classifiers in front of different datasets (and mixes of them), and we show the differences in the obtained accuracies in different contexts. We also study the influence of different descriptors and propose the use of some other new features (like danceability or panning) that have not been traditionally used for genre classification. Results are accompanied by a set of listening experiments presented to a selected group of music students in order to distinguish between the importance of two musical facets in the overall classification process. We also study the pros and cons of different classifiers and propose the use of some other classifiers that have not been used for this task.

Results show how ensembles of dedicated classifiers for each category, instead of traditional global classifiers we find in the state of the art, can help us to cross the glass-ceiling existing in the automatic classification of music genres (Aucouturier & Pachet, 2004). The proposed classifiers provide accuracies over 95% of correct classifications in real datasets but, as demonstrated in the analysis for the cross-datasets, this accuracy can decrease about 20% or more. Here again, we met a trade-off between the performance of a genre classifier and its generality, as expected, but we analyze which are the key points to minimize this problem. On the other hand, we also show how traditional descriptors related to timbre or rhythm provide the best overall results, except for some very specific experiments in which the use of other classifiers (panning, tonality, etc.) can improve the performance rates.

In parallel to all these detailed studies, we present results for some listening experiments which try to complement the output of some classifiers here analyzed, and we also discuss about the results obtained in our submission to the MIREX07[5] competition.

[5] http://www.music-ir.org/mirex/2007


Chapter 1:
Introduction
Chapter 2:
Complementary approaches for
studying music genre
Chapter 3:
Automatic genre classification:
concepts, definitions and methodology
Chapter 4:
Computational techniques for music
genre characterization and classification
Chapter 5:
Contributions and new perspectives for
automatic music genre classification
Chapter 6:
Conclusions and future work

|

Figure 1.2: Organization and structure of the PhD thesis

1.7 Organization of the Thesis

This dissertation is organized as shown in Figure 1.2. We start with a theoretical introduction on musical genres in Chapter 2. We discuss why they were created and how different disciplines deal with them. The evolution of the music industry produces constant changes in the used taxonomies in such a way that, nowadays, each user can create its own taxonomy for audio classification. In Chapter 3, we study the state of the art in automatic classification starting from a conceptual overview of the techniques and methods traditionally used, followed by a more detailed discussion on the different approaches made by the community in the last few years. At this point, the reader is expected to have an overview of the problems presented by the musical genre classification, and how the state of the art tries to solve them. We also extract some preliminary conclusions to find the strong and weak points, and present the specific contexts for our contributions. In Chapter 4, we present the technical background required for the construction of such classifiers. In Chapter 5 we will present our contributions, separated into three main areas: (1) the use of different descriptors, (2) the use of different classifiers, and (3) the use of different datasets. At the end of each part we will draw some preliminary conclusions. Finally, in Chapter 6, we present the overall conclusions and propose the future work.

CHAPTER 2

Complementary approaches for studying music genre

2.1 Introduction

This chapter starts with our own definition of *Musical Genre*. Unfortunately, an universal definition does not exist and authors differ one each other. The term *genre* comes from the Latin word *genus*, which means *kind* or *class*. According to that, genres should be described as a musical category defined by some specific criteria. Due to the inherent personal comprehension of music, this criteria can not be universally established. Then, genres will be differently defined for different people, social groups, countries, etc.

Roughly speaking, genres are assumed to be characterized by the instrumentation, rhythmic structure, harmony, etc., of the compositions that are included into a given category. But there are many external factors which are not directly related to music that influence this classification (performer, lyrics, etc.). The major challenge in our approach for the study of automatic genre classification is to define and set all the musically dependent factors as precise as possible.

2.1.1 Definition

In this section, we will provide our own definition of music genre that will be used from here to the end of this thesis. But we have to start studying the existing ones. According to the Grove Music Online (Samson, 2007) a genre can be defined as:

> a class, type or category, sanctioned by convention. Since conventional definitions derive (inductively) from concrete particulars, such as musical works or musical practices, and are therefore subject to change, a genre is probably closer to an 'ideal type' (in Max Weber's sense) than to a Platonic 'ideal form'.

Samson clarifies that genres are based on the principle of repetition: in a specific musical genre, there exists some well known patterns (coded in the past) that invite to be repeated by the future compositions. According to this definition, genre classification can be reduced to the research of these patterns. Most of these patterns are coded in the music itself as a particular rhythm, melody or harmony, but some others doesn't. In this thesis and for simplicity, we define *genre* as:

> the term used to describe music that has similar properties, in those aspects of music that differ from the others.

What does it means? Some music can be clearly identified by the instruments that belong to the ensemble (i.e. music played with Scottish bagpipes). Other genres can be identified by the rhythm (i.e. tecno music). Of course, both examples can be discussed because the instrument and the rhythm are not the only factors that define these genres. In our proposed definition, *"those aspects of music"* refer to musical and non musical properties that allows a group of works to be identified under a specific criteria. A part of the musicological perspective, this criteria can be defined under geographical, social or technological points of view, among others.

2.1.2 Genre vs Style

Genre and *Style* are often used as synonyms, but it is important to understand the difference between them. The word *style* derives from the word for a greek and roman writing implement (Lat. stilus), a tool of communication, the shaper and conditioner of the outward form of a message. According to the Grove Music Online (Pascall, 2007), the style can be defined as:

> Style is manner, mode of expression, type of presentation. For the aesthetician style concerns surface or appearance, though in music appearance and essence are ultimately inseparable. For the historian a style is a distinguishing and ordering concept, both consistent of and denoting generalities; he or she groups examples of music according to similarities between them

Our work will focus on the first part of the definition concerning to the surface or appearance of a musical piece. In other words, the style describes a 'how to play' or the personal properties of interpretation, independently of the musical genre we are dealing with. The historical definition of style can be completely confused with our previous definition of genre. From here to the end, we will not use the term *style* to refer generalities or groups of music according to a similarity criteria.

Many examples of different styles in a unique musical genre can be found. The theme *Insensatez* from Antonio Carlos Jobim can be played in different styles depending on the performer (Stan Getz, Jobim, etc.) but, in terms of genre, it will always be referred as a bosanova.

2.1.3 The use of musical genres

The use of musical genres has been deeply discussed by the MIR community and reviewed by Mckay & Fujinaga (2006). It has been suggested that music genres are an inconsistent way to organize music and that it is more useful to focus the efforts in music similarity. Under this point of view, music genre is a particular case of music similarity. But musical genres are the only labeling system that takes cultural aspects into account. This is one of the most valuable information for end users when organizing or searching in music collections. In general, the answer to *Find songs that sound like this* refers to a list of songs of the same musical genre.

Other kind of similarities are also useful (Mood, Geography) and the combination of all of them should cover most of the searching criteria in huge databases for general music enthusiasts. The relationship between humans and similarity systems should be established under musical and perceptual criteria instead of mathematical concepts. Users become frustrated when these systems propose mathematically coherent similarities that are far from the musical point of view. Good similarity algorithms must be running in the lower levels of recommending systems but filtering and classification techniques should be applied at the upper levels of recommendation systems

On the other hand, we wonder to what kind of music item the genre classification should apply: to an artist? to an album? to an individual song? Albums from the same artist may contain heterogeneous material. Furthermore, different albums from the same artist can be labelled with different music genres. This makes the genre classification a more unclear task.

2.2 Disciplines studying musical genres

Genres are not exclusive for music. Literature and Film are two disciplines with similar properties that require the classification into different categories also called genres (Duff, 2000; Grant, 2003). Research in these fields addresses issues such as how genres are created, defined, perceived and how they change with new creations that are beyond the limits of the predefined genres. Music genres can be studied from different points of view. Here, we present the most important ones.

2.2.1 Musicology

According to Futrelle & Downie (2003), musicologists included computer based and MIR techniques in their work in the last decades (Cope, 2001; Barthélemy & Bonardi, 2001; Dannenberg & Hu, 2002; Bonardi, 2000; Kornstädt, 2001).

Focusing on music genre, the nature of the studies made by musicologist is quite different: from very specific studies dealing with properties of an author or performer to the influences that specific social and cultural situations can affect to the composers. Fabbri (1981) suggests that music genres can be characterized using the following rules:

- Formal and technical: Content based practices
- Semiotics: Abstract concepts that are communicated
- Behavior: How composers and performers behaves
- Social and Ideological: Links between genres and demographics (age, race...)

- Economical: Economic systems that supports specific genres

Note how only the first rule deals with musical content. Let us remark another interesting research performed by Uitdenbogerd (2004). She discusses about different methodologies for the definition of taxonomies in automatic genre classification systems. This work is based on different experiments and surveys presented to different groups of participants with different skills. Questions like “If you had to categorize music into exactly 7 categories what would they be?” are proposed. The conclusions of this work are the impossibility to establish the exact number of musical genres and a fixed taxonomy for the experiments. The author also presents a methodology to better perform a musical genre classification task, summarized as follows:

1. Acquire a Collection
2. Choose categories for the collection
3. Test and refine categories
    a) Collect two sets of human labels for a small random subset of the collection
    b) Produce a confusion matrix for the two sets
    c) Revise category choices, possibly discarding the worst categories
4. Revise category choices, possibly discarding the worst categories
5. Collect at least two sets of human labels for the entire collection
6. Run the experiment
7. Include in the analysis a human confusion matrix to quantify category fuzziness. Report on the differences between the human and machine confusion matrices.

She also enumerates the common errors made in a categorical tree definition process:

Overextension: This is the case for a concept that is applied more generally than it should, i.e. when a child calls all pieces of furniture “Chair”.

Underextension: This is the case for a concept that is only applied to a specific instance, i.e. when a child uses the term “Chair” to refer its own chair.

Mismatch: This is the case for a concept that is applied to other concepts without any relationship between them, i.e. when a child uses the term “chair” to refer a dog.

In our experiments, we will take into account these recommendations. Other important works study how musicians are influenced by musical genres (Toynbee, 2000) and how are they grouped and how they change (Brackett, 1995).

Store	Genres
Musicovery	Rap, R&B, Jazz, Gospel, Blues, Metal, Rock
www.musicovery.com	Vocal pop, Pop, Electro, Latino, Classical
	Soundtrack, World, Reggae, Soul, Funk, Disco
Amazon	Alternative Rock, Blues, Box Sets, Classic Rock
www.amazon.com	Broadway & Vocalists, Children's Music
	Christian & Gospel, Classical
	Classical Instrumental, Classical: Opera and Vocal
	Country, Dance & DJ, DVDs, DVDs: Musicals
	DVDs: Opera & Classical, Folk, Hard Rock & Metal
	Imports, Indie Music, International, Jazz
	Latin Music, Miscellaneous, New Age, Pop, R&B
	Rap & Hip-Hop, Rock, Soundtracks
iTunes	Alternative, Blues, Children's Music, Classical
www.apple.com	Christian & Gospel, Comedy, Country, Dance
	Electronic, Folk, Hip-Hop/Rap, Jazz, Latino, Pop
	R&B/Soul, Reggae, Rock, Soundtrack, Vocal
Yahoo! Music	Electronic/Dance, Reggae, Hip-Hop/Rap, Blues
music.yahoo.com	Country, Folk, Holyday, Jazz, Latin, New Age
	Miscellaneous, Pop, R&B, Christian, Rock
	Shows and Movies, World, Kids, Comedy, Classical
	Oldies, Eras, Themes, World

Table 2.1: First level taxonomies used by some important on-line stores

2.2.2 Industry

As detailed in Section 2.3, the industry requires musical genres for their business. Buying CDs in a store or dialing a specific radio station requires some previous knowledge about its musical genre. Music enthusiasts need to be guided to the specific music they consume. The idea is to make the search period as short as possible with successful results. In fact, traditional music taxonomies have been designed by the music industry to guide the consumer in a CD store (see below). They are based on parameters such as the distribution of CDs in a physical place or the marketing strategies proposed by the biggest music labels (Pachet & Cazaly, 2000). Nowadays, on-line stores allow less constrained taxonomies but they are also created using marketing strategies. Table 2.1 shows some examples of the first level taxonomies used by some important on-line stores.

2.2.3 Psychology

Some researchers have included the implications of music perception (Huron, 2000; Dannenberg, 2001) or epistemological analysis of music information (Smiraglia, 2001) in the Music Information Retrieval studies. Much research has been done on music perception in psychology and cognitive science (Deliáege & Sloboda, 1997; Cook, 1999).

Focusing on musical genres, research is centered on how human brain perceives the music and categorizes it. Music can not be classified using the same hierarchy schema used for animal classification because of the high amount of overlapping examples and different criteria used at the same time. Psychologists have studied these strategies and create some theories that will be discussed in detail in Section 2.4.

2.2.4 Music Information Retrieval

Finally, the MIR community is another field studying musical genres. In fact, MIR community doesn't discuss about how the musical genres should be defined or how the taxonomies should be constructed. This community tries to mix all the related disciplines to allow computers to distinguish between (in our case) musical genres. But it is not its objective to discuss about music aspects such as whether taxonomies are well defined or not, the prototype songs at each musical genre, etc. We consider that this thesis belongs to the MIR discipline.

2.3 From Taxonomies to Tags

Let's start this section providing a definition of the term *taxonomy*. Merriam-Webster provide two definitions: a) the study of the general principles of scientific classification, and b) orderly classification of plants and animals according to their presumed natural relationships. According to Kartomi (1990), a taxonomy:

> ...consists of a set of taxa or grouping of entities that are determined by the culture or group creating the taxonomy; its characters of division are not arbitrarily selected but are "natura" to the culture or group.

When building taxonomies, we apply one division criterion per step, and then proceed ”downward” from a general to a more particular level. The result is a hierarchy of contrasting division (items at different levels contrast with each other).

On the other hand, *ontology* is also an important concept related to genre classification. Merriam Webster provides two definitions: a) a branch of metaphysics concerned with the nature and relations of being and b) a particular theory about the nature of being or the kinds of existents. According to Gruber (1993), an ontology is a specification of the conceptualization of a term. This is probably the most widely accepted definition of this term (McGuinness, 2003). Controlled vocabularies, glossaries and thesauri are examples of simple ontologies. Ontologies generally describe:

- Individuals: the basic or "ground level" objects
- Classes: sets, collections, or types of objects
- Attributes: properties, features, characteristics, or parameters that objects can have and share
- Relations: ways that objects can be related to one another
- Events: the changing of attributes or relations

Levels	Example 1	Example 2
Global category	Pop	Jazz
Sub-category	General	Live Albums
Artist	Avril Lavigne	Keith Jarrett
Album	Under my skin	Koln Concert

Table 2.2: Two examples of the industrial taxonomy

Due to the automatic classification algorithms try to establish relationships between musical genres, maybe it is more appropriate to talk about ontologies instead of taxonomies. For historical reasons, we will use the term taxonomy even in those cases discussing about attributes or relations between them.

Many taxonomies from different known libraries or web sites can differ a lot (see Table 2.1). All these taxonomies are theoretically designed by musicologists and experts. If they do not coincide, does it mean that all of them are in a mistake? Of course, not. The only problem is that different points of view of music are applied for their definition.

2.3.1 Music Taxonomies

Depending on the application, taxonomies dealing with musical genres can be divided into different groups (Pachet & Cazaly, 2000): taxonomies for the music industry, internet taxonomies, and specific taxonomies.

Music industry taxonomies: These taxonomies are created by big recording companies and CD stores (i.e. RCA, Fnac, Virgin, etc.). The goal of these taxonomies is to guide the consumer to a specific CD or track in the shop. They usually use four different hierarchical levels:

1. Global music categories
2. Sub-categories
3. Artists (usually in alphabetical order)
4. Album (if available)

Table 2.2 shows two examples of albums by using this taxonomy (we will not discuss whether they are right or not). Although this taxonomy has shown its usability by large, some inconsistencies can be found:

- Most of the stores have other *sections* with promotions, collections, etc.
- Some authors have different recordings which should be classified in another Global Category
- Some companies manage the labels according to the copyright management

even so, it is a good taxonomy for music retailers.

Internet Taxonomies: Although these taxonomies are also created under commercial criteria, they are significantly different from the previous ones

#	Path
1	Styles → International → Caribbean&Cuba → Cuba → → Buena Vista Social Club
2	Music for Travelers → Latin Music → Latin Jazz → → Buena Vista Social Club
3	Styles → Jazz → Latin Jazz → Buena Vista Social Club

Table 2.3: Three different virtual paths for the same albumin a Internet taxonomy (Amazon)

because of the multiple relationships that can be established between authors, albums, etc. Their main property is that music is not exposed in a specific physical place. Then, with these multiple relationships, the consumer can browse according to multiple criteria. Table 2.3 shows three different paths or locations for the same album found in Amazon. As in the previous case, some inconsistencies are also found here, specially from the semantic point of view:

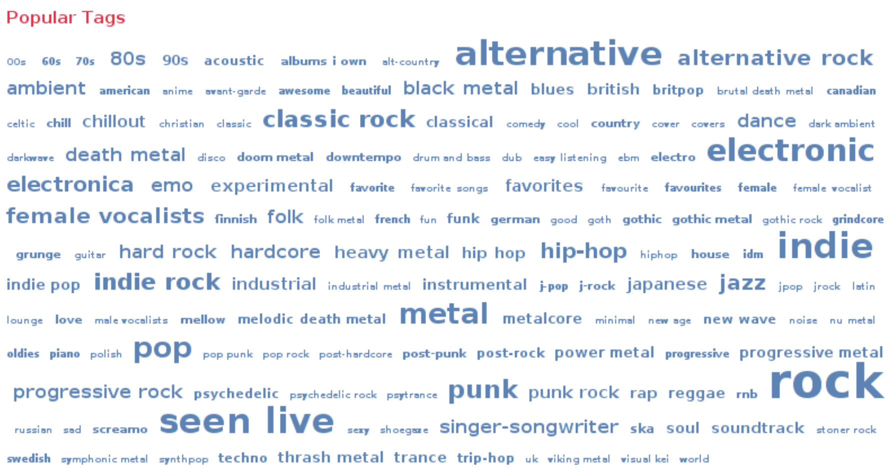
- Hierarchical links are usually genealogical. But sometimes, more than one father is necessary. i.e. both *Pop* and *Soul* are the "fathers" of *Disco*
- In most of the taxonomies, geographical inclusions can be found. It is really debatable whether this classification is correct or not. Some sites propose the category *World Music* (eBay[1]) in which, strictly speaking, one should be able to find some *Folk* music from China, *Pop* music of Youssou N'Dour and *Rock* music of Bruce Springsteen.
- Aggregation is commonly used to join different styles: *Reggae-Ska* → *Reggae* and *Reggae-Ska* → *Ska* (eBay).
- Repetitions can also be found: *Dance* → *Dance* (AllMusicGuide).
- Historical Period labels may overlap, specially in classical music: *Baroque* or *Classical* and *French Impressionist* may overlap.
- Specific random-like dimensions of the sub-genre can create confusion.

Specific Taxonomies: Sometimes, some quite specific taxonomies are needed, even if they are not really exhaustive or semantically correct. A good example can be found in music labelled as *Ballroom*, in which *Tango* can include classical titles from "Piazzolla" as well as electronic titles from "Gotan Project".

2.3.2 Folksonomies

The previous section showed how complicated can be to establish a universal taxonomy. This situation has became more complex in the recent years because of the fast growth of web publishing tools (blogs, wikis, etc.) and music

[1] www.ebay.com



Popular Tags

00s 60s 70s 80s 90s acoustic albums i own alt-country alternative alternative rock
ambient american anime avant-garde awesome beautiful black metal blues british britpop brutal death metal canadian
celtic chill chillout christian classic classic rock classical comedy cool country cover covers dance dark ambient
darkwave death metal disco doom metal downtempo drum and bass dub easy listening ebm electro electronic
electronica emo experimental favorite favorite songs favorites favourite favourites female female vocalist
female vocalists finnish folk folk metal french fun funk german good goth gothic gothic metal gothic rock grindcore
grunge guitar hard rock hardcore heavy metal hip hop hip-hop hiphop house idm indie
indie pop indie rock industrial industrial metal instrumental j-pop j-rock japanese jazz jpop jrock latin
lounge love male vocalists mellow melodic death metal metal metalcore minimal new age new wave noise nu metal
oldies piano polish pop pop punk pop rock post-hardcore post-punk post-rock power metal progressive progressive metal
progressive rock psychedelic psychedelic rock psytrance punk punk rock rap reggae rnb rock
russian sad screamo seen live sexy shoegaze singer-songwriter ska soul soundtrack stoner rock
swedish symphonic metal synthpop techno thrash metal trance trip-hop uk viking metal visual kei world

Figure 2.1: Tag map from Last.fm

distribution (myspace[2], eMule[3], etc.). From that, new strategies have emerged for music classification such as the so-called *folksonomies* that can be defined as user-generated classification schemes specified through bottom-up consensus (Scaringella et al., 2006).

This new schema difficults the design of classification tools but allow the user to organize their music with a better confidence to a personal experience. It is obvious that the music industry can not follow folksonomies, but some examples show how they can influence traditional music taxonomies (i.e. *Reggaeton*).

Folksonomies have emerged due to the growth of internet sites like Pandora[4] or Last.fm[5]. They allow users to tag music according to their own criteria. For instance, Figure 2.1 shows the Tag map in Last.fm. Now, users can organize their music collection using tags like *Late night*, *Driving* or *Cleaning* music. The functionality of music is included in the tags but it is not the unique information that can be included. For instance, specific rhythm patterns, geographical information or musical period tags can also be included.

A particularity of tags is that all the terms in the namespace have no hierarchy, that is, all the labels have the same importance. Nevertheless, we can create clusters of tags based on a specific conceptual criteria. Tags like *Dark* or *Live* or *Female voice* are at the same level than *Classic* or *Jazz*. In general, the use of tags difficult the classification of music into a hierarchical architecture.

2.3.3 TagOntologies and Folk-Ontologies

Taxonomies are conceptually opposed to folksonimies. While taxonomies show a hierarchical organization of terms, folksonomies assume that all the terms

[2] www.myspace.com
[3] www.emule-project.net/
[4] www.pandora.com
[5] www.last.fm

are at the same level of abstraction. Gruber (2007) proposed the use of the *Tagontologies*, that can be defined as:

> TagOntology is about identifying and formalizing a conceptualization of the activity of tagging, and building technology that commits to the ontology at the semantic level.

The main idea behind tagontologies is to allow systems to reason about tags. For instance, to define synonym tags, clusters of tags, etc. This means that, in fact, we are tagging the tags, creating a hierarchical organization. This could be interpreted as the mid-point between folksonomies and taxonomies.

On the other hand, Fields (2007) proposed the use of *Folk-Ontologies* as an alternative to expert ontologies. They focus on more specific types of relationships between things. For instance, a folk-ontology can be defined by the links to other articles included by authors in the wikipedia. All these links point to other articles that belong to a specific ontology, but they are created by users, not by experts.

In our work, we will not use neither folksonomies nor tagontologies nor folk-ontologies. We will focus on the classical taxonomy structure because our datasets are so defined. But this doesn't mean that our conclusions couldn't be applied to a folksonomy problem. Probably, we would get reasonable results, but it is out of the scope of our work.

2.4 Music Categorization

Music Genre Classification is, in fact, a categorization problem with some specifities. In this section we, will introduce the most important theories on human categorization. These theories are not focused on music and they try to explain how classification of different concepts, which sometimes are clearly defined but sometimes not, is performed by humans. As we will see in the forthcoming chapters, all the automatic classification methods imitate some of the main properties of the categorization techniques. In Chapter 6, we will discuss about which categorization theory represents our best approach on genre classification and we will compare with the other algorithms commonly used.

According to Sternberg (1999), a *concept* is a mental representation of a class which includes our knowledge about such things (e.g. dogs, professors). A *category* is the set of examples picked out by a concept. These two terms are often used synonymously. The categorization processes try to define those categories as complete but well defined containers which perfectly explain and represent different mental concepts.

Actually, some of the categories we use date back to 2000 years ago. Musical categories have evolved differently in different musical cultures. Then, these categories, which have been developed to ease musical universe, seem to increase the entropy of it producing more disorder and confusion. The inclusion of the cultural context in which these categories are defined will help us to reduce this disorder.

According to Fabbri (1999), some questions about musical categorization arise:

- Why do we need musical categories?

- How are such categories created?
- Are historical categories like 'genre' or 'style' useful in all contexts?
- What is the status of terms like 'field', 'area', space', 'boundary' and 'frontier'?

The three theories here exposed (a) Classical theory, b) Prototype theory and c) Exemplar theory) try to answer some of these questions.

2.4.1 Classical Theory

According to Aristotle, a category is:

> ...the ultimate and most general predicate that can be attributed to anything. Categories have a logical and ontological function: they allow to define entia exactly, by relating them to their general essence. They are: substance, quality, quantity, relation, place, time, position, condition, action, passion.

If so, categories can be defined by a set of necessary and sufficient features. When a new concept needs to be classified according to the classical theory, the process will start with the extraction of all the features of the instance. Then, the classification is performed by checking whether this new instance has all the required properties to be into one of the categories or not.

Classical theory has been traditionally used until the 20th. century because it can explain most of the categorization scientific problems found. The traditional animal classification is a good example of use. This category model has been studied in depth by Lakoff (1987) and presents the following properties:

1. categories are abstract containers with things either inside or outside the category
2. things are in the same category if and only if they have certain properties in common
3. each category has clear boundaries
4. the category is defined by common properties of the members
5. the category is independent of the peculiarities of any beings doing the categorizing
6. no member of a category has any special status
7. all levels of a hierarchy are important and equivalent

The proposed schema can be interpreted as a multilayer categorization. One example is basic-level categorization made by childs in which the categorization process starts with very simple categories (the level of distinctive action). Then, he proceeds upward to superordinate categories and downward to subordinate categories. The First level of categorization shows the following properties:

- People name things more readily at that level

- Languages have simpler names for things at that level
- Categories at that level have greater cultural significance
- Things are remembered more readily at that level
- At that level, things are perceived holistically, as a single gestalt, while for identification at a lower level, specific details have to be picked out to distinguish

The classical theory shows some weak points that can be summarized as follows:

- Family resemblance: category members may be related to one another without all having properties in common
- Some categories have degrees of membership and no clear boundaries
- Generativity: categories can be defined by a generator plus rules
- Metonymy: some subcategory or submodel is used to comprehend the category as a whole
- Ideals: many categories are understood in terms of abstract ideal cases, which may not be typical or stereotypical
- Radial categories: a central subcategory plus non central extensions. Extensions are based on convention

2.4.2 Prototype Theory

Rosch (1975) was the first to provide a general perspective for the categorization problem. In her studies, she demonstrated the weaknesses of the classical theory of categories in some environments. Her name is mostly associated with the so-called prototype theory.

The prototype view assumes that there is a summary representation of the category, called a prototype, which consists of some central tendency of the features of the category members. All the classification measures will be determined by the similarity of a given instance to the prototype. When new instances are given and the feature vectors are computed, the similarity distance to the prototype is computed. When this similarity is greater than a given threshold, this new instance is considered to be part of the category. In case of multiple categorization options were available, the closest distance to the prototypes will set the categorization decision.

The computational support to the Prototype Theory was given by Hampton (1993). The similarity of a given instance to the prototype can be computed as a weighted sum of features. The weights are selected according to the relevance of that feature for that concept:

$$S(A, t) = \sum_{i=1}^{n} w_i \cdot v_i(t) \tag{2.1}$$

where t is the new instance, A is the prototype, $S(A, t)$ is the similarity of t to the category A, w_i is the weight of the i_{th} feature in the prototype and $v_i(t)$

is the feature itself. This formula provides a similarity measure to the center of the category, weighted by the importance of the features. The similarity measure of a given instance to different categories can be computed using the Luce's Choice Rule (Luce, 1959) as:

$$p(A,t) = \frac{S(A,t)}{S(A,t) + \sum_{j \neq A} S(j,t)} \tag{2.2}$$

where $p(A,t)$ is the probability of assigning t to category A. The prototype theory allows to solve some of the challenges of the Classical Theory mentioned above. First, in a categorization process, there are some differences in the typicality of some of their members. The prototype view uses this information to create the prototype according to the specificities of the most typical members but also taking into account (but with a lower weight) those less typical cases. Second, these differences on the typicality of the members lead to differences in the performance. Members near the prototype will be learned earlier and classified faster (Murphy & Brownell, 1985) even artificial categories (Rosch, 1975). This behavior is quite similar to the categorization process made by humans in a not so evident and easy classification environment. Third, a member that has not been present in the categorization process can be classified with the same performance than those than were present.

From now on, prototype theory seems to solve all the problems of the classical theory described above. But there are some limitations to take into account. First, the human categorization process based on this theory seem to use some kind of additional information to create the clusters, not only a specific distance measurement. Second, from the mathematical point of view, the centers of the clusters are located according to a strict statistic measure and, furthermore, the properties of this center do not depend (in a initial step) on the other clusters. It is the opposite for humans. They use to imagine the prototypes according to neighbors and these prototypes will be defined more or less accurately according to them. Third, humans are also able to distinguish between the properties or attributes that define a specific category while the categorization theory does not. In the prototype theory, the correlation between features and the weight of different attributes do not influence the prototype definition.

This categorization model will solve some of the problems presented in musical genre classification, but it is not sufficient to gather with all its complexity. The Exemplar theory will provide some solutions to these specific problems.

2.4.3 Exemplar Theory

The exemplar models assume that a category is defined by a set of individuals (exemplars). Roughly speaking, the classification of new instances will be defined by their similarity to the stored exemplars. Exemplar models have been studied in detail by many authors (Medin & Schaffer, 1978; Nosofsky, 1986, 1992; Brooks, 1978; Hintzman, 1986).

Initially, a category is represented as a set of representations of the exemplars. In this study, we will consider the context model of the exemplar view which states two important hypothesis:

1. The similarity of a new instance to the stored exemplars is a multiplicative function of the similarity of their features. That means that new instances

whose vectors are quite similar to a stored instance except for only one feature may lead a low similarity measure.

2. The similarity of a new instance is computed against all the existing instances in all the categories and then classifying to the category that has the greater overall similarity.

From the mathematical point of view, the similarity of the item t to the category A is the sum of the similarity to each exemplar of this category :

$$S(A,t) = \sum_{a \in A} S(a,t) \tag{2.3}$$

As a difference with the prototype view, the similarity between new and stored instances is computed as:

$$S(a,t) = \prod_{i=1}^{n} s_i \tag{2.4}$$

where $s_i = 1$ when the i_{th} features of items a and t match, and $s_i = m_i$ when they mismatch.

In some contexts, it is possible that some features have more importance than others. With the formulas shown above, all features are equally weighted. It is possible to add a weighting function to the Equation 2.4 to assign different weights to different attributes in order to obtain weighted s'_i measure.

After this definition, we can derive some properties of the exemplar view: First, as in the prototype view, some instances become more typical than others. Distance measures have lower values between the most typical elements of a category and occur more frequently. Second, differences in classification occur due to related reasons. Third, the exemplar view is able to classify the (missing) prototype correctly. This is due to the high similarity that the prototype shows with the most typical elements in that category.

Furthermore, the exemplar view is able to solve some of the problems shown by the prototype view. It is able to distinguish between the properties (features) that define center or prototype and, as a consequence of that, it allows to save much more additional information. That means that a category can be defined by some specific features while other ones can be defined by other specific features. Second, it takes into account the context to locate the centers in the prototyped space. The definition of the categories depends and, at the same time, influences all the other categories. Then, the location of the center is not based exclusively on a statistical measure as in prototype view.

The exemplar theory of categorization also shows some conceptual limitations. Generally speaking, there is no a clear evidence that the exemplar that define a category should be members of that category. Who defines what is a category or not? Who defines which are the properties that define a category? Detractors of this categorization theory show that the information of these categories is not used in classification.

The exemplar model is usually implemented by using the generalized context model (Ashby & Maddox, 1993; Nosofsky, 1986). It uses a multidimensional scaling (MDS) approach to modeling similarity. In this context, exemplars are represented as points in a multidimensional space, and similarity between exemplars is a decreasing function of their distance. As mentioned above, one of the benefits of the exemplar theory is the possibility of create different categories using different criteria (or descriptors). In this way, we assume that with the experience in a given task, observers often learn to distribute their attention over different descriptors in a manner that tends to optimize performance. Specifically, in an experiment involving multiple categories, the probability that item i is classified into Category J is given by:

$$P(J|i) = \frac{\left(\sum_{j\in J} S_{ij}\right)^{\gamma}}{\left[\sum_{K}\left(\sum_{k\in K} s_{ik}\right)^{\gamma}\right]'} \tag{2.5}$$

where s_{ij} denotes the similarity of item i to exemplar j and the index $j \in J$ denotes that the sum is over all exemplars j belonging to Category J. The parameter g is a response-scaling parameter. When $\gamma = 1$ the observer responds by 'probability matching' to the relative summed similarities. When g grows greater than 1 the observer responds more deterministically with the category that yields the largest summed similarity.

It is common to compute distance between exemplars i and j by using the weighted Minkowski power-model formula:

$$d_{ij} = \left[\sum_{m} w_m ||x_{im} - x_{jm}|^r\right]^{\frac{1}{r}} \tag{2.6}$$

where r defines the distance metric of the space. If $r = 1$ a city block distance metric is obtained and $r = 2$ defines an Euclidean distance metric. The parameters w_m are the attention weights (See Nosofsky & Johansen (2000) for details)

2.4.4 Conclusion

In this section, we have introduced three categorization theories which are, from our perspective, complementary. Human genre classification is performed under more than one criteria at the same time, hence, the classification of a specific song may use more than one theory at the same time. For instance, the classification of our specific song may be set by a rhythmic prototype and many examples of instrumentation. All the classification techniques explained in Chapter 4 are related to these theories. Their results need to be interpreted taking into account which categorization model they follow.

CHAPTER 3

Automatic genre classification: concepts, definitions, and methodology

3.1 Introduction

Music genre classification can be studied from different disciplines, as shown in Section 2.2. In this chapter, we will focus on Music Information Retrieval and, for that, we start with a short description on genre classification performed by humans, automatic classification using symbolic data and automatic classification using collaborative filtering. In Section 3.2, we show the description of the basic schema for automatic genre classification commonly used in MIR, and finally, in Section 3.3, we present a review on the state of the art.

3.1.1 Genre classification by Humans

According to Cook (1999), the musical aspects that humans use to describe music are *Pitch*, *Loudness*, *Duration* and *Timbre* but, sometimes, music is also described with terms such as *Texture* or *Style*. Music genre classification by humans probably involves most of these aspects of music, although the process is far from being fully understood (Ahrendt, 2006). As mentioned in Section 2.3, the cultural listeners' cultural background and how the industry manages musical genres affect the classification process.

An interesting experiment of basic behavior in front of two musical genres was proposed by Chase (2001). Three fish (carps) were trained to classify music into classical or blues. After the training process, carps were exposed to new audio excerpts and they classify with a very low error rate. As reported by Crump (2002), pigeons have demonstrated the ability to discriminate between Bach and Stravinsky (Porter & Neuringer, 1984). These results suggest that

	MAMI1	MAMI2	Weighted rating
Random	26%	30%	35%
Automatic	57%	69%	65%
Human	76%	90%	88%

Table 3.1: Results for the human/non human genre classification experiments proposed by Lippens et al. (2004) using two datasets (MAMI1, MAMI2). The weighted rating is computed according to the number of correct and non correct classifications made by humans for each category

genre is not a cutural issue and this information is intrinsic to the music, and that the music encoding mechanisms are not highly specialized but generalists.

Perrot & Gjerdigen (1999) showed how humans are really good in musical genre classification. Humans only need about 300 milliseconds of audio information to accurately predict a musical genre (with an accuracy above 70%) (Unfortunately, as noticed by Craft et al. (2007), this study is still unpublished but the reader can analyze our own results on human genre classification in Section 5.3). This suggests that it is not necessary to construct any theoretic description in higher levels of abstraction -that require longer analyses- for genre classification, as described by Martin et al. (1998).

Dalla (2005) studies the abilities of humans to classify sub-genres of classical music. The test consists in the classification of 4 genres from baroque to post-romantic, all of them in the *Classical music* category. Authors investigate the so-called "historical distance" in the sense that music which is close in time will also be similar in sound. Results suggest that the subjects use the temporal variability in the music to discriminate between genres.

Another interesting experiment on human classification of musical genres was proposed by Soltau et al. (1998). He exposed 37 subjects to exactly the same training set that a machine learning algorithm. Results show how confusions made by humans were similar to confusions made by the automatic system.

Lippens et al. (2004) perform an interesting comparison test between automatic and human genre classification using different datasets, based on MAMI dataset (See Section 3.3.1 for details). Results show how there is a significant subjectivity in genre annotation by humans and, as a consequence of that, automatic classifiers are also subjective. Results on this research are shown in Table 3.1.

Many other studies can be found in the literature (Futrelle & Downie, 2003; Hofmann-Engl, 2001; Huron, 2000; Craft et al., 2007) and the conclusion is that, although results show that genre classification can be done without taking into account the cultural information, one can find many counter-examples for pieces with similar timbre, tempo or whatever showing that "culturaly free classification" is not possible. Here there is a short list of examples:

- Beethoven sonata and a Bill Evans piece: Similar instrumentations belong to different music genres
- Typical flute sound of Jethro Tull and the overloaded guitars from the Rolling Stones: Different timbres belong to the same music genre

	Leaf Categories
Jazz	Bebop, Jazz&Soul, Swing
Popular	Rap, Punk, Country
Western Classical	Baroque, Modern Classical, Romantic

Table 3.2: Taxonomy with 9 leaf categories used in MIREX05

Then, automatic classifiers should take into account both intrinsic and cultural properties of music to classify it according to a given taxonomy. Although the human mechanisms used for music classification are not perfectly known, a lot of literature can be found. But, according to our knowledge, it doesn't exist any automatic classifier that is capable to include cultural information to the system. In our opinion, this is one of the issues that MIR needs to address.

3.1.2 Symbolic classification

Since the beginning of this thesis we have been discussing about automatic genre classification based on audio data. In many cases, music is represented in a symbolic way such as scores, tablatures, etc. This representation can also provide significant information about the musical genre. The parameters that are represented in this information are related to the intensity (*ppp..fff*, regulators, etc.), timbre (instruments), rhythm (BPM, *ritardando*, *largo*, etc.) melody (notes) and harmony (notes played at the same time). It is possible to study the different historical periods, artists or musical genres according to specific musical resources they use. For instance, the use of the minor seventh in the dominant chords is quite typical in Jazz music, or the instruments represented in a orchestra clearly leads to a specific repertoire in the classical music.

The most common way to represent symbolic data in a computer is using MIDI[1] or XML[2] formats. Scanned scores (in GIF or JPG) or PDF files are not considered symbolic digital formats due to they need a parser to translate their information from image to music notation.

The main advantage when using symbolic representation is that feature extraction is simpler. Instruments, notes and durations are given by the the data itself. From this data, many statistics can be computed (histogram of duration, note distribution, most common intervals, etc.)

Focusing on symbolic audio and genre classification, many successful approaches have been developed by the MIR community. Some interesting proposals were compared in the MIREX competition made in 2005[3]. Two sets of genre categories were used. These categories were hierarchically organized and participants need to train and test their algorithms twice (See Table 3.2 and Table 3.3).

[1] Standard for the Musical Instrument Digital Interface proposed by Dave Smith in 1983

[2] eXtensible Markup Language

[3] www.music-ir.org/mirex2005/index.php/Symbolic_Genre_Classification

	Subcategory	Leaf
Country	Bluegrass	
Country	Contemporary	
Country	Trad. Country	
Jazz	Bop	Bebop, Cool
Jazz	Fusion	Bossanova,Soul,Smooth Jazz
Jazz	Ragtime	
Jazz	Swing	
Modern Pop	Adult Contemporary	
Modern Pop	Dance	Dance Pop,Pop Rap,Techno
Modern Pop	Smooth Jazz	
Rap	Hardcore Rap	
Rap	Pop Rap	
Rhythm and Blues	Blues	Rock, Chicago,Country,Soul
Rhythm and Blues	Funk	
Rhythm and Blues	Jazz Soul	
Rhythm and Blues	Rock and Roll	
Rhythm and Blues	Soul	
Rock	Classic Rock	Blues Rock,Hard Rock,Psycho
Rock	Modern Rock	Alternative,Hard,Metal,Punk
Western Classical	Baroque	
Western Classical	Classical	
Western Classical	Early Music	Medieval,Renaissance
Western Classical	Modern Classical	
Western Classical	Romantic	
Western Folk	Bluegrass	
Western Folk	Celtic	
Western Folk	Country Blues	
Western Folk	Flamenco	
Worldbeat	Latin	Bossanova,Salsa,Tango
Worldbeat	Reggae	

Table 3.3: Taxonomy with 38 leaf categories used in MIREX05

	Accuracy	38 classes	9 classes
McKay	77.17	64.33	90.00
Basili (Alg1)	72.08	62.60	81.56
Li	67.57	54.91	80.22
Basili (Alg2)	67.14	57.61	76.67
Ponce	37.76	24.84	50.67

Table 3.4: Results for the Symbolic Genre Classification at MIREX05

Four authors participated in the contest (Mckay & Fujinaga, 2005; Basili et al., 2005; Ponce & Inesta, 2005). Results for the two datasets are shown in Table 3.4[4]. According to these results, we conclude that accuracies about 75% can be obtained using symbolic genre classification.

Some other interesting approaches on genre classification based on symbolic data can be found in the literature. The first interesting approach was proposed by Dannenberg et al. (1997). This work is considered one of key papers for automatic genre classification by the inclusion of machine learning techniques to audio classification. Authors use 13 low level features (averages and deviations of MIDI key numbers, duration, duty-cycle, pitch and volume as well of notes, pitch bend messages and volume change messages). It is computed over 25 examples of 8 different styles and trained using 3 different supervised classifiers: bayesian classifier, linear classifier and neural networks. The whole dataset is divided in the train (4/5 of the whole database) and test (1/5 of the whole database) subsets. Results show accuracies up to 90% using the 8 musical styles. Variations of these confidence values are also shown as a function of the amount of data used to train the classifier.

Another interesting MIDI-based genre classification algorithm is proposed by Basili et al. (2004). Authors try to create a link between music knowledge and machine learning techniques and they show a brief musical analysis before the computational stuff. They also discuss about the confusions made by humans in the manual annotation task (Pop vs Rock and Blues, Jazz vs. Blues...). After the computation of different features extracted from MIDI data (Melodic Intervals, Instrument classes, Drumkits, Meter and Note Extension) they discuss about the results obtained by the application of different machine learning techniques. Two kinds of models are studied: single multiclass categorization and multiple binary categorization. Discussions about results are shown but global numerical performances are not provided.

McKay & Fuginaga (2004) propose an automatic genre classification using large high-level musical feature sets. The system extracts 109 musical features from MIDI data. Genetic Algorithms (GA) are used to select the best features at the different levels of the proposed hierarchical classifier in order to maximize results in a KNN classifier. The experiment is built over 950 MIDI recordings for training and testing, using a 5 fold cross-validation process. These songs belong to 3 root music genres (Classical, Jazz, Popular) and 9 leaf music genres (Baroque-Modern-Romantic; Bebop-FunkyJazz-Swing; Coutry-Punk-Rap). Results are impressive: 97% of correct classifications at the root level and 88% of accuracy at leaf levels. These results reveal that a good starting point for

[4] www.music-ir.org/evaluation/mirex-results/sym-genre/index.html

genre classification is an exhaustive pool of features, and selecting those which are relevant to our problem using machine learning or genetic algorithms.

Lidy & Rauber (2007) propose the combination of symbolic and audio features for genre classification. Their experiment uses an audio dataset from which it computes a set of descriptors (Rhythm patterns, Spectrum descriptors, etc.), and applies a state of the art transcription system to extract a set of 37 symbolic descriptors (number of notes, number of significant silences, IOI, etc.). The classification is performed over Tzanetakis, Ballroom Dancers and Magntune datasets (See Section 3.3.1 for details) using Support Vector Machines (SVM). Results show how accuracies obtained by combining symbolic and audio features can increase up to 6% the overall accuracy using only audio descriptors, depending on the dataset and the selected descriptors for the combination.

Other interesting approaches combine techniques of selection and extraction of musically invariant features with classification using (1) compression distance similarity metric (Ruppin & Yeshurun, 2006), (2) Hidden Markov Models (HMM) (Chai & Vercoe, 2001), (3) the melodic line (Rizo et al., 2006a), (4) combinations of MIDI related and audio features (Cataltepe et al., 2007), and (5) creating a hierarchical classifiers (De Coro et al., 2007).

3.1.3 Filtering

Having a look to the internet music stores, we observe how they classify music according to genres. They use this information to propose the user new CDs or tracks. In most of these cases, the analysis of music genres is not performed by Music Information Retrieval techniques. In addition to the manual labeling, they use some other techniques such as Collaborative Filtering to group music according to a specific ontology. As cited by Aucouturier & Pachet (2003), Collaborative Filtering term was proposed by Shardanand & Maes (1995) and it is defined as follows:

> *Collaborative Filtering (CF) is based on the idea that there are patterns in tastes: tastes are not distributed uniformly. These patterns can be exploited very simply by managing a profile for each user connected to the service. The profile is typically a set of associations of items to grades. In the recommendation phase, the system looks for all the agents having a similar profile than the user's. It then looks for items liked by these similar agents which are not known by the user, and finally recommends these items to him/her.*

Although CF is out of the scope of this thesis, we think it is necessary to write some words because it helps us to understand which are the limits of genre classification using MIR. Experimental results by using CF show impressive results once a sufficient amount of initial ratings are provided by the user, as reported by Shardanand & Maes (1995). However, Epstein (1996) showed how limitations appear when studying quantitative simulations of CF systems: First, the system creates clusters which provide good results for naive classifications but unexpected results for non typical cases. Second, the dynamics of these systems favors the creation of hits, which is not bad a priori, but difficult the survival of the other items in the whole dataset.

As reported by Celma (2006), many approaches for audio recommendation are based on relevance feedback techniques (also known as community-based systems) which implies that the system has to adapt to the changes of users' profiles. This adaptation can be done in three different ways: (1) manually by the user, (2) adding new information to the user profile, or (3) gradually forgetting the old interests of the user and promoting the new ones. Once the user profile has been created, the system has to exploit the user preferences in order to recommend new items using a filtering method. The method adopted for filtering the information has led to the classification of recommender systems:

Demographic filtering According to Rich (1979), demographic filtering can be used to identify the kind of users that like a certain item. This technique classifies users profiles in clusters according to some personal data (age, marital status, gender, etc.), geographic data (city, country) and psychographic data (interests, lifestyle, etc.).

Collaborative filtering According to Goldberg et al. (1992), collaborative filtering uses the user feedback to the system allowing the system to provide informed guesses, based on ratings that other users have provided. These methods work by building a matrix of users' preferences (or ratings) for items. A detailed explanation of these systems was proposed by Resnick & Varian (1997).

Content-based filtering the recommender collects information describing the items and then, based on the user's preferences, it predicts which items the user will like. This approach does not rely on other users' ratings but on the description of the items. The process of characterizing the item data set can be either automatic or based on manual annotations made by the domain experts. These techniques have its roots in the information retrieval (IR) field. The early systems were focused on text domain, and applied techniques from IR to extract meaningful information from the text.

Many successful applications using CF can be found in the field of music classification and selection, as the work proposed by Pestoni et al. (2001); French & Hauver (2001), but the main problem, according to Pye (2000), is that:

> *it requires considerable data and is only applicable for new titles some time after their release.*

There are some interesting works which are not directly related to the Collaborative filtering, but they also use the textual data available in the Internet. First, Knees et al. (2004) present an artist classification method based on textual data in the web. The authors extract features for artists from web-based data and classify them with Support Vector Machines (SVM). They start by comparing some preliminar results with other methods found in the literature and, furthermore, they investigate the impact on the results of fluctuations over time of the analyzed data from search engines for 12 artists every day for a period of 4 months. Finally, they use all this information to perform genre classification with 14 genres and 224 artists (16 artists per genre). In a more

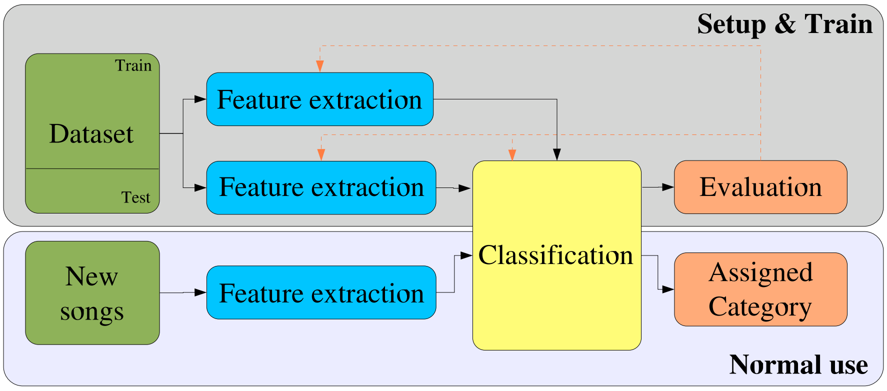


Figure 3.1: Overall block diagram for building a MIR based automatic genre classification system

recent study, Knees shows how accuracies up to 87% are possible. Particular results are obtained with only 2 artist defining a genre, reaching accuracies about 71%.

Bergstra (2006) explores the value of FreeDB[5] as a source of genre and music similarity information. FreeDB is a public and dynamic database for identifying and labeling CDs with album, song, artist and genre information. One quality of FreeDB is that there is high variance in, e.g., the genre labels assigned to a particular disc. Authors investigate the ability to use these genre labels to predict a more constrained set of canonical genres as decided by the curated but private database AllMusic[6].

3.2 Our framework

Since the MIR community started analyzing music for further retrieval, there exists some common properties in their methods and procedures. The basic process of building a MIR classifier is defined by four basic steps: (1) Dataset collection, (2) Feature extraction, (3) Machine learning algorithm and (4) Evaluation of the trained system. Figure 3.1 shows a block diagram for a basic system.

All the approaches proposed by the authors can be characterized by the techniques used at each one of these steps. Some authors focus on a specific part of the whole system, increasing the performance for specific aplications, while others compare the use of different datasets, features or classifiers in a more general environment. In the following sections we will discuss about these four main steps in detail.

3.2.1 The dataset

As shown in Section 2.3.1, it doesn't exist a universal taxonomy for musical genres. There are different parameters that influence the construction of a

[5] www.freedb.org

[6] www.allmusic.com

dataset and, as a consequence of that, the results of the classifier may vary a lot. Here is a list of the most important ones:

Number of genres: The number of musical genres is one of the most important parameters in the design of a dataset. It sets the theoretical baseline for random classification which is, in the case of an equally distributed dataset, computed according to the following formula:

$$accuracy(\%) = \frac{1}{n} \cdot 100 \tag{3.1}$$

where n is the number of musical genres. The accuracies obtained by the automatic classification need to be relative to this value.

Size: There is no universal size for a music genre dataset. There are a few approaches using less than 50 files (Scott, 2001; Xu, 2005) but most of them use larger datasets. A priori, one could assume that the bigger the dataset the better the results, but it is not always true: few representative audio excerpts may better represent the genre space than a large number of ambiguous files. Depending on the goal of the research, maybe it is enough to train a system with a few number of representative audio excerpts per class.

Length of the audio excerpt: Audio excerpts of 10, 30 or 60 seconds extracted from the middle of the song maybe enough to characterize music. This can reduce the size of the dataset and the computational cost without reducing its variability and representativeness. When deciding the size of the dataset, we should take into account that the classification method needs to be tested by using cross-fold validation, splitting the dataset in a train set plus a test set (typically $66\% - 33\%$ respectively), or even better, with an independent dataset. All these techniques will be discussed in Section 3.2.5.

Specificity: The specificity of the selected musical genres will also affect the behavior of the classifier. A priori, general taxonomies produce better results than specialized taxonomies. For instance, it is easier to distinguish between *Classic*, *Jazz* and *Rock* than between *Pop*, *Rock* and *Folk*. Furthermore, some of the datasets found in the literature use a subgroup of songs in a specific musical genre to represent it. This may produce biased results and decreases the performance of the overall system. On the other hand, for some specialized taxonomies, the extracted features and classification algorithms can be tuned to obtain better results than traditional classifiers (i.e. the ballroom music classification proposed by Gouyon & Dixon (2004)).

Hierarchical taxonomy: They are used in datasets with a large number of classes and provide many benefits to the classification. First, specific descriptors or classifiers can be applied to different subgroups of musical genres. Second, post-processing filtering techniques can be applied to increase the overall accuracy at the coarse levels of classification. If the classifier considers all the musical genres at the same level, the hierarchy is not used (i.e. “flat classification“) and the most detailed labels will

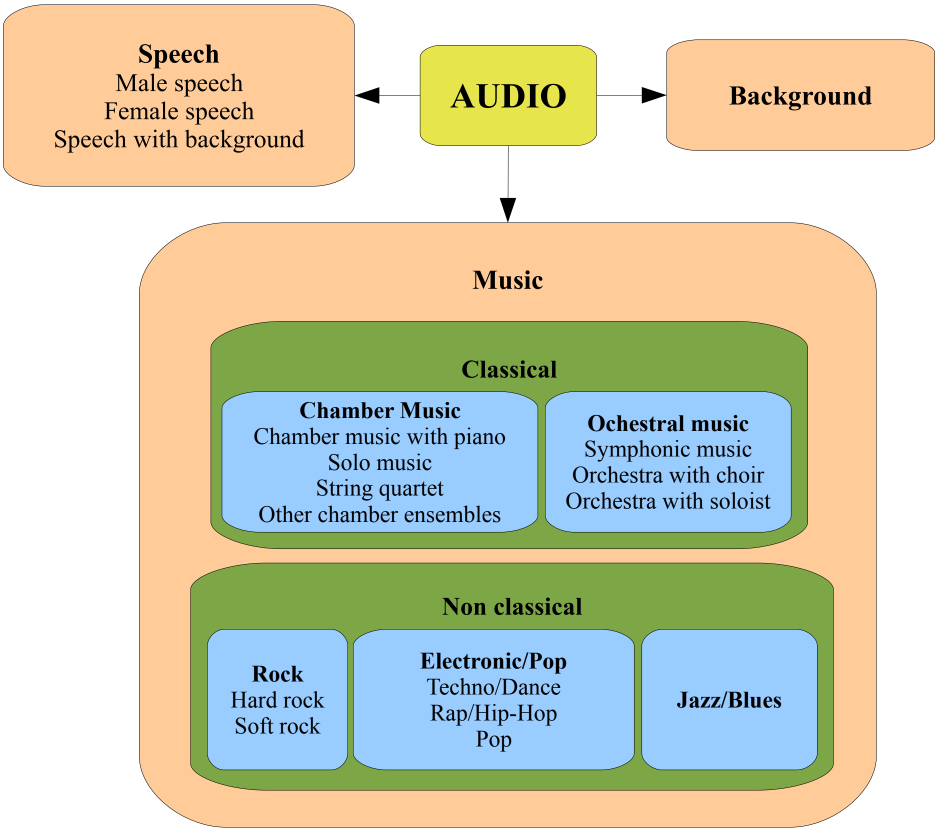


Figure 3.2: Hierarchical taxonomy proposed by Burred & Lerch (2003)

be used. An example of hierarchical taxonomy was proposed by Tzanetakis et al. (2001b). Burred & Lerch (2003) also proposed the use of a hierarchical taxonomy, shown in Figure 3.2.

Variability: Datasets should be built with the maximum variability of music for a specific class by including different artists and sub-genres in it. For that, it is recommended to use only one song per author in the dataset. Moreover, mastering and recording techniques can be similar for all the songs in an album. This phenomena is known as *the producer effect* and it was observed by Pampalk et al. (2005a). In case of using more than one song in an album, the classifier may bias to a specific audio property (timbre, loudness, compression, etc.) that is representative of the album but not of the category. In other words, a given feature might be *over-represented*.

Balance of the dataset: The number of songs for each genre should be similar. There are some well balanced datasets (Tzanetakis & Cook, 2002; Goto et al., 2003) and others which doesn't (Magnatune[7]: see Section

[7] www.magnatune.com

3.3.1). Unbalanced datasets clearly produce biased results although the extracted features and the classification method work properly[8] (Tables 3.11 and 3.14 show the main properties of these datasets).

License: Collections can be built using personal music collections, in which case they will contain basically well known songs and artists or public music from internet repositories (Magnatune, Amazon, LastFM). Public datasets are useful for sharing and comparing results in the research community, but they are not so commonly used than the personal ones. The so called *in-House* collections provide more expandable results, but results they provide can not be shared or compared with the work of other researchers.

3.2.2 Descriptors

As shown in Section 3.1.2 and 3.1.3, automatic genre classification requires some representation of the musical information. This information can be collected from the user profiles (collaborative filtering), from symbolic repositories (XML or MIDI) or, as we will see in this section, from audio files. Most of the music available in the personal collections or in the Internet is stored in digital formats (usually CD, WAV or MP3). Whatever the format is, data can be decoded and transformed to a succession of digital samples representing the waveform. But this data can not be used directly by automatic systems because pattern matching algorithms can not deal with such amount of information.

At this point, automatic classifiers must analyze these samples and extract some features that describe the audio excerpts using a compact representation. These descriptors can be computed to represent some specific facets of music (timbre, rhythm, harmony or melody). But these are not the unique families of descriptors that can be extracted: some descriptors in a higher level of abstraction can also be obtained (mood, danceability, etc.). According to Orio (2006), the most important facets of music related to the MIR community are the following:

Timbre: It depends on the quality of sounds, that is, the used musical instruments and the playing techniques.

Orchestration: It depends on the composers' and performers' decisions. They select which musical instruments are to be employed to play the musical work.

Acoustics: It depends on some characteristics of timbre, including the contribution of room acoustics, background noise, audio post-processing, filtering, and equalization.

Rhythm: It depends on the time organization of music. In other words, it is related to the periodic repetition, with possible small variants, of a temporal pattern.

[8] After many discussions in MIREX2005, a new balanced collection was collected for MIREX2007 genre classification task

	Dimension	Content
short-term	Timbre	Quality of the produced sound
	Orchestration	Sources of sound production
	Acoustics	Quality of the recorded sound
mid-term	Rhythm	Patterns of sound onsets
	Melody	Sequences of notes
	Harmony	Sequences of chords
long-term	Structure	Organization of the musical work

Table 3.5: Facets of music according to the time scale proposed by Orio (2006)

Melody: It is built by a sequence of sounds with a similar timbre that have a recognizable pitch within a small frequency range. The singing voice and monophonic instruments that play in a similar register are normally used to convey the melodic dimension.

Harmony: It depends on the time organization of simultaneous sounds with recognizable pitches. Harmony can be conveyed by polyphonic instruments, by a group of monophonic instrument, or may be indirectly implied by the melody.

Structure: It depends on the horizontal dimension whose time scale is different from the previous ones, being related to macro-level features such as repetitions, interleaving of themes and choruses, presence of breaks, changes of time signatures, and so on.

On the other hand, music is organized in time. It is well known that music has two dimensions: a horizontal dimension that associates time to the horizontal axis and, in the case of polyphonic music, the vertical dimension that refers to the notes that are simultaneously played. Not all the facets of music described above can be computed in the two dimensions, i.e. melody occurs in the horizontal dimension while harmony occur in both horizontal and vertical dimension. All the facets that take place in the horizontal dimension need to be computed at different time scales. Timbre, orchestration, and acoustics are more related to the perception of sounds and can be defined as *short-term* features (humans spend about a few milliseconds to compute them). Rhythm, melody and harmony are related to the time evolution of the basic elements, so, they can be defined as *mid-term* features. Finally, structure is clearly a *long-term* because it depends on the short-term and mid-term features as well as the cultural environment and knowledge of the musician/listener. Table 3.5 summarizes the facets of music according to the horizontal scale.

Similarly, Koelsch & Siebel (2005) propose a neurocognitive model of music perception in which specifies the time required for the human brain to recognize music facets. In his study, they show how cognitive modules are involved in music perception and incorporates information about where these modules might be located in the brain. The proposed time scales are shown in Table 3.6.

Concept	Time(ms)
Feature Extraction (Pitch height, Pitch croma, Timbre, Intensity, Roughness)	10..100
Auditory, Sensory, Memory	100.200
Structure Building (Harmony, Meter, Rhythm, Timbre)	180..400
Meaning	250..500
Structural reanalysis and repair	600..900

Table 3.6: Time required for the human brain to recognize music facets according to the neurocognitive model of music perception proposed by Koelsch & Siebel (2005)

3.2.3 Dimensionality reduction

The feature extraction process can provide the classifier a large amount of data. In most of the cases, not all of the computed descriptors provide useful information for classification. The use of these descriptors may introduce some noise to the system. For instance, it is well known that the MFCC0 coefficient is related to the energy of the input audio data. For automatic genre classification, this descriptor is, a priori, not useful at all because it is more related to the recording conditions than to the musical genre itself. If we avoid the use of this descriptor, the classifier is expected to yield better accuracies.

There exist different techniques to reduce the dimensionality of feature vectors according to their discrimination power. These techniques can be divided into two main groups: the feature selection and the space transformation techniques. Here, we present a short list of the most important ones.

Feature Selection

The feature selection techniques try to discriminate the useless descriptors of the feature vector according to a given criteria, without modifying the other ones. This discrimination is computed for all the given vectors at the same time. Here, we show a brief description of some existing methods:

CFS Subset Evaluation: According to Hall (1998), the CFS Subset Evaluation evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having small intercorrelation are preferred.

Entropy: According to Witten & Frank (2005), the entropy based algorithms selects a subset of attributes that individually correlate well with the class but have small intercorrelation. The correlation between two nominal attributes A and B can be measured using the symmetric uncertainty:

$$U(A,B) = 2\frac{H(A)+H(B)-H(A,B)}{H(A)+H(B)} \tag{3.2}$$

where H is the entropy of the selected descriptor.

Gain Ratio: The information gain is defined as the transmitted information by a given attribute about the object's class (Kononenko, 1995).

$$Gain = H_C + H_A - H_{CA} = H_C - H_{C|A} \tag{3.3}$$

where H_C, H_A, H_{CA} and $H_{C|A}$ are the entropy of the classes, of the values of the given attribute, of the joint events class-attribute value, and of the classes given the value of the attribute, respectively. In order to avoid the overestimation of the multi valued attributes, Quinlan (1986) introduced the gain-ratio:

$$GainR = \frac{Gain}{H_A} \tag{3.4}$$

In other words, it evaluates the worth of an attribute by measuring the gain ratio with respect to the class.

Space Transformation

The feature space transformation techniques try to reduce the dimensionality while improving class representation. These methods are based on the projection of the actual feature vector into a new space that increases the discriminability. Typically, the dimension of the new space is lower than the original one. Here, we show a brief description of three commonly used methods:

Principal Component Analysis (PCA): It finds a set of the most representative projection vectors such that the projected samples retain the most information about original samples (See Turk & Pentland (1991) and Section 4.5.1 for details).

Independent Component Analysis (ICA): It captures both second and higher-order statistics and projects the input data onto the basis vectors that are as statistically independent as possible (See Bartlett et al. (2002) and Draper et al. (2003) for details).

Linear Discriminant Analysis (LDA): It uses the class information and finds a set of vectors that maximize the between-class scatter while minimizing the within-class scatter (See Belhumeur et al. (1996) and Zhao et al. (1998) for details).

3.2.4 Machine Learning

According to the literature, there are many definitions dealing with Machine Learning (ML). Langley (1996) proposes that ML is:

> *a science of the artificial. The field's main objects of study are artifacts, specifically algorithms that improve their performance with experience.*

Mitchell (1997) proposes that ML is:

> *Machine Learning is the study of computer algorithms that improve automatically through experience.*

and Alpaydin (2004) assumes that ML is:

> *programming computers to optimize a performance criterion using example data or past experience.*

From the practical point of view, ML creates programs that optimize a performance criterion through the analysis of data. Many tasks such as classification, regression, induction, transduction, supervised learning, unsupervised learning, reinforcement learning, batch, on-line, generative models and discriminative models can take the advantage of using them.

According to Nilsson (1996), there are many reasons that make the use of ML algorithms useful. Here is a short list of these situations:

- Some tasks can not be perfectly defined but its behavior can be approximated by feeding the system with examples.
- Huge databases can hide some relationship between their members
- Some human designed classification algorithms provide low confidence with expected results. Sometimes it is caused because of the unknown origin of the relationship between members.
- Environments change over time. Some of these algorithms are capable to change according to the change of the input data
- New knowledge is constantly discovered by humans. These algorithms are capable to adapt this new data in these classification schemas.

The MIR community has traditionally used some of these techniques to classify music. Expert systems could be assumed to be the first approaches in music classification. These expert systems, in fact, are not considered machine learning due to they are just an implementation of a set of rules previously defined by humans. For the automatic musical genre classification task this means that we are capable to define a set of properties that uniquely define a specific genre. This assumes a deep knowledge of the data (musical genre in this case) and the possibility to compute the required descriptors from that music, descriptors that the current state of art can not provide. Furthermore, from the engineering point of view, these systems are very expensive to maintain due to the constant changes in music.

Focusing on ML, there are two main groups of algorithms that the MIR community traditionally uses:

Unsupervised learning: The main property of unsupervised classifiers is that the classification emerges from the data itself, based on objective similarity measures. These measures are applied to the feature vectors extracted from the original audio data. The most simple distances between two feature vectors are the euclidiean and the cosine distance. Other sophisticated ways to compute distances between feature vectors are the Kullback-Leibler distance, Earth's mover distance or using Gaussian Mixture Models. Hidden Markov models are used to model the time evolution of feature vectors. Once the distances between all the feature vectors are computed, the clustering algorithms are the responsible to

		Paradigms
Expert systems	1	Uses a Taxonomy
	2	Each class is defined by a set of features
	3	Difficult extraction of required descriptors
	4	Difficult to describe musical genres
Unsupervised Learning	1	No taxonomy is required
	2	Organization according to similarity measures
	3	K-Means, SOM and GHSOM
Supervised Learning	1	Taxonomy is required
	2	Feature mapping without musical description
	3	K-NN, GMM, HMM, SVM, LDA, NN

Table 3.7: Main paradigms and drawbacks of classification techniques, reported by Scaringella et al. (2006)

organize data. K-Means is probably the simplest and most popular clustering algorithm, but its main problem is that the number of clusters (K) must be given a priori and, most of the times, this information is not available. Self Organizing Maps (SOM) and Growing Hierarchical Self Organizing Maps (GHSOM) are used to cluster data and organize it in a 2 dimensional space, which allows an easy visualization of data. The main problem of the unsupervised learning is that, most of the times, the resulting clusters have no musical meaning, which difficult the process of interpreting the results.

Supervised learning: These methods are based on previously defined categories and try to discover the relationship between a set of features extracted from the audio data and the manually labelled dataset. The mapping rules extracted from the training process will be used to classify new unknown data. The most simple technique for supervised learning is the K-Nearest Neighbor but the most widely used technique is probably Decision Trees. Gaussian Mixture Models, Hidden Markov models, Support Vector Machines, Linear Discriminant Analysis and Neural Networks are other widely used supervised algorithms used in the literature. Supervised techniques are the most commonly used techniques for audio classification. The main advantage of using them is that an exact description of (in our case) the musical genre is not required. The provided model will be easily interpretable or not depending on the learning algorithm applied.

Table 3.7, reported by Scaringella et al. (2006), summarizes the main paradigms and drawbacks of the three classification techniques shown above. A more detailed description of the algorithms here presented is provided in Section 4.4.

3.2.5 Evaluation

Evaluation is the last step for building a classifier. Although this part may not be implemented in a real application, it is crucial when designing the classifier. Evaluation will provide information to redefine the classifier for obtaining better results. There are two main parts in the evaluation that will be discussed in the following sections. First, some post-processing techniques may be applied to the preliminary results of the classifier, depending on its output format (frame based or a single value per song). On the other hand, due to the access to datasets with properly labelled audio files is usually limited, there are some techniques to organize and reuse the same data for train and test processes.

Post-processing

Depending on the input descriptors and the architecture of the classifier, the proposed labels for new unknown data may address a whole song or specific frames of it. The musical genre is traditionally assigned to a whole song, so, if the classifier provide different labels for each frame we will need some techniques to obtain a global index. Particularly, genre classification is highly sensitive to this effect because, in a specific song, some frames can be played using acoustic or musical resources from other musical genres.

According to Peeters (2007), we can differentiate between three different techniques:

Cumulated histogram: The final decision is made according to the largest number of occurrences among the frames. Each frame is first classified separately:

$$i(t) = argmax_i \; p(c_i)|f(t)) \tag{3.5}$$

Then, we compute the histogram $h(i)$ of classification results $i(t)$. The bins of this histogram are the different musical genres. The class corresponding to the maximum of the histogram is chosen as the global class.

Cumulated probability: Here, we compute the cumulated probabilities $p(c_i|f(t))$ over all the frames:

$$p(c_i) = \frac{1}{T}\sum_t p(c_i|f(t) \tag{3.6}$$

and select the class i with the highest cumulated probability:

$$i = argmax_i \; p(c_i) \tag{3.7}$$

Segment-statistical model: This technique was proposed by Peeters (2007). It learns the properties of the cumulated probability described above by using statistical models, and perform the classification using them. Let s be the whole audio data and $p_s(c_i)$ its cumulated probability. Let S_i be the set of audio segments belonging to a specific class i in the training set. Then, for each class i, we compute the cumulated probabilities $p_{s \in S_i}(c_i)$. Then, we model the behavior of the bins c_i over all the $s \in S_i$ for a specific class i. $\hat{p}_i(c_i)$ is this segment-statistical model. For the indexing procedure, first we compute its accumulated probability $p_s(c_i)$ and classify it using the trained segment-statistical method. The statistical models to be considered can be based on means and deviations or on gaussian modeling. In other words, it learns the *patterns* of the cumulated probabilities for all the categories and classifies according to them.

These indexing techniques for evaluation here proposed should not be confused with the time domain feature integration process described in Section 3.2.3. Now, the features are independently computed (using time integration or not), but the output of the classifier may require global indexing method to compare results with the output of other classifiers.

Validation

Once system is trained, different methods to evaluate its performance can be used. It doesn't make sense to test the system with the same audio dataset used for training because this will reduce the generality of the system and we will not provide any idea on the behavior of the trained system in front of new data. The following techniques are used to train and test the classifier with the same dataset:

K-fold cross-validation: The original dataset is splitted into K equally distributed and mutually exclusive subsamples. Then, a single subsample is retained as the validation data for testing, and the remaining $K-1$ subsamples are used as training data. This process is repeated K times (the folds), with each of the K subsamples used exactly once as the validation data. The K results from the folds then can be averaged (or otherwise combined) to produce a single estimation.

Leave-one-out: It uses a unique observation from the original dataset to validate the classifier, and the remaining observations as the training data. This process is repeated until each sample in the original dataset is used once as the validation data. This process is similar than K-fold cross-validation but setting K as the number of observations in the original dataset. Sometimes, Leave-one-out is also called Jacknife.

Holdout: This method reserves a certain number of samples for testing and uses the remainder for training. Roughly speaking, it is equivalent to randomly split the dataset into two subsets: one for training and the other for testing. It is common to hold out one-third of the data for testing. From the conceptual point of view, Holdout validation is not cross-validation in the common sense, because the data is never crossed over.

Bootstrap estimates the sampling distribution of an estimator by sampling the original sample with replacement with the purpose of deriving robust estimates of standard errors of a population parameter (mean, median, correlation coefficient, etc.).

The Leave-one-out method tends to include unnecessary components in the model, and has been provided to be asymptotically incorrect (Stone, 1977). Furthermore, the method does not work well for data with strong clusterization (Eriksson et al., 2000) and underestimates the true predictive error (Martens & Dardenne, 1998). Compared to Holdout, cross-validation is markedly superior for small data sets; this fact is dramatically demonstrated by Goutte (1997) in a reply to Zhu & Rohwer (1996). For an insightful discussion of the limitations of cross-validatory choice among several learning methods, see Stone (1977).

	Subcategories (# tracks)
Ballroom:	Waltz(323), Tango(189), Viennese Waltz(137), Foxtrot(242), Quickstep(263)
Latin:	Cha Cha(215), Samba(194), Int'l Rumba and Bolero(195), American Rumba(8), Paso Doble(24), Salsa(17), Mambo(8)
Swing:	EC Swing(3), WC Swing(5), Lindy(8), Jive(140)

Table 3.8: Summary of the BalroomDancers dataset

3.3 State of the art

In this section, we analyze the art in automatic music genre classification based on audio data. For that, we start with a short description of the datasets commonly used in the MIR community to test the algorithms. Then, we introduce the MIREX contests which provide an excellent benchmark to compare the algorithms proposed by different authors and, finally, we also introduce many other interesting approaches that have been developed independently.

3.3.1 Datasets

Here we describe different datasets which are relevant to our work for different reasons: some of them are widely known and used by the whole community while the others are specially collected for this work. Our idea is to make results of our tests independent of the selected dataset, so, we need to design our experiments by combining them.

According to Herrera (2002), there are some general requirements and issues to be clarified in order to set up a usable dataset:

- Categories of problems: multi-level feature extraction, segmentation, identification, etc.
- Types of files: sound samples, recordings of individual instruments, polyphonic music, etc.
- Metadata annotation: MIDI, output of a MIR algorithm, etc.
- Source: personal collections, internet databases, special recordings, etc.

Some of the following datasets have been collected to address specific problems, but all of them can be used for genre classification.

Ballroom Dancers: This dataset is created by the 30*sec* preview audio excerpts available in the Ballroom Dancers website[9].The most interesting property of this dataset is that the BPM value is given for each song. It is created by 3 coarse cateogires and many leaf subcategories, as shown in Table 3.8.

This dataset is very specialized, unbalanced but available to the community. It has been built using a hierarchical taxonomy and it has been used

[9] secure.ballroomdancers.com/Music/style.asp

Entries	Albums	Artists	Styles	Genres
8764	706	400	251	10

Table 3.9: Summary of the USPOP dataset

	# Songs
Popular Music Database	100
Royalty-Free Music	15
Classical Music	50
Jazz Music	50
Music Genre	100
Musical Instrument Sound	50

Table 3.10: Summary of the RWC dataset

for rhythm detection and ballroom music classification by Dixon et al. (2004) and Gouyon & Dixon (2004).

USPOP: This dataset was created on 2002 by a set of full-length songs and their corresponding AllMusic meta-information (Berenzweig et al., 2004). Due to legal issues, this dataset is not freely available and only the precomputed Mel-Scale Frequency Cepstral Coefficients can be distributed. The aim of this dataset is to represent popular music using 400 popular artists. The distribution of songs, artists and genres is shown in Table 3.9. This dataset is specialized in western pop music and the number of songs at each musical genre is balanced. It has been used for many authors and also in the MIREX05 competition.

RWC: The RWC (Real World Computing) Music Dataset is a copyright-cleared music dataset that is available to researchers as a common foundation for research (Goto et al., 2003; Goto, 2004). It contains six original collections, as shown in Table 3.10.

The Music Genre subset is created by 10 main genre categories and 33 subcategories: 99 pieces (33 subcategories * 3 pieces). Finally, there is one piece labelled *A cappella*. See Table 3.11 for details. This dataset is available on request in CD format[10].

Tzanetakis: This dataset was created by Tzanetakis & Cook (2002). It contains 1000 audio excerpts of 30sec distributed in 10 musical genres (*Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae, Rock*). Audio files for this dataset are *mono*, *wav* format, using a $sr = 22050Hz$. This dataset has been used for many authors (Li & Ogihara, 2005; Holzapfel & Stylianou, 2007).

MAMI: This dataset was collected with a focus on *Query by humming* research, but also provided a good representation of western music to the whole community. It contains 160 full length tracks based on the sales

[10] staff.aist.go.jp/m.goto/RWC-MDB/

	Subcategories (# tracks)
Popular	Popular(3), Ballade(3)
Rock	Rock(3), Heavy Metal(3)
Dance	Rap(3), House(3), Techno(3), Funk(3), Soul/RnB(3)
Jazz	Big Band(3), Modern Jazz(3), Fusion(3)
Latin	BossaNova(3), Samba(3), Reggae(3), Tango(3)
Classical	Baroque(3), Classic(3), Romantic(3), Modern(3)
March	Brass Band(3)
Classical(Solo)	Baroque(5), Classic(2), Romantic(2), Modern(1)
World	Blues(3), Folk(3), Country(3), Gospel(3), African(3)
	Indian(3), Flamenco(3), Chanson(3), Canzone(3)
	Popular(3), Folk(3), Court(3)
Cappella	Cappella(1)

Table 3.11: Summary of the Music Genre - RWC dataset

	# Songs		# Songs
Blues	100	Jazz	100
CLassical	100	Metal	100
Country	100	Pop	100
Disco	100	Reggae	100
Hip-Hop	100	Rock	100

Table 3.12: Summary of the Tzanetakis dataset

information from the IFPI (International Federation of the Phonographic Industry) in Belgium for the year 2000. The songs belong to 11 musical genres but some of them are very poorly represented. Lippens et al. (2004) conducted a manual labeling process obtaining a subset formed by only 6 representative and consistent musical genres (Pop, Rock, Classical, Dance, Rap and Other). This new dataset is known as MAMI2 and some works are based on it (Craft et al., 2007; Lesaffre et al., 2003).

Garageband: Garageband[11] is a web community that allows free music download from artists that upload their work. Visitors are allowed to downoad music, rate it and write comments to the authors. Although it is a continuously changing collection, many works derived from this dataset. First, some students downloaded it and gathered some metadata. They manually classified music (1886 songs) into 9 musical genres (Pop, Rock, Folk / Country, Alternative, Jazz, Electronic, Blues, Rap / Hip-Hop, Funk / Soul) and compute some descriptors to perform audio classification experiments (Homburg et al., 2005; Mierswa & Morik, 2005). All this information is currently available in the web[12]. On the other hand, a recent work proposed by Meng (2008) updates and redefines the dataset based on Garageband. The author fuse the original 16706 songs distributed into 47 categories into a smaller 18 genre taxonomy. After that, Jazz became

[11] www.garageband.com

[12] www-ai.cs.uni-dortmund.de/audio.html

Categories(18)	Garageband genres (47)
Rock	Alternative pop, Pop, Pop rock, Power pop Alternative Rock, Indie Rock, Hard Rock Modern Rock, Rock
Progressive Rock	Instrumental Rock, Progressive Rock
Folk/Country	Acoustic, Folk, Folk Rock, Americana, Country
Punk	Emo, Pop Punk, Punk
Heavy Metal	Alternative Metal, Hardcore Metal, Metal
Funk	Funk, Groove Rock, R&B
Jazz	Jazz
Electronica	Ambient, Electronica, Electronic Experimental Electronica, Experimental Rock
Latin	Latin, World, World Fusion
Classical	Classical
Techno	Dance, Techno, Trance
Industrial	Industrial
Blues	Blues, Blues Rock
Reggae	Reggae
Ska	Ska
Comedy	Comedy
Rap	Hip-Hop, Rap
Spoken word	Spoken word

Table 3.13: Fusion of Garageband genres to an 18-terms taxonomy proposed by Meng (2008).

the smallest category (only 250 songs) and rock became the bigger one (more than 3000 songs). The rest of the categories gather about 1000 songs (See Table 3.13).

Magnatune: It is a record label founded in April 2003, created to find a way to run a record label in the Internet. It helps artists get exposure, make at least as much money they would make with traditional labels, and help them to get fans and concerts. Visitors can download individual audio files from the Internet and it has been used as a groundtruth in the MIREX05 competition (See Section 3.3.3). Details on this dataset are shown in Table 3.14.

Mirex05: Music Genre Classification was the most popular contest in the MIREX05. Two datasets were used to evaluate the submissions: Magnatune and USPOP, described above. In fact, two simpler versions of these databases were used, following the properties shown in Table 3.15.

STOMP (Short Test of Music Preferences): This dataset was proposed by Rentfrow & Gosling (2003) to study social aspects of music. It contains 14 musical genres according to musicological (a set of experts were asked) and commercial criteria (taxonomies in online music stores were consulted) and it is considered to be representative of the western music. A list of 10 songs for each of these genres is proposed, assuming that

	Subcategories(# albums)
Classical	Classical(178), After 1800(12)
Electronica	Electronica(106), Ambient(53)
Jazz & Blues	Jazz & Blues (18)
Metal & Punk Rock	Metal & Punk Rock(35)
New Age	New Age(110)
Pop/Rock	Pop/Rock(113)
World	World(85)
Other	Ambient(53)

Table 3.14: Summary of the Magnatune dataset

	Entries	Artists	Genres
USPOP	1515	77	6
Magnatune	1414	77	10

Table 3.15: Summary of the MIREX05 simplified dataset

	# Songs		# Songs
Alternative	10	Rap & Hip-Hop	10
Blues	10	Jazz	10
Classical	10	Pop	10
Country	10	Religious	10
Electronica/Dance	10	Rock	10
Folk	10	Soul/Funk	10
Heavy Metal	10	Soundtrack	10

Table 3.16: Summary of the STOMP dataset

they are clear prototypes for each one of the genres (See Table 3.16 for details).

Radio: This can be considered as our *in-house* dataset. It was created by collecting the most common music broadcasted by spanish radio stations in 2004. It was defined by musicologists and uses 8 different musical genres and 50 full songs per genre without artist redundancy. Each musical genre has associated a set of 5 full songs for test. One of the particularities of this database is that includes the *Speech* genre which includes different families of spoken signal such as advertisements, debates or sports transmission.

3.3.2 Review of interesting approaches

One of the earliest approaches in automatic audio classification was proposed by Wold et al. (1996). The author proposes the classification for different families of sounds such as animals, music instruments, speech and machines. This method extracts the loudness, pitch, brightness and bandwidth from the

	# Train songs	# Test songs
Classical	50	10
Dance	50	10
Hip-Hop	50	10
Jazz	50	10
Pop	50	10
Rhythm&blues	50	10
Rock	50	10
Speech	50	10

Table 3.17: Summary of the Radio dataset

original signals and compute the statistics such as mean, variance and auto correlation over the whole sound. The classification is made using a gaussian classifier. Although this system is not centered in musical genres, it can be considered the starting point for this research area.

Five years later, one of the most relevant studies in automatic genre classification is proposed by Tzanetakis & Cook (2002). In this paper, authors use timbre related features (Spectral Centroid, Spectral Rolloff, Spectral Flux, MFCC and Analysis and Texture Window), some derivatives of the timbric features, rhythm related features based on Beat Histogram calculation (Tzanetakis et al., 2001a) and pitch related features based on the multipitch detection algorithm described by Tolonen & Karjalainen (2000). For classification and evaluation, authors propose the use of simple gaussian classifiers. The dataset is defined by 20 musical genres with 100 excerpts of 30 seconds per genre. Many experiments and evaluations are discussed and the overall accuracy of the system reaches a 61% of correct classifications, using 10-fold cross validation, over the 20 musical genres.

In the recent years, the activity has been centered in the improvement of both descriptors and classification techniques. Table 3.18, Table 3.19, and Table 3.20 shows a non exhaustive list for the most relevant papers presented in journals and conferences for the last years. Although accuracies are not completely comparable due to the different datasets the authors use, similar approaches have similar results. This suggests that music genre classification, as it is known today, seems to reach a 'glass ceiling' (Aucouturier & Pachet, 2004) in the used techniques and algorithms.

	Author	Year	Accuracy	Gen.
A Statistical Approach to Musical Genre Classification using Non-Negative Matrix Factorization	Holzapfel, A	2007	86.7%	10
Temporal Feature Integration for Music Genre Classification	Meng, A	2007	40-44%	11
Optimal filtering of dynamics in short-time features for music organization	Arenas, J	2006	61%	11
Inter Genre similarity modeling for automatic music genre classification	Bagci, U	2006	64%	9
Meta-Features and AdaBoost for Music Classification	Bergstra, J	2006	85%	6
Probabilistic Combination of Features for Music Classification	Flexer , A	2006	60%	8
Evaluation of MFCC Estimation Techniques for Music Similarity	Jensen, JH	2006	80%	6
A Genre Classification Plug-in for Data Collection	Lehn-Schioler, T	2006	37.5 %	15
An investigation of feature models for music genre classification using the support vector classifier	Meng, A	2006	43%	11
Independent Component Analysis for Music Similarity Computation	Pohle, T	2006	68.5%	5
A Study on Music Genre Classification Based on Universal Acoustic Models	Reed, J	2006	69.3%	5
Co-occurrence models in music genre classification	Ahrendt, P	2005	–	11
Combination of homogeneous classifiers for musical genre classification	Koerich, A.L.	2005	>85%	2
Musical Genre Classification Enhanced by Improved Source Separation Techniques	Lampropoulos, AS	2005	75.8%	4
Music genre classification with taxonomy	Li, T	2005	75-83%	10

Table 3.18: Non exhaustive list for the most relevant papers presented in journals and conferences

	Author	Year	Accuracy	Gen.
Genre Classification via an LZ78-Based String Kernel	Li, M	2005	67%	10
Sound Re-synthesis from feature sets	Lidy, T	2005	–	–
Improving Music Genre Classification by short-time feature integration	Meng, A	2005	98-92%	5-6
Improvements of audio-based music similarity and genre classification	Pampalk, E	2005	64%	6..16
On the Modeling of Time Information for Automatic Genre Recognition Systems in Audio Signals	Scaringella, N	2005	69.9%	7
Fast Recognition of Musical Genres Using RBF Networks	Turnbull, D	2005	71.5%	10
Decision time horizon for music genre classification using shorttime features	Ahrendt, P	2004	80%	5
Towards characterization of music via rhythmic patterns	Dixon, S	2004	50-96%	8
Dance music classification: A tempo-based approach	Gouyon, F	2004	67%	8
Mixture of experts for audio classification: an application to male female classification and musical genre recognition	Harb, H	2004	60%	6
A comparison of human and automatic musical genre classification	Lippens, S	2004	<60%	6
Unsupervised Classification of Music Genre Using Hidden Markov Model	Shao, X	2004	89%	4
Musical style perception by a linear auto-associator model and human listeners	Tillman, B	2004	–	3

Table 3.19: Non exhaustive list for the most relevant papers presented in journals and conferences (Cont.)

	Author	Year	Accuracy	Gen.
Features and classifiers for the automatic classification of musical audio signals	West, C	2004	<70%	6
A hierarchical approach to Automatic Musical Genre classification	Burred, JJ	2003	60%	17
Factors in Automatic Musical Genre Classification of Audio Signals	Li, T	2003	71%	10
A Comparative Study on Content-Based Music Genre	Li, T	2003	75%	10
Semi-Automatic Approach for Music Classification	Zhang, T	2003	–	10
Finding Songs That Sound The Same	Aucouturier, JJ	2002	–	–
Music Classification using Neural Networks	Scott, P	2001	94.8%	4
Automatic musical genre classification of audio signals	Tzanetakis, G	2001	–	15
Genre Classification Systems of TV sound signals based on a spectrogram analysis	Han, P	1998	55%	4
Recognition of music types	Soltau, H	1998	79-86%	4

Table 3.20: Non exhaustive list for the most relevant papers presented in journals and conferences (Cont. 2)

If we plot the obtained accuracies shown in Tables 3.18, 3.19, and 3.20 with respect to the number of genres used in the approach, we obtain a plot as shown in Figure 3.3. The regression line we plot is that corresponding exponential function:

$$\%\ of\ accuracy = A \cdot r^{\#\ of\ genres} \tag{3.8}$$

where $A = 93.36$ and $r = 0.9618$. We also plot the corresponding theoretical baseline for an equally distributed dataset following the $1/n$ curve, where n is the number of genres in the collection. These values will be useful for the conclusions extracted in the forthcoming chapters.

Focussing on the works centered on the descriptors, authors study how to integrate all the extracted descriptors (which have no time-related information at all) into bigger chunks of descriptors with some time relative information. First, West & Cox (2005) presented a study comparing results between different approaches that include individual short frames (23 ms), longer frames (200 ms), short sliding textural windows (1 sec) of a stream of 23 ms frames, large fixed windows (10 sec) and whole files. The conclusions of this work showed how onset detection based segmentations of musical audio provide better features for classification than the fixed or sliding segmentations examined. These features produced from onset detection based segmentations are both simpler to model and produce more accurate models.

Scaringella & Zoia (2005) investigated means to model short-term time structures from context information in music segments to consolidate classification consistency by reducing ambiguities. The authors compared 5 different methods taking low-level, short-term time relationships into account to classify audio excerpts into musical genres. SVMs with delayed inputs prove to give the best results with a simple modeling of time structures providing accuracies near 70% over 7 musical genres.

According to Meng et al. (2007), mean and variance along the temporal dimension were often used for temporal feature integration, but they didn't capture neither the temporal dynamics nor dependencies among the individual feature dimensions. The authors proposed a multivariate autoregressive feature to solve this problem for music genre classification. This model gave two different feature sets, the diagonal autoregressive (DAR) and multivariate autoregressive (MAR) features which were compared to the baseline mean-variance as well as two other temporal feature integration techniques.

Li et al. (2003); Li & Tzanetakis (2003) proposed the use of the Daubechies Wavelet Coefficient Histogram (DWCH) to capture local and global information of music at the same time. Authors also showed an exhaustive comparative analysis between different descriptors and classifiers, as shown in Table 3.21.

Ahrendt et al. (2005) proposed the use of co-occurrence models which, instead of considering the whole song as an integrated part of the probabilistic model, considered it as a set of independent co-occurrences. All the proposed models had the benefit of modeling the class-conditional probability of the whole song instead of just modeling short time frames.

In our study, we will use basic statistics (mean, variance, skewness and kurtosis) to collapse the computed descriptors among time. It is our goal to compare the performance of different families of descriptors and their combinations, assuming that, in those cases where necessary, the descriptor itself

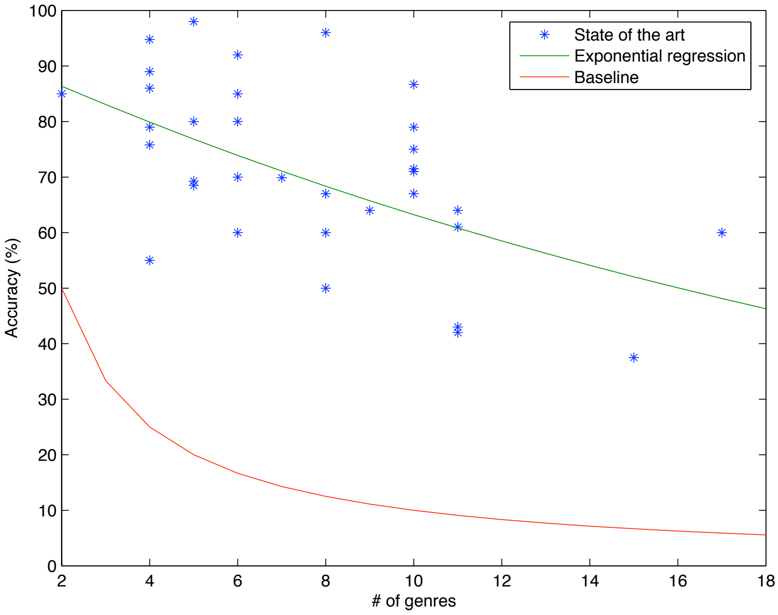


Figure 3.3: Accuracies of the state of the art with respect to the number of genres, and the exponential regression curve.

contains this information (rhythm transform, complexity, etc.). See Section 4.3 for details.

Other approaches proposed different techniques to improve the accuracy of the descriptors. Lampropoulos et al. (2005) proposed a sound source separation method to decompose the audio signal into a number of component signals, each of which corresponds to a different musical instrument source. The extracted features were used to classify a music clip, detecting its various musical instruments sources and classifying them into a musical dictionary of instrument sources or instrument teams. Accuracies about 75% in a 4 genres database were obtained.

Bagci & Erzin (2006) investigated inter-genre similarity modeling (IGS) to improve the performance of automatic music genre classification. Inter-genre similarity information was extracted over the miss-classified feature population.

Ellis (2007) proposed the use of beat-synchronous chroma features, designed to reflect melodic and harmonic content and be invariant to instrumentation which improved about 3% the accuracies of his experiments.

Hierarchical classifiers were also developed by many authors. They used the information of hierarchical taxonomies to classify genres in a higher category level. For instance, Zhang & Kuo (1999) proposed a Hierarchical classification of audio data for archiving and retrieving. The system divided the classification in three main steps: (1) a coarse level classification to discriminate between speech, music, environmental audio and silence, (2) fine level step to discriminate between different environmental sounds like *applause*, *birds*, *rain*, etc.,

	SVM1	SVM2	MPSVM	GMM	LDA	KNN
DWCH's	**74.9**	**78.5**	**68.3**	**63.5**	**71.3**	62.1
All the rest	70.8	71.9	66.2	61.4	69.4	61.3
Beat + FFT + MFCC	71.2	72.1	64.6	60.8	70.2	**62.3**
Beat + FFT + Pitch	65.1	67.2	56.0	53.3	61.1	51.8
Beat + MFCC + Pitch	64.3	63.7	57.8	50.4	61.7	54.0
FFT + MFCC + Pitch	70.9	72.2	64.9	59.6	69.9	61.0
Beat + FFT	61.7	62.6	50.8	48.3	56.0	48.8
Beat + MFCC	60.4	60.2	53.5	47.7	59.6	50.5
Beat + Pitch	42.7	41.1	35.6	34.0	36.9	35.7
FFT + MFCC	70.5	71.8	63.6	59.1	66.8	61.2
FFT + Pitch	64.0	68.2	55.1	53.7	60.0	53.8
MFCC + Pitch	60.6	64.4	53.3	48.2	59.4	54.7
Beat	26.5	21.5	22.1	22.1	24.9	22.8
FFT	61.2	61.8	50.6	47.9	56.5	52.6
MFCC	58.4	58.1	49.4	46.4	55.5	53.7
Pitch	36.6	33.6	29.9	25.8	30.7	33.3

Table 3.21: Classification accuracies proposed by Li et al. (2003) for different descriptors and classifiers. SVM1 refers to pairwise classification and SVM2 refers to one-versus-the-rest classification. "All the Rest" features refers to: Beat + FFT + MFCC + Pitch

and (3) a Query-By-Example audio retrieval system. The author reported an accuracy of 90% in a coarse-level classification when using a dataset of 1500 sounds. As shown in Section 3.2.1, Tzanetakis et al. (2001b); Burred & Lerch (2003) also obtained interesting results by using hierarchical taxonomies.

Pampalk et al. (2005b) presented an improvement to audio-based music similarity and genre classification based on the combined spectral similarity proposed by Aucouturier & Pachet (2004) with three additional similarity measures based on fluctuation patterns. He presented two new descriptors and a series of experiments evaluating the combinations. This approach increased about 14% other state of the art algorithms, but the author reported the presence of a glass ceiling in genre classification.

Finally, some improvements in classification algorithms were proposed by some authors. Bagci (2005) investigated discriminative boosting of classifiers to improve the automatic music genre classification performance. He used two different classifiers, the boosting of a GMM and another one that used inter-genre similarity information. In the first classifier (Boosting Gaussian Mixture Models) the author used the first multi-class extension of the AdaBoost algorithm and proposed a modification to the expectation-maximization algorithm (Redner & Walker, 1984) based training of mixture densities to adapt the weighting approach of the boosting algorithm. In the second classifier (Boosting with Inter-genre Similarity information) two novel approaches were proposed: The first approach addressed capturing the inter-genre similarities to decrease the level of confusion across similar music genres (the inter-genre similarity modeling was closely related with the boosting idea) and the second one presented an automatic clustering scheme to determine similar music genres in a hierarchical classifier architecture. Many tests were performed for each classifier and many results are shown for all of them obtaining accuracies up to 80% which are quite similar to results obtained by 3s or 20s human classification.

Soltau et al. (1998) proposed a Neural Network based music classifier with the inclusion of time-domain information. The authors used a new approach to temporal structure by the inclusion of the Explicit Time Modeling with Neural Network (ETM-NN) and compared results with a more classical approach using Hidden Markov Models (HMM). After the computation of the acoustical features of the input signal (based on MFCC), the results obtained with a left-to-right structure in HMM achieved a performance of 79.2% while the ETM-NN approaches achieved a 86.1% of performance.

Li et al. (2003) used Support Vector Machines and Linear Discriminant Analysis to improve previously published results using identical data collections and features. After some experiments, they showed the relative importance of feature subsets (FTT,MFCC, Pitch, Beat) in order of decreasing accuracy. Best results were obtained by Linear Discriminant Analysis (71% of accuracy) which is indirectly comparable to the 70% of accuracy in other human classification tests.

Flexer et al. (2005) presented a clear discussion on the benefits of using Hidden Markov Models (HMM) in front of Gaussian Mixture Models (GMM) for music genre classification. Authors based their study on the audio data available in the Magnatune dataset. After the introduction of HMMs and GMMs concepts, the authors created different models for the comparison of their likelihoods. Results and conclusions show how HMMs better describe spectral similarity for individual songs but HMMs perform at the same level as GMMs when used for spectral similarity in musical genres.

As mentioned in Section 2.3.1, music genre classification depends on some elements which are extrinsic to the actual music. Whitman & Smaragdis (2002) tried to solve this problem by combining musical and cultural features which are extracted from audio and text. The cultural features were found from so-called community metadata (Whitman & Lawrence, 2002) based on textual information from the Internet. The author concludes that this mix can be useful for some specific cases, i.e. the confusion between Rap and Rhythm'n'Blues classification using only audio data.

Finally, only one work have been found focused on automatic genre classification of non western music. In this study, Noris et al. (2005) investigated the factors affecting automated genre classification using eight categories: Dikir Barat, Etnik Sabah, Inang, Joget, Keroncong, Tumbuk Kalang, Wayang Kulit, and Zapin. A total of 417 tracks from various Audio Compact Discs were collected and used as the dataset. Results show how accuracies near 75% can be obtained using spectral features and J48 classifier.

3.3.3 Contests

In the following section, we will discuss about the contests, on different MIR tasks, organized in the context of the International Symposium of Music Information Retrieval (ISMIR) for the last few years. These contests became a *must* reference for the whole MIR community because the presented algorithms can be directly compared under the same testing conditions. We will focus on the audio genre classification task which has been carried out for three years

	Lab	Accuracy
Dan Ellis & Brian Whitman	Columbia University, MIT	51.48%
Elias Pampalk	OFAI	78.78%
George Tzanetakis	Univ. of Victoria	58.60%
Kris West	Univ. of East Anglia	67.22%
Thomas Lidy & Andreas Rauber	Vienna University of Tech.	55.70%

Table 3.22: Participants and obtained accuracies for the Audio Description Contest (ISMIR2004)

(2004, 2005 and 2007).

Audio Description Contest

The Audio Description Contest is considered the starting point for further competitions between algorithms in MIR community[13]. It was organized by the Music Technology Group at Universitat Pompeu Fabra, Barcelona, which hosted the International Conference of Music Information Retrieval (ISMIR) in 2004. It was the first time that comparisons between algorithms instead of results was carried out. It forced researchers to test and document their algorithms and to build a pre-defined i/o format. Six tasks were defined for this contest:

Genre Classification: label unknown songs according to one of 6 given musical genres

Artist Identification: identify one artist given three songs of his repertoire

Artist Similarity: mimic the behavior of experts in suggesting an artist similar to a given one

Rhythm Classification: label audio excerpts with one out of eight rhythm classes

Tempo Induction: main beat detection from polyphonic audio

Melody Extraction: main melody detection from polyphonic audio

Five authors participated to the genre classification contest. The used dataset was based on Magnatune dataset (See Section 3.3.1 for details) and the obtained results are shown in Table 3.22.

The big difference between the obtained accuracies can be explained by different reasons. First, the lack of previous initiatives to compare algorithms produced that the authors focused on their work, and the performance of their algorithms could decrease in this new scenario. Second, the Magnatune dataset is clearly an unbalanced dataset which may deviate the good performance of some algorithms. Although problems here described and the small number of participants, this contest fixed the starting point for further editions.

[13] ismir2004.ismir.net/ISMIR_Contest.html

MIREX 2005

As a consequence of the difficulties in the organization of the contests in 2004, the International Music Information Retrieval System Evaluation Laboratory (IMIRSEL)[14] was created. They organized the Music Information Retrieval Evaluation eXchange (MIREX) in the context of ISMIR 2005. As in the Audio Description Contest, many different tasks were proposed:

Audio Artist Identification: recognize the singer or the group that performed a polyphonic audio recording

Audio Drum Detection: detect the onsets of drum sounds in polyphonic pop song

Audio Genre Classification: label unknown songs according to one of 10 given musical genres

Audio Melody Extraction: extract the main melody, i.e. the singing voice in a pop song or the lead instrument in a jazz ballad

Audio Onset Detection: detect the onsets of any musical instrument

Audio Tempo Extraction: compute the perceptual tempo of polyphonic audio recordings

Audio and Symbolic Key Finding: extract the main key signature of a musical work

Symbolic Genre Classification: label unknown songs according to one of 38 given musical genres in MIDI format. See Section 3.1.2 for further details

Symbolic Melodic Similarity: retrieve the most similar documents from a collection of monophonic incipits.

Focusing on audio genre classification, all the tests were carried out using two independent databases (See Section 3.3.1 for details) and the participants had to make independent runs on the two collections. The overall result for each submitted algorithm was computed as the average of the two performances. Results are shown in Table 3.23.

Bergstra et al. (2005) propose an algorithm that compensates the large discrepancy in temporal scale between feature extraction (47 milliseconds) and song classification (3-5 minutes). They classify features at an intermediate scale (13.9 seconds). They also decompose the input song into contiguous, non-overlapping segments of 13.9s, and compute the mean and variance in standard timbre features over each segment. For classification, the authors use an extension of Adaboost called Adaboost.MH (first algorithm) and 2-level trees (second algorithm).

Mandel & Ellis (2005b) use support vector machines to classify songs based on features calculated over their entire lengths. They model songs as single Gaussians of MFCCs and use a KL divergence-based kernel to measure the distance between songs.

[14] www.music-ir.org/evaluation

	Accuracy
Bergstra, Casagrande & Eck (2)	82.34%
Bergstra, Casagrande & Eck (1)	81.77%
Mandel & Ellis	78.81%
West, K.	75.29%
Lidy & Rauber (SSD+RH)	75.27%
Pampalk, E.	75.14%
Lidy & Rauber (RP+SSD)	74.78%
Lidy & Rauber (RP+SSD+RH)	74.58%
Scaringella, N.	73.11%
Ahrendt, P.	71.55%
Burred, J.	62.63%
Soares, V.	60.98%
Tzanetakis, G.	60.72%

Table 3.23: Participants and obtained accuracies for the Audio Genre Classification task in MIREX2005

West (2005) uses a novel feature called Mel-band Frequency Domain Spectral Irregularity which is computed from the output of a Mel-frequency scale filter bank and is composed of two sets of coefficients, half describing the spectrum and half describing the irregularity of the spectrum. He also uses rhythmic descriptors based on onset detection. The classifier is based on a modified version of classification and regression trees, replacing the normal single variable with single Gaussian distributions and Mahalanobis distance measurements. Finally, he applies Linear Discriminant Analysis to weight the extracted features.

Lidy & Rauber (2005) submitted a system that uses combinations of three feature sets (Rhythm Patterns, Statistical Spectrum Descriptor and Rhythm Histogram). All feature sets are based on fluctuation of modulation amplitudes in psychoacoustically transformed spectrum data. For classification, the authors applies Support Vector Machines.

Pampalk (2005) uses cluster models of MFCC spectra, fluctuation patterns and two descriptors derived from them: Gravity and Focus. For each piece in the test set, the distance to all pieces in the training set is computed using a nearest neighbor classifier. There is no training other than storing the features of the training data. Each piece in the test set is assigned the genre label of the piece closest to it.

Scaringella & Mlynek (2005) parameterize audio content by extracting 3 sets of features describing 3 different dimensions of music: timbre, energy and rhythm. Once features extracted, an ensemble of Support Vector Machines (SVMs) is used for classification into musical genres. The underlying idea is to use separate models to approximate different parts of the problem and to combine the outputs from the experts with probabilistic methods.

Ahrendt & Meng (2005) model an audio segment (1.2s) of short time features (statistical moments of MFCC) for creating a multivariate autoregressive model. This data feeds a generalized linear model with softmax activation function is trained on all the MAR-feature vectors from all the songs. To reach a final decision for a 30s music clip, the sum-rule is used over all the frames.

Burred (2005) computes the mean, the variance and the derivatives of many timbre a rhythm related descriptors. In order to maximize genre separability, author uses a sequential forward feature selection algorithm based on an objective measure of class separability. The system is designed to work in the two environments of the contest: (1) Hierarchical: The feature selection is repeated for each subset of classes of the taxonomy tree, so that only the features that are most suitable for separating that particular subset are retained. When classifying an unknown input signal, the appropriate features are selected and computed at each level of the hierarchy. (2) Parametric classification: The classes are modeled as 3-cluster Gaussian Mixture Models. The classification is performed on a Maximum Likelihood basis.

Soares (2005) computes a wide group of more than 400 signal processing features including some transformations to the original features, such as derivatives and correlation between features. Then, the "wrapper" methodology is followed to select the most important ones. For classification, the author proposes a Multivariate Time Series datasets method (also called taxoDynamic), which dynamically adjusts the subset of features at each node of the taxonomy.

Finally, Tzanetakis & Murdoch (2005) use the $3s$ mean and variances of 18 timbre features, and Support Vector Machine for classification. The outputs of the classifier are mapped to probabilities using logistic regression and the classification decision over the entire song is done by taking weighted (by the classifier outputs) sums for each class and selecting the one with the highest sum.

MIREX 2007

Following with the MIREX, in 2007 the people of the IMIRSEL and the organizers of the 8th International Conference on Music Information Retrieval (ISMIR 2007) proposed a set of tasks for a new competition:

- Audio Artist Identification
- Audio Classical Composer Identification
- Audio Cover Song Identification
- Audio Genre Classification
- Audio Music Mood Classification
- Audio Music Similarity and Retrieval
- Audio Onset Detection
- Multiple Fundamental Frequency Estimation and Tracking
- Query-by-Singing/Humming
- Symbolic Melodic Similarity

For the Audio Genre Classification task, five authors submitted their algorithms. The proposed dataset was built using 7000 clips (length=$30s$) covering

Main Category	Genre
JazzBlues	Jazz
	Blues
CountryWestern	Country
GeneralClassical	Baroque
	Classical
	Romantic
Electronica	Electronica
Hip-Hop	Hip-Hop
GeneralRock	Rock
	HardRockMetal

Table 3.24: Hierarchical taxonomy for the Audio Genre Classification task in MIREX 2007

the following musical genres (700 tracks per genre): Baroque, Blues, Classical, Country, Dance, Jazz, Metal, Rap/Hip-Hop, Rock'n'Roll and Romantic.

A hierarchical taxonomy was also provided to participants in order to compare results between raw and hierarchical classification (See Table 3.24). The evaluation procedure is detailed in the Wiki of Audio Genre Classification task[15].

Lidy et al. (2007) proposed the use of both audio and symbolic descriptors for genre classification. For that, the authors proposed a new transcription system to get a symbolic representation from audio signals. The audio features used in this work were the rhythm patterns, rhythm histograms, statistical spectrum descriptors and onset features. The set of extracted symbolic descriptors was based on the work proposed by Ponce & Inesta (2007); Rizo et al. (2006b), and included the number of notes, number of significant silences, the number of non-significant silences, note pitches, durations and Inter Onset Intervals, among others. Support Vector Machines were used for classification.

Mandel & Ellis (2007) proposed an algorithm that was used for several tasks in the contest (audio similarity, composer and artist identification and genre and mood classification). This work computes spectral and time domain features based on the work of Mandel & Ellis (2005a) and Rauber et al. (2002) respectively. The classification is performed by a DAG-SVM which allows n-way classification based on support vector machines using a directed acyclic graph (DAG). See Platt et al. (2000) for details.

Tzanetakis (2007) proposed an algorithm that covered the audio artist identification, audio classical composer identification, audio genre classification, audio music mood classification, and audio music similarity and retrieval results tasks using Marsyas[16]. The features used were spectral centroid, rolloff, flux and MFCC. To capture these features, the author computed a running mean and standard deviation over the past M frames. All this data was collapsed to a single vector that represents the overall behavior of each audio clip. The author also used support vector machines for classification.

[15] www.music-ir.org/mirex2007/index.php/Audio_Genre_Classification

[16] marsyas.sness.net/

Participant	Hierarchical	Raw
IMIRSEL(1)	76.56%	68.29%
Lidy	75.57%	66.71%
Mandel(1)	75.03%	66.60%
Tzanetakis	74.15%	65.34%
Mandel(2)	73.57%	65.50%
Guaus	71.87%	62.89%
IMIRSEL(2)	64.83%	54.87%

Table 3.25: Obtained accuracies for the MIREX 2007 Audio Genre Classification tasks for all submissions

Participant	Runtime (sec)	Folds
IMIRSEL(1)	6879	51
Lidy	54192	147
Mandel(1)	8166	207
Tzanetakis	—	1442
Mandel(2)	8018	210
Guaus	22740	194
IMIRSEL(2)	6879	1245

Table 3.26: Time for feature extraction and # folds for train/classify for the MIREX 2007 Audio Genre Classification tasks for all submissions

Finally, the IMIRSEL lab (organizers of the contest) also participated using a simple feature extraction and classification techniques but, undortunately, they did not document their approaches. Our own approach will be described in detail in Section 5.4. We used spectral and rhythmic features to train a SVM algorithm.

The summary of the results for all the participants is shown in Table 3.25 and the runtimes for the different submissions are shown in Table 3.26. A more detailed discussion about these results is given in Section 5.4.5.

3.4 Conclusions from the state of the art

As described above, music genre classification can be considered as one of the traditional challenges in the music information retrieval field. Roughly speaking, the literature can be divided into two main groups: the first one dealing with works focused on audio descriptors and their compact representation, and the second one focused on machine learning algorithms that improve the performance of the genre classification systems. Fortunately, the MIREX has became a reference point for the authors providing a benchmark to compare algorithms and descriptors with exactly the same testing conditions.

From the point of view of many authors (Aucouturier & Pachet, 2004; Pampalk et al., 2005b) whose ideas are also supported by us, all these efforts are near to reach a glass-ceiling on accuracies. Algorithms become more and more complicated but, from our knowledge, only few works compare these results with real listening experiments on genre classification (Soltau et al., 1998; Huron, 2000; Futrelle & Downie, 2003). We think that *Less is more*, and we have played with simple concepts and techniques, thinking with music and how humans perceive it. Working in that direction, we propose the listening of the errors provided by the automatic classifiers and use this information to design a better classifier. Moreover, we wonder whether algorithms should imitate classical, prototype or exemplar categorization theory. Although some authors did, (Fujinaga et al., 1998; Jäkel et al., 2008), these kind of works are traditionally far from the MIR channels. From here to the end, we will study de behavior of descriptors and classifiers and, in the last chapter, we will establish the relationships between all the elements. For that, we first introduce the scientific background required for our analysis, we introduce a set of listening experiments, we compare the performance of different families of descriptors, and finally, deduce the architecture of our proposed classifier.

CHAPTER 4

Computational techniques for music genre characterization and classification

4.1 Introduction

In this chapter, we introduce the scientific concepts that will be used in our experiments. All the techniques here exposed are not new, so, they are not part of the contribution of this thesis. In some cases, these algorithms have never been applied to Music Information Retrieval. In this chapter, we explain the techniques themselves and we study their application to music genre classification in the forthcoming chapters.

4.2 Basic Statistics

In this section, we define some statistical concepts that will be used in almost all the experiments. We start with an overview of the statistical models that will be used for generating descriptors and for computing the obtained accuracies provided by the classifiers. Then, we define the periodogram that will be used in the computation of rhythmic descriptors.

4.2.1 Statistic moments

Mean

The *mean* or *expected value* of a discrete random variable X, can be computed as Montgomery & Runger (2002):

$$\mu = E(X) = \sum_x x f(f) \tag{4.1}$$

where x are the obtained values of the experiment and $f(x)$ is the weight of each for these values. Typically, $f(x) = \frac{1}{N}$ where N is the number of occurrences in the experiment.

Variance

The *variance* of X is a measure of the dispersion of the samples around the mean value, and it can be computed as (Montgomery & Runger (2002)):

$$\sigma^2 = V(X) = E(X-\mu)^2 = \sum_x (x-\mu)^2 f(x) = \sum_x x^2 f(x) - \mu^2 \tag{4.2}$$

or, by using Matlab nomenclature [1]:

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{4.3}$$

Finally, the *Standard Deviation* of X can be computed as:

$$\sigma = [V(X)]^{\frac{1}{2}} = \big(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2\big)^{\frac{1}{2}} \tag{4.4}$$

Skewness

The skewness of a distribution is defined as:

$$y = \frac{E(x-\mu)^3}{\sigma^3} \tag{4.5}$$

where μ is the mean of x, σ is the standard deviation of x and $E(x)$ is the expected value of x. Skewness is a measure of the asymmetry of the data around the sample mean. If skewness is negative, the data are spread out more to the left of the mean than to the right. If skewness is positive, the data are spread out more to the right. The skewness of the normal distribution (or any perfectly symmetric distribution) is zero.

Kurtosis

The kurtosis of a distribution is defined as:

$$k = \frac{E(x-\mu)^4}{\sigma^4} \tag{4.6}$$

where μ is the mean of x, σ is the standard deviation of x and $E(x)$ is the expected value of x. Kurtosis is a measure of how outlier-prone a distribution is. The kurtosis of the normal distribution is 3. Distributions that are more flat-shaped than the normal distribution have kurtosis greater than 3; distributions that are peak-shaped have kurtosis smaller than 3.

[1] Matlab calculates the variance with $\frac{1}{n-1}$ instead of $\frac{1}{n}$ as an approximation for small-size samples ($N < 30$)

4.2.2 Periodogram

In this section, we will present an overview of the periodogram that will be used in following chapters. The Periodogram was introduced by Schuster in 1898 to study periodicity of sunspots.

The sample mean and variance statistics are unbiased and asymptotically unbiased estimators respectively, and they are both consistent estimators (Oppenheim & Schaffer, 1989). Sometimes, in digital signal processing, the estimation of the power density spectrum $P_{ss}(\Omega)$ of a continuous stationary random signal $s_c(t)$ is needed. After the anti-aliasing filtering, another discrete-time stationary random signal $x[n]$ will be created, and its power density spectrum $P_{xx}(\omega)$ will be proportional to $P_{ss}(\Omega)$ over the whole new bandwidth of $x[n]$:

$$P_{xx}(\omega) = \frac{1}{T} P_{ss}\left(\frac{\Omega}{T}\right) \qquad |\omega| < \pi \tag{4.7}$$

where T is the sampling period. Then, a good estimation of $P_{xx}(\omega)$ will provide a reasonable estimation of $P_{ss}(\Omega)$.

Let $v[n]$ be the windowed input signal:

$$v[n] = x[n] \cdot w[n] \tag{4.8}$$

where $w[n]$ is the windowing function. Then, the Fourier Transform of $v[n]$ can be computed as:

$$V(e^{j\omega}) = \sum_{n=0}^{L-1} w[n]x[n]e^{-j\omega n} \tag{4.9}$$

where L is the length (in samples) of the windowing function. Now, let $I(\omega)$ be the estimation of the power density spectrum:

$$I(\omega) = \frac{1}{LU} |V(e^{j\omega})|^2 \tag{4.10}$$

where U is the normalization factor for removing the bias in the spectral estimate. Depending on the windowing function, this estimator can be:

- If $w[n]$ is the rectangular window → $I(\omega)$ is the *periodogram*
- If $w[n]$ is NOT the rectangular window → $I(\omega)$ is the *modified periodogram*

Furthermore, note that the periodogram can also be computed as:

$$I(\omega) = \frac{1}{LU} \sum_{m=-(L-1)}^{L-1} c_{vv}[m]e^{-j\omega m} \tag{4.11}$$

where

$$c_{vv}[m] = \sum_{n=0}^{L-1} x[n]w[n]x[n+m]w[n+m] \tag{4.12}$$

As $c_{vv}[m]$ is the aperiodic correlation sequence for the finite-length sequence $v[n]$, the periodogram can be interpreted as the Fourier Transform of the aperiodic correlation of the windowed input data.

Finally, as we are in discrete domain, the periodogram can only be obtained at discrete frequencies. Then, discrete periodogram is computed as:

$$I(\omega_k) = \frac{1}{LU}|V[k]|^2 \tag{4.13}$$

where $V[k]$ is the N-point DFT of $w[n]x[n]$.

4.3 Descriptors

Descriptors are, from our point of view, the key point in audio classification. They are the responsible to extract the required information from raw data and, as a consequence of that, to decide which aspects of music will participate in the classification. Descriptors can be classified according to different criteria (Pohle, 2005). First, they cover different facets of music such as rhythm, timbre, melody, etc. In the following sections we will introduce them according to this criteria. Moreover, descriptors can describe low level features of music (p.e. energy, spectra, etc.) or high level (p.e. mood, genre, etc.), which are more related to the perception of the music by humans. Low level features are usually easier to compute, but they provide less musical information than high level descriptors.

Focusing on the goal of this thesis, the major challenge in music genre classification is to choose an appropriate bag of descriptors and the corresponding machine learning algorithm that provides high accuracies in a given dataset. For the author, music genre should reach a status of high level descriptor instead of a specific application of machine learning algorithms. We will include as many facets of music as possible in our classifier in order to work for this goal.

4.3.1 Previous considerations

Length of the audio excerpt

There are some properties which are shared for the computation of most of the descriptors. First, we have to decide which part of the audio data we will use. Some datasets provide access to the full song, but many authors prefer to compute their descriptors on a short audio excerpt of the whole song. Different configurations can be found in the literature: from 5 seconds at the middle of the song to 2 minutes (See Pampalk et al. (2003) for further comparisons). In the context of the Audio Genre Classification contest proposed for the MIREX 2007 [2], the results of the survey asking to the participants whether to use audio excerpts or not, results were clear: all of them preferred to use audio excerpts (See Table 4.1 and Table 4.2 for details).

The format of the input file is also a non standardized property. Omitting the compressed formats which are usually not used for feature computation (compressed data is converted to WAV or RAW data format), the sampling

[2] http://www.music-ir.org/mirex2007

	Yes	No	Votes
Use clips?	100%	0%	11

Table 4.1: Use of clips for the MIREX 2007 (data collected in August 2007)

	30s	60s	90s	120s	Votes
Length	83%	17%	0%	0%	8

Table 4.2: Preferred length of clips for the MIREX 2007 (data collected in August 2007)

	22KHz		44KHz	
	Mono	Stereo	Mono	Stereo
WAV	59%	0%	29%	6%
MP3	6%	0%	0%	0%

Table 4.3: Preferred formats for the MIREX 2007 participants (data collected in August 2007, votes=17)

rate and the number of channels can affect some descriptors (p.ex. MFCC, Panning, etc.). Here again, the preferred formats for the participants in the Genre Classification contest at the MIREX 2007 are shown in Table 4.3.

As will be discussed in Section 5.3, our listening experiments are based on $5sec$ audio excerpts while the computation of descriptors is carried out, in most of the cases, using $30sec$ audio excerpts.

Frame size and hop-size

Another parameter that needs to be chosen a priori is the length of the frame and its hop-size. Typical frames are selected from 512 to 4096 frames (that is $23.22ms$ and $186ms$ respectively at sampling rate $sr = 22050Hz$), and the hop-size is typically fixed at 50%. These parameters affect the time and frequency resolution of the analysis and there is no a universal rule to select them. Rhythmic descriptors need longer frames than timbre ($> 2s$) which can be obtained by summing results over short frames or computing them directly in a longer frame.

Velocity and Acceleration

Finally, sometimes we also compute the derivative and the second derivative of the original descriptors. The aim of computing them, also referred as the *velocity* or the *acceleration* of a given descriptor, is to capture how this descriptor evolves in time.

The first order differentiation can be computed in different ways, but many authors use the Causal FIR filter implementation:

$$\dot{p}[n] = \frac{\partial p[n]}{\partial t} = \sum_{m=n}^{m=N} p[n-m] \tag{4.14}$$

where N is the depth of the differentiation. The descriptor we get by applying differentiation is denoted as *delta* descriptor (∇). Finally, if we have to apply the second-order differentiation, we will compute the Eq. 4.14 recursively. Then, we have the *delta-delta* descriptor (∇^2).

4.3.2 Time domain descriptors

Energy

From a mathematical point of view, the time-domain energy of the input signal can be defined as:

$$E = \sum_{n=0}^{N} x[n]^2 \tag{4.15}$$

where $x[n]$ is the input time-domain data and N is the length of $x[n]$ (in samples).

The energy is not a representative descriptor at all. It depends on many not fixed parameters of the experiment such as the mic/line-in amplifier level while recording, the used codification, and so on. But in this thesis, we will use the derivative of the time domain energy here defined to compute the rhythmic descriptors.

Zero Crossing Rate

As defined by Kedem (1986) and Saunders (1996), the Zero Crossing Rate (ZCR) of the time domain waveform provides a measure of the weighted average of the spectral energy distribution. This measure is similar to the spectral center of mass or Spectral Centroid of the input signal (see Section 4.3.3). From a mathematical point of view, it can be computed as:

$$ZCR = \frac{1}{2} \sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])| \tag{4.16}$$

where *sign* function is 1 for positive arguments and 0 for negative arguments, $x[n]$ is the input time-domain data and N is the length of $x[n]$ (in samples).

4Hz Modulation

The 4Hz Modulation Energy Peak is a characteristic feature of speech signals due to a near 4Hz syllabic rate. It is computed by decomposing the original waveform into 20 (Karneback, 2001) or 40 (Scheirer, 1998), depending on the accuracy, mel-frequency bands. The energy of each band is extracted and a second band pass filter centered at 4 Hz is applied to each one of the bands. Of course, this 4Hz value depends on the language: catalan or spanish languages, this value is near the $6[Hz]$ instead of the $4[Hz]$ value for English.

4.3.3 Timbre descriptors

Spectral Centroid

The Spectral Centroid is defined as the *balancing point* of the spectral power distribution (Scheirer, 1998). The Spectral Centroid value rises up, specially for percussive sounds, due to the high density of harmonics in the upper bands of the spectrum. This concept has been introduced by psychoacoustic and music cognition research. It can be interpreted as a measure of the average frequency, weighted by amplitude, of a spectrum, that is, a measure related with the brightness of the signal. Be careful of confusing the Spectral Centroid and the Fundamental Frequency: while the Spectral Centroid can be higher for a trumpet sound than for a flute sound, both instruments can play exactly the same note. From a mathematical point of view, the Spectral Centroid can be calculated as:

$$SC = \frac{\sum f_i a_i}{\sum a_i} \tag{4.17}$$

where f_i is the frequency value of each bin of the FFT and a_i is its amplitude. In many applications, it is averaged over time. This $\bar{SC}$ value can be averaged into different time-domain frames as shown in next equation:

$$\bar{SC} = \frac{1}{N} \sum SC_i \tag{4.18}$$

where N is the number of frames and SC_i is the Spectral Centroid value for each frame. Finally, the spectral centroid is sometimes normalized with the fundamental frequency, making this value adimensional:

$$SC = \frac{\sum f_i a_i}{f_1 \sum a_i} \tag{4.19}$$

Spectral Flatness

The Spectral Flatness is defined as the ratio of the geometric mean to the arithmetic mean of the power spectral density components in each critical band for the input signal. According to Izmirli (1999), it can be computed according to the following steps: the signal should be sampled at $f_s = 22050[Hz]$ and the 2048-points FFT should be performed after the Hanning windowing. The windows should be 30% overlapped. A Pre-emphasis filter should be applied in order to compensate the behavior of the human ear. The bark-band filter output should be calculated from the FFT and the power spectral density should be computed for each critical bands. All these values are used to compute the arithmetical and geometrical means.

$$SFM = \frac{G_m}{A_m} \tag{4.20}$$

where G_m and A_m are the arithmetical and geometrical means of the spectral power density function respectively. Sometimes, the Spectral Flatness Measure is converted to decibels as follows (Johnston, 1998):

$$SFM_{dB} = 10log_{10}\frac{G_m}{A_m} \tag{4.21}$$

and, furthermore, it can be used to generate a coefficient of tonality α as follows:

$$\alpha = min\left(\frac{SFM_{dB}}{SFM_{dB_{max}}}, 1\right) \tag{4.22}$$

For instance, a SFM of $SFM_{dB_{max}} = -60dB$ is used to estimate that the signal is entirely tone-like, and an SFM of $0dB$ to indicate a signal that is completely noise-like.

Spectral Flux

The Spectral Flux is also known as Delta Spectrum Magnitude. It is defined as (Tzanetakis & Cook, 2002):

$$F_t = \sum_{n=1}^{N} \left(N_t[n] - N_{t-1}[n]\right)^2 \tag{4.23}$$

where N_t is the (frame-by-frame) normalized frequency distribution at time t. It is a measure for the rate of local spectral change: if there is much spectral change between the frames $t-1$ and t then this measure produces high values.

Spectral Roll-Off

There are many different definitions for the Roll-off frequency, but more or less all of them express the same concept. It can be defined as (Tzanetakis & Cook, 2002):

$$SR_t = max\left\{ f \middle| \sum_{n=1}^{f} M_t[n] < TH \cdot \sum_{n=1}^{N} M_t[n] \right\} \tag{4.24}$$

where M_t is the magnitude of the Fourier transform at frame t and frequency bin n. Typical values of the threshold TH are between 0.8 and 0.95.

Mel Frequency Cepstrum Coefficient

The Cepstrum of an input signal is defined as the Inverse Fourier Transform of the logarithm of the spectrum of the signal (Picone, 1993):

$$c[n] = \frac{1}{N} \sum_{k=0}^{N-1} \log_{10} |X[k]|^{j\frac{2\pi}{N}kn}, \quad 0 < n < N-1 \tag{4.25}$$

where $X[k]$ is the spectrum of the input signal $x[n]$ and N is the length of $x[n]$ (in samples). The process is an Homomorphic Deconvolution because it is able to separate the excitation part of the input signal for further manipulations. For the Mel-Cepstrum computation, some few modifications have to be done. The *mel* scale tries to map the perceived frequency of a tone onto a linear scale:

$$mel\ frequency = 2595 \cdot \log_{10}\left[1 + \frac{f}{700}\right] \tag{4.26}$$

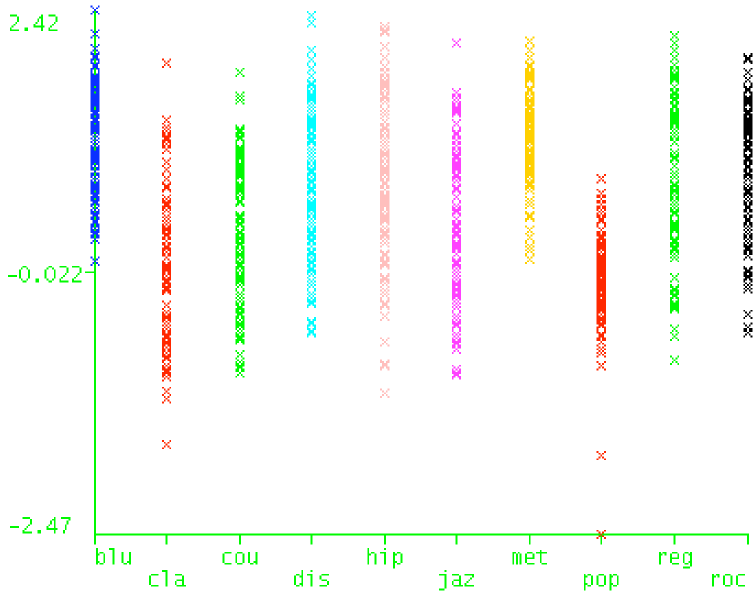


Figure 4.1: Behavior of the MFCC5 coefficient for different musical genres. Each point corresponds to the mean for all the short-time MFCC5 descriptor computed over 30 seconds audio excerpts in the Tzanetakis dataset (Tzanetakis & Cook, 2002). The musical genres are represented in the x axis

The Mel scale can be used as a rough approximation to estimate the bandwidths of human auditory filters expressed as Barks (Zwicker & Terhardt, 1980):

$$bark = 13 \cdot \arctan\left(\frac{0.76 \cdot f}{1000}\right) + 3.5 \cdot \arctan\left(\frac{f^2}{7500^2}\right) \tag{4.27}$$

Figure 4.1 shows an example of the behavior of a MFCC coefficient for different musical genres. For this case, we only show the MFCC5 coefficient. Each point represents the mean for all the short-time MFCC5 coefficients computed over the 30 seconds audio excerpts included in the Tzanetakis dataset (Tzanetakis & Cook, 2002). In this case, some differences between metal a pop music can be found. On the other hand, Figure 4.2 shows the plot for MFCC6 vs MFCC10 coefficients for the two musical genres mentioned above: Metal and Pop. In summary, we can combine the information from different coefficients to discriminate between different musical genres.

In some cases, the descriptor we use has not the whole information we need. Then, the first-order or the second-order differentiation of the original parameter is used (see Section 4.3.1).

4.3.4 Rhythm related descriptors

Inter Onset Interval

According to Allen & Dannenberg (1990), the Inter Onset Interval (IOI) is the time difference between two successive onsets but, according to Dixon (2001) it

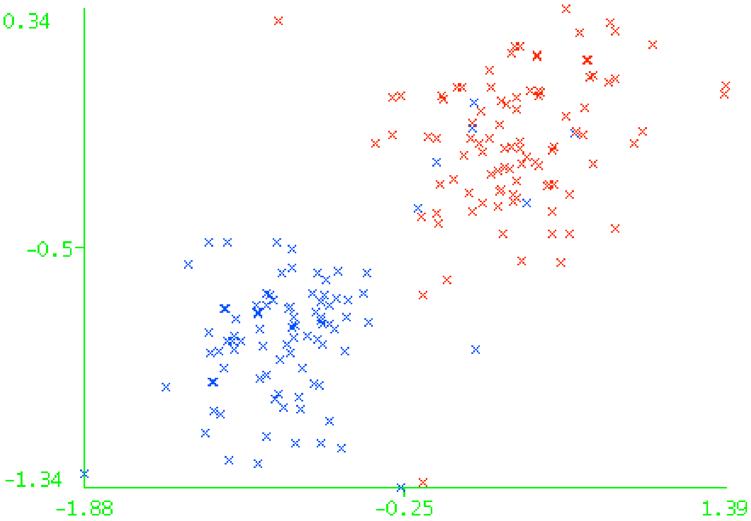


Figure 4.2: MFCC6 (x axis) vs MFCC10 (y axis) values for Metal (in blue) and Pop (in red). Each point represents the mean for all the short-time MFCC coefficients computed over the 30 seconds audio excerpts included in the Tzanetakis dataset (Tzanetakis & Cook, 2002).

can also be defined as the difference between any two onsets. Many algorithms can be found for IOI computations, but only one of them will be explained here. According to Gouyon (2003), the IOI histogram can be computed as:

1. Onset detection: First of all, the energy of each non-overlapping frames is calculated. The onset will be detected when the energy of the current frame is superior to a specific percentage (i.e. 200%) of a fixed number (i.e. 8) of the previous frames energy average. It is assumed that there is a gap of 60$[ms]$ between onsets, and a weighting factor is applied to each onset according to the number of consecutive onsets whose energy satisfies the threshold condition mentioned above.

2. IOI computations: In this algorithm proposed by Gouyon (2003), the time differences between any two onsets is taken. Each IOI has an associated weight according to the smallest weight among the two onsets used for this IOI computation

3. IOI histogram computation: With all these computed IOI, a histogram is created. This histogram is smoothed by the convolution of a Gaussian function. The parameters of this Gaussian function are fixed empirically.

At this point, the histogram of IOI is available. This data can be used for tick induction computations, rhythm classification, automatic BPM detection, and so on.

Beat Histogram

This concept was proposed by Tzanetakis et al. (2001a) as part of his automatic genre classification system. Some additional techniques such as Wavelet Transform are also used(Vidakovic & Müller, 1991; D. & L., 1992). The Wavelet decomposition of a signal can be interpreted as successive high-pass and low-pass filtering of the time domain signal. This decomposition is defined by:

$$y_{high}[k] = \sum_n x[n]g[2k-n] \tag{4.28}$$

$$y_{low}[k] = \sum_n x[n]h[2k-n]$$

where $y_{high}[k]$ and $y_{low}[k]$ are the output of high-pass and low-pass filters respectively, and $g[n]$ and $h[n]$ are the filter coefficients for the high-pass and low-pass filters associated to the scalar and wavelet functions for 4th. order Daubechies Wavelets. The main advantage of using the Wavelet Transform deals with the similarity of the decomposed signal to a $1/3$ octave filter bank in a similar way than the human ear does. Once the signal is decomposed, some additional signal processing (in parallel for each band) is needed:

1. Full Wave Rectification (FRW):

   $$z[n] = abs(y[n]) \tag{4.29}$$

   where $y[n]$ is the output of the Wavelet decomposition at that specific scale (or octave)

2. Low-pass filtering (LPF): One pole filter with $\alpha = 0.99$:

   $$a[n] = (1-\alpha)z[n] - \alpha \cdot a[n] \tag{4.30}$$

3. Downsampling(↓) by k=16:

   $$b[n] = a[kn] \tag{4.31}$$

4. Normalization (Noise removal NR):

   $$c[n] = b[n] - E\left[b[n]\right] \tag{4.32}$$

5. Autocorrelation (AR):

   $$d[n] = \frac{1}{N}\sum_n c[n]c[n+k] \tag{4.33}$$

   This autocorrelation is computed by using the FFT for efficiency.

Figure 4.3 shows a screenshot of the block diagram for beat histogram computation proposed by Tzanetakis et al. (2001a).

At this point, the first five peaks of the autocorrelation function are detected and their corresponding periodicities in beats per minute(BPM) are calculated and added to the beat histogram. Finally, when the beat histogram is computed, some features can be used:

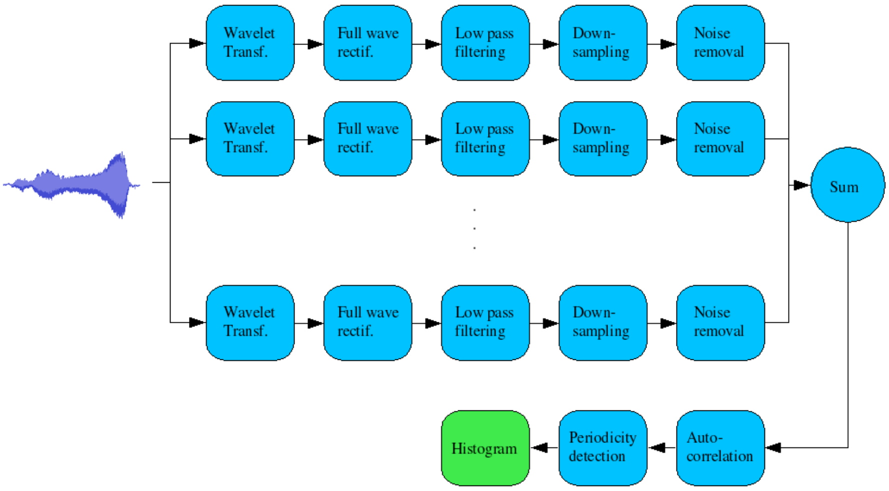


Figure 4.3: Screenshot of the block diagram for Beat Histogram computation proposed by Tzanetakis et al. (2001a)

1. Period0: Periodicity in BPM of the first peak
2. Amplitude0: Relative amplitude of the first peak
3. Ratio Period1: Ratio of the periodicity of the second peak to the first one
4. Amplitude1: Relative amplitude of the second peak
5. Ratio Period2, Amplitude2...

Other authors use the number of peaks, their distribution, *max* and *min* operations over the peaks, etc. Grimalidi (2003) uses the beat histogram as an input feature to his classification system.

Beat Spectrum

Beat Spectrum was introduced by Foote & Uchihashi (2001). It is a measure of the acoustic self-similarity as a function of time lag. The goal of this method is that it doesn't depend on fixed thresholds. Hence, it can be applied to any kind of music and, furthermore, it can distinguish between different rhythms at the same tempo. The *Beat Spectrogram* is also introduced in this work as the time evolution of the rhythm representation and it can be computed according the following steps:

1. Audio parameterization: The FFT of the windowed input data is computed. Then, by using any known filtering technique (i.e MFCC), the vector of the log energy for each band is obtained.
2. Frame similarity computation: Data derived from previous parameterization is embedded in a 2D representation. A dissimilarity measure between

two vectors i and j is computed as:

$$D_C(i,j) = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|} \tag{4.34}$$

3. Distance Matrix Embedding: The similarity matrix S contains all the measures for all the i and j as shown in Eq. 4.34. In this matrix, audio similarities can easily be observed.

4. The Beat spectrum: Periodicities and rhythmic structure can be derived from this similarity matrix. An estimation of the Beat Spectrum can be found by summing S along the diagonal as follows:

   $$B(l) = \sum_{k \subset R} S(k, k+l) \tag{4.35}$$

   where $B(0)$ is the sum for all the elements of the main diagonal over some continuous range R, $B(1)$ is the sum of all the elements along the first super-diagonal, and so on. A more robust estimation of the Beat Spectrum can be computed as:

   $$B(k,l) = \sum_{i,j} S(i+k, j+l) \tag{4.36}$$

   where the autocorrelation of S is computed. Some applications like onset detections can be computed by using this Beat Histogram.

In Section 5.2.1, we will present our own approach to rhythm description and their advantages and inconveniences with respect to the techniques here presented.

4.3.5 Tonal descriptors

The tonal features used in this thesis are based on the work made by Gomez (2006). They are inspired on the Harmonic Pitch Class Profile (HPCP) proposed by Krumhansl (1990). The HPCP measures the intensity of each of the twelve semitones in a diatonic scale. The basic idea is to map each frequency bin of the spectrum of the input signal to a given pitch class. For doing that, it uses a weighting function for feature computation, it considers the presence of harmonics from a fundamental frequency, and it uses high resolution grids in the HPCP bins (less than a semitone). The bandwidth that is considered is in between $100Hz$ and $5KHz$. Then, the HPCP vector can be computed as:

$$HPCP(n) = \sum_{i=1}^{nPeaks} \omega(n, f_i) \cdot a_i^2 \qquad n = 1 \dots N \tag{4.37}$$

where a_i and f_i are the linear magnitude and frequency values of the peak number i, $nPeaks$ is the number of spectral peaks that we consider, n is the HPCP bin, N is the size of the HPCP vector (i.e. number of bins: 12, 24, 36, . . .), and $\omega(n, f_i)$ is the weight of the frequency f_i when considering the HPCP bin n.

The Transposed Harmonic Pitch Class Profile (THPCP) proposed by Gomez (2006) is one of the main contributions in her thesis. They are considered as

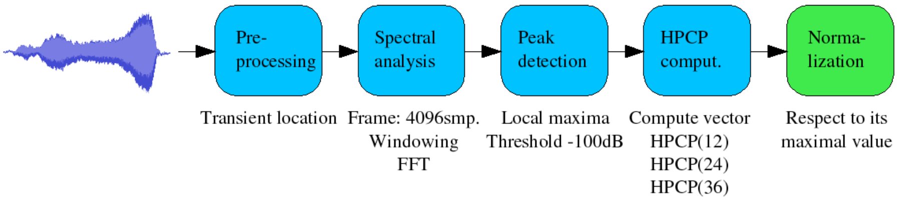


Figure 4.4: General diagram for computing HPCP features

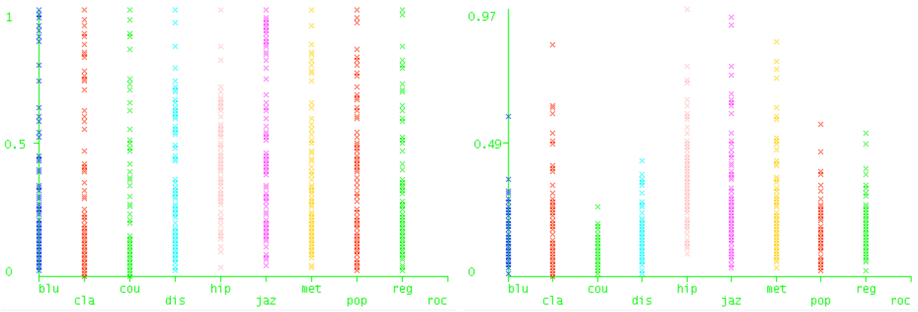


Figure 4.5: Comparison of the behavior of the 7th. coefficient of the HPCP vs THPCP for different musical genres. Information provided for THPCP seem to provide more discriminative power than the HPCP for the different genres

an enhanced HPCP which is invariant to transposition, that is, a transposed version of the HPCP. The THPCP can be computed as a shifted version of the HPCP according to the following formula:

$$THPCP(n) = HPCP(mod(n - shift, N)) \quad n = 1 \ldots N \tag{4.38}$$

where n is the size of the HPCP vector and the index $shift$ can be defined in different ways. First, $shift$ can be assigned to a certain key in order to analyze the tonal profile given by the HPCP features. On the other hand, $shift$ can be automatically assigned according to the estimated key or, finally, according to the maximum value of the HPCP vector.

Figure 4.5 shows the distribution of the 7th HPCP and THPCP coefficients for different musical genres. Note how the information provided for THPCP seem to provide more discriminative power than the HPCP for the different genres. This property is also present in many of the other 11 coefficients.

Some other features can be derived from the HPCP or THPCP. They will be included in our analysis of musical genres in addition to THPCP described above, and we will compare their performance with the timbre and rhythm descriptors. In the following subsections we briefly describe them.

Equal-tempered deviation

The *Equal-tempered deviation* feature is proposed by Gomez & Herrera (2008). It measures the local maxima deviation of the HPCP from equal tempered bins. First, we need to extract a set of local maxima from the HPCP:

$$\{pos_i, a_i\} \quad i = 1 \ldots N \tag{4.39}$$

and then we compute their deviations from closest equal-tempered bins, weighted by their magnitude and normalized by the sum of peak magnitudes:

$$ETD = \frac{\sum a_i \cdot abs(pos_i - equalTempered_i)}{\sum a_i} \tag{4.40}$$

Non tempered to tempered energy ratio

The *Non tempered to tempered energy ratio* is also a descriptor derived from the HPCP and proposed by Gomez (2007). It represents the ratio between the HPCP amplitude for non-tempered bins and the total amplitude:

$$ER = \frac{\sum HPCP_{iNT}}{\sum HPCP_i} \quad i = 1 \ldots HPCPsize \tag{4.41}$$

where $HPCPsize = 120$ and $HPCP_{iNT}$ are given the HPCP positions related to the equal-tempered pitch classes.

Ditaonic Strength

Here again, this descriptor was proposed by Gomez (2007). It represents the maximum correlation of the HPCP vector and the diatonic major profile ring-shifted in all possible positions. We suppose that, as western music somehow characterized by the diatonic major scale, this descriptor should provide higher values than non-western music.

Octave Centroid

The octave centroid descriptor was also proposed by Gomez (2007). All the other tonal related descriptors do not take into account the octave location. The octave centroid finds the geometry center of the played pitches using the following steps: (1) finding the pitches by applying the multi-pitch estimation method proposed by Klapuri (2004), (2) computing the centroid of its representation for each frame, and (3) considering different statistics of frame based values (mean, deviation, etc.) as global descriptors for the analyzed song.

Tonal Roughness

This is the last tonal related descriptor proposed by Gomez (2007). It attempts to be a measure of sensory dissonance. The computation starts with the obtention of the roughness according to the estimation model proposed by Vassilakis (2001) and Vassilakis (2005). The roughness of a frame is obtained by summing the roughness of all pairs of components in the spectrum. The frequency components whose spectral magnitude is higher than 14% of the maximum spectral amplitude are considered as the main frequency components. Then, the global roughness is computed as the median of the instantaneous values.

4.3.6 Panning related descriptors

The Panning Coefficients here presented are based on the work of Gómez et al. (2008). His work is inspired by the results obtained by Barry et al. (2004) and the use of them in the MIR is inspired by Jennings et al. (2004). The extraction of spatial features or Panning Coefficients is defined using the following steps: First, the input audio is divided into frames and windowed by the windowing function $w(n)$. Then, we compute the FFT for each channel in order to obtain the two power spectra functions $S_L(t, f)$ and $S_R(t, f)$. Now, the ratio of the power spectra is computed for each frame:

$$R[k] = \frac{2}{\pi} arctan \left(\left| \frac{S_L[k]}{S_R[k]} \right| \right) \tag{4.42}$$

The resulting sequence represents the spatial localization of each frequency bin k, The range of the panning factor is $[-45, +45]$. This process can be carried out using all the frequency bins k or separately for different frequency bands, providing a more detailed information about the localization of sound sources from different nature.

In order to have a spatial description more related to the human perception (Mills, 1958), non linear functions are applied to the azimuth angle: a warping function will provide more resolution near the $azimuth = 0^0$ as shown below:

$$R_W[k] = f(R[k]) \tag{4.43}$$

where:

$$\begin{aligned} f(x) &= -0.5 + 2.5x - x^2 \quad x \geq 0.5 \\ f(x) &= 1 - (-0.5 + 2.5(1-x) - (1-x)^2) \quad x < 0.5 \end{aligned} \tag{4.44}$$

and $x \in [0, 1] \leftrightarrow Az \in [-45^0, +45^0]$. After the computation of warped ratios sequence, its histogram is deduced by weighting each bin of the histogram i_k with the energy of the frequency bin of the STFT,

$$\begin{aligned} i_k &= floor(M \cdot R_W[k]) \\ H_W(i_k) &= \sum_k = 0^N |S_L[k] + S_R[k]| \end{aligned} \tag{4.45}$$

where $S_L[k] = S_L(t, f_k)$ and $S_R[k] = S_R(t, f_k)$, M is the number of bins of the histogram and N is the size of the spectrum (half of the STFT size).

At this point, the panning image histograms are available. In real performances, they can differ a lot from one frame to the next. A $2s$ averaging is made using a first order low pass filter for each of the M values obtaining:

$$\bar{H}_{w,n} = (1 - a) \cdot \bar{H}_{w,n-1} + a \cdot \bar{H}_{w,n} \tag{4.46}$$

where $a = 1/A$ and A is the number of frames in the averaging. Finally, the panning coefficients need to be independent of the energy of the input signal. We will apply a normalization by the sum of the energy in the bins:

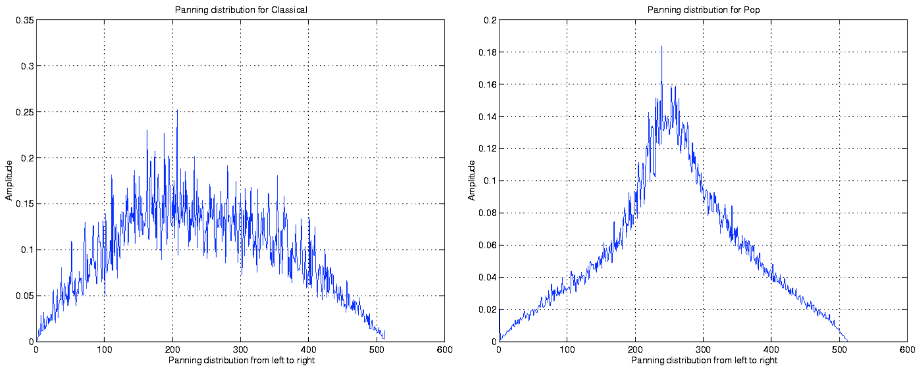


Figure 4.6: Panning distribution for a classical and a pop song (vertical axes are not the same)

$$\bar{H}^{norm}_{w,n} = \frac{\bar{H}_{w,n}}{\sum \bar{H}_{w,n}} \tag{4.47}$$

Figure 4.6 shows two examples of the panning distribution for two different songs (classical and pop musical pieces). We can observe the 512 *bins* which provide the information from left (bin 0) to right (bin 512). The center of the stereo image is located at the bin 256. Note how classical music has a wider stereo image than pop music in which the information is basically centered at the middle of the image.

The information of this normalized and smoothed histogram holds all the spatial information contained in the input signal but it can become a bit hard to manage. The panning coefficients p_l are defined using a cepstrum-based computation over the whole data. It computes the logarithm of the normalized panning histogram $\bar{H}^{norm}_{w,n}$ and applies the IFFT to the result. By taking the real part of the L first coefficients we get a compact description of the panning histogram of the audio signal. A good trade-off between spatial resolution and size/compactness can be achieved with $L = 20$. Figure 4.7 shows two examples of the time evolution of the panning coefficients for two pieces of classical and pop music.

4.3.7 Complexity descriptors

The music complexity descriptors used in our experiments are taken from the set proposed by Streich (2007). They were designed to form a compact representation of music complexity properties such as variability, repetitiveness, regularity and redundancy on the music track level. For our experiments we use the algorithms developed for the facets of timbre, rhythm, dynamics and spatial range.

Dynamic complexity

The dynamic complexity component relates to the properties of the loudness evolution within a musical track. We refer to it in terms of abruptness and rate of changes in dynamic level. The author proposes two different techniques to

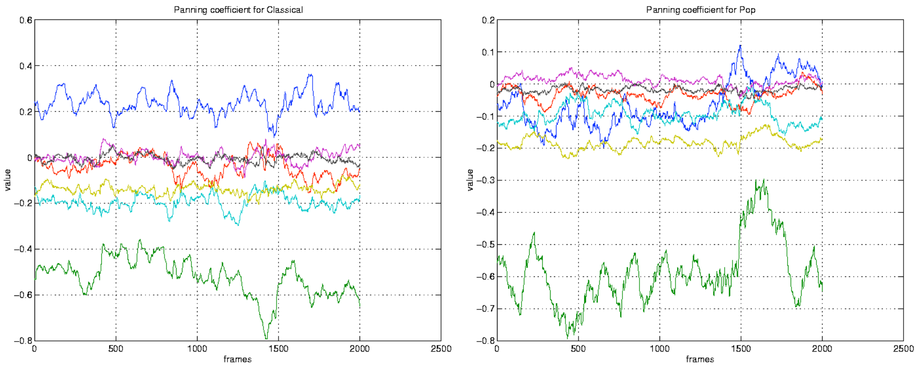


Figure 4.7: Time evolution of panning coefficients for a classical and a pop song (vertical axes are not in the same scale).

compute the dynamic complexity: the first one was proposed by Vickers (2001) and the second one is based on the Pampalk (2001) implementation of Stevens' method (Stevens, 1956, 1962). In our experiments, we use the Pampalk's implementation because it shares many computation steps with other complexity descriptors. Roughly speaking, first we compute the estimation of the loudness of the audio signal and, after that, we compute the time-domain fluctuations.

The process starts with the computation of the power spectrum of the input data in frames of $16ms$. The power spectrum $P(k)$ is then weighted by a curve that is inspired by the frequency response of the outer ear proposed by Terhardt (1979):

$$A(f) = 10^{(0.5*e^{0.6(f-3.3)^2} - 5f^4 \cdot 10^{-5} - 2.184 \cdot f^{-0.8})} \tag{4.48}$$

where $A(f)$ is the weighting value assigned to each frequency f. From this weighted power spectrum we compute then the bark band energies of the 24 bark bands as defined by Zwicker & Fastl (1990):

$$P_{cb}(l) = \sum_{k=low\ B(l)}^{high\ B(l)} P_w(k) \tag{4.49}$$

where $P_w(k) = P(k) \cdot A(k)^2$, and $low\ B(l)$ and $high\ B(l)$ are the high and low limits of each bark band. At this point, a heuristic spreading function $s(l)$ is applied to account for spectral masking effects, according to Schroeder et al. (1979):

$$s(l) = 10^{(0.75 \cdot l + 1.937 - 1.75\sqrt{l^2 + 0.948 \cdot l + 1.225})} \tag{4.50}$$

resulting in the spread energy distribution:

$$P_{spread}(m) = \sum_{l=1}^{24} P_{cb}(l) \cdot s(m-l) \qquad 1 \leq m \leq 24 \tag{4.51}$$

Now, we convert the energy values of each band from the decibel scale ($P_{dB}(m) = 10 \cdot log(p_{spread}(m))$) to the sone scale proposed by Stevens (1956):

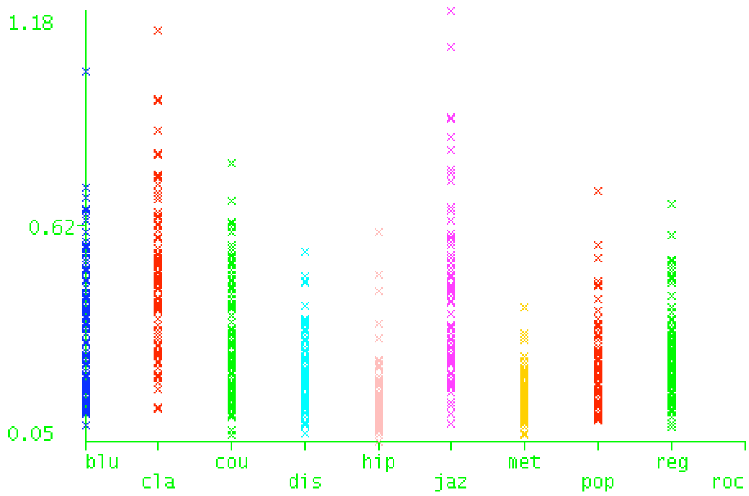


Figure 4.8: Dynamic complexity descriptors computed over 10 musical genres

$$L(m) = \begin{cases} 2^{0.1(P_{dB}(m)-40)} & for P_{dB}(m) \geq 40 \\ (P_{dB}(m)-40)^{2.642} & else \end{cases} \tag{4.52}$$

Finally, we obtain the total loudness estimate for each frame by evaluating

$$L_{total} = 0.85 \cdot L(m_{max}) + 0.15 \sum_{i=1}^{24} L(i) \tag{4.53}$$

where m_{max} is the index of the band with the biggest loudness in the frame. After some empirical smoothness and decimation processes, the fluctuation of the total loudness is computed as follows:

$$C_{dyn} = \frac{1}{N-1} \sum_{i=1}^{N-1} |log_{10}(L_{totsd}(n)) - log_{10}(L_{totsd}(n-1))| \tag{4.54}$$

with N being the number of decimated loudness values for the entire track and L_{totsd} is the smoothed and decimated loudness estimation value.

Figure 4.8 provides some examples of dynamic complexity descriptor computed over 10 musical genres according to the dataset proposed by Tzanetakis & Cook (2002). The Lowest values correspond to Hip-Hop, Disco and Metal music in which the amplitude of the recorded signal is highly compressed (therefore, more constant). Higher values are obtained for classical and jazz music, two musical genres that maximally exploit the dynamics of the musical instruments.

Timbre complexity

Timbre complexity is not a well defined term and it refers to many concepts at the same time. For instance, timbre complexity can be pointed by the number of distinguishable instruments, sound textures present in the music, the rate at which the leading instruments change, or the amount of modulation of the sound sources. According to Streich (2007), we use the LZ77 algorithm proposed by Ziv & Lempel (1977). The assumption is that it is possible to apply techniques like entropy estimation or compression algorithms for measuring timbre complexity.

The basic idea is to apply a sliding window on the sequence of symbols that is to be encoded. The window is split into two parts, the memory buffer with a substring A of fixed length that has been encoded already, and the look-ahead buffer with substring B of fixed length that still needs to be encoded. In each step the algorithm searches for the longest prefix of B beginning in A. This prefix is then encoded as a code word composed of three parts. It contains the offset (number of symbols between prefix and match), the length (number of matching symbols), and the terminating symbol of the prefix in B (the first symbol that doesn't match anymore). If there is no match found, offset and length are set to zero in the code word. The encoded symbols are then shifted into the memory buffer and the procedure starts again until the entire string is encoded.

In our context of the estimation of perceived timbre complexity, we use the compression gain r_c of LZ77 applied to timbre sequences:

$$r_c = \frac{n_c \cdot l_c}{n_s} \tag{4.55}$$

where l_c is the length of the code words relative to the length of the symbols in the original source alphabet, and n_c and n_s are respectively the number of code words and the number of symbols that are needed to represent the string. A low compression factor means that a lot of redundancy was removed in the compression process, and thus the source entropy is low. A compression factor close to one means that the compression algorithm was not able to take much advantage of redundancy, thus the source entropy is supposed to be high. The required compact timbre representation is made up by four timbre descriptors:

Bass: The intensity ratio of spectral content below 100 Hz to the full spectrum. This feature reflects the amount of low frequency content in the signal (originating for example from bass drums, bass guitars, or humming noises)

Presence: The intensity ratio of spectral content between 1.6 and 4 kHz to the full spectrum. This feature reflects a sensation of "closeness" and brilliance of the sound, especially noticeable with singing voices and certain leading instruments.

Spectral Roll-Off: The frequency below which 85% of the spectral energy are accumulated. This feature is related with the perceived bandwidth of the sound. In music it reacts to the presence of strong drum sounds, which push the spectral roll-off up.

Spectral Flatness: The spectral flatness between $250Hz$ and $16kHz$ reflects whether the sound is more tonal or noise-like, as defined in 4.3.3

In this case, Figure 4.9 does not provide relevant differences of timbre complexity for any of the musical genres.

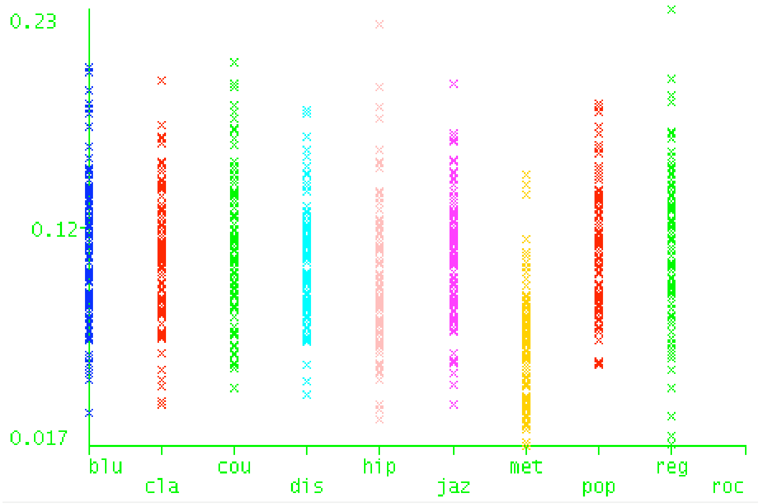


Figure 4.9: Timbre complexity descriptors computed over 10 musical genres

Rhythmic complexity

In our study, rhythmic complexity is addressed in terms of danceability. This descriptor doesn't need any previous assumptions about metrics or tempo, hence it will be useful for all the possible scenarios. Danceability has proved to be a good descriptor in large datasets (Streich & Herrera, 2005) and it is based on the research proposed by Jennings et al. (2004) using the Detrended Fluctuation Analysis (DFA) technique proposed by Peng et al. (1994).

Danceability measure is computed as follows. First, the original audio data is segmented into non-overlapping blocks of $10ms$. For each block, the standard deviation $s(n)$ of the amplitude values is computed. Then, we compute the unbounded time series $y(m)$:

$$y(m) = \sum_{n=1}^{m} s(n) \tag{4.56}$$

The series $y(m)$ can be thought of as a random walk in one dimension. $y(m)$ is now again segmented into blocks of τ elements length. This time, we advance only by one sample from one block to the next. From each block we remove the linear trend $\hat{y}_k$ and compute the mean of the squared residual:

$$D(k,\tau) = \frac{1}{\tau}\sum_{m=0}^{\tau-1}(y(k+m) - \hat{y}_k(m))^2 \tag{4.57}$$

Now, we compute the detrended fluctuation $F(\tau)$ of the time series:

$$F(\tau) = \sqrt{\frac{1}{K}\sum_{k=1}^{K} D(k,\tau)} \tag{4.58}$$

As F is a function of τ, we repeat the same process for different values of τ that are in our range of interest (from $310ms$ to $10s$). The DFA exponent

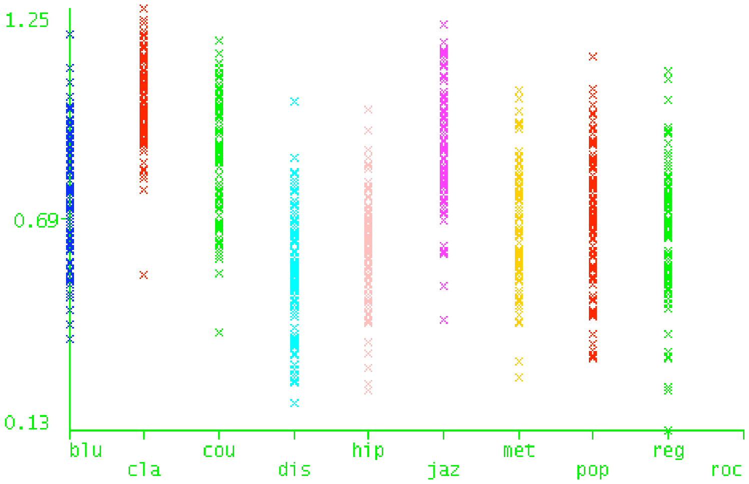


Figure 4.10: Danceability descriptors computed over 10 musical genres. High values correspond to low danceability and low values correspond to high danceability

α is defined as the slope of the double logarithmic graph of F over τ. Then, according to Buldyrev et al. (1995) , the DFA exponent can be computed as:

$$\alpha(i) = \frac{log_{10}(F(\tau_{i+1})/F(\tau_i)}{log_{10}(\tau_{i+1}+3)/(\tau_i+3)} \tag{4.59}$$

White noise will produce a DFA exponent equals to 0.5. The behavior of α is different for genres like classical, where α is near to be constant, in contrast to techno music that produces more abrupt results. According to Streich (2007), the *danceability* descriptor is computed as the lowest local minimum in the scaling curve for time scales below $1.5s$.

Figure 4.10 shows the differences of danceability descriptors between Classical and Jazz music with respect to Disco and Hip-Hop music, as expected.

Spatial complexity

The aim of the spatial complexity is to measure the amount of fluctuation in the stereo image. Although there exists different methods to compute the position of a single sound in the stereo image (Silverman et al., 2005), we will focus on the differences between the information of two channels without taking into account the exact location of the sound source. The method proposed by Streich (2007) is based on the research made by Barry et al. (2004) and Vinyes et al. (2006), described in Section 4.3.6. The process starts with the windowing of the input signal with $w(n)$ with frames of $93ms$ and 50% overlap. Then, we will compute a different power spectra for each channel:

$$P_L(k) = |X_L(k)|^2 \quad and \quad P_R(k) = |X_R(k)|^2 \tag{4.60}$$

where $X_L(k)$ and $X_R(k)$ are the Fourier Transform from the windowed input signal for the left and right channels respectively. Now, we estimate the angle $\tan(\alpha(k))$ corresponding to the direction of the acoustic energy source in a horizontal 180^o range:

$$\tan(\alpha(k)) = \frac{P_R(k) - P_L(k)}{2\sqrt{P_R(k) \cdot P_L(k)}} \tag{4.61}$$

Let $\alpha_q(k)$ be a natural number between the limits $1 \leq \alpha_q(k) \leq 180$. It is a measure of the acoustic energy being concentrated in the left channel ($\alpha_q(k) = 1$) or the right channel ($\alpha_q(k) = 180$) for each bin k. The instantaneous spatial energy distribution $E(k)$ is computed by:

$$E(k) = log_{10}(P_R(k)) + log_{10}(P_L(k)) + 14 \tag{4.62}$$

where 14 is an empirical threshold used for noise suppression. The overall spatial energy distribution $E_{spatial}(m)$ is obtained summing all values $E(k)$ of one frame that have the same value $\alpha_q(k)$:

$$E_{spatial}(m) = \sum_{k \in \{k|\alpha_q(k)=m\}} E(k) \tag{4.63}$$

with $1 \leq m \leq 180$ is the panning in degrees. After some energy normalization and smoothing using a median filter (with the corresponding decrease in accuracy, obtaining a resolution of 3^o and $230ms$) we get the $E^{[i]}_{norm}(m)$, where i is the frame number. Now, we are ready to compute the spatial fluctuation as a measure of the complexity in terms of changes in the stereo image over time:

$$C_{spatFluc} = \frac{1}{N-1} \sum_{i=0}^{N-1} g^{[i]} g^{[i+1]} \cdot \sum_{j=1}^{180} |E^{[i]}_{norm}(j) - E^{[i+1]}_{norm}(j)| \tag{4.64}$$

where N is the total number of frames. But the Spatial Fluctuation measure does not take into account the wideness of the acoustical scene. Streich (2007) also proposes the computation of the Spatial Spread Complexity as:

$$C_{spatSpread} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{m=1}^{180} E^{[i]}_{norm}(m) \cdot |m - c^{[i]}_g|}{E^{[i]}_{sum}} \tag{4.65}$$

where $c^{[i]}_g$ is the center of mass of each distribution:

$$c^{[i]}_g = \frac{E^{180}_{i=1} m \cdot E^{[i]}_{norm}(m)}{E^{[i]}_{sum}} \tag{4.66}$$

and:

$$E^{[i]}_{sum} = \sum_{m=1}^{180} E^{[i]}_{norm}(m) \tag{4.67}$$

Figures 4.11 and 4.12 show the spatial flux complexity and spatial spread complexity descriptors for computed over the STOMP dataset (Rentfrow & Gosling, 2003) respectively. In the case of the spatial flux, we will focus on the variance of the obtained results: Alternative, Electronic and Heavy-metal

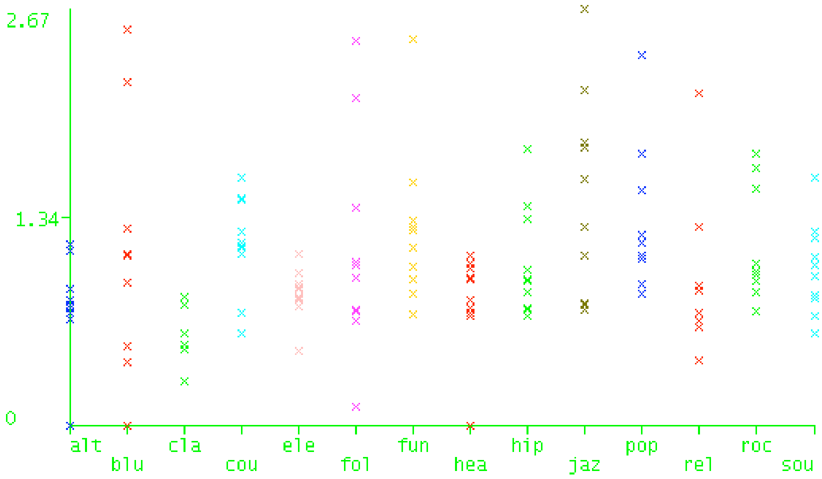


Figure 4.11: Spatial flux complexity descriptors computed over 14 musical genres

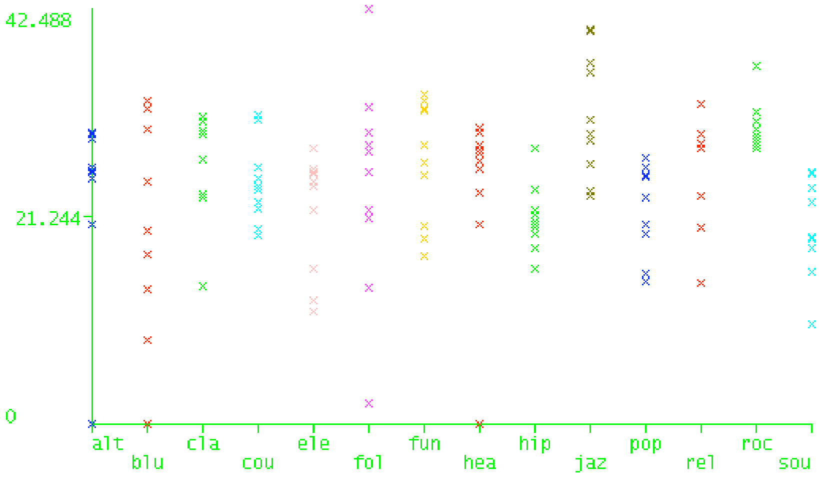


Figure 4.12: Spatial spread complexity descriptors computed over 14 musical genres

music concentrate all the spatial flux values in a very narrow margin while Blues, Folk and Jazz music present a wider margin. It is curious to see how, in this case, Classical music does not follow the same distribution than other *acoustical* musical genres. On the other hand, spatial spread does not show, a priori, any other interesting behavior.

4.3.8 Band Loudness Intercorrelation

Feature computation for musical purposes has been dominated by Mel-Frequency Cepstrum Coefficients and FFT-derived spectral descriptors such as the spectral centroid or the spectral flatness (Pohle et al., 2005a). Most of them also base on the critical-band energy model (Fletcher, 2940), and use, as a kind of pre-processing, rough simplifications of it, as Bark bands or ERB. Contrastingly, we explore here the idea of exploiting the information that arises from the covariational aspects of our environment. The Band Loudness Intercorrelation descriptor (BLI) tries to encode covariations of information at the output of the auditory filters (Aylon, 2006).

McAuley et al. (2005) studied sub-band correlation in robust speech recognition. In order to isolate noisy bands, they created a complex data structure consisting of all possible combinations between sub-bands (i.e., $\{C_{n_1 \ldots n_{N-M}}\}$ where N is the number of sub-bands and M the number of supposedly corrupted bands). To capture correlation between sub-bands, they treated each of these combinations as a single band and calculated a single feature vector for it. Shamma (2008) also studied the correlation between bands in speech processing, from the physiological point of view.

In this thesis, we use these descriptors to find different correlations between the energy of different critical bands for different genres. From the mathematical point of view, BLI is composed by a series of values which correspond to the cross-correlations of the specific loudness at each band weighted by the contribution of the total loudness. The specific loudness can be computed as:

$$L_{ik} = \left\{ \begin{array}{cc} 2^{S_{ik}-40)/10} & id\ S_{ik} \geq 40dB \\ (S_{ik}-40)^{2.643} & otherwise \end{array} \right. \tag{4.68}$$

where S_{ik} is the weighted sonogram in dB at frame k and band i. The total loudness at frame k is then computed as:

$$L_k = \max\{S_{ik}\}_{i=0}^{nBands-1} - 0.15 \left(\sum_{i=0}^{nBands-1} s_{ik} - = \max\{S_{ik}\}_{i=0}^{nBands-1} \right) \tag{4.69}$$

Then, we can express the weighted cross-correlations as:

$$IC_{ij} = \frac{(\omega_i + \omega_j) \cdot CC_{ij}}{W} \tag{4.70}$$

where CC_{ij} is the cross-correlation between bands i and j:

$$CC_{ij} = \frac{COV(\{L_{ik}\}_{k=0}^{N-1}, \{L_{jk}\}_{k=0}^{N-1})}{\sigma_i \cdot \sigma_k} \tag{4.71}$$

and:

$$W = \sum \omega_i \quad and \quad \omega_i = \sum_{k=0}^{N-1} L_{ik} \tag{4.72}$$

where k is the frame number. The resulting matrix has dimension $nBands \cdot nBands$. The higher the number of critical bands is, the higher the size of the resulting matrix. However, CC_{ij} is symmetric and we only take into account one half of the matrix plus the diagonal. Then, the dimension of the final vector is:

$$Dim = \frac{nBands^2 + nBands}{2} \tag{4.73}$$

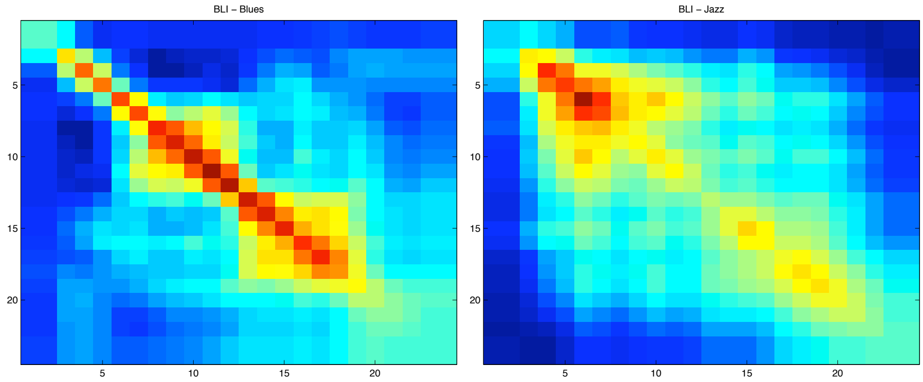


Figure 4.13: BLI matrices for a Blues and a Jazz songs

Finally, $\{IC_{ik}\}_{i=j}$ represents the relative loudness at band i. In our implementation, the frame size is set to 92.9ms with 50% overlap computed over 24 frequency bands.

Figure 4.13 shows two examples of the BLI coefficients for a Blues and a Jazz songs. In the first case, there is no high correlation between the energy bands providing high level values in the diagonal. In the second case, the higher correlation between bark bands produces smaller values in the diagonal and a more spread distribution (the two computed matrices are not representative of the behavior of all the songs in Blues or Jazz musical genres).

4.3.9 Temporal feature integration

Time feature integration is the process of combining all the short-time feature vectors, traditionally computed over frames from 10 to 50ms, into a unique feature vector comprising a larger time frame (from 0.5 to 10s). According to Ahrendt (2006), this process is required for a better description of musical facets such as loudness, rhythm, etc. or other musical properties like tremolo, swing, etc.

As humans use temporal information for genre classification, it seems that classifiers should include this information in some way. Depending on the classification technique used, this process is performed by the classifier while, in other cases, time integration should be specifically computed from the data. In this dissertation, we study the differences between integrating of all the features in a song with respect to the use of the descriptors without time integration (See Section 5.5).

Some of the audio descriptors described in the previous sections are built using temporal feature integration (p.ex. rhythm related descriptors, panning or danceability) but some others doesn't (MFCC or Zero crossing Rate). The integration process provide a unique feature vector for a larger time scale. Roughly speaking, it can be defined as:

$$Z_n = T(x_{n-(N-1)}, \ldots, x_n) \tag{4.74}$$

where Z_n is the new feature vector, x_n is the original feature vector at high resolution for frame n and N is the length of the new feature vector. T symbolizes the feature transformation. The new integrated feature Z_n normally has higher dimensionality than the x_n features. This is necessary to capture all the relevant information from the N frames. For instance, the common Gaussian Model uses the mean and variance of each element. Hence, the dimensionality of the vector Z_n will be twice large as x_n.

Different techniques can be used for feature integration. Here is a short description for some of them:

Basic statistics: These methods use simple statistics of the short-time features such as the mean, variance, skewness and kurtosis, or the autocorrelation coefficient. They are easy to compute and provide a compact representation of the features in a mid-term time scale. Some examples using this technique can be found in Lippens et al. (2004); Li et al. (2003).

Gaussian Model: The use of simple statistics assumes the independence of short-time features in time and among coefficients, but this assumption is not always true. The extension of the basic feature integration model allows to use a full covariance matrix to capture correlations between the individual feature dimensions, as shown by Meng et al. (2007). The main problem using a covariance matrix is the increase of dimensionality in the long-time feature vector: for an original dimension of short-time features d, new feature vectors are $d(d+1)/2$ long.

Multivariate autoregressive model: Gaussian models allow to model the correlation between features but they are not able to model their correlation along time. The multivariate autoregressive model, proposed by Ahrendt & Meng (2005), models the multivariate time series of feature vectors with an autoregressive model. Mathematically, the model can be written in terms of the random process x_n as

$$x_n = \sum_{p=1}^{P} A_p x_{n-p} + v + u_n \tag{4.75}$$

where P is the model order, the A_i's are (deterministic) autoregressive coefficient matrices, v is the so-called (deterministic) intercept term and u_n is the driving noise process. It is found that

$$v = (I - \sum_{p=1}^{P} A_p)\mu \tag{4.76}$$

where μ is the mean of the signal process x_n. Hence, the intercept term v is included explicitly to allow a (fixed) mean value of the feature signal. The noise process u_n is here restricted to be white noise (ie. without temporal dependence) with zero mean and covariance matrix C.

Dynamic principal component analysis: The use of this method for music genre classification was proposed by Ahrendt et al. (2004). The main idea is to first perform a time stacking of the original signal (the short-time feature vectors) which provides results in a high dimensional feature space. Principal component analysis (PCA) is then used to project the stacked features into a new feature space of (much) lower dimensionality. From the mathematical point of view, it can we written:

$$y_n = \begin{bmatrix} x_{n-(N-1)} \\ \vdots \\ x_n \end{bmatrix} \tag{4.77}$$

where N is the frame-size. The long-term feature vector can be computed as:

$$Z_n = \tilde{U}^T (y_n - \hat{\mu}) \tag{4.78}$$

where each row of $\tilde{U}$ is the estimated k first eigenvectors of the covariance matrix of y_n. The eigenvectors belong to the k largest eigenvalues and represent the directions with the greatest variance. $\hat{\mu}$ is the estimate of the mean of y_n. The use of this technique for genre classification aims to detect the strongest correlation between both the individual features and at different times.

Other: Another approach is proposed by Cai et al. (2004). He models the temporal evolution of the energy contour using a polynomial function. Another temporal feature integration method for genre classification is proposed by Esmaili et al. (2004). His algorithm uses the entropy energy ratio in frequency bands, brightness, bandwidth and silence ratio.

4.4 Pattern Recognition

In this section, we introduce the pattern recognition techniques that are used in this thesis. Most of them are widely known and commonly used by the MIR community. Here, we present a short overview of all them. We make a special effort in the explanation of the SIMCA algorithm which, after some analysis and decisions explained in Section 5.6, is one of the most important contributions of this thesis. Although the technique itself is not new, it is the first time that SIMCA is applied to solve a MIR problem.

4.4.1 Nearest Neighbor

Nearest Neighbor is the simplest classification algorithm. Its basic idea consists in the fact that an object is classified by a majority vote of its neighbors. It is classified according to the class most common amongst its k nearest neighbors, where k is a positive integer, typically small. For instance, if $k = 1$, the object is assigned to the class of its nearest neighbor (See Figure 4.14 for details).

The neighbors are taken from a set of objects for which the correct classification is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. In order to identify neighbors, the objects are represented by position vectors in a multidimensional feature space (p.e. 2 dimensional vector). The full process can be summarized as follows:

Training: For each training example $x, f(x)$, add the example to the list *training examples.*

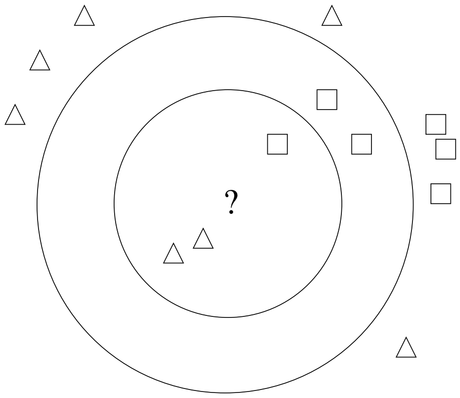


Figure 4.14: Example of a Nearest Neighbor classification. Categories (triangles and squares) are represented in a 2D feature vector and plotted in the figure. The new instance to be classified is assigned to the *triangles* category for a number of neighbors $N = 3$ but to the *squares* category for $N = 5$.

Classification: Given a query instance x_q to be classified, let $x_1 \dots x_k$ denote the k instances from *training examples* that are nearest to x_q and return:

$$\bar{f}(x_q) \leftarrow argmax \sum_{i=1}^{k} \delta(v, f(x_i)) \tag{4.79}$$

where $\delta(a, b) = 1$ if $a = b$ and $\delta(a, b) = 0$ otherwise

4.4.2 Support Vector Machines

Support Vector Machines(SVMs) have proved to be a useful technique for data classification. Their study started in the late seventies by Vapnik (1972) but it is not until mid nineties that they received the attention from researchers. According to Burges (1998), SVMs have provided notorious improvements in the fields of handwritten digit recognition, object recognition and text categorization, among others.

SVMs try to map the original training data into a higher (maybe infinite) dimensional space by a function ϕ. For that, SVMs create a linear separating hyper-plane with the maximal margin in this higher dimensional space (See Figure 4.15 for the visualization of the hyper-plane w and support vectors, reduced in a 2D space). From the mathematical point of view (and following the nomenclature proposed by Hsu et al. (2008)), given a training set of instance-label pairs $(x_i, y_i), i = 1 \dots l$, where $x_i \in R^n$ and $y \in \{-1, 1\}^l$, SVMs search the solution of the following optimization problem:

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \\ \textit{subject to}: \; &y_i \left(w^T \phi(x_i) + b \right) \geq \; 1 - \xi_i; \; \xi_i > 0 \end{aligned} \tag{4.80}$$

Here training vectors x_i are mapped into the higher dimensional space by the function ϕ. $C > 0$ is the penalty parameter of the error term. Furthermore,

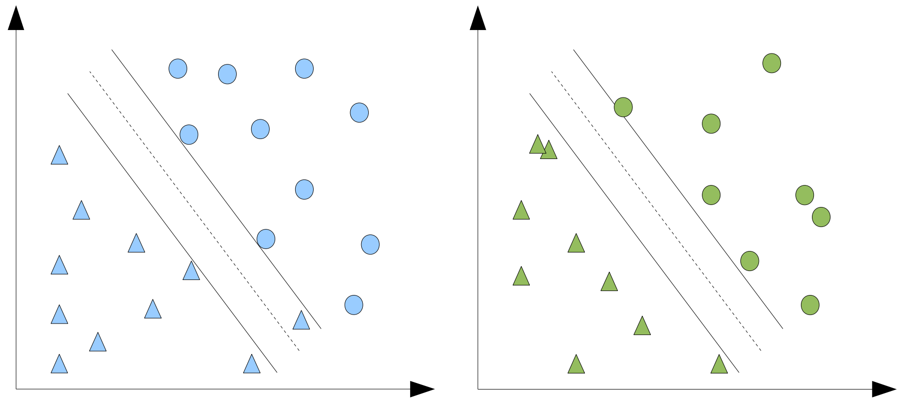


Figure 4.15: Hyper-planes in a SVM classifier. Blue circles and triangles belongs to training data; Green circles and triangles belongs to testing data (Figure extracted from Hsu et al. (2008))

$K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is called the kernel function. There are four basic kernel functions: linear, polynomial, radial basis function and sigmoid but, for the music classification problems, the Polynomial (Poly) and Radial Basis Function (RBF) are the most commonly used.

SVMs are able to deal with two-class problems, but there exists many strategies to allow SVMs work with a larger number of categories. Finally, SVMs use to provide better results working with balanced datasets. Further information on SVMs can be found in Vapnik (1995); Burges (1998); Smola & Schölkopf (2004).

4.4.3 Decision Trees

According to Mitchell (1997), decision tree learning is a method for approximating discrete-valued target functions in which the learned function is represented by a decision tree. Learned trees can also be represented as sets of *if-then* rules to improve the human readability.

Decision trees classify instances by sorting them down the tree from the root to some lead node which provides the classification of the new instance, and each branch descending from that node corresponds to one of the possible values for this attribute. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, the moving down the tree branch corresponding to the value of the attribute in a given example. This process is repeated for the subtree rooted at the new node. Here is a possible algorithm to train a decision tree:

1. Select the best decision attribute for next node. The selected attribute is that one that, according to a threshold, best classifies the instances in the dataset.

2. Assign the selected attribute as the decision attribute for that node

3. For each value of the selected attribute, create new descendant of node

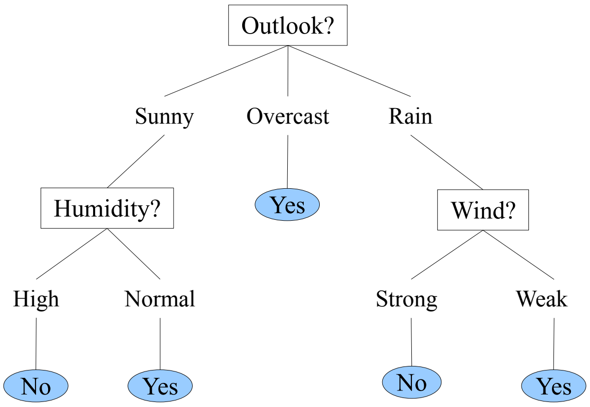


Figure 4.16: Typical learned decision tree that classifies whether a Saturday morning is suitable for playing tennis or not, using decision trees (Figure extracted from Mitchell (1997))

4. Sort training examples to leaf nodes
5. If training examples perfectly classified, Then STOP, Else iterate over new leaf nodes

Figure 4.16 show a typical learned decision tree that classifies whether a Saturday morning is suitable for playing tennis or not.

4.4.4 Ada-Boost

The AdaBoost algorithm was introduced by Freund & Schapire (1997). AdaBoost is an algorithm for constructing a *strong* classifier as a linear combination of *weak* classifiers. The algorithm takes as input a training set $(x_1, y_1) \ldots (x_m, y_m)$ where each x_i belongs to some domain or instance space X, and each label y_i is in some label set Y . Assuming $Y = -1, +1$, AdaBoost calls a given weak algorithm (that is, a simple classification algorithm) repeatedly in a series of rounds $t = 1 \ldots T$. One of the main ideas of the algorithm is to maintain a distribution or set of weights over the training set. The weight of this distribution on training example i on round t is denoted $D_t(i)$. Initially, all weights are set equally, but on each round, the weights of incorrectly classified examples are increased so that the weak learner is forced to focus on the hard examples in the training set.

The weak learner's job is to find a weak hypothesis $h_t : X \rightarrow -1, +1$ appropriate for the distribution D_t. The goodness of a weak hypothesis is measured by its error

$$\epsilon_t = Pr_{i \sim D_t}[h_t(x_u) \neq y_i] = \sum_{i:h_t(x_i) \neq y_i} D_t(i) \tag{4.81}$$

Notice that the error is measured with respect to the distribution D_t on which the weak learner was trained. In practice, the weak learner may be an algorithm that can use the weights D_t on the training examples. Alternatively, when this is not possible, a subset of the training examples can be sampled according to D_t, and these (unweighted) resampled examples can be used to train the weak learner.

4.4.5 Random Forests

As mentioned in Section 3.2.4, Machine learning methods are often categorized into supervised and unsupervised learning methods. Interestingly, many supervised methods can be turned into unsupervised methods using the following idea: one creates an artificial class label that distinguishes the observed data from suitably generated synthetic data. The observed data is the original unlabeled data while the synthetic data is drawn from a reference distribution. Breiman & Cutler (2003) proposed to use random forest (RF) predictors to distinguish observed from synthetic data. When the resulting RF dissimilarity is used as input in unsupervised learning methods (e.g. clustering), patterns can be found which may or may not correspond to clusters in the Euclidean sense of the word.

The main idea of this procedure is that for the k_{th} tree, a random vector $\odot_k$ is generated, independent of the past random vectors $\odot_1, \ldots, \odot_{k-1}$ but with the same distribution; and a tree is grown using the training set and $\odot_k$, resulting in a classifier $h(x, \odot_k)$ where x is an input vector. For instance, in bagging the random vector $\odot$ is generated as the counts in N boxes resulting from N darts thrown at random at the boxes, where N is number of examples in the training set. In random split selection $\odot$ consists of a number of independent random integers between 1 and K. The nature and dimensionality of $\odot$ depends on its use in tree construction. After a large number of trees is generated, they vote for the most popular class.

4.5 Statistical Methods

In this section, we will introduce the PCA and SIMCA methods. These techniques are not new and have been used in many research topics for the last decades. To our knowledge, the SIMCA algorithm has never been tested in MIR problems, and we decided to include it in our study because of the conclusions extracted in all our previous analysis. We will discuss the obtained results in Section 5.7 but we include the technical explanation in this section because the algorithm itself is not new. The principal component analysis is part of the SIMCA. Hence, we also include its technical explanation here.

4.5.1 Principal Components Analysis

Here, we will show a brief explanation on Principal Components Analysis (PCA). PCA is a powerful statistical technique that tries to identify patterns in our data representing it in such a way as to reinforce their similarities and differences. One of the main advantages using PCA is the data compression by reducing the number of dimensions without much loss of information.

Let $Z^{(i)}$ be the feature vector of dimension $i = 1..N$,where N is the number of features sampled at frame rate f_r:

$$Z^{(i)} = \left[Z_1^{(i)}, Z_2^{(i)}, \ldots, Z_K^{(i)} \right] \tag{4.82}$$

where K is the number of frames extracted from the audio file. Now, we randomly select r feature vectors from $Z^{(i)}$[3]:

$$Z_r = \left[Z_r^{(0)}, Z_r^{(1)}, \ldots, Z_r^{(N-1)}\right] \tag{4.83}$$

Now, we compute the squared covariance matrix $Z_r^T Z_r$ and perform the eigenvalue decomposition as:

$$Z_r^T Z_r = U_r \wedge_r U_r^T \tag{4.84}$$

where $\wedge_r$ is the squared symmetric matrix with ordered eigenvalues and U_r contains the corresponding eigenvectors. Next, we compute the outer product eigenvectors of $Z_r Z_r^T$ using the relationship between inner and outer products from the singular value decomposition. At this point, we only retain the p eigenvectors with largest eigenvalues:

$$Z_r = V_p S_p U_p^T \quad \rightarrow \quad V_p = Z_r U_p S_p^{-1} \tag{4.85}$$

where V_p has dimension $K \times p$, S_p is the singular values of dimension $p \times p$ and U_p is of dimension $rN \times p$. V_p is computed by calculating $Z_r U_p$ and normalizing the columns of V_p for stability reasons.

With the selection of only the p largest eigenvalue/eigenvector pairs, the eigenvectors can be considered as an approximation to the corresponding p largest eigenvector/eigenvalue pair of the complete matrix $ZZ^T = V \wedge V^T$. Then,

$$V_p \wedge_r V_p^T \approx ZZ^T \tag{4.86}$$

New data Z_{test} can be projected into the p leading eigenvectors as:

$$Z'_{test} = V_p^T Z_{test} \tag{4.87}$$

From our point of view, this mathematical procedure can be interpreted as a linear combination of the existing features into a new feature space. For dimension reduction we will work only with the most important linear combinations of the projection. The reader will find more information about PCA in Jolliffe (2002); Shlens (2002); Meng (2006).

4.5.2 SIMCA

The SIMCA method (Soft Independent Modeling of Class Analogies) was proposed by Wold (1976). It is specially useful for high-dimensional classification problems because it uses PCA for dimension reduction, applied to each group or category individually. By using this simple structure, SIMCA also provides information on different groups such as the relevance of different dimensions and measures of separation. It is the opposite than applying PCA to the full set of observations because the same reduction rules are applied through all the original categories. SIMCA can be robustified in front of the presence of outliers by combining robust PCA method with a robust classification rule

[3] The random selection of feature vectors is performed according to a specific validation method detailed in Section 3.2.5

based on robust covariance matrices (Hubert & Driessen, 2004), defining the RSIMCA.

As mentioned above, the goal of SIMCA is to obtain a classification rule for a set of m known groups. Using the nomenclature proposed by Vanden & Hubert (2005), let X^j be the m groups where j indicates the class membership ($j = 1 \dots m$). The observations of group X^j are represented by x_i^j, where $i = 1 \dots n_j$ and n_j is the number of elements in the group j. Now, let p be the number of variables for each element providing $x_i^j = (x_{i1}^j, x_{i2}^j, \dots, x_{ip}^j)'$. This number of variables p can be really high, up to some hundreds or thousands of different variables for each element. Finally, let Y^j be the validation set, with $j = 1 \dots m$.

The goal of SIMCA is not only the classification itself but also to enhance the individual properties of each group. Then, PCA is performed on each group X^j independently. This produces a matrix of scores T^j and loadings P^j for each group. Let $k^j << p$ be the retained number of principal components for group j. At this point, let's define OD as the orthogonal distance (Euclidean distance) from a new observation to the different PCA models. Let y be a new observation to be classified, and let $\tilde{y}^{(l)}$ represent the projection of this observation on the PCA model of group l:

$$\tilde{y}^{(l)} = \bar{x}^l + P^l(P^l)'(y - \bar{x}^l) \tag{4.88}$$

where $\bar{x}^l$ is the mean of the training observations in group l. The OD to group l is then defined as the norm of the deviation of y from its projection $\tilde{y}^{(l)}$:

$$OD^{(l)} = \|y - \tilde{y}^{(l)}\| \tag{4.89}$$

The classification of new data is performed by comparing the deviation $(OD^{(l)})^2$ when assigned to a specific class l with the variance of the l^{th} training group s_l^2. Specifically, it is computed using the F-test[4] by looking at $(s^{(l)}/s_l)^2$:

$$(s^{(l)})^2 = \frac{(OD^{(l)})^2}{p - k_l} \tag{4.90}$$

and:

$$s_l^2 = \frac{\sum_{i=1}^{n_l}(OD_i^l)^2}{(p - k_l)()n_l - k_l - 1} \tag{4.91}$$

If the observed F-value is smaller than the critical value $F_{p-k_l,(p-k_l)(n_l-k_l-1);0.95}$, the 95% quantile of the F-distribution with $p - k_l, (p - k_l)(n_l - k_l - 1)$ degrees of freedom, the new observation y is said to belong to the l_{th} group. Thus, an observation can be classified in many different groups at the same time.

This approach does not completely exploit the benefit of applying PCA in each group separately. Wold (1976) and Albano et al. (1978) suggested to include another distance in the classification rule. It was defined as the distance to the boundary of the disjoint PCA models. For each of the m groups, a multidimensional box is constructed by taking into account the scores t_i^l for $i = 1, \dots, n_l$, where $t_i^l = (t_{i1}^l, t_{i2}^l, \dots, t_{ik_l}^l)'$ represents the k_l-dimensional score

[4] the F Test provides a measure for the probability that they have the same variance

of the i_{th} observation in the training set X^l. The boundary for each set of scores is defined by looking at the minimal and maximal value of the scores componentwise:

$$\min_{i=1,\dots,n_l} t^l_{ij} - cd^l_j, \max_{i=1,\dots,n_l} t^l_{ij} - cd^l_j \tag{4.92}$$

where d^l_j is the standard deviation of the j_{th} component of the t^l_i:

$$d^l_j = \sqrt{\frac{1}{n_l - 1}\sum_{i=1}^{n_l}\left(t^l_{ij} - \frac{1}{n_l}\sum_{a=1}^{n_l} t^l_{aj}\right)^2} \tag{4.93}$$

where c is usually taken equal to 1.

The new Boundary Distance $BD^{(l)}$ is defined as the distance of a new observation y to the boundary of the l_{th} PCA model. If the observation falls inside the boundaries, $BD^{(l)} = 0$. Finally, assigning y to any of the m classes is again done by means of an F-test based on a linear combination of $(BD^{(l)})^2$ and $(OD^{(l)})^2$. The reader will find more information about SIMCA in Vanden & Hubert (2005).

4.6 Conclusions

In this chapter, we introduced the scientific background that will be used in our experiments. We started with the statistical concepts required for the definition of our own rhythmic descriptor. Then, we discussed some previous considerations when computing audio descriptors (length of the audio excerpt, hop-size, etc.) and showed different descriptors grouped into different families. Some of these descriptors are widely known by the community (MFCC, IOI, etc) while others are recently developed at the Music Technology Group for different purposes (THPCP, Danceability, etc.). Although the definition of these descriptors is not in the scope of this thesis, we will evaluate its behavior in front of the automatic genre classification problem. Then, we discussed different machine learning techniques traditionally used in classification techniques with an special effort on the PCA and SIMCA techniques that, although they are not new, it is the first time that this last one is used to solve an audio classification problem.

As a conclusion, we presented all the individual parts required to start analyzing automatic classification of musical genres. Although some of them are relatively new (i.e. danceability descriptor) or never used in the MIR community (i.e. SIMCA classification technique) we preferred to separate its description from the study in genre classification. So, all the discussions presented in Chapter 5 can be considered as contributions of this thesis for automatic music genre classification.

CHAPTER 5

Contributions and new perspectives for automatic music genre classification

5.1 Introduction

In this chapter, we describe the main contributions of this thesis. We start presenting a rhythmic descriptor designed to solve some specific gaps we found in the literature. Specifically, we present the *rhythm transform* as an alternative way to show rhythm in a similar way we can compute the spectrum of a signal, so, we are able to compute al the derived descriptors of the spectrum (spectral centroid, spectral flatness, MFCC) but using rhythmic information.

Next, we show the configuration and the results of a set of listening experiments developed to guide us in the research process. These experiments allow us to determine the importance of two musical facets (timbre and rhythm) in genre classification. Results of these experiments are contrasted with the output of automatic classifiers and they allow us to decide how to build our ideal genre classifier.

After that, we describe our participation to the MIREX'07 contest which serve us to establish a baseline for music genre classification, with respect to the whole community, proposing a state of the art algorithm. Starting from that, we present a quite exhaustive comparison of classification using different descriptors, classifiers and datasets: all the available families of descriptors are computed for different datasets, and we build genre classifiers using different machine learning techniques. We also evaluate the behavior of classifiers in front environments (i.e. mixing datasets) and analyze the obtained accuracies to extract some preliminary conclusions.

The results of the listening experiments and the conclusions extracted from the evaluation of automatic genre classifiers drive us to think about the architecture of a conceptually different genre classifier. We try to get closer to the human classification process instead of obtaining better accuracies in the classification. We show the reasoning process and the obtained results with the new proposed classifier in different scenarios. Finally, we perform some additional tests to show how this classifier can be also useful to other classification tasks.

5.2 Descriptors

5.2.1 The rhythm transform

Many rhythmical descriptors can be shown in the literature. As shown in Section 4.3.4, some of them depend on manually fixed parameters or experimental thresholds, or they only give a partial point of view about the whole rhythmic information. The so called *Rhythm Transform* offers a rhythmic representation that is able to represent the rhythm content for all kinds of music without using thresholds based on experimental values.

Note that we call *rhythm transform* from a conceptual point of view. It is not a real transform from a mathematical point of view since the inverse transform can not be defined. But the obtained data could be interpreted as data in the so called *rhythm domain.*

Rhythm transform

Some rhythmic descriptors compute the frequency analysis of the input signal and search for the common energy periodicities through different (linear o mel-frequency based) sub-bands. The energy's periodicity search is implemented as a bank of resonators and represented as a Beat Spectrum or as a Beat Histogram. The Rhythm Transform is slightly different: the periodogram is calculated for the energy derivative of each sub-band of the input data and a weighted sum is implemented for a global rhythm representation. Next, we show each step in the process:

Frequency decomposition: The input data $x(t)$ is filtered with the anti-alias filtering and sampled with $f_s = 22050[Hz]$. The length of the frames is $l = 300[ms]$, the hop-size is $h = 30[ms]$ and Hamming windowing is applied. Digital windowed data $x_w[n]$ is decomposed into different sub-bands with a 1/3 octave filter bank to simulate the perceptual behavior of the human ear. At this point, different digital signals are obtained:

$$\begin{array}{rcl} x_{f_c=20[Hz]}[n] & = & \frac{1}{N_1}\sum_{f=17.8[Hz]}^{f=22.4[Hz]} |x_w[n]| \\ x_{f_c=25[Hz]}[n] & = & \frac{1}{N_2}\sum_{f=22.4[Hz]}^{f=28.2[Hz]} |x_w[n]| \\ & \cdots & \\ x_{f_c=10000[Hz]}[n] & = & \frac{1}{N_{28}}\sum_{f=8913[Hz]}^{f=11220[Hz]} |x_w[n]| \end{array} \tag{5.1}$$

where N_i is the number of points of the FFT inside each 1/3 octave band.

Energy Extraction: The log of the energy is obtained for each band:

$$\begin{array}{rcl} e_{f_c=20[Hz]}[n] & = & log_{10}\left(x_{f_c=20[Hz]}[n]\right) \\ e_{f_c=25[Hz]}[n] & = & log_{10}\left(x_{f_c=25[Hz]}[n]\right) \\ & \cdots & \\ e_{f_c=10000[Hz]}[n] & = & log_{10}\left(x_{f_c=10000[Hz]}[n]\right) \end{array} \tag{5.2}$$

Derivative of the Energy: The derivative of the energy is computed:

$$\begin{array}{rcl} d_{i,f_c=20[Hz]}[n] & = & e_{i,f_c=20[Hz]}[n] - e_{i-1,f_c=20[Hz]}[n] \\ d_{i,f_c=25[Hz]}[n] & = & e_{i,f_c=25[Hz]}[n] - e_{i-1,f_c=25[Hz]}[n] \\ & \dots & \\ d_{i,f_c=10000[Hz]}[n] & = & e_{i,f_c=10000[Hz]}[n] - e_{i-1,f_c=10000[Hz]}[n] \end{array} \tag{5.3}$$

Periodogram calculations: The periodogram is computed for each buffer, as explained in Sec. 4.2.2. Then,

$$\begin{array}{rcl} I_{f_c=20[Hz]}[\omega] & = & \frac{1}{LU}\sum_{m=-(L-1)}^{L-1} c_{vv,f_c=20[Hz]}[m]e^{-j\omega m} \\ I_{f_c=25[Hz]}[\omega] & = & \frac{1}{LU}\sum_{m=-(L-1)}^{L-1} c_{vv,f_c=25[Hz]}[m]e^{-j\omega m} \\ & \dots & \\ I_{f_c=10000[Hz]}[\omega] & = & \frac{1}{LU}\sum_{m=-(L-1)}^{L-1} c_{vv,f_c=10000[Hz]}[m]e^{-j\omega m} \end{array} \tag{5.4}$$

where c_{vv,f_i} is the aperiodic correlation sequence of each d_{f_i} sequence, and $L = 6[s]$, which is the worst case for a full 4/4 bar at 40 BPM.

Weighted sum: Finally, the weighted sum for all the periodograms for each band is computed. The weighting vector is:

$$r[1..nBands] = \left\{\frac{1}{nBands}, \frac{1}{nBands}, \dots\right\} \tag{5.5}$$

but it can be manually modified in order to emphasize some frequency bands. For general pop music, where the rhythm is basically played by Bass and Bass drums, it can be any decreasing sequence, i.e.:

$$r[1..nBands] = \left\{\frac{1}{1}, \frac{1}{2}, \dots\right\} \tag{5.6}$$

and for some kind of Latin music, where some high-frequency instruments are usually played, the weighting vector could be:

$$r[1..nBands] = \left\{\frac{1}{1}, \frac{1}{2}, \dots, \frac{1}{9}, \frac{1}{10}, \frac{1}{9}, \dots, \frac{1}{2}, \frac{1}{2}\right\} \tag{5.7}$$

But in a general way, the weighting vector described in Eq. 5.5 is good enough. By performing the weighted sum of the squared values of all the periodograms, data in the *rhythm domain* is obtained:

$$T(\omega) = \sum_{j=1}^{nBands} r(j)I_{f_j}[\omega] \tag{5.8}$$

Figure 5.1 shows the block diagram to compute the Rhythm Transform.

Interpretation of data in *Rhythm Domain*

Which information is available from data in the *Rhythm Domain*? The BPM information can be found as the greatest common divisor for all the representative peaks since the beat can be defined as the common periodicity of the

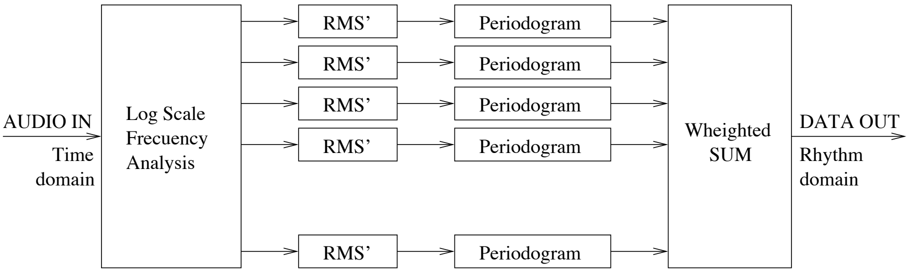


Figure 5.1: Block diagram for *Rhythm Transform* calculation

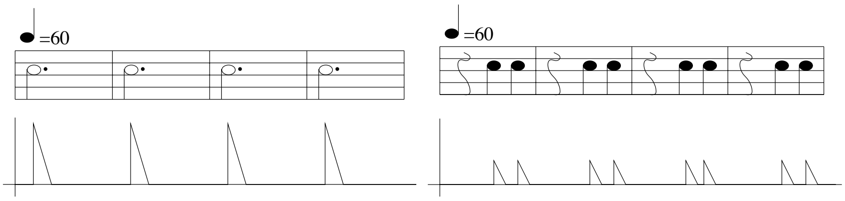


Figure 5.2: Two examples of different periodicities of a musical signal with the same tempo, and their corresponding temporal evolution of the energy. The first one corresponds to a strong beat (with lower periodicity) and the second one corresponds to the weak beats (with higher periodicity)

energy peaks for all the instruments in a song. For BPM detection, any peak detection algorithm across data in rhythm domain data can be used.

But the major advantage of this representation is that it gives some time domain information too. Let's see this duality from a conceptual point of view: It is well known that music is structured in bars, in a given meter. It is well known that the strongest beat in a bar is usually the first one. This means that this strongest beat in a bar appears less frequently: it has the smaller periodicity. On the other hand, a weak beat will appear more frequently since it has a higher periodicity (see Figure 5.2)

In rhythm domain, weak beats appear at higher BPM than strong beats. Furthermore, in time domain, weak beats appear later than strong beats too. This correspondence allows to interpret data in rhythm domain *as* data in time domain. This is what we call *duality* of data in rhythm domain and time domain.

Assuming this duality, the time signature from audio data can easily be deduced. Data between two higher peaks can be seen as the distribution of the beats in a bar. If data between two maximum peaks is divided by twos, a simple meter is assumed as it is shown in Figure 5.3. If data between two maximum peaks is divided by threes, a compound meter is assumed as it is shown in Figure 5.4. On the other hand, if data in a simple or compound bar is sub-divided by twos, a duple meter is assumed as it is shown in the first Figure of 5.3 and 5.4. If this data is sub-divided by threes, a triple meter is assumed as it is shown in the first Figures of 5.3 and 5.4. Finally, in Figure 5.5 if a simple duple meter is sub-divided by twos, the presence of swing is assumed. Let the swing structure as a dotted quarter-note and a eight-note, fact that

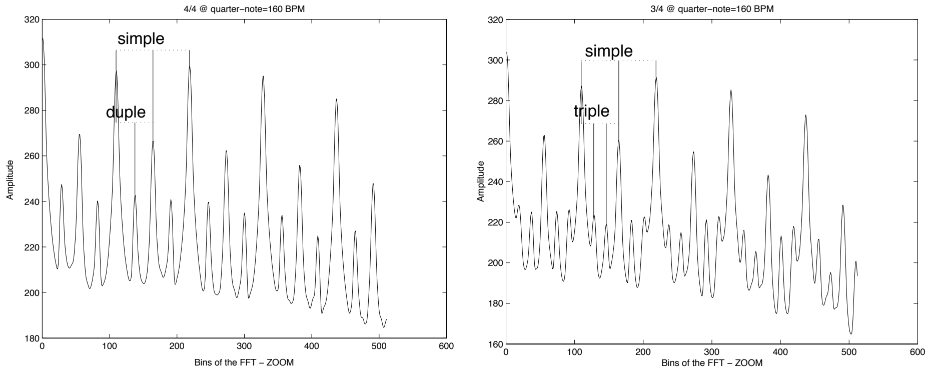


Figure 5.3: Examples of data in Rhythm Domain for a simple duple meter and a simple triple meter

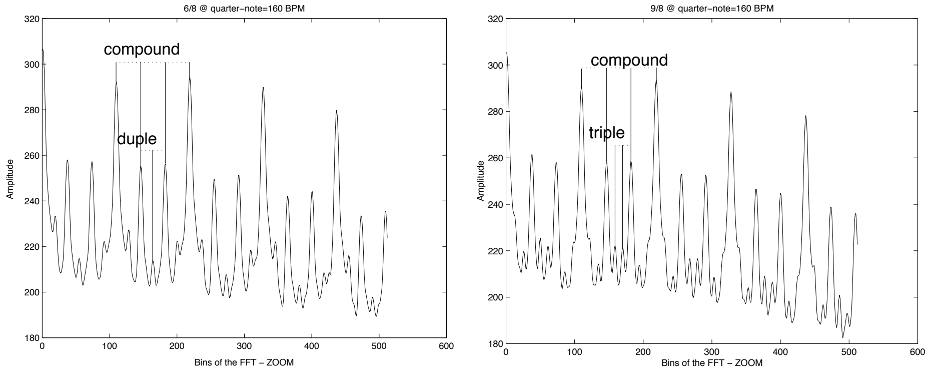


Figure 5.4: Examples of data in Rhythm Domain for different compound duple meter and compound triple meter

can be debatable (Gouyon, 2003).

In conclusion, the main advantage of this method is that we have much more information than the information available at the output of a set of resonators tuned to different BPM typical values and a unique frequency value is related to a unique BPM value. The BPM resolution is higher than other methods, and we have not only the BPMs but all the existing periodicities as well according to different human aspects of music. Furthermore, different rhythms with the same meter and structure but played with different *feeling* can be distinguished by using the Rhythm Transform, as shown in both the swinged and real audio examples.

Limitations

The Rhythm Transform is limited by FFT resolution. For low BPM values, the periodicity is low, then the subdivisions by twos or threes will be much closer than the distance between two bins. On the other hand, this descriptor may fail in those cases with music performed by non-attack instruments that may define a rhythm with pitch or timbre variations (some strings, choirs, synthetic

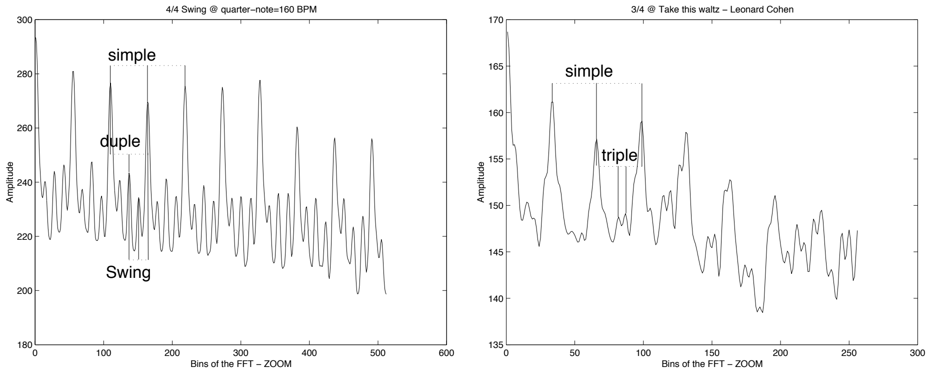


Figure 5.5: Examples of data in Rhythm Domain for simple duple meter (swing) and a real case: Take this Waltz by Leonard Cohen

Genre	BPM_{mean}	BPM_{var}	Beatedness
Dance	140	0.49	3.92
Pop	108	0.71	5.23
Soul	96	0.43	3.48
Jazz	132	1.26	3.54
Classic	90	32	0.95
Voice	66	46.9	0.36

Table 5.1: BPM and Beatedness for different musical genres

pads, etc.).

5.2.2 Beatedness descriptor

The *beatedness* calculation is an application on the use of data in *Rhythm domain*. This concept was introduced by Foote et al. (2002) and evaluated by Tzanetakis et al. (2002). The Beatedness is a measure of how strong are the beats in a musical piece. The beatedness is computed as the Spectral Flatness of the sequence but in the rhythm domain. Spectral Flatness is a measure of the tonality components in a given spectrum, and it is defined as:

$$SF_{dB} = 10 \cdot \log \frac{G_m}{A_m} \tag{5.9}$$

where G_m and A_m are the geometric an arithmetic mean values from all the bins of the Fourier Transform of the signal, respectively. In the case of the Beatedness computation, G_m and A_m are the geometric an arithmetic mean values from all the bins of data in rhythm domain.

High beatedness values are due to very rhythmic compositions as it happens in Dance or Pop whereas low beatedness values are due to non rhythmical compositions as it happens in some Jazz Solo, classical music or speech.

Some BPM & Beatedness measures for different musical genres are shown in Table 5.1. All these measures belong to one frame of "No Gravity" by

DJ Session One for Dance music, “Whenever,Wherever” by Shakira for Pop music, “Falling” by Alicia Keys for Soul music, “Summertime” by Gershwin for Jazz music, “Canon” by Pachelbel for Classic music and one minute of radio recording for voice. Note that in *Dance*, *Pop* and *Soul* music, the the system shows low BPM_{var} values. That means that the BPM measure is successful. Not the same for *Classic* and *Voice*, but this is not an error: the selected excerpts of classic music and speech don’t have a clear tempo. Focusing on the Beatedness, high values are due to rhythmic music and low values are due to *Classic* music or *Voice*.

5.2.3 MFCC in rhythm domain

As described in Section 4.3.3, the Cepstrum of an input signal is defined as the inverse Fourier transform of the logarithm of the spectrum of the signal. The use of the Mel scale is justified as a mapping of the perceived frequency of a tone onto a linear scale, as a rough approximation to estimate the bandwidths of human auditory filters. As a result of the MFCC computation we obtain a compact representation of the spectrum of the input signal that can be isolated from the original pitch.

Our goal is to obtain a compact representation of the rhythm of the input signal independent of a specific BPM value. For that, we use the same algorithm that transforms a spectrum to a MFCC but using data in rhythm domain instead of the spectrum. We know we are applying the mel scale conversion in a completely outstanding context, but it is a clear example of the flexibility of data in rhythm domain. The resulting descriptors are a compact representation of the whole rhythm of the input audio without a clear relationship with the rhythmic parameters such as BPM, tempo or the presence of swing, and independent of the BPM value.

5.3 Listening Experiments

In the previous section, we have described a specific set of audio descriptors that can be used for genre classification. These proposed descriptors, in addition to the state of the art descriptors described in Section 4.3, are the key point in the construction of an automatic classifier. But, which are the relevant descriptors for genre classification? Maybe not all of them contribute to genre decisions and, if so, maybe they contribute at different levels.

In this section, we present a set of listening experiments specially developed to determine the importance of different musical facets on genre decisions. As described in Section 3.1.1, there are many works dealing on genre classification by humans and, as shown in Section 3.3.2, there are also many interesting works dealing with the classifiers and sets of descriptors that provide best performances in automatic classification.

The aim of the listening experiments here proposed is to establish the relationship between two musical facets of music used by humans (timbre and rhythm) and the classifier and features used in the automatic process. For that, we present a series of listening experiments where audio has been altered in order to preserve some properties of music (rhythm, timbre) but at the same time degrading other one. It was expected that genres with a characteristic timbre provide good classification results when users deal with rhythm modified audio excerpts, and vice-versa. We also want to study whether the different levels of distortion affect the classification o not.

5.3.1 The dataset

Our experiment uses music from 6 genres (Alternative, Classic, Electronic, Jazz, Pop, Rock) taken from the STOMP dataset proposed by Rentfrow & Gosling (2003). This dataset is made up of 14 musical genres according to musicological criteria (a set of experts were asked), commercial criteria (taxonomies in online music stores were consulted) and also the familiarity of participants with the proposed genres (see Section 3.3.1 for details). In our experiment, we discarded some of these genres to avoid possible confusions to participants due to several reasons (e.g.. Religious). Furthermore, we intentionally kept genres with widely accepted boundaries (classic, jazz, electronic) in addition to some other ones with more debatable limits (alternative, pop, rock).

5.3.2 Data preparation

We selected 5 seconds-long audio excerpts. According to the main goal of this work, some rhythm and timbre modifications were applied to the audio, in order to create excerpts where the timbre or rhythm information of the music was somehow degraded.

In one hand, rhythmic modifications were designed to preserve timbre avoiding the participant to extract any temporal information from the audio excerpt. This modification was based on the scrambling of the original audio. Short segments were randomly selected to create a new audio segment with the same length as the original. The length of the scrambled segments varied among 3 values: 125ms, 250ms and 500ms. It was expected that genres with a particular timbre provide good classification results when users exploited these audio excerpts. We also wanted to study whether the different levels of distortion affect the classification. On the other hand, timbre modifications were designed to preserve rhythm while avoiding the participant to easily extract any timbre information from the audio excerpt. This modification was based on the filtering of the input signal into frequency bands. The energy for each log-scale band was used to modulate gaussian noise centered in that specific frequency band. The energies were computed for each frame, then, this process was similar to basic vocoding. Three different filter-bank bandwidths were applied (3rd. Octave, 6th. Octave and 12th. Octave) to study the discrimination power affected by this parameter in the classification results. It was expected that genres with a particular rhythm provide good classification results when users exploited these audio excerpts.

In summary, we used excerpts from 6 different genres, sometimes *distorted* with either timbre or rhythm alterations, and in some cases *clean* (i.e., with no alteration). We had 3 levels for each modification (125ms, 250ms or 500ms for the length of the presented segments for rhythmic modification; 3rd. octave band, 6th. octave band or 12th. octave band for timbre modification). The task presented to the subjects was a dichotomic decision (yes/no) task where a genre label was presented in the screen, a 5 seconds excerpt was played and they had to decide whether it belongs to the genre which label was presented in the computer screen or not. In order to keep balanced the proportion of

		Timbre																		Ritme																		Res					
		1/3 octava						1/6 octava						1/12 octava						125ms						250ms						500ms											
Estudi:	Sona:	a	b	c	d	e	f	a	b	c	d	e	f	a	b	c	d	e	f	a	b	c	d	e	f	a	b	c	d	e	f	a	b	c	d	e	f	a	b	c	d	e	f
Alternative	Alternative	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1
	Classic		1						1						1						1						1						1						1				
	Electronic			1						1						1						1						1						1						1			
	Jazz				1						1						1						1						1						1						1		
	Pop					1						1						1						1						1						1						1	
	Rock						1						1						1						1						1						1						1
Classic	Alternative						1						1						1						1						1						1						1
	Classic	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1
	Electronic		1						1						1						1						1						1						1				
	Jazz			1						1						1						1						1						1						1			
	Pop				1						1						1						1						1						1						1		
	Rock					1						1						1						1						1						1						1	
Electronic	Alternative					1						1						1						1						1						1						1	
	Classic						1						1						1						1						1						1						1
	Electronic	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1
	Jazz		1						1						1						1						1						1						1				
	Pop			1						1						1						1						1						1						1			
	Rock				1						1						1						1						1						1						1		
Jazz	Alternative				1						1						1						1						1						1						1		
	Classic					1						1						1						1						1						1						1	
	Electronic						1						1						1						1						1						1						1
	Jazz	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1
	Pop		1						1						1						1						1						1						1				
	Rock			1						1						1						1						1						1						1			
Pop	Alternative			1						1						1						1						1						1						1			
	Classic				1						1						1						1						1						1						1		
	Electronic					1						1						1						1						1						1						1	
	Jazz						1						1						1						1						1						1						1
	Pop	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1
	Rock		1						1						1						1						1						1						1				
Rock	Alternative		1						1						1						1						1						1						1				
	Classic			1						1						1						1						1						1						1			
	Electronic				1						1						1						1						1						1						1		
	Jazz					1						1						1						1						1						1						1	
	Pop						1						1						1						1						1						1						1
	Rock	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1

Table 5.2: Details of the presented audio excerpts to the participants: The experiment was divided in 6 blocks (corresponding to 6 musical genres). A total of 70 audio excerpts were presented in each block. 35 excerpts belonged to the musical genre that defines the block). 15 excerpts had timbre distortion (splitted into 3 different levels), 15 excerpts had rhythmic distortion (splitted into 3 levels) and 5 excerpts without musical distortion

answers, half of the excerpts belonged to the targeted genre and a half of *fillers* was used, according to the schema depicted in Table 5.2, which provides an overview of the trials, events and blocks that were used.

The collected data was divided in two groups: First, two independent variables such as the type of distortion (timbre or rhythm) and the degree of distortion (one of the three described above). Second, two dependent variables such as the correctness of answers and the response time. This data will be analyzed in Section 5.3.5

5.3.3 Participants

In the experiment participated 42 music students from the High School of Music of Catalonia[1](ESMUC), 27 males and 15 females. The age of participants was between 18 and 43 years old (Mean=25.43; Standard Dev=5.64). All of them were students of the first two years in different specialities, as shown in Table 5.3.

[1] www.esmuc.cat

	# Students	%
Early Music	6	14.3%
Classical	25	59.5%
Jazz	11	26.2%

Table 5.3: Summary of the students that participate in the listening experiment

	Mean	Std. Dev.
Alternative	2.16	1.09
Classical	3.83	1.12
Electronic	2.20	0.93
Jazz	3.47	0.94
Pop	3.03	0.85
Rock	3.04	0.96

Table 5.4: Familiarization degree with musical genres for the participants in the listening experiments

Students spent a daily mean of 2.2 hours (Standard Dev=1,4) listening to music. The associated activities in this period of time are usually traveling or doing homework. Rehearsals and instrument training are excluded from these statistics.

Participants were also asked to define the familiarization degree from 1 (I've never heard about this kind of music) to 5 (I'm an expert in this type of music) for all the selected genres. Results are shown in Table 5.4.

We were also interested in detecting how participants classified their CD collection. The proposed options were Alphabetical Order (22.6%), Genre (47.2%), Chronologic (13.2%) and Other (17%) which includes "No Order", "Recently Bought" or "Favorites on the top". This test shows how important are genre labels in classification process for CD collections.

Finally, we proposed to repeat the experience proposed by Uitdenbogerd (2004). Participants were asked to caregorize music into exactly 7 categories. Results are shown in Figure 5.6. Genres only proposed once are not shown in this table (Ambient, Bossa-Nova, Songwriter, Flamenco, etc.). Proposed genres may be affected by those used in the test, as this question was asked after the instructions for the experiment had been presented. See Figure 5.6 for details.

5.3.4 Procedure

The experiment was carried out using the SuperLab 4.0 software. The instructions of the experiment informed about its goals and its general structure (one block per genre, expected binary responses "Yes" or "No" for all the audio excerpts, time information relevant but not crucial, etc.). Then, a training block was presented, divided into three parts: 1) Participants were asked to familiarize with the audio buttons used during the entire test. No distinction between left-handed and right-handed people was applied. 2) Participants were invited to listen to some audio excerpts for each one of the musical genre in order to

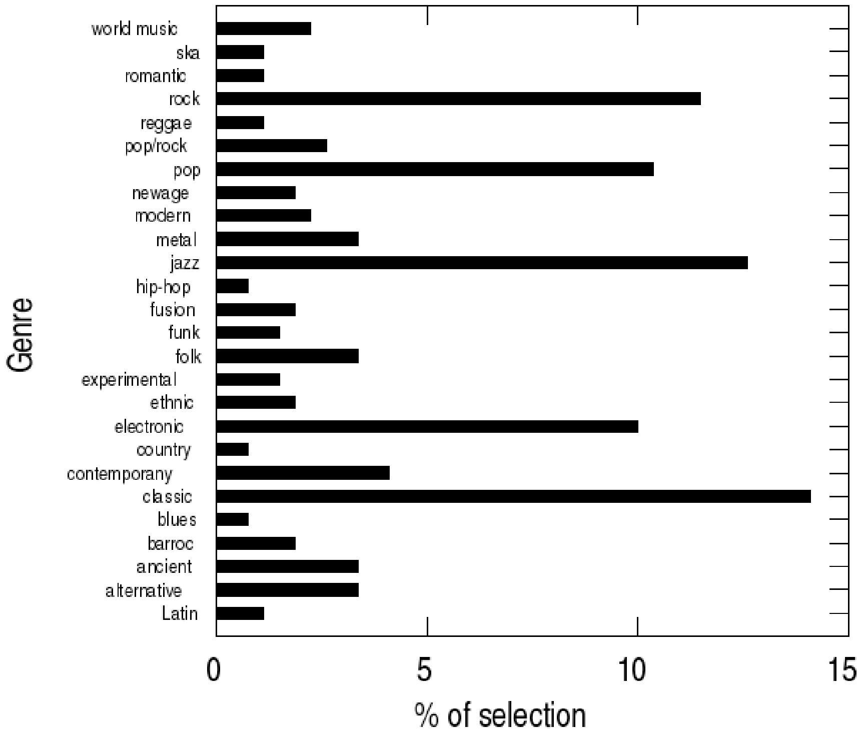


Figure 5.6: Percentage of proposed musical genres by participants in the listening experiment when they were asked to categorize music into exactly 7 categories, as proposed by Uitdenbogerd (2004)

adjust the genre boundaries. 3) Participants were invited to listen to some rhythmic and timbre modifications of the original excerpts in order to familiarize with the modifications used. At this point, the experiment began and the first block (Alternative) started until the last one (Rock) finished. Participants could rest for a short period between blocks. The presentation of different audio excerpts inside a block was randomized according to the Table 5.2. The overall required time for completing the experiment was about 30 minutes.

5.3.5 Results

General overview

Figure 5.7 shows the percentage of correct classified instances for different genres. The figure on the left shows results for rhythm modifications (preserving timbre) and the figure on the right shows results for timbre modifications (preserving rhythm). Figure 5.8 shows the corresponding averaged response times for the same conditions. All numerical results are shown in Table 5.5

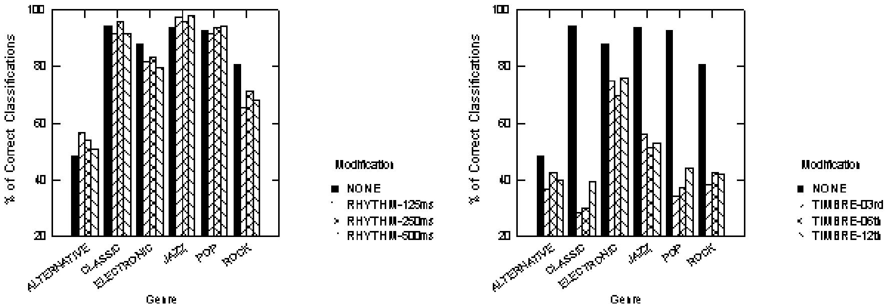


Figure 5.7: Percentage of correct classified instances for different genres. The figure on the left shows results for rhythm modifications (preserving timbre) and the figure on the right shows results for timbre modifications (preserving rhythm)

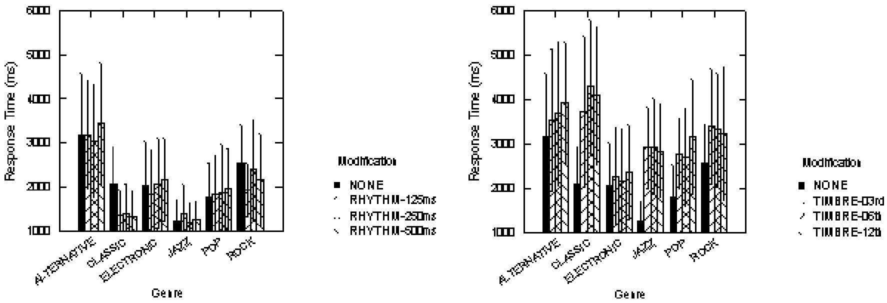


Figure 5.8: Averaged response times for different genres. The figure on the left shows averaged times for rhythm modifications (preserving timbre) and the figure on the right shows averaged times for timbre modifications (preserving rhythm)

			Timbre			Rhythm		Orig.
		1/3rd	1/6th	1/12th	125ms	250ms	500ms	
Alternative	Hits	1.85	2.125	2.0	2.825	2.7	2.55	2.425
	Resp. Time	3522.4	3682.3	3918.1	3177.2	3049.7	3458.6	3161.0
Classical	Hits	1.425	1.5	1.975	4.575	4.8	4.6	4.725
	Resp. Time	3722.1	4284.6	4083.9	1362.3	1385.7	1328.8	2073.0
Electronica	Hits	3.75	3.5	3.8	4.1	4.175	3.975	4.4
	Resp. Time	2270.0	2156.1	2349.5	1822.5	2077.4	2176.7	2050.6
Jazz	Hits	2.8	2.575	2.65	4.875	4.8	4.9	4.7
	Resp. Time	2915.3	2930.7	2838.3	1404.6	1206.5	1253.4	1242.8
Pop	Hits	1.725	1.875	2.2	4.575	4.7	4.725	4.65
	Resp. Time	2773.2	2684.5	3164.4	1851.4	1881.6	1958.5	1780.4
Rock	Hits	1.925	2.125	2.1	3.275	3.575	3.4	4.05
	Resp. Time	3394.3	3311.3	3228.2	1936.1	2411.1	2167.7	2553.5

Table 5.5: Numerical results for listening experiments. Response Time is expressed in ms and the hits can vary from 0 to 5.

	# Hits				Resp. Time			
	Modification		Degree		Modification		Degree	
	F	p	F	p	F	p	F	p
Alternative	6,567	0,014	0,508	0,604	4,27	0,053	4,831	0,01
Classic	169,319	0,000	4,115	0,020	45,35	0,000	2,216	0,13
Electronic	5,124	0,029	0,451	0,639	4,53	0,040	3,067	0,05
Jazz	121,268	0,000	1,450	0,241	190,42	0,000	0,506	0,61
Pop	156,465	0,000	4,166	0,019	37,81	0,000	1,028	0,37
Rock	21,644	0,000	3,243	0,044	28,05	0,000	1,6	0,21

Table 5.6: Results for the ANOVA tests for distortion analysis. The ANOVA results are presented for the analysis on the number of hits and the response time. For each case, we specify results (1) for the type of distortion and (2) for the degree of the distortion

Observing genres individually, alternative music shows similar results for all kind of distortions. The number of correct classifications is a slightly higher for rhythm distortion as well as the response time is slightly lower, but no clear conclusions can be extracted. Classical and jazz music show good classification results with low response times for rhythmic distortion, but the opposite for timbre distortion. The conclusion is that these two musical genres are clearly defined by particular timbres.

In contrast, electronic music is the only one that presents good classification results and low response times with timbre distortion. This musical genre is equally defined by rhythm and timbre due to results with two distortions are similar. Pop music also presents better results for timbre identification and, finally, rock music is in between pop and alternative. The conclusion is that, according to the selected taxonomy, alternative music has no clear difference in rhythm or timbre with pop and rock. Maybe "alternative" music is an artificial genre without musical fundament, or maybe the difference lies in another musical component like the harmony or the lyrics.

Analysis of Variance

One-way Analysis of variance (ANOVA) is used to test the null-hypothesis within each genre block, assuming that sampled population is normally distributed. ANOVA is computed on both response time and # of hits, taking into account only the correct answers. We will not discriminate between two tests due to results are comparable.

- First, we test whether the distortion degree for both modifications had real influence in classification results. The null-hypothesis is defined as follows:

  H_0 = *Presented distortions, in one genre, do not influence classification results*

  Results are shown in Table 5.6. Concerning the modification analysis on the # of hits and response time, we have to reject the null-hypothesis,

	# Hits		Resp. Time	
	F	p	F	p
1/3rd. Octave	20,72	0,000	4,723	0,001
1/6th. Octave	11,41	0,000	7,023	0,008
1/12th. Octave	13,36	0,000	6,585	0,008
125ms	33,39	0,000	27,903	0,000
250ms	38,52	0,000	18,551	0,000
500ms	42,24	0,000	32,371	0,000

Table 5.7: Results for the ANOVA tests for overall classification, independent from the musical genre

that is, different modifications do affect classification results. In contrast, the distortion degrees provide some significance to the experiment: the distortion degree show some confidence for Alternative, Electronic and Jazz music, that is, we have to accept the null hypothesis and distortion degrees do not affect the classification results. The distortion degree for the response time show some confidence Classic, Electronic, Jazz, Pop and Rock. Comparing these two cases, we realize that behaviors are not the same but similar, then, the null-hypothesis can not be rejected. As a conclusion, the applied distortion affects the classification results while the distortion degree doesn't.

- Now, we test whether the distortions have the same influence in different genres. The null-hypothesis is defined as:

  $$H_0 = \textit{Presented genres are equally affected for each specific distortion}$$

  Results in Table 5.7 show how the null-hypothesis can be rejected with a high level of confidence, thus, proposed distortions affect genres in different ways.

Overall classification

Finally, results for overall classification independent of the genre are shown in Figure 5.9. It shows the results of classification for all genres as a function of the presented distortion when the presented audio excerpt is that which belongs to the block (figure on the left) and the presented audio excerpt is that which does not belong to the block (presentation of "fillers", figure on the right). Black columns correspond to the number of hits and dashed columns correspond to the response time. Roughly speaking, rhythm modifications (preserving timbre) provide better classification results and lower response times than timbre modifications (preserving rhythm). Furthermore, it is easier to recognize when a given audio excerpt does not belong to a musical genre than when it belongs.

Collecting all the information provided above, we can conclude that, according to the configuration of this experiment, the easiest musical genre classification for humans is to detect when a specific timbre does not belong to classical music, and the more difficult is to detect whether a given rhythm belongs to (again) classical music.

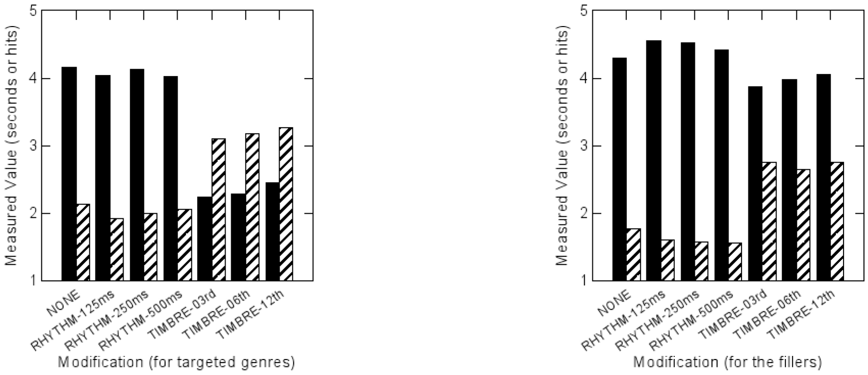


Figure 5.9: Results of classification for all genres as a function of the presented distortion when the presented audio excerpt is that which belongs to the block (figure on the left) and the presented audio excerpt is that which does not belong to the block (presentation of "fillers", figure on the right). Black columns correspond to the number of hits and dashed columns correspond to the response time.

5.3.6 Comparison with automatic classifiers

In this section, we study the behavior of an automatic classification system for musical genre in similar conditions that the listening experiments have been performed. This study is not focused on the performance of the classifier itself but on the differences on results provided by humans and machines.

Datasets

Two different datasets have been used for these experiments. First, we use the Magnatune dataset (see Section 3.3.1 for details), which is used to verify that descriptors and classification schemes we propose are not so far than those used in the Genre Classification contest organized in the context of the International Symposium on Music Information Retrieval - ISMIR 2004 (see Section 3.3.3 for details). Note how the Magnatune dataset has completely unbalanced categories that will affect the overall classification performance. Second, the STOMP dataset is used to compare results with the Listening Experiments described above (see Section 3.3.1 for details).

Descriptors

Audio descriptors proposed for automatic classification are divided in two main groups, as in the listening experiments: timbre and rhythm.

Timbre: Our timbre description is defined by a compact set of 39 descriptors which include: Zero Crossing Rate (1); Spectral Centroid (1), Spectral Flatness (1), MFCC (12), derivative (12) and acceleration (12) . Means and variances of these descriptors are computed for the whole song.

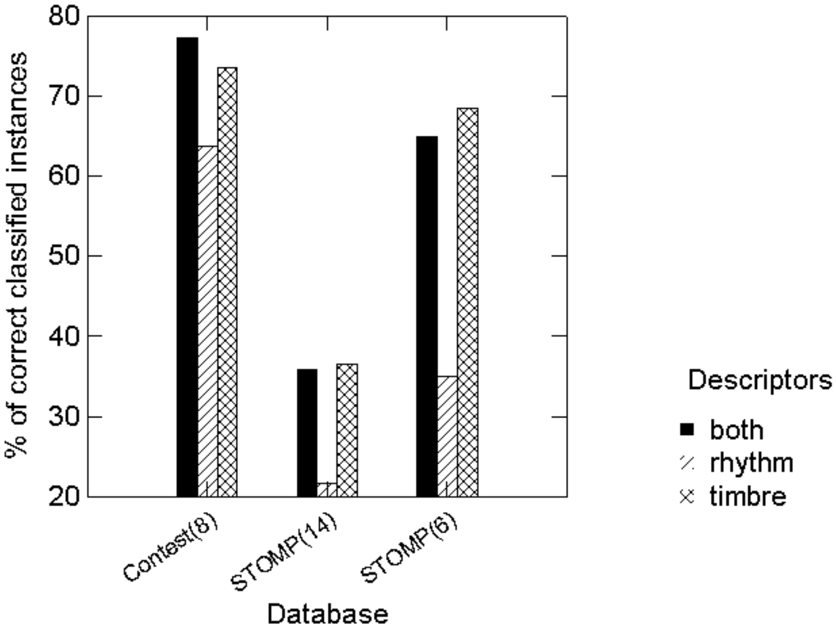


Figure 5.10: Results for the automatic classification experiments

Rhythm: The rhythm description is also defined by a compact set of 39 descriptors which include: Zero Crossing Rate (1), Spectral Centroid (1), Spectral Flatness (1), MFCC (12), derivative (12) and acceleration of data in Rhythm domain. Means and variances of these descriptors are computed for the whole song.

Classification

All the classification experiments have been made using WEKA[2]. After some initial tests using Support Vector Machines (SMO), Naive Bayes (IBk), Nearest Neighbours (kNN) and Decision Trees (J48), the classification algorithm finally used is Support Vector Machine with a polynomial kernel with the exponential parameter set to 2. We applied CFS feature selection to avoid over-fitting. The evaluation has been performed using 10 fold cross-validation.

Results

Results for this experiment are shown in Figure 5.10, According to the musical aspects discussed above, the used descriptors are grouped in timbre, rhythm and both. Results are shown independently for these three configurations. Although different train-set and test-set were provided for the Magnatune dataset, 10-fold cross validation method has been used for all the cases, then, results are more comparable.

[2] http://www.cs.waikato.ac.nz/ml/weka

classified as →	a	b	c	d	e	f	g	h
classical (a)	311	0	0	0	0	0	0	9
electronic (b)	3	87	0	0	0	0	15	10
jazz (c)	2	0	12	0	0	0	6	6
metal (d)	0	2	0	6	0	0	21	0
pop (e)	0	1	0	0	0	0	5	0
punk (f)	0	0	0	0	0	7	9	0
rock (g)	8	12	0	2	0	0	67	6
world (h)	28	11	0	0	0	0	10	73

Table 5.8: Confusion matrix for classification on the Magnatune database using both timbre and rhythm descriptors

Results for Magnatune dataset show accuracies up to 80% in classification using both timbre and rhythm descriptors. Roughly speaking, these results are comparable to results obtained by Pampalk in the MIREX contest in 2004. Timbre related descriptors provide more classification accuracy than rhythm ones with differences about 15% but the inclusion of rhythm information improves results in all cases. Applying the same classification conditions to the STOMP dataset, accuracies decrease because of the high number of musical genres in the taxonomy (14) and the size of the database (10 songs/genre). But the pattern of results depending on the timbre and rhythm facets is similar to that obtained for the Magnatune dataset. Finally, accuracies near 70% are obtained with the reduced version of STOMP, which is the dataset used for the listening experiments. Here again, the contribution of timbre is more important than rhythm in the classification process. Timbre classification can yield better results than those obtained with both timbre and rhythm descriptors.

Tables 5.8, 5.9 and 5.10 show confusion matrices for classification results of three datasets using both timbre and rhythmic descriptors. Note how classical music is correctly classified, pop and rock have some kind of confusion between them, and jazz, electronic and alternative music are worse classified.

5.3.7 Conclusions

Assuming that both experiments are not identical, results are quite similar in such a way that timbre features provide better accuracies than rhythm features. Even so, genres like electronica require some rhythmic information for better results. As discussed above, for humans it is easier to identify music that does not belong to a given genre than to recognize wether an audio excerpt belongs to a specific genre (see Figure 5.9). These results suggest that automatic classification could be based on expert systems for specific genres instead of global systems. On the other hand, listening experiments show how the selected taxonomy also affects directly to classification results: confusions between alternative and rock music appear as well as for automatic classifier when different subgroups of taxonomies provide different results in genre classification.

For the listening experiments, the two proposed distortions provide differences in the classification results depending on the musical genre. Results show

classified as →	a	b	c	d	e	f	g	h	i	j	k	l	m	n
alternative (a)	5	0	0	1	0	0	0	0	0	1	0	1	2	0
blues (b)	1	1	0	0	0	4	0	0	0	1	0	0	0	2
classical (c)	0	1	8	0	0	0	0	0	0	0	0	0	0	0
country (d)	2	1	0	2	0	2	1	0	0	0	1	0	1	0
electronica (e)	1	1	0	0	3	0	2	1	1	0	0	1	0	0
folk (f)	0	5	0	2	0	1	0	0	0	0	0	0	0	2
funk (g)	0	0	0	0	1	1	4	1	0	0	0	2	1	0
heavymetal (h)	2	0	0	0	0	0	0	4	2	0	0	0	2	0
hip-hop (i)	1	0	0	1	0	0	0	0	5	0	0	1	0	2
jazz (j)	1	2	0	0	0	2	0	0	1	4	0	0	0	0
pop (k)	0	0	0	0	0	0	1	1	1	0	3	1	0	2
religious (l)	1	1	0	2	0	0	0	0	0	0	1	3	0	0
rock (m)	3	0	0	1	0	0	1	1	0	0	0	1	2	0
soul (n)	0	2	0	1	0	1	0	0	3	0	0	0	0	3

Table 5.9: Confusion matrix for classification on the STOMP database using both timbre and rhythm descriptors

classified as →	a	b	c	d	e	f
alternative (a)	6	0	0	1	1	2
classical (b)	0	9	0	0	0	0
electronica (c)	1	0	6	2	1	0
jazz (d)	1	1	1	6	1	0
pop (e)	1	0	1	0	5	2
rock (f)	1	0	1	0	2	5

Table 5.10: Confusion matrix for classification on the reduced STOMP database using both timbre and rhythm descriptors used in the listening experiments

how the distortion degree has not a direct relationship with the obtained accuracies. The response time for non distorted audio excerpts can be a measure of how assimilated and musically defined are the musical genres. Alternative music provides response times higher than 3 seconds while classic or jazz music provide response times between 1.5 and 2 seconds. Maybe “Alternative” label was created under some commercial criteria while the jazz music can be defined exclusively by musical properties.

Finally, results of this listening experiment need to be extended to non musician participants. The inclusion of other facets of music like harmony or tonal information as well as other high level semantic descriptors could be crucial for a full characterization of musical genre discrimination in humans.

5.4 MIREX 2007

5.4.1 Introduction

As explained in Section 3.3.3, in the context of the 8th. International Conference on Music Information Retrieval (ISMIR 2007), the organizers of the conference and the IMIRSEL lab organized the MIREX07 competition. The goal of our submission was to compare our work with other state of the art genre classification algorithms. For that, we submitted a classifier that was supposed to be the baseline for further developments. This classifier was designed to deal with different taxonomies, to be fast enough for real applications and to work independently from other external software. The algorithm was built as a C++ library and it was tested in different environments.

This section shows the details of the classifier, the results of our previous tests in our datasets and the results in the MIREX competition.

5.4.2 Description

The algorithm has been developed as a set of C++ classes. The final implementation uses three well known libraries: libsndfile[3] for i/o of audio files, FFTW[4] for FFT computations and libSVM[5] for Support Vector Machine train and test processes. Three bash scripts have been developed to provide compatibility with MIREX specifications[6].

5.4.3 Features

As mentioned in Section 5.4.1 a set of tests using different audio descriptors have been performed. Having a look to the preliminary results shown in Table 5.11, we observe that timbre related features provide best results in different environments, followed by rhythmic descriptors. Other descriptors related to musical facets (melody, tonality, tempo, etc.) seem to provide worse accuracies in the presented datasets (this part will be discussed in detail in Section 5.5). Although the accuracies obtained by the rhythm features are about 5..10% lower than those obtained with timbre features, the combination of both descriptor sets increases about 1 to 4% points the overall performance.

Timbre descriptors

The timbre descriptors we mainly use are the MFCC. According to Logan (2000), they have proven to be quite robust in automatic classification. Specifically, we use a set of timbre descriptors comprising: 12 $MFCC$, 12 $\Delta MFCC$, 12 $\Delta^2 MFCC$, Spectral Centroid, Spectral Flatness, Spectral Flux and Zero Crossing Rate. The frame size we use is 92.9ms and 50% overlap. For each audio excerpt we compute basic statistics (mean, variance, skewness and kurtosis) for the used descriptors.

[3] http://www.mega-nerd.com/libsndfile/

[4] http://www.fftw.org/

[5] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

[6] http://www.music-ir.org/mirex2007/index.php/Audio_Genre_Classification

Rhythm descriptors

The rhythmic description used in the experiments is based on the *Rhythm Transformation* described in Section 5.2.1. Although many successful approaches on rhythmic description can be found in literature, this algorithm has proved to be a good and compact representation of rhythm, even for signals such as speech, or for some excerpts of classical music where rhythm is not present at all. The frame size we use is $92.9ms$ and 50% overlap, $1/3rd$ filterbanks and a $3s$ window size to compute rhythm. Here again, we compute basic statistics (mean, variance, skewness and kurtosis) for the used descriptors.

5.4.4 Previous evaluation: two datasets

In order to build a classifier that is capable of dealing with different music collections, our algorithm has been tested on two datasets: Radio and Tzanetakis (see Section 3.3.1 for details).

For the experiments, 4 different classifiers have been used: Nearest Neighbours, Support Vector Machines, AdaBoost and RandomForest. All the proposed tests have been made on Weka. Results are computed using a 10-fold cross-validation. k-NN and SVM are probably the most popular classifiers in music description problems. We fixed k=2 for k-NN experiments. We used a SVM with a polynomial kernel and after performing grid search we set the exponent to "2". For the RanfomForest experiments, the number of features randomly selected for each tree was set to 1, and the number of trees was set to 10. Experiments done with ADABoost used the default parameters proposed by Weka.

The most representative classification algorithms we have tested are compared in Table 5.11, using different descriptors and datasets. The best results are obtained using Timbre and Rhythm features and a Support Vector Machine with exp=2 classifier. This is the approach we have implemented for the MIREX. Concrete results for these experiments are shown in Table 5.12 and Table5.13 for Radio and Tzanetakis datasets, respectively.

Dataset	Descriptors	IB1	SVM1	SVM2	AdaBoost	Random Forest
Radio	Timbre	63.342%	80.299%	81.296%	71.820%	75.062%
Radio	Rhythm	53.117%	59.850%	62.594%	58.354%	56.608%
Radio	Timbre + Rhythm	69.576%	82.294%	83.791%	77.057%	74.564%
Tzanetakis	Timbre	80.578%	90.030%	90.030%	83.484%	37.361%
Tzanetakis	Rhythm	45.619%	52.467%	60.020%	57.905%	57.301%
Tzanetakis	Timbre + Rhythm	84.390%	91.239%	90.533%	83.685%	80.765%

Table 5.11: Results for preliminary experiments on genre classification for 2 datasets (Radio and Tzanetakis) and 2 sets of descriptors (Timbre and Rhythm) using 4 classification techniques. Accuracies are obtained using 10-fold cross validation

	a	b	c	d	e	f	g	h
Classic(a)	47	0	0	2	0	1	0	0
Dance(b)	0	45	1	0	1	0	3	0
Hip-Hop(c)	0	0	43	0	1	4	1	0
Jazz(d)	2	1	1	39	0	7	1	0
Pop(e)	0	1	2	0	36	10	2	0
R'n'B(f)	1	1	7	4	8	28	1	0
Rock(g)	0	1	0	1	6	0	42	0
Speech(h)	0	0	0	0	0	0	0	50

Table 5.12: Results for preliminary experiments using timbre and rhythmic descriptors and SVM (exp=2) for the Radio dataset

5.4.5 Results

The results we obtained in the MIREX evaluation are the following: Average for Hierarchical Classification Accuracy: 71.87% (best obtained accuracy: 76.56%). Average Raw Classification Accuracy: 62.89% (best obtained accuracy: 68.29%). Runtime for feature extraction: 22740s (fastest submission: 6879s). The results for all the participants are shown in Section 3.3.3, in Table 3.25 and Table 3.26, and summarized here in Table 5.14.

The detailed analysis of the results obtained by our implementation provide a confusion matrix as shown in Figure 5.11, which is numerically detailed in Table 5.15. The most relevant confusions in our implementation are (in descending order): (1) Baroque, Classical and Romantic, (2) Blues and Jazz, (3) Rock'n'Roll and Country and (4) Dance and Rap-HipHop. All these confusions are musically coherent with the selected taxonomy, which is not the same taxonomy used in our previous experiments. Results are quite close to the best submission using this basic approach (less than 5% below). The Rap-HipHop category is our best classified genre (as in the other approaches) while Rock'n'Roll is our worst classified genre (as in other 3 approaches, but others show worse results for Dance, Romantic, Classical).

	a	b	c	d	e	f	g	h	i	j
Blues(a)	98	1	0	0	0	0	0	0	0	1
Classical(b)	0	92	0	0	0	0	0	0	0	1
Country(c)	0	0	99	1	0	0	0	0	0	0
Disco(d)	0	1	3	90	6	0	0	0	0	0
Hip-Hop(e)	1	0	2	12	81	0	0	2	1	1
Jazz(f)	0	0	0	0	0	98	0	1	1	0
Metal(g)	0	0	0	0	0	0	94	0	2	4
Pop(h)	0	0	0	0	1	3	0	85	5	6
Reggae(i)	1	0	0	0	0	2	2	8	83	4
Rock(j)	1	0	0	0	0	2	3	3	5	86

Table 5.13: Results for preliminary experiments using timbre and rhythmic descriptors and SVM (exp=2) for the Tzanetakis dataset

Participant	Hierarchical	Raw	Runtime (sec)	Folds
IMIRSEL(1)	76.56%	68.29%	6879	51
Lidy	75.57%	66.71%	54192	147
Mandel(1)	75.03%	66.60%	8166	207
Tzanetakis	74.15%	65.34%	—	1442
Mandel(2)	73.57%	65.50%	8018	210
Guaus	71.87%	62.89%	22740	194
IMIRSEL(2)	64.83%	54.87%	6879	1245

Table 5.14: Summary of the results for the MIREX 2007 Audio Genre Classification tasks for all submissions

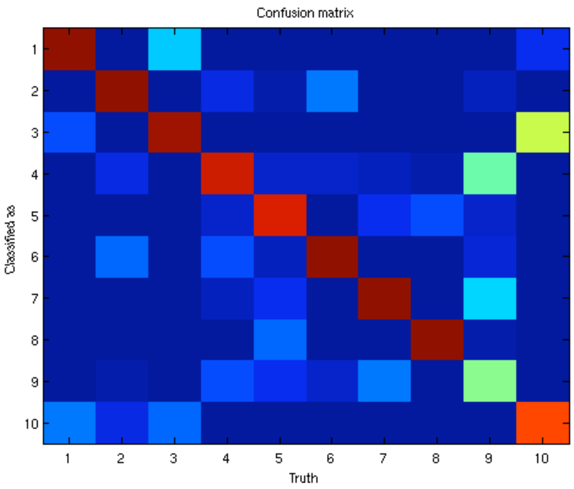


Figure 5.11: Confusion matrix of the classification results: 1:Baroque, 2:Blues, 3:Classical, 4:Country, 5:Dance, 6:Jazz, 7:Metal, 8:Rap-HipHop, 9:Rock'n'Roll, 10:Romantic

	a	b	c	d	e	f	g	h	i	j
a	64.29%	0.57%	15.00%	0.43%	0.43%	1.43%	0.00%	0.00%	0.29%	15.14%
b	2.00%	71.00%	0.29%	4.00%	1.14%	12.14%	0.29%	1.00%	2.43%	1.00%
c	15.43%	0.14%	57.43%	0.00%	0.14%	1.14%	0.14%	0.00%	0.14%	24.57%
d	0.86%	5.43%	0.57%	61.29%	4.14%	6.71%	3.43%	2.29%	23.43%	0.57%
e	0.29%	1.57%	0.14%	3.29%	65.14%	3.57%	6.14%	11.57%	4.14%	0.43%
f	1.86%	14.86%	1.43%	9.29%	2.43%	67.14%	1.29%	1.00%	4.71%	2.00%
g	0.14%	0.57%	0.00%	2.29%	7.29%	0.57%	66.57%	1.14%	23.57%	0.00%
h	0.14%	1.00%	0.00%	1.57%	12.00%	1.29%	1.57%	81.86%	2.43%	0.00%
i	1.14%	2.57%	0.00%	17.71%	5.86%	4.71%	20.00%	1.14%	38.43%	0.57%
j	13.86%	2.29%	25.14%	0.14%	1.43%	1.29%	0.57%	0.00%	0.43%	55.71%

Table 5.15: Numerical results for our MIREX submission. Genres are: (a) Baroque, (b) Blues, (c) Classical, (d) Country, (e) Dance, (f) Jazz, (g) Metal, (h) Rap/Hip-Hop, (i) Rock'n'Roll, (j) Romantic.

	Forming moods
Cluster 1	passionate, rousing, confident, boisterous, rowdy
Cluster 2	rollicking, cheerful, fun, sweet, amiable/good natured
Cluster 3	literate, poignant, wistful,bittersweet, autumnal, brooding
Cluster 4	humorous, silly, campy, quirky, whimsical, witty, wry
Cluster 5	aggressive, fiery, tense/anxious, intense, volatile, visceral

Table 5.16: clustered mood tags selected for the contest

Participant	Accuracy
IMIRSEL M2K knn	47.17%
IMIRSEL M2K svm	55.83%
Cyril Laurier, Perfecto Herrera	60.50%
Kyogu Lee 1	49.83%
Kyogu Lee 2	25.67%
Lidy, Rauber, Pertusa, Inesta	59.67%
Michael Mandel, Dan Ellis	57.83%
Michael Mandel, Dan Ellis spectral	55.83%
George Tzanetakis	61.50%

Table 5.17: Obtained accuracies of all the participants in the Mood Classification contest.

5.4.6 Cross experiment with Audio Mood Classification task

Mood classification for musical signals is another active topic in MIR. Although it is completely out of the scope of this thesis, we performed an experiment on mood classification using our submitted algorithm for genre classification, and compare the results with the submitted algorithm by Laurier & Herrera (2007) in the contest[7]. The idea of comparing results of different classification tasks was presented at the ISMIR conference by Hu & Downie (2007), who compared mood with genre, artist and recommendation algorithms using metadata available in the web. On the other hand, many authors proposed the same algorithm for different tasks (Tzanetakis, 2007; Mandel & Ellis, 2007) and results are not far from the best approaches.

In this context, we compare two algorithms that are conceptually similar (SVM) but using different statistics extracted from the descriptors and different parameters of the classifier. We want to thank Andreas Ehmann, from the University of Illinois, who spend some time running these experiments and collecting results.

Table 5.16 shows the clustered mood tags selected for the contest. Table 5.17 shows the obtained accuracies for all the participants. Best performances are achieved by Tzanetakis, Laurier and Lidy with differences lower than 2%. The results after the cross experiment are the following: mood classification obtained by genre classifier presents an accuracy of 50.8% after 3-fold cross validation. The confusion matrix is shown in Table 5.18. The overall accuracy

[7] http://www.music-ir.org/mirex/2007/index.php/Audio_Music_Mood_Classification

Mood	A	B	C	D	E
A	31.67%	10.83%	4.17%	10.00%	20.00%
B	15.00%	34.17%	17.50%	20.83%	3.33%
C	9.17%	14.17%	72.50%	10.83%	6.67%
D	17.50%	35.83%	4.17%	55.00%	9.17%
E	26.67%	5.00%	1.67%	3.33%	60.83%

Table 5.18: Confusion matrix of mood classification using genre classifier

is not so far from the best approaches in the contest (about 55..60%) and the confusion matrix has a clear diagonal.

This experiment shows how different problems studied by the MIR community can be afforded using similar techniques. We wonder if it is conceptually coherent to use the same techniques for those problems that humans doesn't. We will discuss about this at the end of this chapter.

5.4.7 Conclusions

After these results, we can support the idea of the presence of a glass-ceiling in the traditional automatic genre classification systems, idea that was introduced by Aucouturier & Pachet (2004). The complexity of the algorithms proposed in this contest vary from the most generic to the most specialized ones but results differ only in 5%. As shown by our previous experiments and the accuracies obtained in the contest, results have a high dependence on the selected taxonomy and dataset. This means that the most specialized algorithms will fail in general aplications and the final user will not obtain as good results if new musical genres or songs are included. The classifier we propose, with acceptable accuracies, is compact and easily configurable. This allows to be implemented in real environments of music reccomendation and classification.

5.5 The sandbox of Genre Classification

In this section, we present a set of experiments that are the starting point for further developments. We compare different descriptors, datasets and classifiers in order to evaluate their importance in genre classification. The main idea is to test many combinations of descriptors with different classifiers. It is our goal to check how the selection of datasets also affects the classification process too. For that, we start comparing the obtained accuracies, using different state of the art classifiers, for a set of frame based descriptors versus a set of segment-based descriptors. Then, we include some non traditional descriptors in the experiment and, we combine them to evaluate which are the most relevant ones. Finally, we will combine different datasets and discuss about the reliability of the presented classifiers. The goal of this test is to analyze the behavior of the classifier in front of new unknown data, which is near to possible real applications.

	time (ms)
Alternative	3161.0
Classical	2073.0
Electronica	2050.6
Jazz	1242.8
Pop	1780.4
Rock	2553.5
Mean	*2143.5*

Table 5.19: Overview of time responses for the listening experiments when subjects are presented unprocessed audio excerpts

5.5.1 Frame based vs segment based classification

We start discussing whether the classification should be performed using frame based descriptors (whose lengths can vary from few milliseconds up to some seconds, depending on the descriptors) or using statistics computed over longer time audio excerpts. As discussed previously in Section 3.2.5, there exist many different techniques to collapse the frame-based information. For simplicity, we will use basic statistics computed over the whole audio excerpt (mean, variance, skewness and kurtosis). According to Perrot & Gjerdigen (1999), humans perform accurate genre classification based on $250ms$ of audio. This suggests that automatic classifications should not need other high level structures to perform this task (Martin et al., 1998). In opposition of that, our listening experiments show how, for unprocessed audio excerpts, the mean time response for genre classification is $2.143[s]$ from the presentation of the audio excerpt to the response action to the computer (See Table 5.19 for details).

Then, how the genre classification should be performed? Should we use frame-based classifiers or compute other compact representations representing longer audio excerpts?

Description

For this experiment, we use 3 different datasets described previously in Section 3.3.1: STOMP, Radio and Tzanetakis. We use these three datasets to ensure that the results can be generalized to any real environment because of their properties: STOMP database is built using a large number of classes (14 musical genres) with few examples for each one (10 full songs) while Radio and Tzanetakis datasets are built using a smaller number of genres (8 or 10 musical genres respectively) and more audio examples for each one (50 and 100 respectively).

We will also use different families of descriptors in order to separate the different musical facets in our analysis. Here is a short description of them:

MFCC: This set of descriptors uses 12 MFCC coefficient and their derivatives, as described in Section 4.3.3. It is the most used and compact timbre description of audio. We use a $frame = 2048$, $hopsize = 1024$, $sr = 22050Hz$ from an audio excerpt with $length = 60sec$ centered at the middle of the song.

Spectral: This set of descriptors includes the MFCC coefficients and their derivatives, as in the previous configuration, but also many other spectral related features such as Zero-Crossing rate, Spectral Flatness, Spectral Centroid, Spectral Flux and Spectral Roll-off. We use a $frame = 2048$, $hopsize = 1024$, $sr = 22050Hz$ and $length = 60sec$ of audio centered at the middle of the song.

Rhythm: This set of descriptors include the same features described in the spectral set, but computed over the data in the rhythm domain instead of the traditional spectrum. See Section 5.2.1 for a detailed description of the rhythm transformation. We use a $frame = 2048$, $hopsize = 1024$, $sr = 22050Hz$ and $length = 60sec$ of audio centered at the middle of the song, $1/3rd$ octave filter bands and a sliding window of $3sec$.

Tonality: For this set of descriptors, we have used the THPCP tonal coefficients proposed by Gomez (2006) and described in Section 4.3.5. We use a $frame = 2048$, $hopsize = 1024$, $sr = 44100Hz$ and $length = 60sec$ of audio centered at the middle of the song.

Panning: For this set of descriptors, we have used the Panning coefficients proposed by Gómez et al. (2008) and described in Section 4.3.6. We use a $frame = 2048$, $hopsize = 1024$, $sr = 22050Hz$ and $length = 60sec$ of audio centered at the middle of the song, only 1 frequency band and a sliding window of $2sec$.

Complexity: The complexity descriptors used in this set are those proposed by Streich (2007) and described in Section 4.3.7. The panning related complexity could not be computed for the Tzanetakis dataset because it is build up with mono files. Furthermore, the complexity descriptors, as presented by Streich, can not be computed frame by frame, so we should not compare the behavior of these descriptors. The details for configuration are also explained in Section 4.3.7 and we can not modify them. This descriptor can not be computed frame to frame.

BLI: The Band Loudness Intercorrelation descriptors were developed at the Music Technology Group, inspired by the work of McAuley et al. (2005), and described in Section 4.3.8. These descriptors propose an alternative point for timbre description of audio based on the codification of covariation of information at the output of the auditory filters. We use a $frame = 2048$, $hopsize = 1024$, $sr = 22050Hz$ and $length = 60sec$ of audio centered at the middle of the song, 24 frequency bands and a $8KHz$ as the maximum analysis frequency This descriptor can not be computed frame to frame.

All the descriptors have been computed on a $60sec$ audio excerpt centered at the middle of the song to avoid introductions, *fade-in* and *fade-out* effects, applause, etc. Finally, we have used different classifiers to make results independent of classification techniques: Nearest Neighbours, Support Vector Machines, Ada-boost and Random forests. All these classifications have been done using Weka[8] using the default parameters for each classifier except for

[8] http://www.cs.waikato.ac.nz/ml/weka

Dataset	Descr	IB1	SVM1	SVM2	AdaBoost	R.Forest
STOMP	MFCC	49.6%	28.3%	31.0%	6.4%	42.6%
	Spectral	48.0%	36.6%	39.3%	25.0%	34.1%
	Rhythm	53.0%	26.6%	29.0%	35.9%	59.4%
	Tonality	15.8%	15.7%	17.3%	11.7%	14.8%
	Panning	96.8%	8.9%	16.6%	56.3%	94.5%
	Complexity	—	—	—	—	—
	BLI	—	—	—	—	—
Radio	MFCC	58.0%	42.0%	44.9%	43.7%	55.3%
	Spectral	50.8%	55.2%	60.7%	45.1%	54.5%
	Rhythm	57.2%	49.3%	49.7%	50.0%	61.5%
	Tonality	22.8%	25.0%	25.0%	22.5%	25.2%
	Panning	95.9%	25.0%	35.9%	70.3%	92.9%
	Complexity	—	—	—	—	—
	BLI	—	—	—	—	—
Tzanetakis	MFCC	51.7%	35.6%	37.1%	35.6%	42.3%
	Spectral	44.0%	43.8%	45.5%	35.9%	42.5%
	Rhythm	46.9%	35.1%	35.8%	37.2%	51.5%
	Tonality	17.9%	17.8%	16.7%	16.8%	18.2%
	Panning	—	—	—	—	—
	Complexity	—	—	—	—	—
	BLI	—	—	—	—	—

Table 5.20: Classification results for frame based descriptors using different descriptors, classifiers and datasets

the support vector machine that we have included another test using a polynomial kernel with the exponential parameter set to 2. All these classifiers are described in Section 4.4. The evaluation procedure is set to 10 fold cross-validation for all the experiments.

Results

The results for frame based classification are shown in Table 5.20. Some of these descriptors can not be computed for the discussed reasons.In some cases, a 5% of resampling has been needed to avoid memory problems. Results for segment-based descriptors are shown in Table 5.21.

Conclusions

First of all, we have to assume that the results obtained in Tables 5.20 and 5.21 are not representative of the state of the art in automatic musical genre classification. Both descriptors and classifiers are not tuned to obtain best accuracies but comparable results for the different configurations are obtained. According to these results, segment-based descriptors provide better accuracies than frame based analysis (up to 90% using a simple SVM classifier using spectral descriptors). Panning descriptors are the exception of this general rule providing extremely high accuracies using some specific classifiers (96.8%, 8.9%, 16.6%, 56.3%, 94.5% for IB1, SVM1, SVM2 , AdaBoost and Random

Dataset	Descr	IB1	SVM1	SVM2	AdaBoost	R.Forest
STOMP	MFCC	—	—	—	—	—
	Spectral	25.7%	41.4%	39.3%	32.9%	34.3%
	Rhythm	27.9%	25.0%	24.3%	25.7%	27.9%
	Tonality	15.0%	19.3%	17.1%	15.0%	16.4%
	Panning	15.0%	8.6%	17.9%	20.0%	20.7%
	Complexity	25.7%	22.1%	25.0%	20.7%	22.1%
	BLI	24.3%	37.1%	35.7%	25.7%	25.7%
Radio	MFCC	—	—	—	—	—
	Spectral	63.3%	80.3%	81.3%	71.8%	75.1%
	Rhythm	53.1%	59.9%	62.6%	58.4%	56.6%
	Tonality	34.6%	44.4%	44.1%	38.6%	37.5%
	Panning	20.2%	25.3%	31.6%	41.3%	40.1%
	Complexity	52.8%	55.9%	59.4%	55.9%	57.7%
	BLI	57.2%	75.6%	75.3%	63.2%	60.2%
Tzanetakis	MFCC	—	—	—	—	—
	Spectral	80.6%	90.0%	90.0%	83.5%	37.4%
	Rhythm	45.6%	52.5%	60.0%	57.9%	57.3%
	Tonality	40.8%	43.9%	45.7%	37.6%	37.7%
	Panning	—	—	—	—	—
	Complexity	26.1%	30.7%	31.0%	28.4%	30.8%
	BLI	51.9%	65.0%	65.9%	49.8%	50.8%

Table 5.21: Classification results for segment-based descriptors using different descriptors, classifiers and datasets

Forest, respectively) using frame based classification.. The unstable behavior of these descriptors claims for a detailed study which is shown in Section 5.5.3. Focusing on the collapsed descriptors, note how spectral descriptors provide best accuracies (Using the *Spectral* or *BLI* sets) followed by the Rhythm related descriptors. The third place is for Complexity descriptors which computes the complexity of the input signal for different musical facets of music (energy, spectra, rhythm or stereo image). Here again, a detailed study of the behavior of these descriptors is shown in Section 5.5.2. These results agree with those obtained in the listening experiments shown in Section 5.3, in which humans perform better classification using timbral information. It is also relevant to note how frame based descriptors are too short to describe musical genres, and long-term descriptors (built using basic statistics) describes them better. The exceptions to these rule can be found in the rhythmic descriptors where the differences are lower: as described above, these descriptors uses audio frames of (at least) 3 seconds, in order to capture all the rhytmic information in a whole bar. Then, using these *frame based* descriptors we are, in fact, taking more than the minimum time lag required by humans to identify musical genres (about 2.14s).

The classifier that produces best accuracies is Support Vector Machines in any of the two configurations detailed above. The dataset that produces best results is Tzanetakis (using 10 different musical genres). This difference is relevant for spectral descriptors but, focusing on rhythmic descriptors, the Radio dataset provide better accuracies. This phenomena shows the high dependence on the selected descriptors and taxonomies in genre classifiers: although Radio dataset uses only 8 different musical genres, spectral classification is worst than Tzanetakis dataset. In contrast, if we use rhythmic descriptors, Radio dataset produces better accuracies. The STOMP dataset produces worst results due to the high number of musical genres (14) and the low number of examples per genre (only 10). From now on, we will use this dataset only for testing, not for training.

5.5.2 Complexity descriptors

Description

As described in Section 4.3.7, complexity descriptors cover different musical facets. In this section we will discuss about their use in genre classification but computing the complexity descriptors for these musical facets. From the 8 computed descriptors we can create three groups:

Compl-12: Includes the dynamic and timbre complexity described in Sections 4.3.7 and 4.3.7

Compl-3456: Includes the rhythmic complexity described in Section 4.3.7

Compl-78: Includes the spatial complexity (that is panning information) described in Section 4.3.7. These descriptors can not be computed on the Tzanetakis database because we do not have the stereo files.

As in the previous experiments, we have used different classifiers to make results independent of classification techniques: Nearest neighbours, Support Vector Machines, Ada-boost and Random forests (See Section 4.4). All these classifications have been done using Weka. The evaluation procedure is set to 10 fold cross-validation for all the experiments.

Results

Results of these experiments are shown in Table 5.22. The used classifiers are configured in the same conditions than previous experiments in order to provide comparable results. These *Compl-78* descriptors can not be computed on the Tzanetakis dataset because we have not access to the stereo files.

Conclusion

In the three tested datasets, rhythmic complexity descriptors have proved to be the best of the three groups. This can be explained from two different points of view. First, the number of descriptors for rhythmic complexity is higher than the dynamic or panning complexity descriptors. Hence, the classifier deals with more information to classify. Second, maybe the rhythmic complexity is more relevant for genre classification than the others. This does not contradicts the previous results in which timbre (or spectral) descriptors provide better results in classification. According to that, genres may be clearly identified by different timbres, but its variability and evolution into a genre is not relevant. The spatial complexity also provides good accuracies in the classification process,

Dataset	Descr	IB1	SVM1	SVM2	AdaBoost	R.Forest
STOMP	Compl-12	17.9%	13.6%	16.7%	12.1%	16.4%
	Compl-3456	22.1%	17.9%	15.7%	15.0%	18.6%
	Compl-78	22.1%	16.4%	18.6%	13.6%	20.7%
Radio	Compl-12	26.8%	30.9%	24.2%	30.1%	27.0%
	Compl-3456	37.0%	43.9%	42.6%	37.3%	33.4%
	Compl-78	30.1%	32.7%	33.4%	37.2%	33.4%
Tzanetakis	Compl-12	17.9%	22.0%	18.9%	21.6%	18.0%
	Compl-3456	25.6%	24.3%	26.6%	25.4%	27.5%
	Compl-78	—	—	—	—	—

Table 5.22: Classification results for complexity descriptors using different classifiers and datasets

Dataset	Descr	IB1	SVM1	SVM2	AdaBoost	R.Forest
Radio	Compl-12	26.8%	30.9%	24.2%	30.1%	27.0%
	Compl-3456	37.0%	43.9%	42.6%	37.3%	33.4%
	Compl-78	30.1%	32.7%	33.4%	37.2%	33.4%
	All	52.8%	55.9%	59.4%	55.9%	57.7%

Table 5.23: Comparison of accuracies using individual or composite complexity descriptors for the Radio dataset

better than dynamic complexity. Furthermore, classification using all complexity descriptors is always higher than using only one of the proposed groups individually, as shown in Table 5.23 for the Radio dataset.

Then, we conclude that rhythmic complexity provide better results because of a) the higher number of descriptors and b) rhytmic complexity is more relevant for genre classification. More over, all the three proposed subgroups contribute positively to the genre classification, increasing about 10 . . . 15% the overall accuracy for each subset individually.

5.5.3 Panning descriptors

As shown in Section 5.5.1, panning descriptors are the unique that provide best classification results for the frame-based approach. These results need to be studied in detail in order to find the right environment in which they provide best discriminability. We will focus this study in the Radio dataset, using SVM and KNN classifiers and we will compare results with those obtained using spectral descriptors (which provide best results for segment based classification).

Description

The study is divided into two main groups:

Frame based descriptors: For the spectral descriptors we use the following configuration: $framesize = 2048$, $hopsize = 1024$, $sr = 22050$, audio

		SVM		KNN	
Dataset	Descr	Mean	StdDev	Mean	StdDev
Radio	Spectral	55.9%	0.37	63.1%	0.51
	Panning	42.4%	0.24	96.0%	0.25
	Both	64.9%	0.23	69.2%	0.50

Table 5.24: Comparison for frame-based classification using spectral and panning descriptors. Mean and Standard Deviations are shown for the 5 resamplings performed to the dataset

Database	Descr	SVM	KNN
Radio	Spectral	75.7%	61.2%
	Panning	33.8%	25.3%
	Both	74.4%	56.6%

Table 5.25: Comparison for segment-based classification using spectral and panning descriptors

$length = 30sec.$ centered at the middle of the song. Because of the high amount of collected data for spectral descriptors, we had to resize the dataset using only the 20% of the available frames. We performed 5 resamplings using different seeds and compute the mean and standard deviation for all the obtained results. The panning descriptors are computed using the same configuration with a historic time lag of $H = 2s$. In this case, we also include the resampling process, to obtain comparable results.

Segment based descriptors: Basic statistics for the descriptors detailed above (Mean, Variance, Skewness and Kurtosis).

We performed three experiments for each configuration using a) only spectral features, b) only panning features and c) a mix of them. We used a SVM classifier because it produced best results for most of the cases, and KNN (with K=1) which produced best accuracies using panning descriptors, as shown in Table 5.20. Some preliminary experiments were done to fix the number of neighbours to 1. All the experiments were carried out using Weka and the evaluation procedure was 10-fold cross validation.

Results

Results of the panning analysis are shown in Tables 5.24 and 5.25 for frame-based and segment based classification respectively. Results for the Radio dataset may be slightly different from shown in Section 5.5.1 because we change the length of the audio excerpt (from $60sec$ to $30sec$) and the resampling process is more exhaustive here.

Conclusion

Extremely high accuracies are obtained using frame-based panning descriptors with a K-NN classifier. But this accuracy dramatically decreases when using the SVM classifier which has proved to provide best results in most of the scenarios presented above. Under our point of view, these results are not coherent and the architecture of the classifier is the main cause: we use the 10-fold cross validation method to evaluate. The train and test samples are randomly selected from all the available samples. Then, we are training with the 90% of the samples but, with a high probability, we are training with frames belonging to all the audio excerpts available in the experiment. It is quite probable that, for a given *new* instance, there exists a temporally near sample (maybe the previous one or the following one) that has been used for training.

To solve this problem, we have manually splitted de dataset into train and test subsets before extracting the features, and repeat the whole process. Now, the test samples are completely separate of the training ones. We use the Holdout evaluation method and, because of the high amount of data in the experiment, we repeat the experiment 5 times with different random subsets of the original dataset, resampling it to a 20% of the original one.

Although this method is not comparable to that one explained above, results are more coherent: Accuracies using KNN provide a $mean = 43.8\%$ ($std = 0.53$) and accuracies using SVM provide a $mean = 35.7\%$ ($std = 2.0$). We realize that we should repeat the manual splitting process to assure that results do not depend on that selection but obtained results are coherent and the experiment is not relevant at all. To conclude, frame based panning coefficients do not provide best classification than segment based spectral or rhythm descriptors.

5.5.4 Mixing Descriptors

In this section we will discuss about how the mix of different families of descriptors can better explain genre classification. There are many studies in the literature dealing with descriptors (See section 3.3.2 for details). The main problem is that these studies are not comparable because they use different configuration of descriptors, classifiers and datasets. It is our goal to provide a set of (non exhaustive) comparable classification results depending on the families of descriptors but independent of the classifiers and datasets. For that, we work with the segment based set of descriptors which have proved to give best accuracies. We will use the Radio and Tzanetakis datasets (as we mentioned above, the STOMP database will be only used for testing because of its short number of examples per genre).

Description

The experiments proposed in this section are divided in two groups:

- First, we take the spectral group of descriptors as a reference and we combine them with the other groups (Rhythm, Tonal, Complexity, BLI and Panning). Then, we perform the classification using the same conditions used for segment-based descriptors a shown in Section 5.5.1. We repeat this procedure by taking the Rhythmic group of descriptors as a reference and combining them with the Tonality and BLI groups. All the other combinations have been examined in preliminary experiments producing worse results.

- Second, we build a big bag with all the computed descriptors and we apply PCA covering a 95% of variance (See Section 4.5.1 for a details on PCA). We compare the results from the best individual approach using spectral descriptors with the use of a) all the descriptors together, b) applying PCA and c) applying only those descriptors that have been selected to perform a PCA (without the linear transformation).

Results

Results for mixed descriptors for manual and automatic (PCA covering 95% of variance) feature selection are shown in Tables 5.26 and 5.27 respectively. Note how results for individual groups of descriptors marked as *ref* are the same than those shown in Table 5.21. This is the consequence of using the same sets of descriptors and classifiers in order to provide comparable results.

Having a look to the results, we can observe how the combination of spectral plus Rhythm, Complexity or BLI descriptors can increase the overall accuracy of the classifier. There is neither a fixed rule through different classifiers nor through datasets (Spectral+BLI combination provides best results for K-NN classifier while Spectral+Complexity provide best results for SVM classifier, using the Radio dataset. Spectral+Tonality provide best accuracies for AdaBoost, using the Tzanetakis dataset). Taking the Rhythmic descriptors as a reference, note how the combination with the BLI descriptors provides best accuracies (as explained in Section 4.3.8, the BLI descriptors are also considered timbre related descriptors). This combination can be as high as the combinations made using Spectral descriptors as a reference, as shown for the Radio dataset using SVM configuration.

Going further the combination of spectral and rhythmic descriptors, which is the more stable combination, spectral and complexity descriptors generally provide better accuracies for Radio dataset while the combination of spectral and tonal descriptors provide better accuracies for Tzanetakis dataset. That means that more detailed combinations of descriptors depend on the dataset.

Results on the second analysis show that PCA slightly improves the obtained accuracies using spectral descriptors using SVM. Because of that, we will come back to the PCA later in Section 5.6.

Dataset	Descr	IB1	SVM1	SVM2	AdaBoost	R.Forest
Radio	*Spectral(ref)*	*63.3%*	*80.3%*	*81.3%*	*71.8%*	*75.1%*
	Spectral + Rhythm	69.6%	82.3%	83.8%	77.1%	74.6%
	Spectral + Tonality	60.4%	80.0%	80.8%	75.3%	70.8%
	Spectral + Complexity	68.0%	84.0%	84.5%	77.1%	75.1%
	Spectral + BLI	68.6%	83.0%	84.5%	76.3%	72.1%
	Spectral + Panning	61.8%	80.3%	82.8%	75.3%	68.6%
	Rhythm(ref)	*57.2%*	*49.3%*	*49.7%*	*50.0%*	*61.5%*
	Rhythm + Tonality	50.6%	67.6%	63.1%	64.1%	57.9%
	Rhythm + BLI	64.3%	82.8%	81.5%	73.8%	64.8%
Tzanetakis	*Spectral(ref)*	*80.6%*	*90.0%*	*90.0%*	*83.5%*	*37.4%*
	Spectral + Rhythm	84.4%	91.2%	90.5%	83.7%	80.8%
	Spectral + Tonality	76.3%	90.4%	90.6%	85.8%	78.3%
	Spectral + Complexity	83.0%	89.3%	90.1%	83.5%	81.4%
	Spectral + BLI	77.3%	90.7%	90.7%	84.3%	72.7%
	Spectral + Panning	—	—	—	—	—
	Rhythm(ref)	*46.9%*	*35.1%*	*35.8%*	*37.2%*	*51.5%*
	Rhythm + Tonality	48.6%	63.2%	64.0%	62.5%	58.1%
	Rhythm + BLI	55.3%	73.2%	74.5%	65.4%	60.7%

Table 5.26: Results for mixed descriptors experiments manually selected. Results marked as *ref* are the same than those shown in Table 5.21

Dataset	Descr	IB1	SVM1	SVM2	AdaBoost	R.Forest
Radio	*Spectral(ref)*	*63.3%*	*80.3%*	*81.3%*	*71.8%*	*75.1%*
	All	69.6%	86.3%	86.3%	23.9%	74.1%
	PCA (0.95)	44.9%	75.1%	75.1%	24.9%	50.9%
	Descr. used in PCA	69.3%	84.0%	84.8%	24.4%	71.3%
Tzanetakis	*Spectral(ref)*	*80.6%*	*90.0%*	*90.0%*	*83.5%*	*37.4%*
	All	77.2%	92.2%	92.0%	20.0%	76.4%
	PCA (0.95)	71.4%	90.2%	90.3%	19.4%	55.3%
	Descr. used in PCA	78.1%	92.3%	91.5%	20.0%	78.1%

Table 5.27: Results for mixed descriptors experiments using PCA. Results marked as *ref* are the same than those shown in Table 5.21

Conclusions

As a conclusion, there is no a fixed rule on how to combine the extracted features to provide best accuracies. Only the combination of spectral and rhythmic descriptors seems to be consistent in all the environments. In fact, this is the main reason to select the Timbre and Rhythm facets of music for the listening experiments described in Section 5.3.

The second experiment mixes all the available descriptors and apply PCA to perform the attribute selection. PCA assures that we will work in a new feature space, maximizing the differences between classes, using only a reduced set of features computed automatically. But results are not clearly better than the traditional approach.

Although accuracies can grow up to 85% and 92% for Radio and Tzanetakis databases using SVM, we have not much information on the behavior of each musical genre. Other tested combinations for descriptors and classifiers (computed over different databases), not shown in this work, do not produce better accuracies. At this point, we should think we have reached the glass-ceiling of automatic genre classification, as introduced by Aucouturier & Pachet (2004). We need to change the philosophy of classifiers, as will be explained below in Sections 5.6 and 5.7.

5.5.5 Mixing Datasets

According to Livshin & Rodet (2003), evaluation using *self-classification* is not necessary a good statistic for the ability of a classification algorithm. In this paper, the authors demonstrated several important points:

- Evaluation using self-classification is not a good measure for the generalization abilities of the classification process
- Self classification results do not reflect the classifier ability, after learning the specific dataset, to deal with new audio excerpts
- Enriching the learning dataset with diverse samples from other datasets improves the generalization power of the classifier
- Evaluation using self-classification of a classification process where specific instruments are being classified does not necessary reflect the suitability of the feature descriptors being used for general classification of these instruments

In this section, we discuss about the generalization power of our proposed classifiers according to the items presented above. For that, we find the shared musical genres between datasets. Then, by limiting the analysis to this reduced taxonomy, we use one of the datasets for training and the other one for testing, and we compare results with the 10-fold cross validation evaluation.

Description

For these experiments, we use three datasets: Radio and Tzanetakis for train and STOMP for test. As mentioned in Section 5.5.1, the STOMP dataset is defined by only 10 full songs for each category. This low number does not

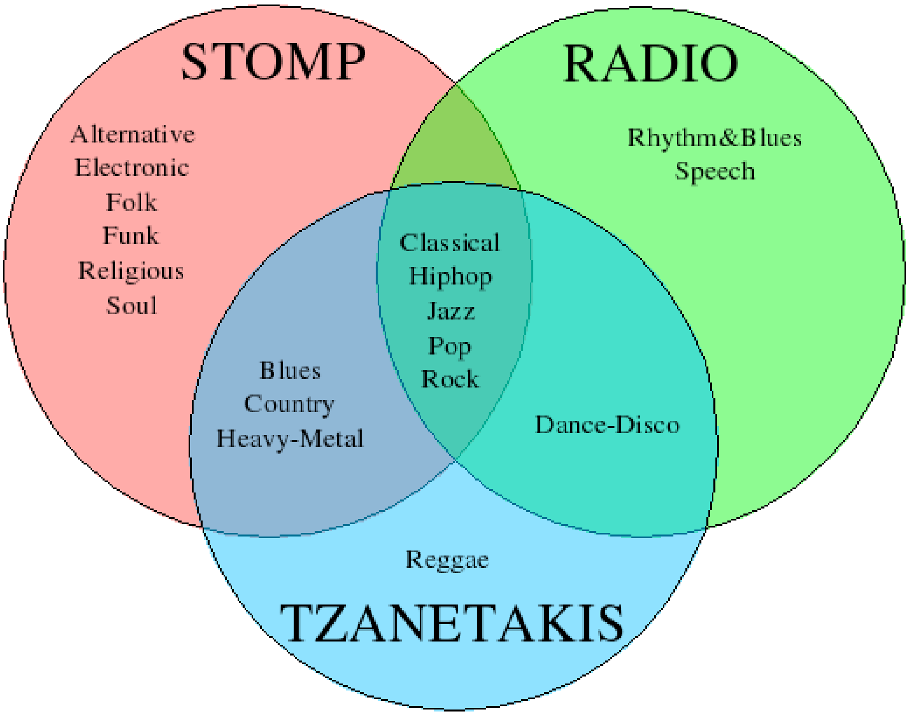


Figure 5.12: Shared taxonomies between Radio, Tzanetakis and STOMP datasets

recommend its use for training. Hence, we use the Radio and Tzanetakis for both training and testing, and the STOMP dataset only for testing. We use the shared taxonomies between them which are presented in Figure 5.12. We use the most discriminative descriptors (Spectral, Rhythm, Tonal and Complexity) and the most important combinations of them (Spectral + Rhythm, Tonality or Complexity). The classifier we use is Support Vector Machines , omitting the results obtained using other classifiers which provide worse results.

We present here three sets of experiments. The first one uses only two of the three datasets, that is Radio or Tzanetakis datasets for training STOMP dataset for testing. For the first case, Radio and STOMP datasets share 5 musical genres (classical, hiphop, jazz, pop and rock) while for the second case, Tzanetakis and STOMP datasets share 7 musical genres (blues, classical, country, hiphop, jazz, pop, rock). For the second experiment, we will not use the STOMP dataset for testing, and we will compare the two *big* datasets one against the other, which share 5 musical genres (classical, hiphop, jazz, pop and rock). For the third experiment, we will use again the 3 datasets, sharing 5 musical genres, and we will use two of them for training and the other one for testing in all the possible configurations (this will be the unique experiment in which we use the STOMP dataset for training). These combinations are the following:

- Train: Radio + Tzanetakis; Test: STOMP
- Train: Radio + STOMP; Test: Tzanetakis

	#	Descr	10-fold	STOMP
Radio	5	Spectral	91.2%	72.0%
	5	Rhythm	59.4%	68.0%
	5	Tonality	60.2%	42.0%
	5	Complexity	59.9%	54.0%
	5	Spectral + Rhythm	88.0%	75.5%
	5	Spectral + Tonality	89.2%	78.0%
	5	Spectral + Complexity	93.2%	74.0%
		Mean	*77.3%*	*66.2%*
Tzanetakis	7	Spectral	93.8%	60.0%
	7	Rhythm	61.3%	35.7%
	7	Tonality	52.1%	22.9%
	7	Complexity	34.5%	41.4%
	7	Spectral + Rhythm	95.1%	47.1%
	7	Spectral + Tonality	94.4%	39.1%
	7	Spectral + Complexity	95.4%	61.4%
		Mean	*75.2%*	*43.9%*
Tzanetakis	5	Spectral	93.1%	74.0%
	5	Rhythm	70.6%	54.0%
	5	Tonality	63.3%	30.0%
	5	Complexity	46.7%	64.0%
	5	Spectral + Rhythm	94.5%	58.0%
	5	Spectral + Tonality	94.5%	55.1%
	5	Spectral + Complexity	94.3%	74.0%
		Mean	*79.6%*	*58.4%*

Table 5.28: Results for the mixed datasets experiments using a) 10-fold cross validation and b) STOMP database

- Train: Tzanetakis + STOMP; Test: Radio

This experiment aims to compare results with the most heterogeneous dataset available by combining two them (the other one is used for testing). Although the collections are supposed to be representative of the musical genres, the more diverse the datasets are built, the more realistic results will be obtained.

Finally, we have also included an additional experiment in the first set using the Tzanetakis dataset for training and STOMP for testing, which share 7 musical genres as described above, but using only the 5 musical genres that are shared by the 3 datasets. The goal of this test is to compare results with the same number of categories than the experiment using Radio dataset for training and STOMP dataset for testing.

Results

Results of the first experiments are shown in Table 5.28. In this table, we include results of the classification using 10-fold cross validation and using the test set extracted from STOMP. Note how results using 10-fold cross validation can seem extremely high. This is because of the reduced number of musical

	#	Descr	10-fold	Other
Radio	5	Spectral	91.2%	65.3%
	5	Rhythm	59.4%	51.3%
	5	Tonality	60.2%	32.0%
	5	Complexity	59.9%	40.2%
	5	Spectral+Rhythm	88.0%	66.9%
	5	Spectral+Tonality	89.2%	58.8%
	5	Spectral+Complexity	93.2%	76.5%
		Mean	*77.3%*	*55.9%*
Tzanetakis	5	Spectral	93.1%	78.9%
	5	Rhythm	70.6%	52.2%
	5	Tonality	63.3%	26.8%
	5	Complexity	46.7%	50.6%
	5	Spectral+Rhythm	94.5%	77.7%
	5	Spectral+Tonality	94.5%	55.8%
	5	Spectral+Complexity	94.3%	78.1%
		Mean	*79.6%*	*60.0%*

Table 5.29: Results for the mixed databases experiments using a) 10-fold cross validation and b) Other *big* dataset: Tzanetakis when training with Radio and vice-versa

genres used, according to Figure 5.12. Note how we repeat the experiment using the Tzanetakis dataset for 7 and 5 categories to make results comparable with those obtained using Radio dataset.

As mentioned above, we also compare results by using the two *big* datasets (Radio and Tzanetakis), one against the other. Table 5.29 shows the obtained results for different groups of descriptors. Finally, we also perform classification by training with two of the databases and testing with the other one. Results are shown in Table 5.30

Conclusions

According to the presented results, the evaluation process made using the STOMP dataset provides lower accuracies than cross validation, as expected. Note how this difference is higher for the Tzanetakis dataset: differences up to 55% between cross-fold validation and mixing datasets strategies are found when combining spectral and tonality descriptors, and up to 45% when combining spectral and rhythm descriptors. Although the experiment using this dataset has more musical genres than the experiment using the Radio dataset, these results may reflect the musicological criteria applied when building this last one. Remember that the STOMP dataset was manually built by selecting the most representative songs for each genre by a musicologist. Differences of results between Tzanetakis and Radio dataset may also show that the last one covers a wider range of music inside each musical genre. The different numbers of songs per genre (100 for Tzanetakis and 50 for Radio) may affect this result: the classifier is more exhaustively trained using the Tzanetakis dataset. As a consequence of that, classification with completely new and unknown data provide worse accuracies. What it is clear is that experiments using 10-fold

	#	Descriptor	10-fold	Other
Tzanetakis+Radio	5	Spectral	92.9%	78.0%
	5	Rhythm	69.1%	60.0%
	5	Tonality	56.6%	58.0%
	5	Complexity	50.5%	58.0%
	5	Spectral+Rhythm	92.7%	76.0%
	5	Spectral+Tonality	92.7%	75.5%
	5	Spectral+Complexity	93.4%	66.0%
		Mean	*78.3%*	*67.3%*
Tzanetakis+STOMP	5	Spectral	91.5%	80.9%
	5	Rhythm	69.4%	55.0%
	5	Tonality	58.9%	44.3%
	5	Complexity	49.2%	52.6%
	5	Spectral+Rhythm	91.7%	80.9%
	5	Spectral+Tonality	90.6%	75.7%
	5	Spectral+Complexity	91.7%	76.9%
		Mean	*77.5%*	*66.6%*
Radio+STOMP	5	Spectral	86.7%	68.2%
	5	Rhythm	63.8%	53.3%
	5	Tonality	59.1%	31.0%
	5	Complexity	63.6%	42.4%
	5	Spectral+Rhythm	87.3%	72.4%
	5	Spectral+Tonality	88.0%	58.4%
	5	Spectral+Complexity	89.0%	77.8%
		Mean	*76.7%*	*57.6%*

Table 5.30: Results for the classification by training with two of the databases and testing with the other one

cross validation provide very optimistic results which are far of the possible real scenario in which these classifiers should work.

The second experiment here presented show differences about 20% when combining our two *big* datasets. Results are similar for the two reciprocal configurations. This reinforces the fact that these datasets do not cover the overall spectra for each specific musical genre or that selected descriptors do not accurately represent musical genres. If so, results for 10-fold cross validation or split evaluation methods should be similar than those obtained when presenting a completely new dataset. Due to the reciprocal behavior, results are independent of the number of songs for each category. It makes us to conclude that both datasets have their own area in the genre space which is quite well defined by their elements.

Finally, when combining all the three available datasets in the third experiment, the mean values show how the differences between the 10-fold cross validation and the use of different datasets is about 10% for the first two cases, and about 20% for the last one. This last case corresponds to a training set formed by Radio and STOMP datasets, and a test set formed by Tzanetakis. Because of that, we conclude that the Radio dataset covers a genre space with similar properties than the STOMP one, while the Tzanetakis dataset is not

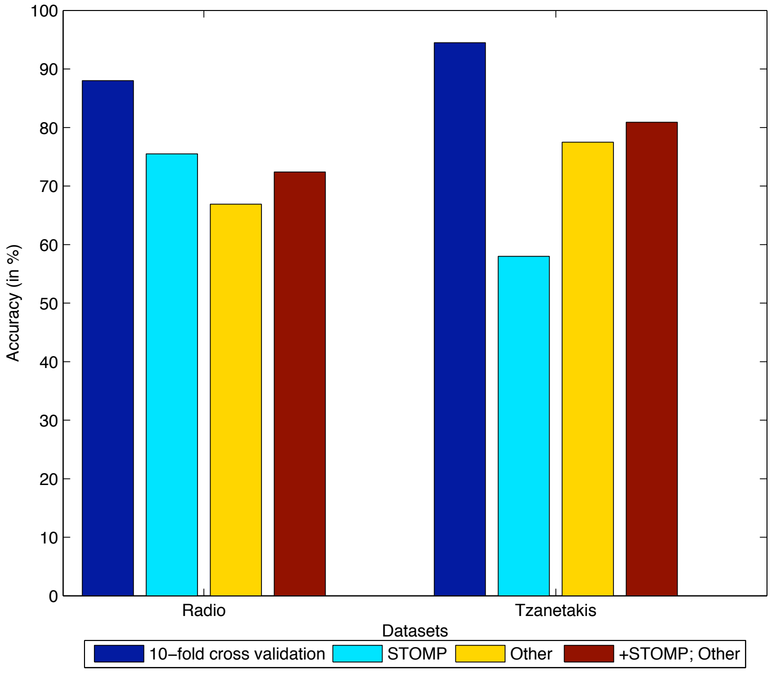


Figure 5.13: Summary of the obtained accuracies for Radio and Tzanetakis datasets using spectral and rhythmic descriptors for a) using 10-fold cross validation, b) evaluating with the Stomp dataset, c) evaluating with the *other* dataset and d) including the STOMP dataset to the training set and evaluating with the *other*

so overlapped (Let us remark that it does not mean that this last database is not correctly built!).

Figure 5.13 shows the summary of the obtained accuracies for Radio and Tzanetakis datasets, using spectral and rhythmic descriptors, for a) using 10-fold cross validation, b) evaluating with the Stomp dataset, c) evaluating with the *other* dataset[9] and d) including the STOMP dataset to the training set and evaluating with the *other*. This figure shows differences from 10% to 30% between 10-fold cross validation and the other combinations that should not be ignored in our study.

As a conclusion, these test could be interpreted as a measure on how universal a dataset is. Biased collections built without artist pre-filter should provide lower performances in this set of experiments. And the opposite: datasets built using artists that cover the whole space for that specific genre will provide better results. In our case, it seems that STOMP and Radio datasets are more universal than the Tzanetakis collection. It is also important to compare the number of available songs for each category: the 10 songs/genre in STOMP dataset contrasts with the 100 songs/genre in Tzanetakis. This will intrinsi-

[9] The *other* dataset is Tzanetakis when training with Radio, and Radio when training with Tzanetakis

cally affect the overall accuracy when combining them and, of course, it is more difficult to find 100 different songs per genre that musicologically represents the overall category than a simple list of 10.

Note how we express *universal* datasets without entering in the discussion whether they are correctly defined or not. For instance, Religious music (included in the STOMP dataset) can be a perfectly defined category for commercial purposes although it includes a great ensemble of styles and artists. The differences between datasets can be produced by some of the following reasons:

- Different goals of the dataset: For instance, it can be collected for commercial purposes, musicological research or a simple recompilation of our favorite artists
- Availability of musical audio excerpts: Although the Internet provides many different options for downloading audio files, it is not always easy to find some specific songs. This can be an important factor when collecting a dataset.
- Other reasons like storage availability or the use of specific datasets for comparing results with other researchers may also affect the collected dataset.

5.6 Single-class classifiers

In the previous sections, we have studied in depth the behavior of genre classifiers for different databases, sets of descriptors and classifiers. According to Aucouturier & Pachet (2004), all these approaches for genre classification reach a glass-ceiling which seems difficult to cross. In this section, we propose other methods that change the traditional point of view by focusing on an ensemble of individual classifiers. For that, we will create a specific classifier for each musical genre.

5.6.1 Justification

Up to now, all the classifiers deal with all the categories at the same time. The only exception could be found with Support Vector Machines. As described in Section 4.4.2, SVMs try to find, in a higher dimensional space, an hyper-plane that maximally separates the distance between training data. This process is repeated for all the pairwise possible combinations of categories belonging to the proposed problem. In our case, SVMs find the maximum distance between blues and classical, blues and hip-hop, blues and rock, etc., but also the maximum distance between classical and hip-hop, classical and rock, etc. As SVMs has proved to be the best tested classifier for most of the cases, we extract the idea of building an ensemble of individual classifiers.

The idea of ensemble of classifiers is not new. Ponce & Inesta (2005) propose the use of a set of k-nearest neighbors and Bayesian classifiers for genre classification based on MIDI data. The classifiers are independently trained for different groups of descriptors and the resulting models are combined using a majority vote scheme. Scaringella & Zoia (2005) propose the use of an ensemble of three SVM classifiers focused on different features of audio describing different dimensions of music. Resulting models are combined computing the average of posterior probabilities over all input feature vectors and selecting the highest averaged posterior probability class. Harb et al. (2004) proposes the use of an ensemble of neural networks and Piecewise Gaussian Modeling for audio signal representation. One of the important contributions of this work is the comparison of different combining strategies for the individual classifiers. One of the simplest ways for combining classifiers is the majority vote schema but sometimes it is necessary to apply some mathematical functions (addition, multiplication, etc.) to the output of the classifiers. The choice of the output weights is not a trivial task and it depends on the specific problem to solve. Harb et al. (2004) proposes a list of different weighting strategies:

Equal weights: Each individual classifier contributes to the final classification with the same weight

Gate network: Each individual weight is defined by the output of an expert trained to classify an (many) observation(s) from the training dataset into that specific class.

Novelty of experts: This method proposes to assign less weight to the expert classifiers that do not provide a novelty on the mixture of experts. The novelty can be computed according to the training time or the number of operations.

Error estimation: The weight is inversely proportional to the error rate of this expert in front of a development database.

In our proposal, the classifiers are not built using pairwise configuration but the *1-against-all* strategy. That means that we build as many classifiers as categories and each one will be trained for discriminating between the selected category and all the rest (i.e. blues against not blues, classical against not classical, etc.). Figure 5.14 shows an overview of the proposed method. The main advantage of using this configuration is that we can perform attribute selection independently for each classifier and train it using only the relevant features that define that musical genre. We are killing two birds in one shot. First, we have a set of specialized classifiers that are experts in one category (divide and conquer!) and, second, we have an idea of the most representative descriptors that define each specific genre, which maybe can be interpreted from a musicological point of view, following the ideas emerged in the listening experiments described in Section 5.3.

The main problem we found is the classification of new data. The flow for new data is shown in with a wide arrow in Figure 5.14. As we don't know at which category it belongs, we have to compute as many attribute selection and classifications as the number of different categories we have in our problem. The output for each classifier will provide a probability of new data belonging to that category. We will assume that the assigned category is that one with higher probability, but the obtained probabilities could also be useful for building hierarchical classifiers (which are not in the scope of this thesis).

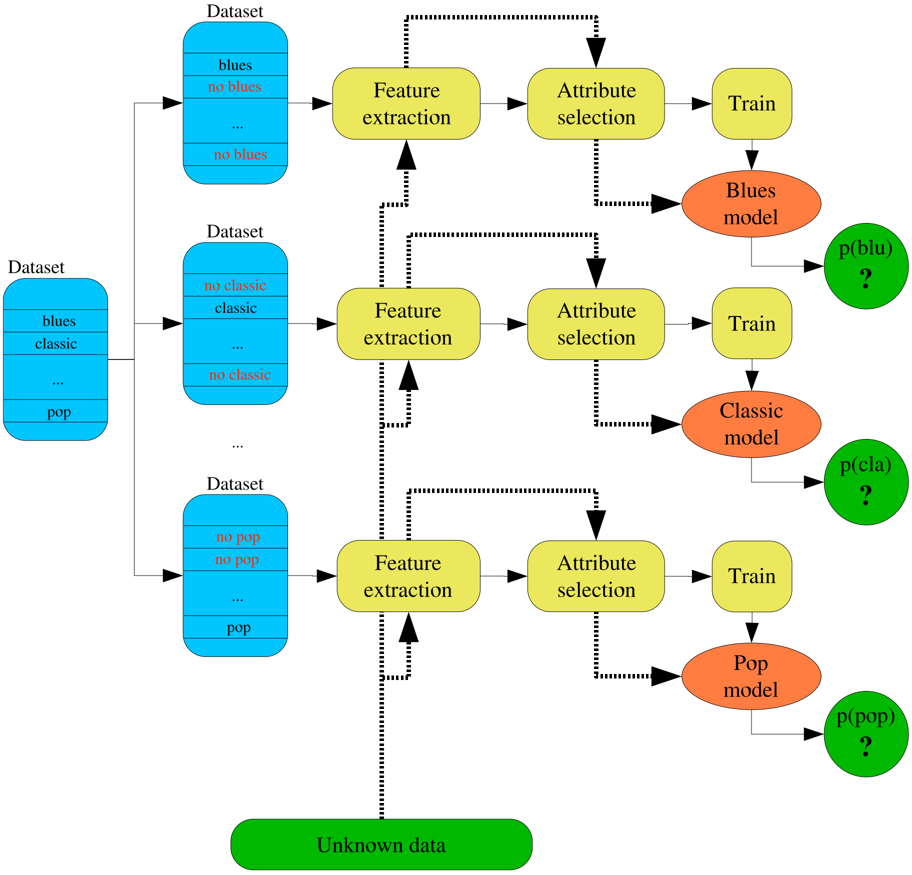


Figure 5.14: Block diagram for single-class classifiers

5.6.2 Attribute selection

We have created a new pairwise dataset for all the genres in the Tzanetakis dataset. For each genre we have merged all the others into a unique category, obtaining a Genre/non-Genre pair. It is obvious that this new data is completely unbalanced. For classification using SVM, we have applied multiple resampling with different seeds (we have not applied the resampling for tree classification using J48). With all these new pairwise datasets we have selected the most relevant descriptors for each case, as shown in Tables 5.31, 5.32, 5.33 and 5.34.

	Blues	Classical	Country
Attr 1	KurtosisRHY_Energy	VarDELTADELTAMFCC3	MeanSpectralVariance
Attr 2	MeanSpectralVariance	VarDELTADELTAMFCC2	MeanTONALITY_5
Attr 3	SkewnessSpectralVariance	VarSpectralCentroid	MeanSpectralMean
Attr 4	KurtosisRHY_Mean	MATRIX_10_10	VarSpectralCentroid
Attr 5	MeanSpectralFlatness	VarDELTADELTAMFCC4	KurtosisRHY_Energy
Attr 6	MeanDELTADELTAMFCC3	MeanDELTADELTAMFCC1	MeanSpectralKurtosis
Attr 7	SkewnessMFCC6	MeanDELTADELTAMFCC3	VarDELTAMFCC8
Attr 8	KurtosisRHY_SpectralCentroid	MATRIX_1_11	MATRIX_21_23
Attr 9	MeanMFCC8	VarDELTAMFCC3	SkewnessTONALITY_5
Attr 10	VarDELTADELTAMFCC2	MeanSpectralSkewness	MeanSpectralFlatness
J48	94.0%	94.8%	87.0%
SVM	93.5%	98.4%	92.0%

Table 5.31: 10 most representative descriptors for each musical genre (in descending order), computed as 1-against-all categories, and the obtained accuracies when classifying this specific genre independently (Table 1/4)

	Disco	HipHop	Jazz
Attr 1	MeanSpectralMean	MeanTONALITY_18	SkewnessSpectralMean
Attr 2	MeanDELTADELTAMFCC2	KurtosisTONALITY_12	SkewnessRHY_SpectralCentroid
Attr 3	MeanDELTADELTAMFCC8	VarDELTADELTAMFCC4	VarDELTADELTAMFCC4
Attr 4	MeanSpectralVariance	MeanTONALITY_2	VarDELTADELTAMFCC0
Attr 5	BEAT_RESULT	MeanTONALITY_30	MATRIX_12_16
Attr 6	MeanDELTAMFCC5	MeanRHY_Mean	VarSpectralCentroid
Attr 7	MeanDELTADELTAMFCC3	VarMFCC6	VarTONALITY_4
Attr 8	MATRIX_6_7	KurtosisTONALITY_18	VarDELTADELTAMFCC3
Attr 9	MeanSpectralKurtosis	VarTONALITY_28	KurtosisDELTAMFCC0
Attr 10	VarDELTAMFCC3	LONG_RESULT	KurtosisMFCC4
J48	92.5%	85.0%	88.5%
SVM	93.5%	89.5%	93.0%

Table 5.32: 10 most representative descriptors for each musical genre (in descending order), computed as 1-against-all categories, and the obtained accuracies when classifying this specific genre independently (Table 2/4)

	Metal	Pop	Reggae
Attr 1	SkewnessSpectralSkewness	VarSpectralFlatness	VarSpectralFlatness
Attr 2	MeanSpectralSkewness	MeanMFCC11	VarDELTADELTAMFCC4
Attr 3	VarSpectralSkewness	MeanSpectralSkewness	KurtosisSpectralMean
Attr 4	MATRIX_7_10	VarSpectralCentroid	KurtosisRHY_SpectralFlatness
Attr 5	VarSpectralVariance	MeanSpectralVariance	VarMFCC6
Attr 6	SkewnessMFCC10	VarDELTAMFCC5	SkewnessMFCC0
Attr 7	MATRIX_19_19	VarRHY_Energy	MATRIX_4_9
Attr 8	VarMFCC12	SkewnessSpectralFlatness	KurtosisRHY_Energy
Attr 9	MeanSpectralMean	VarSpectralKurtosis	VarMFCC2
Attr 10	VarMFCC8	MeanMFCC7	MATRIX_14_22
J48	93.5%	92.0%	90.0%
SVM	97.5%	92.0%	87.0%

Table 5.33: 10 most representative descriptors for each musical genre (in descending order), computed as 1-against-all categories, and the obtained accuracies when classifying this specific genre independently (Table 3/4)

	Rock
Attr 1	MeanSpectralMean
Attr 2	VarSpectralCentroid
Attr 3	MeanDELTADELTAMFCC4
Attr 4	MeanDELTADELTAMFCC0
Attr 5	MeanDELTADELTAMFCC3
Attr 6	MeanDELTADELTAMFCC1
Attr 7	MeanSpectralFlatness
Attr 8	SkewnessRHY_Mean
Attr 9	MeanDELTADELTAMFCC9
Attr 10	VarDELTADELTAMFCC12
J48	88.0%
SVM	86.0%

Table 5.34: 10 most representative descriptors for each musical genre (in descending order), computed as 1-against-all categories, and the obtained accuracies when classifying this specific genre independently (Table 4/4)

	Family		Statistic	
	Timbre	Rhythm	Mean	Var
Blues	2	1	1	0
Classical	3	0	0	3
Country	2	0	3	0
Disco	3	0	3	0
Hip-Hop	1	0	1	1
Jazz	3	0	0	1
Metal	3	0	1	1
Pop	3	0	2	1
Reggae	3	0	0	2
Rock	3	0	2	1

Table 5.35: Predominant descriptor family (timbre/rhythm) and statistic (mean/variance) for each musical genre

According to the results presented in Tables 5.31, 5.32, 5.33 and 5.34, we can assume that spectral descriptors present best discriminative power, followed by rhythmic a tonal descriptors. Rhythm descriptors are relevant in Blues (first and four position in the ranking of descriptors), Jazz (second position) and reggae (fourth position). On the other hand, tonal descriptors are relevant in Country (second position) and Hip-Hop music (first, second and fourth positions). As expected, Blues, Jazz and Reggae music can be described by their rhythmic properties (swing or syncopations for Blues or Jazz and Reggae, respectively). These results partly agree with the results of the listening experiments presented in Section 5.3 in which jazz music is the second classified using only rhythmic information, after the electronic music. Unfortunately, we don't have the information of human classification for Blues and Reggae genres using rhythmic descriptors, but we assume that they should be similar to Jazz music. On the other hand, the relevance of tonal descriptors is not so evident. The tonal behavior in Country music is not so different from Rock or Pop music and, furthermore, Hip-Hop music can be characterized by its low relation with tonal information (in the sense of the characteristic voice used in this style). Maybe this *non-use-of-tonality* can be the property that makes this style different from the others.

Focusing on the three most important attributes proposed by the feature selection algorithm for each genre, we can study two properties:

1. Whether the selected attributes correspond to timbre or rhythm families of descriptors

2. Whether the selected attributes are the mean or the variance of a specific descriptor for all the song

This analysis is not exhaustive: Tonal or BLI descriptors are not taken into account and, in a similar way, skewness and kurtosis statistics are neither taken into account. These properties are summarized in Table 5.35.

All the proposed musical categories show timbre descriptors as the most relevant ones. The statistics to be used do not show a clear behavior instead.

The computation of the variance of the descriptors for the whole audio excerpt is relevant for Classical, Hip-Hop, Jazz, Metal and Reggae. We didn't find any musicological explanation for that.

5.6.3 Classification

New data is tested against all the classifiers with the corresponding attribute selection (manually applied according to results in Tables 5.31, 5.32, 5.33 and 5.34). The new data flow is shown in with a wide arrow in Figure 5.14. We use a majority voting schema for the overall classification. For classification using trees, we have selected the minimum number of instances per leaf equals to 10 (some previous tests have been done to experimentally fix this value). Initial experiments have been done using this technique providing quite interesting results, using both SVM and Trees (see Tables 5.31, 5.32, 5.33 and 5.34). But we can not show exhaustive and comparative results with all the previous experiments because of the following reasons:

- The dataset is clearly unbalanced and this affects the performance of the classification using SVMs (we don't think resampling it is a good solution)

- 10-fold cross validation can not be computed for all the experiments with the tools we use for all the previous experiments because of the complexity of the architecture

The solution to these problems is to use a specific classifier that follows this philosophy, that is, the Simca classifier described in Section 4.5.2. We will use the Simca classifier to perform all these tests automatically and results will be shown in Section 5.7.

5.6.4 Conclusion

In this section we have presented the use of individual experts for genre classification. Starting from the idea of pairwise classification used by SVM, we build an ensemble of expert classifiers represented by the most discriminative descriptors for each specific category. Although spectral descriptors present the best discriminative power, rhythmic and tonal features can help in classification of Blues, Jazz and Reggae, and Country and Hip-Hop respectively. We could not provide exhaustive results for this experiment, but we put a special emphasis in these tests because they are the main reason of using the Simca classifier. This technique provides all the mathematical support to perform the classification experiments proposed in this section.

5.7 The SIMCA classifier

In the previous section, we explained the motivations of using Simca. We explained our intention of building an ensemble of expert classifiers represented by the most discriminative descriptors for each specific category. For that, we computed attribute selection independently for each classifier and trained it using only the relevant features that define that musical genre. We killed two birds in one shot: First, we got a set of specialized classifiers that are experts in one category (divide and conquer!), and second, we had an idea of the most representative descriptors that define each specific genre, which maybe can be interpreted from a musicological point of view. In the previous section, we also argued that Simca classification could be more similar to human reasoning for genre classification than other traditional classifiers, inspired by the results of the listening experiments presented in Section 5.3. These two arguments drive us to propose this classification algorithm for classifying music genres. The main differences between the technique previously exposed and the Simca method are:

1. The attribute selection previously described is based on the ranking list of the most discriminative descriptors while Simca classifier uses PCA.

2. The classifiers previously proposed are SVM or Decision Trees while the Simca classifier compares the distance from new data to each transformed space for specific categories.

3. The previous overall classification method is based on a voting schema while the Simca classifier computes the assigned class by means of a F-test on the computed distances.

A detailed explanation of the Simca algorithm is provided in Section 4.5.2.

5.7.1 Description

For the experiments here proposed, we use the Simca implementation provided by the LIBRA toolbox in Matlab, developed at the Katholieke Universiteit Leuven and the University of Antwerp[10]. This toolbox requires formated input data (descriptors and categories) and provide the classification results for the new data in terms of category labels. Some additional code have been developed to adapt our datasets to the required format. On the other hand, some modifications to the toolbox have been made:

- We adapted the algorithm for PCA in order to manually set the percentage of variance covered by the principal components instead of the number of components

- We adapted the algorithm to provide distance measures from the new data to all the existing categories instead of the category labels.

Moreover, our code is the responsible to perform a random split of the whole database into a (configurable) 66% for train and 33% for test. All the experiments have been repeated 10 times with different seeds in the splitting process to avoid possible biasing effects. The conditions for the SVM experiments are the same than shown above in order to provide comparable results.

[10] http://wis.kuleuven.be/stat/robust/LIBRA.html

Database	#Genres	Descriptor	SVM	SIMCA	diff
Radio	8	Spectral	76.8	95.5	18.7
		Rhythm	60.3	83.7	27.4
		Tonality	42.9	83.3	40.4
		Complexity	61.2	49.4	-11.8
		Spectral+Rhythm	80.1	97.0	16.9
		Spectral+Tonality	77.2	94.9	17.7
		Spectral+Complexity	78.8	96.6	17.8
		mean	*68.2*	*85.8*	*17.6*
Tzanetakis	10	Spectral	89.0	95.8	6.8
		Rhythm	54.6	79.5	24.9
		Tonality	39.2	96.6	57.4
		Complexity	31.2	37.1	5.9
		Spectral+Rhythm	89.2	97.2	8
		Spectral+Tonality	88.8	99.0	10.2
		Spectral+Complexity	89.5	99.0	9.5
		mean	*68.8*	*86.3*	*17.5*

Table 5.36: Results for SVM and SIMCA classifiers for different datasets and sets of descriptors, presented as the mean of accuracies for 10 experiments with different random splits for train and test subsets.

5.7.2 Results

Initial Experiments

In this section, we present and discuss results for genre classification. The experiments, can be divided into two main groups. First, we show the performance of Simca in comparison to Support Vector Machine under the same conditions of datasets, descriptors and classifier parameters (splitting databases, standardization, etc.). Second, we show the performance of SIMCA classifier when combining different datasets. All these experiments have been repeated for different sets of descriptors and datasets in order to ensure that the obtained accuracies are not dependent on specific data, as we made in all the previous tests.

Results for the first set of experiments are shown in Table 5.36. Note how results obtained using SVMs differ from those obtained in Section 5.5.5 because we use a split ($66\% - 33\%$) evaluation method instead of 10-fold cross validation. Here, we compare the use of SVM's with SIMCA using different sets of descriptors but the same dataset for training and testing. Remind that the Radio dataset contains 8 categories and the Tzanetakis dataset contains 10 categories. All the accuracies here presented represent the mean of 10 experiments with different random splits between train and test subsets. Note how we report accuracies about 99%. We will discuss about its reliability in Section 5.7.4. The averaged difference between the accuracies obtained using SVMs or Simca is about 17.5% for the two datasets. In this environment, the Simca classifier clearly outperforms SVMs, one of the best traditional classifiers.

Dataset	# Genres	Descriptor	Split	STOMP	diff	Other	diff
Radio	5	Spectral	94.8	96	1.2	68.6	-26.2
		Rhythm	88.3	94	5.7	62.8	-25.5
		Tonality	92.9	100	7.1	81.5	-11.4
		Complexity	62.8	56	-6.8	32.7	-30.1
		Spectral+Rhythm	98.6	98	-0.6	72.9	-25.7
		Spectral+Tonality	94.6	100	5.4	75.7	-18.9
		Spectral+Complexity	95.0	98	3	50.1	-44.9
		mean	*89.6*	*91.7*	*2.1*	*67.5*	*-22.1*
Tzanetakis	5	Spectral	97.4	98	0.6	99.2	1.8
		Rhythm	89.0	66	-23	80.0	-9
		Tonality	98.7	86	-12.7	82.9	-15.8
		Complexity	57.3	20	-37.3	20.4	-36.9
		Spectral+Rhythm	99.2	94	-5.2	95.9	-3.3
		Spectral+Tonality	99.4	92	-7.4	92.2	-7.2
		Spectral+Complexity	98.9	52	-46-9	55.9	-43.0
		mean	*91.4*	*72.5*	*-18.9*	*75.2*	*-16.2*

Table 5.37: Results for SIMCA classifier mixing datasets. The *Split* accuracies are presented as the mean of accuracies for 10 experiments with different random splits for train and test subsets. The *Other* dataset corresponds to *Tzanetakis* when training with *Radio* and viceversa. The *diff* column shows the difference between the accuracies using separate datasets and the $66-33\%$ split experiment.

Dataset	Other(SVM)	Other(SIMCA)
Radio	55.9	67.5
Tzanetakis	60.0	75.2

Table 5.38: Comparison between SVMs and Simca classifier for cross datasets experiments. The *Other* dataset corresponds to *Tzanetakis* when training with *Radio* and viceversa.

On the other hand, Table 5.37 presents results of mixing datasets, training with Radio and Tzanetakis and testing with STOMP and Tzanetakis or Radio respectively (shown in the column labelled as *Other*). Note how the number of genres has been reduced to five which is the number of shared categories for all the datasets. We also present the behavior of the classifier for different sets of descriptors. Differences between 16.2% and 22.1% are observed, for the mean of all the families of descriptors, for the accuracies between the $66-33\%$ splitted tests and separate datasets. The only exception is the increasing of 2.1% of the accuracy using the STOMP dataset for test when training with Radio. Comparing this experiment with the equivalent one carried out using SVM (See Table 5.29), we clearly observe that Simca outperforms SVMs, as summarized in Table 5.38.

Scalability

The goal of this section is to determine, for the case of classification of musical genres using rhythmic and timbre descriptors described in Section 5.5, the behavior of the Simca classifier in front of a) a large number of musical genres and b) a large number of instances per category. We want to verify that the performance does not dramatically fail in these conditions which are near to a real environment for genre classification.

For that, we use a huge collection based on the previews of the iTunes store with more than one million of songs. We use the taxonomy proposed by iTunes as ground-truth, selecting 12 of the most representative musical genres: Alternative, Classical, Country, Dance, Electronic, Folk, Jazz, Pop, Rap, Rock, Soul, Soundtrack. We performed two experiments:

Large scale: In this experiment, we grow from 10 to 10000 songs per genre in 4 decades.

Detailed scale: In this experiment, we grow in 10 linear steps between the 2 decades with higher accuracies.

All the experiments are the result of 10 random splits of the dataset into 66% for training and 33% for testing. Then, experiments performed with 100 songs are, in fact, experiments that only use 66 songs for training. This deviation is not crucial because we are looking for an order of magnitude, instead of a specific number.

Figure 5.15 shows the evolution of the 12 musical genres from 10 to 10000 instances per category. Classifications performed with less than 100 instances show bad results. After that, all of the genres slightly increase except for *Classical*, *Alternative* and *Jazz*. After that, between 1000 and 10000 instances

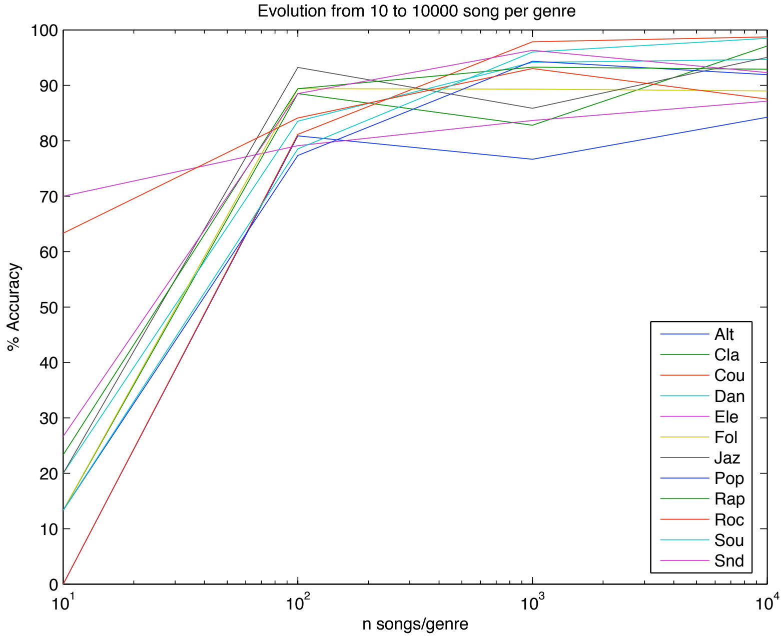


Figure 5.15: Evolution from 10 to 10000 song per genre

per category, the behavior is not clear. Thus, if we look for a compromise between the accuracy and the number of songs per genre, it seems necessary to perform a detailed experiment between 100 and 1000 songs per genre.

Figure 5.16 show these results for the detailed scale. Here, we can see the real measured values (as means of results of 10 random splits in the dataset) as well as the result of a 3rd. order polynomial regression. The maximum of these regression curves are mainly located between 300 and 700. If we compute the mean for all the genres we got a maximum near 500 instances per genre, as shown in Figure 5.17

We can conclude that, for automatic classification of 12 musical genres using an ensemble of timbre a rhythm descriptors using the SIMCA classifier, the number of instances per category that maximizes the accuracies is near 500. Furthermore, classification results for experiments using 100 or 10000 instances per category show the same order of magnitude. Then, Simca classifier is flexible enough to deal with real scenarios.

5.7.3 Other classification tasks

Mood or Western/Non-western music classification are clearly out of the scope of this thesis but, as shown in the previous sections, the good results on classification using Simca suggest that this technique can be applied to other classification problems of the MIR community. In this section, we present a set of very simple experiments using SIMCA and compare their results with other state of the art algorithms. We designed these experiments just to test whether SIMCA can be a good technique to be applied or not, so, don't interpret these

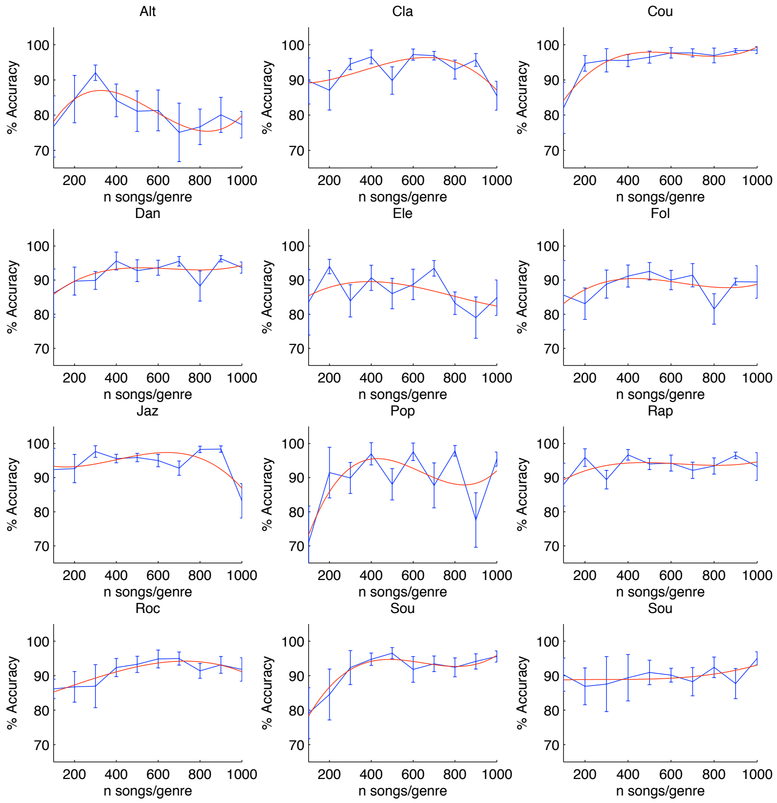


Figure 5.16: Evolution from 100 to 1000 song for genre. The real measured values are shown in blue and the 3rd. order polynomial regression are shown in red

results as if they were contrasted with the state art in their respective areas.

Mood evaluation

According to the literature, mood/emotion is an important criterion in music organization of huge databases (Cunningham et al., 2004, 2006). As a consequence of that, the MIR community started his research providing many interesting approaches (Lu et al., 2006; Pohle et al., 2005b). The actual state of the art on Mood classification can be summarized in the Audio Music Mood Classification task from the MIREX contest. As described in Section 5.4, the goal of MIREX is to provide the ideal environment to compare algorithms of different authors performing the same task. The complete description of this task can be found in the MIREX Wiki[11].

[11] http://www.music-ir.org/mirex/2007/index.php/Audio_Music_Mood_Classification

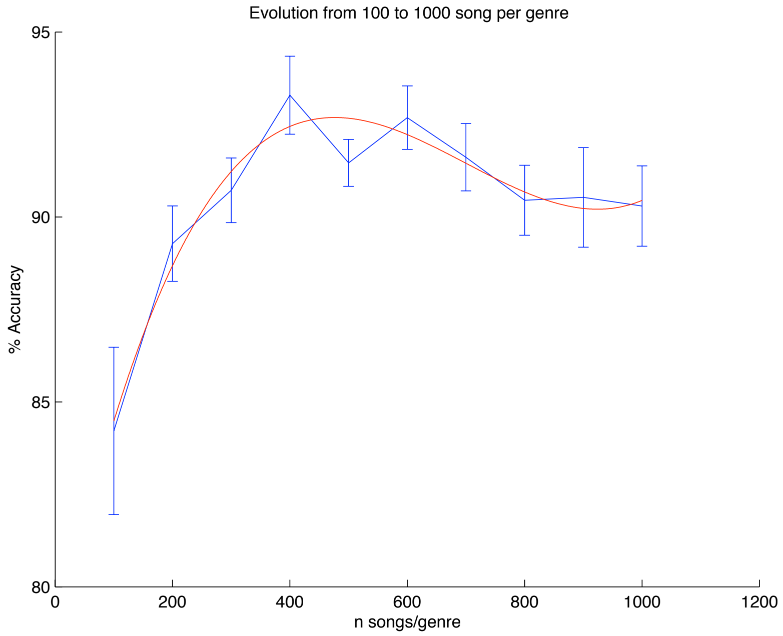


Figure 5.17: Mean of evolution from 100 to 1000 song for all genres. The real measured values are shown in blue and the 3rd. order polynomial regression are shown in red.

Label	# songs
Aggresive	132
happy	109
sad	133
relax	224

Table 5.39: Number of instances per category in the mood dataset

Unfortunately, we have no access to the dataset used in the MIREX. Our comparison is based on in-house dataset collected by Cyril Laurier, participant and organizer of the MIREX contest (Laurier & Herrera, 2007) and ranked in second place in that event. The algorithm is considered to be one of the state of the art. The authors use a set of 133 descriptors and a Support Vector Machine classifier to predict the mood. The features are spectral, temporal, tonal but also include loudness and danceability. The SVM classifier is optimized using a grid search algorithm. The dataset is created by 4 moods distributed as shown in Table 5.39.

Due to that dataset is not splitted in train and test, we performed 10 experiments with different 66−33% random separation of the train and test datasets. Then, we provided the Mean and the Std accuracies for all the experiments. We performed two experiments: a) using the 4 proposed categories and b) using only three of them (skipping the Relax category) because of the overlap

	SVM		SIMCA	
# of categories	Mean	StdDev	Mean	StdDev
4	62.8	2.2	93.1	2.0
3	81.3	3.7	95.0	2.1

Table 5.40: Obtained accuracies for the SVM and SIMCA classifiers using the mood dataset

Category	Subcategory(# of files)
Western (1133)	Alternative(10), Blues(110), Classical(103), Country(110), Disco(100), Electronica(10) Folk(10), Funk(10), Hiphop(110), Jazz(110), Metal(110), Pop(110), Reggae(100), Religious(10), Rock(110), Soul(10).
non-Western (508)	Africa(74), Arabic(72), Centralasia(87), China(63), India(74), Japan(40), Java(98).

Table 5.41: Overview of the dataset used for Western/non-Western classification

from this category with the other ones. The SIMCA has been configured to perform PCAs covering the 90% of the variance. The obtained accuracies are shown in Table 5.40.

Results show how SIMCA clearly performs better classification than the state of the art algorithm using SVM: differences up to 30 and 15 points are found for 4 and 3 categories respectively. Then, the use of a SIMCA classifier is a good option for mood classification. We encourage the experts on this field to perform further research on this area.

Western vs non-Western music

In this section we present a short test concerning the Western/non-Western music classification. One of the particularities of these two big groups is the use of an equal tempered scale against many others (we will not take into account early music and other non tempered compositions although they have been composed in the *western* influence region). Assuming that, we will use the tonal descriptors proposed by Gomez (2006) described in Section 4.3.5.

The dataset was manually collected using audio files from the existing datasets and personal collections from the researchers in our lab, summarized in Table 5.41.

The idea is to compare results between SVM and SIMCA classifiers. Because of our system is able to classify balanced datasets and to evaluate using the Holdout technique (66%-33%), we propose to compute results as the mean of 5 random subsamples of the Western subset, applying the Holdout evaluation technique for each case. Under these conditions, we perform classification using SVMs and the SIMCA classifier providing the results shown in Table 5.42.

Just for fun, we also compute country classification for the non-western

	SVM	SIMCA
Western/non-Western	82.8	97.0
Country	38.5	53.0

Table 5.42: Comparison of obtained accuracies for Western/non-Western and Country classification using SVMs and SIMCA

subset, randomly balancing the dataset to 40 instances and computing the mean of the results for both SVMs and SIMCA classifiers, as described above. Results are also shown in Table 5.42

Here again, SIMCA classifier improves the performance obtained by using SVMs up to a 15% for Western-nonWestern classification. Although results in Country classification are not good, SIMCA also improves the accuracy about another 15% with respect to SVMs. We encourage the experts on this field to perform further research.

5.7.4 Conclusions

In this section, we have presented an extensive comparison of SVM and SIMCA classifiers in different environments. First, we have studied the behavior of the SIMCA classifier dealing with a unique dataset. Second, we have studied its behavior in front of completely independent datasets in order to study whether it is able to deal with real scenarios. Then, we have proposed a brief study on scalability and, finally, we have presented a set of experiments dealing with completely new problems.

There exists a special behavior in Table 5.36 in both datasets for tonal descriptors: while SVM provide low accuracies (42.9% and 39.2% for Radio and Tzanetakis datasets respectively) more reasonable results are obtained using Simca (83.3% and 96.6%). We assume this the consequence of the single class classifiers philosophy: although tonal descriptors are not the best ones for genre classification, the separation of the problem into a set of multiple individual problems increases the accuracy (each classifier can be focused on its own problem).

Furthermore, some interesting results are found in Table 5.37. As explained in Section 3.3.1, the STOMP, Radio and Tzanetakis databases are built using 10, 50 and 100 songs per genre respectively. Training with Radio, acceptable results are obtained when testing with STOMP but not testing with Tzanetakis. This can be explained by the different sizes of the datasets: STOMP is five times smaller than Radio. Thus, the classifier trained with Radio database can predict STOMP but it can't predict Tzanetakis, which is two times bigger. On the other hand, the classifier trained with Tzanetakis can reasonably explain both STOMP and Radio databases, which are smaller by a factor of ten and two respectively. As expected, for cross experiments, SIMCA classifier performs good classification when the size of the train dataset is big enough to extract the essence of musical genres. In particular, as described in Section 5.7.2, the number of instances per category that maximizes the results is near 500 for train and test, which corresponds to 333 instances for training. The Tzanetakis dataset consists in 100 instances per genre, which is about one third of the ideal case. Having a look to Figure 5.17, we observe that this number is far from the ideal case, but this effect is compensated by a lower number of categories (only 5).

Roughly speaking, Simca classifier provides better classification results than SVM for individual and mixed datasets. As deduced in Section 5.6, the use of individual and independent attribute selection for each category allows better performances.

On the other hand, in Table 5.37 we have presented accuracies over the 99%. Initially, we though they were produced by the over-fitting in the classifier. As we are using only rhythm and timbre descriptors, the ratio between instances and accuracies seem reasonable and all the rest of experiments, including those mixing datasets, are in a similar range. The reader is reminded that these values are obtained using a specific dataset and splitting it by random points. As the main dataset we use has not the list of songs/artists available, we assume the main reason for these extremely high results is because of the dataset. Then, these values higher than 99% should be interpreted as *quite good results* without taking the specific number as a reference.

To conclude, we observe that best results are obtained using spectral and rhythmic descriptors, or spectral and tonality. Then, we assume that spectral, rhythm and tonality related descriptors can better describe musical genres, and the use of Simca classifier may be the technique that can better classify them.

CHAPTER 6

Conclusions and future work

6.1 Introduction

In this thesis we proposed an exhaustive study on automatic music genre recognition. First, we introduced some theoretical concepts about music genre such as taxonomies, the most extended theories on human categorization, and the concept of music content processing. We also introduced the state of the art of automatic music genre classification and compared different successful approaches. We also dedicated some pages for showing the importance of the MIREX evaluation initiative in the whole MIR community. From the technical point of view, we detailed the overall schema for classification, that is (1) selection of the taxonomy and datasets, (2) design and computation of musical descriptors, and (3) algorithms for automatic classification. We described each part individually and we proposed a set of tests by manipulating these three variables.

Moreover, we have been interested in the meaning of each part of the process too. Descriptors are usually representative of a specific musical facet (timbre, rhythm, etc.). We introduced new families of descriptors that traditionally have not been exploited for genre classification (panning, complexity, tonal). Rhythmic descriptors were developed by the author in order to achieve a coherent rhythmic representation also in those cases where rhythm was no present (some excerpts of classic music, speech, etc). This is a good example of one of the aims of this work which is to study the behavior of classifiers in a wide range of situations. Detailed tests show how timbre and rhythm features provide the best classification results for a generic classification of musical genres. This doesn't mean that all the other descriptors we tested are useless. They can (and need to) be computed for some specific classification problems (ballroom, speech/music, etc.) but this is out of the scope of this thesis.

Next, we to compared results from classification using a single dataset (that share some implicit information as codification, sampling rate, origin of some files) against the combination of two or more of them (between-collection gen-

Algorithm	Classical	Prototype	Exemplar
Dedicated	x		
Decision Trees	x		x
GMM		x	
HMM		x	
SVM			x
K-NN			x

Table 6.1: Associated categorization theory to the most important machine learning techniques

eralization). The big differences in the presented scenarios drove us to detect and analyze possible gaps in the application of the current state of the art machine learning algorithms. If selected descriptors and classification techniques were able to extract the essence of musical genres, results should not depend so much on the selected dataset. We addressed some efforts on this problem, as explained above.

Finally, we analyzed results of using different machine learning techniques. The state of the art proved that, generally, the best classification results are obtained using Support Vector Machines (SVM). But SVMs are not able to extract the essence of musical genres as shown in the experiments performed by mixing datasets. This, in addition to the results obtained by a set of listening experiments, lead us to develop a new approach for classification. The Soft Independent Modeling of Class Analogies (SIMCA) method here presented gathers all the required conditions to perform classification accomplishing all of our requirements. We showed results obtained using this technique and compared them to the state of the art methodologies.

In this chapter, we present the main conclusions for the overall study. Then, we summarize the contributions presented in this work and present some ideas for future work.

6.2 Overall discussion

In Chapter 3, we presented different categorization theories and we promised to contrast them with the different classification techniques used for automatic genre classification. Table 6.1 shows this association for the most well known algorithms.

Decision Trees can be interpreted as a representation of the classical or exemplar based theory depending on whether the decisions are musicologically meaningful or not, respectively. Support Vector Machines is the general purpose machine learning technique that provide best results in different environments (not only in audio genre classification but in other MIR-specific problems), as described in Section 5.5. Under our point of view, SVMs can be associated to the exemplar based theory where, in fact, only some of the exemplars are used to define the separation hyperplanes. At this point, we wonder how these methods can extract the real essence of musical genres. We don't want to open the never-ending discussion about how humans categorize musical genres, but it seems clear that, using exemplar based methods, this

Train	Descriptor	Split	STOMP		Other	
			SVM	SIMCA	SVM	SIMCA
Radio (5)	Spectral	94.8	72.0	96	65.3	68.6
	Rhythm	88.3	68.0	94	51.3	62.8
	Tonality	92.9	42.0	100	32.0	81.5
	Complexity	62.8	54.0	56	40.2	32.7
	Spectral+Rhythm	98.6	75.5	98	66.9	72.9
	Spectral+Tonality	94.6	78.0	100	58.8	75.7
	Spectral+Complexity	95.0	74.0	98	76.5	50.1
Tzanetakis (5)	Spectral	**97.4**	**74.0**	**98**	**78.9**	**99.2**
	Rhythm	**89.0**	**54.0**	**66**	**52.2**	**80.0**
	Tonality	**98.7**	**30.0**	**86**	**26.8**	**82.9**
	Complexity	57.3	64.0	20	50.6	20.4
	Spectral+Rhythm	**99.2**	**58.0**	**94**	**77.7**	**95.9**
	Spectral+Tonality	**99.4**	**55.1**	**92**	**55.8**	**92.2**
	Spectral+Complexity	98.9	74.0	52	78.1	55.9

Table 6.2: Comparison of SVM and SIMCA classifiers for mixing databases experiments. The *Split* accuracies are presented as the mean of accuracies for 10 experiments with different random splits for train and test subsets. The *Other* dataset corresponds to *Tzanetakis* when training with *Radio* and viceversa.

essence is not modeled by the computer, but only their boundaries. In our experiments, we have observed how, using SVM classifier and crossing datasets, the accuracies fall about 20% in comparison to the 10-fold cross-validation for single collections (See Table 5.29). If SVM's were able to model the essence of musical genres, this decrease on the accuracies should be lower.

Let's discuss the results of the listening experiments presented in Section 5.3. They suggest that, for humans, it is easier to distinguish a given excerpt that does not belong to a specific musical genre than the given excerpts that do. This fact drives us to think about other ways of classification, and we found that the SIMCA method perfectly accomplishes our requirements. In addition to the independent 1-against-all classification for all the given categories, what is interesting in SIMCA is the fact that, by using PCA, we reinforce the relevant descriptors while discarding some non relevant information. Transforming data to the new feature space (and what is more important, performing independent transformations for each category) and comparing them with the residual variances of the model, we are measuring a distance from a new instance to a given category: the residual variance of the non-targeted category can be interpreted as a direct measure of a given expert to the targeted category.

Conceptually, this classifier uses only the relevant information (linear combination of best audio descriptors) that best explains this category which is, under our point of view, coherent with the definition of a prototype. We can compare the sets of experiments mixing datasets for SIMCA and SVM classifiers, described in Section 5.7.2 and Section 5.5.5 respectively. This comparison is shown in Table 6.2. The use of Tzanetakis dataset for training (100 songs/genre) and STOMP or Radio dataset for testing (10 or 50 songs/genre respectively), and the SIMCA classifier, produces lower differences between ac-

Database	#Genres	Descriptor	SVM	SIMCA
Tzanetakis	10	Spectral	89.0	95.8
		Rhythm	54.6	79.5
		Tonality	39.2	96.6
		Complexity	31.2	37.1
		Spectral+Rhythm	89.2	97.2
		Spectral+Tonality	88.8	99.0
		Spectral+Complexity	89.5	99.0

Table 6.3: Results for SVM and SIMCA classifiers for Tzanetakis dataset and for different sets of descriptors, presented as the mean of accuracies for 10 experiments with different random splits for train and test subsets.

curacies than using the SVM classifier[1]. Then, we conclude that SIMCA is a classification technique that more accurately models the essence of the musical genres in a sill far but similar way that humans do.

On the other hand, we also tested many different descriptors. We compared frame-based descriptors against collapsed descriptors for the whole audio excerpt using basic statistics (mean, standard deviation, skewness and kurtosis). Only panning descriptors showed to perform better with a frame-based description with respect to the collapsed descriptors (mean, standard deviation, skewness and kurtosis), but we have discarded the inclusion of such amount of data due to practical issues (computational constrains, availability of stereo datasets, etc.). Complexity and Band Loudness Intercorrelation showed interesting results, but it is not possible to extend these results to other datasets and configurations for genre classification. As seen in Table 6.2, spectral descriptors, and the combination of spectral descriptors with rhythmic or tonal descriptors, showed the best accuracies. These combinations provide the best results for SVM classifiers as well as for SIMCA classifiers (See Table 6.3) for all the tested scenarios.

Our participation in MIREX-2007 was designed to verify how a traditional and simple approach for genre classification deals with unknown data. The main idea was to establish a baseline for all our experiments and allow to compare results obtained by the SIMCA classifier with respect to the approaches of other authors. Our MIREX approach, based on timbre and rhythm descriptors, and a SVM classifier, obtained an overall accuracy of 71.87% for hierarchical classification and 62.89% for raw classification. The best presented approach obtained 76.56% and 68.29% accuracies for hierarchical and raw classification respectively. Both results are about a 5% above our state of the art implementation. Then, assuming a linear relationship between accuracies (which could be debatable), our proposed classifiers should increase about this 5% the accuracy obtained by our SVM approach, using our own datasets and descriptors, which is to be tested, probably, in MIREX2009.

As discussed previously in Section 5.7.4, the 99% accuracy is shocking, but results obtained using the SIMCA classifier provide a higher accuracy with re-

[1] As seen in the tests for scalability (See Section 5.7.2), we suggest to use a minimum of 100 songs/genre for training to extract a correct model. As a consequence of that, we do not take into account results of Table 6.2 training with the Radio dataset

spect to the traditional approaches based on SVMs in most of the presented scenarios. As discussed in Section 3.3.2, some authors reported the presence of a glass ceiling in genre classification (Aucouturier & Pachet, 2004; Pampalk et al., 2005b). In our opinion, this glass ceiling should be comparable to the accuracies obtained by exhaustive human experiments, which unfortunately are not available. Taking the scalability experiments as example, the dataset of which is fixed to 500 songs for each one of the 12 musical genres, it seems reasonably to assume that an expert could classify these 6000 audio excerpts according to the given taxonomy, providing high accuracies. So, the glass ceiling of automatic classifiers should be comparable to this accuracy obtained by the experts (I guess it could be greater than 95%). If automatic classification does not reach this values, maybe the classifier or the descriptors are not properly selected. In other words, if a SVMs classifier using timbre and spectral descriptors is not able to improve an accuracy about 70% (See our contribution to MIREX 2007 with a flat classifier) it is not because the automatic genre classification has a natural glass ceiling related to the complexity or structure of the problem. This is, otherwise, because of new descriptors and classification techniques need to be developed and combined. Under our point of view, SIMCA can be a starting point for a new concept of audio classifiers that better capture human classification.

6.3 Summary of the achievements

Proposal of new descriptors for automatic genre classification. Traditionally, genre classifiers are based on spectral and rhythmic descriptors of audio. While spectral features of audio are well represented by MFCC, spectral centroid, spectral flatness and many others, there is a variety of rhythm descriptors available in the literature. As explained in Section 4.3.4, some of them fail or are redundant describing non rhythmic audio (speech, classical music, etc.). Here, we propose some new descriptors which imitate the cepstrum representation but in the so called *rhythm domain*.

Evaluation of other non traditional descriptors that have not been previously used in automatic genre classification. In addition to timbre and rhythm descriptors, we analyze classification results using THPCP descriptors proposed by Gomez & Herrera (2008), complexity descriptors proposed by Streich (2007), Panning descriptors proposed by Gómez et al. (2008) and Band Loudness Intercorrelation descriptors proposed by Aylon (2006). Tonal features seem to provide meaningful information to the system, as shown in Table 6.2 and Table 6.3.

Generalization of the classifier in front of different datasets, number of instances per category, number of categories, etc. Our idea is that a good classifier must be able to capture the essence of a given category and, as a consequence of that, it should be robust to changes produced by its application to different scenarios. Classifiers only reaching high accuracy for specific environments are not assumed to be good classifiers.

Behavior of classifiers in different situations: This is the mix of the items presented above. First, we have compared different techniques (Decision trees, SVM, etc.) by using only one dataset and 10-fold cross validation and combining different families of descriptors. Support Vector Machines have proved to provide the best results. Then, we use SVMs to compare results when mixing datasets and, finally, with the best (conceptual and numerical) combinations, we compare results with our proposed SIMCA classification method.

Identification of the machine learning techniques with the categorization theories in order to design a "human-like" classifier. The adoption of SIMCA is the result of the interpretation of many experiments focused on the study of the confusions, accuracies and the structure of classifiers.

Extension of SIMCA classifier to other problems of Music Information Retrieval. We have tested SIMCA classifiers with other problems of MIR such as western/non-western and mood classification. We have not contrasted these tests with the actual state of the art, so, results can not be compared with other approaches. We designed these experiments just to test that SIMCA can be a good technique to be applied to other music classification problems.

6.4 Open Issues

There are many open research paths in automatic genre classification that derive from this dissertation. In general, the community is focusing on the research of new descriptors with more semantic meaning that should be included in our study. In a similar way, new classification techniques including non linear processing and fuzzy logics are gaining attention in papers and presentations. As new descriptors and classifiers are developed to solve MIR problems, they should be included in the genre classification task and compare their results with other existing approaches.

Concerning our study, we draw some specific future work that should be followed by the community

Detailed analysis of the confusions: It is quite interesting to listen to the misclassified audio excerpts. In some of the cases, it could be discussed whether the groundtruth is correct or not (p.e. a *pop* song that, under the musicological point of view, could be labelled as *funk*). Other cases show misclassifications that could not be explained. There are many confusions within audio excerpts with high spectral content (*metal, rock*) but the inclusion of rhythmic descriptors, among others, minimize them. Finally, there are some *acceptable* confusions such as *blues* vs. *jazz*, *pop* vs *rock*, etc. The detailed study of these confusions could be useful to (1) improve the accuracy of classifiers, (2) help musicologists to find specific properties of musical genres, (3) explain the evolution of musical genres, and (4) find for a systematic method for doing that.

Listening experiments: Listening experiments presented in this dissertation need to be extended to a bigger audience. We also propose a deeper analysis of the results in which, for instance, we can correlate the familiarity degree of the presented musical genres (detailed in a questionnaire) with the response time in the experiment. Moreover, the number of examples and categories must be extended, and other descriptors should also be included.

Inclusion of new genres: In this study we have not taken into account how new musical genres are included in the proposed system. New genres may require to be included because of different reasons: (1) they were not included initially in the system because of requirement constrains or (2) new genres appear as a subset or of an existing one, or as a fusion of two existing genres. In both cases, as the SIMCA classifier uses a *1-against-all* strategy, the whole system has to be retrained with the new category. As argued in the previous chapters, we try to design a classifier that, somehow, mimics the human behavior in front of genre classification task. This means that SIMCA technique is still far from simulating brain processes. Variations of SIMCA or the research of new techniques is still required. In our opinion, these techniques should include online learning principles.

Hierarchy classifier: Related to the problem of the inclusion of new musical genres detailed above, the SIMCA technique is not able to deal with hierarchical classification. The main problem is that training a *parent* category and their *sons* at the same time create a big confusion in the *all* subset for each category (remember that SIMCA uses the *1-against-all* strategy). It is possible that the parent category does not include the audio excerpts from the sons into the dataset, or that the not belonging audio excerpts of the son category includes the audio excerpts from his father. From now on, the unique solution is the manual creation and labeling of the datasets.

Inclusion of metadata: This thesis is focused in music content processing. Nevertheless, the musical genre is a social and cultural phenomena. So, the results here obtained should be complemented with other available metadata such as tags, dates, etc.

6.5 Final thoughts

Personally, I would like to think this thesis can be a good reference for those who are interested in audio classification. Not only musical genre or the methodologies here presented can be interesting, but also all the comparisons of classification using different descriptors and datasets, and their multiple combinations. I have compared the size of datasets, their overlap and how classification techniques are able to extract the essence of the categories. I also presented some listening experiments that support the decisions I took in the whole process.

It was no my intention to find the best classifier for a specific problem. I wanted to present a broad point of view in audio classification using general techniques. I also wanted to focus on the music instead of the algorithms themselves. In my opinion, simple things work better. The only problem is to really know the problem and how to combine the available pieces to solve it.

My best reward is that, after reading this thesis, the reader keeps thinking on how to improve one of their algorithms inspired by one idea here presented. If so, good luck!

Bibliography

Ahrendt, P. (2006). *Music Genre Classification Systems - A Computational Approach.* Tesi Doctoral, Technical University of Denmark.

Ahrendt, P., Larsen, J., & Goutte, C. (2005). Co-occurrence models in music genre classification. *Machine Learning for Signal Processing, 2005 IEEE Workshop on*, ps. 247–252.

Ahrendt, P. & Meng, A. (2005). Music genre classification using the multivariate ar feature integration model. Dins *Proc. ISMIR - Mirex.*

Ahrendt, P., Meng, A., & Larsen, J. (2004). Decision time horizon for music genre classification using shorttime features. Dins *Proc. EUSIPCO*, ps. 1293–1296.

Albano, C., Dunn III, W., Edlund, U., Johansson, E., Nordenson, B., Sjostrom, M., & Wold, S. (1978). Four levels of pattern recognition. *Anal. Chim. Act.*, 103:429–443.

Allen, P. & Dannenberg, R. (1990). Tracking musical beats in real time. Dins *Oric. ICMC.*

Alpaydin, E. (2004). *Introduction to Machine Learning.* Cambridge, Massachusetts: The MIT Press.

Ashby, F. G. & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, 37:372–400.

Aucouturier, J. & Pachet, F. (2003). Representing musical genre: A state of art. *Journal of New Music Research*, 32(1):83–93.

Aucouturier, J. J. & Pachet, F. (2004). Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1).

Aylon, E. (2006). *Automatic Detection and Classification of drum kit sounds.* Projecte F. de Carrera o Tesina de L., Universitat Pompeu Fabra, Barcelona.

Bagci, E., U.; Erzin (2005). *Lecture Notes in Computer Science*, volum 3733, capítol Boosting Classifiers for Music Genre Classification, ps. 575–584. Springer-Verlag.

Bagci, U. & Erzin, E. (2006). Inter genre similarity modelling for automatic music genre classification. *Proc. DAFx*, ps. 153–156.

Barry, D., Lawlor, B., & Coyle, E. (2004). Sound source separation: Azimuth discrimination and resynthesis. Dins *Proc. of the 7th International Conference on Digital Audio Effects*.

Barthélemy, J. & Bonardi, A. (2001). Figured bass and tonality recognition. Dins *Second International Symposium on Music Information Retrieval*, ps. 129–136. Bloomington, USA.

Bartlett, M., Movellan, J., & Sejnowski, T. (2002). Face recognition by independent component analysis. *IEEE Trans. on Neural Networks*, 13(6):1450–1464.

Basili, R., Serafini, A., & Stellato, A. (2004). Classification of musical genre: A machine learning approach. Dins *Proc. of ISMIR*.

Basili, R., Serafini, A., & Stellato, A. (2005). Extracting music features with midxlog. Dins *Proc. ISMIR*.

Belhumeur, P., Hespanha, J., & Kriegman, D. (1996). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. Dins *Proc. of the Fourth European Conference on Computer Vision*, volum 1, ps. 45–58.

Berenzweig, A., Logan, B., Ellis, D., & Whitman, B. (2004). A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, 28(2):63–76.

Bergstra, J. (2006). Meta-features and adaboost for music classification. *Machine Learning*.

Bergstra, J., Casagrande, N., & Eck, D. (2005). Two algorithms for timbre and rhythm based multiresolution audio classification. Dins *Proc. ISMIR - Mirex*.

Bonardi, A. (2000). Ir for contemporary music: What the musicologist needs. Dins *International Symposium on Music Information Retrieval*. Plymouth, USA.

Brackett, D. (1995). *Intepreting Popular Music*. New York: Canbridge University Press.

Breiman, L. & Cutler, A. (2003). Random forests manual. Report tècnic, UC Berkeley.

Brooks, L. (1978). Nonanalytic concept formation and memory for instances. *Cognition and Categorization*, ps. 169–211.

Buldyrev, S. V., Goldberger, A. L., Havlin, S., Mantegna, R. N., Matsa, M. E., Peng, C.-K., Simons, M., & Stanley, H. E. (1995). Long-range correlation properties of coding and noncoding dna sequences: Genbank analysis. *Physical Review E*, 51(5):5084–5091.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Burred, J. & Lerch, A. (2003). A hierarchical approach to automatic musical genre classification. Dins *Proc. DAFx-03.*

Burred, J. J. (2005). A hierarchical music genre classifier based on user-defined taxonomies. Dins *Proc. ISMIR - Mirex.*

Cai, R., Lu, L., Zhang, H., & Cai, L. (2004). Improve audio representation by using feature structure patterns. Dins *IEEE Proc. of ICASSP.*

Cambridge International Dictionary (2008). content analysis. Http://dictionary.cambridge.org.

Cataltepe, Z., Yaslan, Y., & Sonmez, A. (2007). Music genre classification using midi and audio features. *EURASIP Journal on Advances in Signal Processing.*

Celma, O. (2006). *Music Recommendation: a multi-faceted approach.* Tesi Doctoral, Universitat Pompeu Fabra.

Chai, W. & Vercoe, B. (2001). Folk music classification using hidden markov models. Dins *Proc. of International Conference on Artificial Intelligence.*

Chase, A. (2001). Music discriminations by carp. *Animal Learning & Behavior*, 29(4):336–353.

Cook, P. (1999). *Music, Cognition, and Computerized Sound: An Introduction to Psychoacoustics.* The MIT Press.

Cope, D. (2001). Computer analysis of musical allusions. Dins *Second International Symposium on Music Information Retrieval*, ps. 83–84. Bloomington, USA.

Craft, A., Wiggins, G., & Crawford, T. (2007). How many beans make five? the consensus problem in music-genre classification and a new evaluation method for single-genre categorisation systems. Dins *Proc. ISMIR.*

Crump, M. (2002). *A principal components approach to the perception of musical style.* Projecte F. de Carrera o Tesina de L., University of Lethbridge.

Cunningham, S. J., Bainbridge, D., & Falconer, A. (2006). More of an art than a science': Supporting the creation of playlists and mixes. Dins *Proc. ISMIR.*

Cunningham, S. J., Jones, M., & Jones, S. (2004). Organizing digital music for use: An examination of personal music collections. Dins *Proc. ISMIR.*

D., R. A. & L., B. J. (1992). *Wavelets.* Acta Numer., pages 1-56. Cambridge Univ. Press, Cambridge.

Dalla (2005). Diferentiation of classical music requires little learning but rhythm. *Cognition*, 96(2).

Dannenberg, R., Thom, B., & Watson, D. (1997). A machine learning approach to musical style recognition. Dins *Proc. ICMC.*

Dannenberg, R. B. (2001). Music information retrieval as music understanding. Dins *Second International Symposium on Music Information Retrieval*, ps. 139–142. Bloomington, USA.

Dannenberg, R. B. & Hu, N. (2002). Pattern discovery techniques for music audio. Dins *Third International Conference on Music Information Retrieval*, ps. 63–70. Paris, France.

De Coro, C., Barutcuoglu, Z., & Fiebrink, R. (2007). Bayesian aggregation for hierarchical genre classification. Dins *Proc. ISMIR*.

Deliáege & Sloboda, J. A. (1997). Perception and cognition of music. *Hove, East Sussex: Psychology Press.*

Dixon, S. (2001). Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30(1).

Dixon, S., Gouyon, F., & Widmer, G. (2004). Towards characterisation of music via rhythmic patterns. Dins *Proc. ISMIR*.

Downie, S. (2003). Music information retrieval. *Annual Review of Information Science and Technology*, 32:295–340.

Draper, B., Baek, K., Bartlett, M., & Beveridge, J. (2003). Recognizing faces with pca and ica. *Computer Vision and Image Understanding (Special Issue on Face Recognition)*, 91(1-2):115–137.

Duff, D., editor (2000). *Modern Genre Theory*. New York: Longman.

Ellis, D. (2007). Classifying music audio with timbral and chroma features. Dins *Proc. ISMIR*.

Epstein, J. (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. The MIT Press.

Eriksson, L., Johansson, E., Muller, M., & Wold, S. (2000). On the selection of the training set in environmental qsar analysis when compounds are clustered. *J. Chemometrics*, 14:599–616.

Esmaili, S., Krishnan, S., & Raahemifar, K. (2004). Content based audio classification and retrieval using joint time-frequency analysis. Dins *IEEE Proc. of ICASSP*.

Fabbri, F. (1981). A theory of musical genres: Two applications. *Popular Music Perspectives*.

Fabbri, F. (1999). Browsing music spaces: Categories and the musical mind.

Fields, K. (2007). Ontologies, categories, folksonomies: an organised language of sound. *Organised Sound*, 12(2):101–111.

Fingerhut, M. & Donin, N. (2006). Filling gaps between current musicological practice and computer technology at ircam. Report tècnic, IRCAM.

Fletcher, H. (2940). Auditory patterns. Dins *Rev. Mod. Phys*, volum 12, ps. 47–65.

Flexer, A., Pampalk, E., & Widmer, G. (2005). Hidden markov models for spectral similarity of songs. Dins *Proc. of the 8th. Conference on Digital Audio Effects, DAFx'05.*

Foote, J., Cooper, M., & Nam, U. (2002). Audio retrieval by rhythmic similarity. Dins *Proc. ISMIR.*

Foote, J. & Uchihashi, S. (2001). The beat spectrum: A new approach to rhythm analisis. Dins *Proc. International Conference on Multimedia and Expo.*

French, J. C. & Hauver, D. B. (2001). Flycasting: On the fly broadcasting. *Proceedings of the WedelMusic Conference.*

Freund, Y. & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Fujinaga, I., Moore, S., & Sullivan, D. (1998). Implementation of exemplar-based learning model for music cognition. Dins *Proc. of the International Conference on Music Perception and Cognition*, ps. 171–179.

Futrelle, J. & Downie, S. (2003). Interdisciplinary research issues in music information retrieval: Ismir 2000-02. *Journal of New Music Research*, 32(2):121–131.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Collaborative filtering to weave and information tapestry. *Communications of the ACM*, 35(12):61–70.

Gomez, E. (2006). *Tonal Description of Music Audio Signals.* Tesi Doctoral, Universitat Pompeu Fabra.

Gomez, E. (2007). Comparative analysis of music recordings from western and non-western traditions by automatic tonal feature extraction and data mining. Report tècnic, Universitat Pompeu Fabra, Barcelona.

Gomez, E. & Herrera, P. (2008). Comparative analysis of music recordings from western and non-western traditions by automatic tonal feature extraction. *Empirical Musicology Review*, 3(3).

Gómez, E., Herrera, P., Cano, P., Janer, J., Serrà, J., Bonada, J., El-Hajj, S., Aussenac, T., & Holmberg, G. (2008). Music similarity systems and methods using descriptors. United States patent application number 12/128917.

Goto, M. (2004). Development of the rwc music database. *Proc. 18th. International Congress on Acoustics.*

Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2003). Rwc music database: Music genre database and musical instrument sound database. Dins *Proc. ISMIR.*

Goutte, C. (1997). Note on free lunches and cross-validation. *Neural Computation*, 9:1211–1215.

Gouyon, F. (2003). *Towards Automatic Rhythm Description of Musical Audio Signals. Representations, Computational Models and Applications.* Tesi Doctoral, Universitat Pompeu Fabra, Barcelona.

Gouyon, F. & Dixon, S. (2004). Dance music classification: A tempo-based approach. Dins *Proc. of ISMIR.*

Grant, B., editor (2003). *Film Genre Reader III.* Austin TX: University of Texas Press.

Grimalidi, M. (2003). Classifying music by genre using a discrete wavelet transform and a round-robin ensemble. Report tècnic, Trinity College, University of Dublin.

Gruber, T. (2007). Folksonomy of ontology: A mash-up of apples and oranges. *International Journal on Semantic Web & Information Systems*, 3(2).

Gruber, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199–220.

Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning.* Hamilton, New Zeland.

Hampton, J. (1993). *Prototype Models of Concept Representation.* Cognitive Science Series. Academic Press, London.

Harb, H., Chen, L., & Auloge, J. (2004). Mixture of experts for audio classification: an application to male female classification and musical genre recognition. *Proc. ICME.*

Heittola, T. (2003). Automatic classifcation of music signals. *Master's Thesis.*

Herrera, P. (2002). Setting up an audio database for music information retrieval benchmarking. Dins *Proc. ISMIR.*

Hintzman, D. (1986). 'schema abstraction' in a multiple-trace-model. *Psychological Review*, 93:411–428.

Hofmann-Engl, L. (2001). Towards a cognitive model of melodic similarity. Dins *Proc. ISMIR.*

Holzapfel, A. & Stylianou, Y. (2007). A statistical approach to musical genre classification using non-negative matrix factorization. Dins *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volum 2, ps. II–693–II–696.

Homburg, H., Mierswa, I., Morik, K., Möller, B., & Wurst, M. (2005). A benchmark dataset for audio classification and clustering. Dins *Proc. ISMIR*, ps. 528–531.

Hsu, C., Chang, C., & Lin, C. (2008). *A Practical guide to Support Vector Classification.* National Taiwan University.

Hu, X. & Downie, S. (2007). Exploring mood metadata: Relationships with genre, artist and usage metadata. Dins *Proc. ISMIR.*

Hubert, M. & Driessen, K. (2004). Fast and robust discriminant analysis. *Computational Statististics and Data Analysis*, 45:301–320.

Huron, D. (2000). Perceptual and cognitive applications in music information retrieval. Dins *Proc. ISMIR*.

Izmirli, O. (1999). Using spectral flatness based feature for audio segmentation and retrieval. Report tècnic, Department of Mathematics and Computer Science, Connectucut College.

Jäkel, F., Schölkopf, B., & Wichman, F. A. (2008). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin and Review*, 15(2):256–271.

Jennings, H. D., Ivanov, P. C., Martins, A. M., da Silva P. C., & Viswanathan, G. M. (2004). Variance fluctuations in nonstationary time series: a comparative study of music genres. *Physica A: Statistical and Theoretical Physics*, 336(3-4):585–594.

Johnston, J. D. (1998). Transform coding of audio signals using perceptual noise criteria. *IEEE Journal on Selected Areas in Communications*, 6(2):314–323.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, second edició.

Karneback, S. (2001). Discrimination between speech and music based on a low frequency modulation feature. *Proc. Eurospeech*.

Kartomi, M. J. (1990). *On Concepts and Classifications of Musical Instruments*. University Of Chicago Press.

Kassler, M. (1966). Toward musical information retrieval. *Perspectives of New Music*, 4:59–67.

Kedem, B. (1986). Spectral analysis and discrimination by zero-crossings. *Proc. of the IEEE*, 74.

Klapuri, A. (2004). *Signal processing methods for the automatic transcription of music*. Tesi Doctoral, Tampere University of Technology, Finland.

Knees, P., Pampalk, E., & Widmer, G. (2004). Artist classification with web-based data. Dins *Proc. of ISMIR*.

Koelsch, S. & Siebel, W. (2005). Towards a neural basis of music perception. *Trends in Cognitive Sciences*, 9(12):578–584.

Kononenko, I. (1995). On biases in estimating multi-valued attributes. Dins *IJCAI*, ps. 1034–1040.

Kornstädt, A. (2001). The jring system for computer-assisted musicological analysis. Dins *Second International Symposium on Music Information Retrieval*, ps. 93–98. Bloomington, USA.

Krumhansl, C. L. (1990). *Cognition Foundations of Musical Pitch. New York: Oxford*. University Press.

Lakoff, G. (1987). *Women, Fire And Dangerous Things. What Categories Reveal about the Mind.* University of Chicago Press.

Lampropoulos, A., Lampropoulou, P., & Tsihrintzis, G. (2005). Musical genre classification enhanced by improved source separation techniques. *Proc. ISMIR.*

Langley, P. (1996). *Elements of Machine Learning.* Morgan Kaufmann Publishers, Inc.

Laurier, C. & Herrera, P. (2007). Audio music mood classification using support vector machine. Dins *Proc. ISMIR.*

Lesaffre, M., Leman, M., Tanghe, K., De Baets, B., De Meyer, H., & Martens, J. P. (2003). User dependent taxonomy of musical features as a conceptual framework for musical audio-mining technology. Dins *Proceedings of the Stockholm Music Acoustics Conference.*

Li, T. & Ogihara, M. (2005). Music genre classification with taxonomy. Dins *Proc. ICASSP*, volum 5, ps. 197–200.

Li, T., Ogihara, M., & Li, Q. (2003). A comparative study on content-based music genre. *Proc. SIGIR.*

Li, T. & Tzanetakis, G. (2003). Factors in automatic musical genre classification of audio signals. Dins *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).* New Paltz, NY.

Lidy, T. & Rauber, A. (2005). Mirex 2005: Combined fluctuation features for music genre classification. Dins *Proc. ISMIR - Mirex.*

Lidy, T. & Rauber, A. (2007). Improving genre classification by combination of audio and symbolic descriptors using a transcription system. Dins *Proc. ISMIR.*

Lidy, T., Rauber, A., Pertusa, A., & Inesta, J. M. (2007). Combining audio and symbolic descriptors for music classification from audio. Dins *Proc. ISMIR - Mirex.*

Lincoln, H. (1967). *Some criteria and techniques for developing computerized thematic indeces.* Electronishe datenverarbeitung in der Musikwissenschaft. Regensburg: Gustave Bosse Verlag.

Lippens, S., Martens, J. P., & De Mulder, T. (2004). A comparison of human and automatic musical genre classification. Dins *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, volum 4, ps. iv–233–iv–236 vol.4.

Livshin, A. & Rodet, X. (2003). The importance of cross database evaluation in sound classification. Dins *Proc. ISMIR.*

Logan, B. (2000). Mel frequency cepstral coeficients for music modeling. Dins *Proc. ISMIR.*

Lu, D., Liu, L., & Zhang, H. (2006). Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing*, 14(1).

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis.* New York, Wiley.

Mandel, M. & Ellis, D. (2005a). Songlevel features and support vector machines for music classification. Dins *Proc. ISMIR*, ps. 594–599.

Mandel, M. & Ellis, D. (2005b). Song-level features and svms for music classification. Dins *Proc. ISMIR - Mirex.*

Mandel, M. & Ellis, D. P. (2007). Labrosa's audio music similarity and classification submissions. Dins *Proc. ISMIR - Mirex.*

Martens, H. A. & Dardenne, P. (1998). Validation and verification of regression in small data sets. *Chemometrics and Intelligent Laboratory Systems*, 44:99–121.

Martin, K. D., Scheirer, E. D., & Vercoe, B. L. (1998). Musical content analysis through models of audition. *Proceedings of the ACM Multimedia Workshop on Content-Based Processing of Music.*

McAuley, J., Ming, J., Stewart, D., & Hanna, P. (2005). Subband correlation and robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13:956–964.

McGuinness, D. L. (2003). Ontologies come of age. Dins D. Fensel, J. A. Hendler, H. Lieberman, & W. Wahlster, editors, *Spinning the Semantic Web*, ps. 171–194. The MIT Press.

McKay, C. & Fuginaga, I. (2004). Automatic genre classification using large high-level musical feature sets. Dins *Proc. of ISMIR.*

Mckay, C. & Fujinaga, I. (2005). The bodhidharma system and the results of the mirex 2005. Dins *Proc. ISMIR.*

Mckay, C. & Fujinaga, I. (2006). Musical genre classification: Is it worth pursuing and how can it be improved? Dins *Proc. ISMIR.*

Medin, D. & Schaffer, M. (1978). Context theory of classification learning. *Psychological Review*, 85:207–238.

Meng, A. (2006). *Temporal Feature Integration for Music Organisation.* Tesi Doctoral, Technical University of Denmark.

Meng, A. (2008). General purpose multimedia dataset - garageband 2008. Report tècnic, Technical University of Denmark.

Meng, A., Ahrendt, P., Larsen, J., & Hansen, L. K. (2007). Temporal feature integration for music genre classification. *Audio, Speech and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on]*, 15(5):1654–1664.

Merriam-Webster Online (2008). content analysis. Http://www.merriam-webster.com.

Mierswa, I. & Morik, K. (2005). Automatic feature extraction for classifying audio data. *Machine Learning Journal*, 58:127–149.

Mills, A. W. (1958). On the minimum audible angle. *The Journal of the Acoustical Society of America*, 30(4):237–246.

Mitchell, T. (1997). *Machine Learning*. The McGraw-Hill Companies, Inc.

Montgomery & Runger (2002). *Probabilidad y Estadistica aplicadas a la ingenieria*. Limusa Wiley.

Murphy, G. & Brownell, H. (1985). Category differentiation in object recognition: Typicality constraints on the basic category advantage. *Experimental Psychology: Learning, Memory, and Cognition*, 11(1):70–84.

Nilsson, N. (1996). *Introduction to Machine Learning*. Department of Computer Science,Stanford University.

Noris, M., Shyamala, D., & Rahmita, W. (2005). Factors affecting automatic genre classification: an investigation incorporating non-western musical forms. *Proc. ISMIR.*

Nosofsky, R. (1986). Attention, similarity and the identification-categorization relationship. *Experimental Psychology: General*, 115:39–57.

Nosofsky, R. (1992). Exemplars, prototypes and similarity rules. *From learning theory to connectionist theory: Essays in honor of William K. Estes*, 1:149–167.

Nosofsky, R. M. & Johansen, M. K. (2000). Exemplar-based accounts of 'multiple-system' phenomena in perceptual categorization. *Psychonomic Bulletin and Review*, 7(3):375–402.

Oppenheim, A. & Schaffer, R. (1989). *Discrete-Time Signal Processing*. Prentice Hall International Inc.

Orio, N. (2006). Music retrieval: A tutorial and review. *Foundations and Trends in Foundations and Trends in Information Retrieval*, 1(1):1–90.

Pachet, F. & Cazaly, F. (2000). A taxonomy of musical genres. Dins *Proc. Dontent-Based Multimedia Information Access.*

Pampalk, E. (2001). Islands of music: Analysis, organization, and visualization of music archives. Dins *Master's thesis, Vienna University of Technology, Vienna, Austria.*

Pampalk, E. (2005). Speeding up music similarity. Dins *Proc. ISMIR - Mirex.*

Pampalk, E., Dixon, S., & Widmer, G. (2003). On the evaluation of perceptual similarity measures for music. *Proc. DAFx.*

Pampalk, E., Flexer, A., & Widmer, G. (2005a). Improvements of audiobased music similarity and genre classification. Dins *Proc. ISMIR.*

Pampalk, E., Flexer, A., & Widmer, G. (2005b). Improvements of audio-based music similarity and genre classification. *Proc. ISMIR*.

Pascall, R. (2007). Style. Grove Music Online ed. L. Macy. Http://www.grovemusic.com.

Peeters, G. (2007). A generic system for audio indexing: Application to speech/music segmentation and music genre recognition. Dins *Proc. DAFx*. Bordeaux.

Peng, C., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., & Goldberger, A. L. (1994). Mosaic organization of dna nucleotides. *Physical Review E*, 49:1685–1689.

Perrot, D. & Gjerdigen, R. O. (1999). Scanning the dial: An exploration of factors in the identification of musical style. *Proceedings of the Society for Music Perception and Cognition*.

Pestoni, F., Wolf, J., Habib, A., & Mueller, A. (2001). Karc: Radio research. Dins *Proceedings WedelMusic Conference*.

Picone, J. W. (1993). Signal modeling techniques in speech recognition. *Proc. of the IEEE*, 81(9).

Platt, J. C., Cristianini, N., & Shawe-Taylor, J. (2000). Large margin dags for multiclass classification. *Advances in Neural Information Processing Systems*, 12:547–553.

Pohle, T. (2005). *Extraction of audio descriptors and their evaluation in Music Classification Tasks*. Projecte F. de Carrera o Tesina de L., Osterreichisches Forschungsinstitut für Artificial Intelligence (OFAI).

Pohle, T., Pampalk, E., & Widmer, G. (2005a). Evaluation of frequently used audio features for classification of music into perceptual categories. Dins *Fourth International Workshop on Content-Based Multimedia Indexing*.

Pohle, T., Pampalk, E., & Widmer, G. (2005b). Generating similarity-based playlists using traveling salesman algorithms. Dins *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx'05)*, ps. 20–22.

Polotti, P. & Rocchesso, D., editors (2008). *Sound to Sense, Sense to Sound: A state of the art in Sound and Music Computing*. to appear.

Ponce, P. J. & Inesta, J. (2005). Mirex 2005: Symbolic genre classification with an ensemble of parametric and lazy classifiers. Dins *Proc. ISMIR*.

Ponce, P. J. & Inesta, J. M. (2007). A pattern recognition approach for music style identification using shallow statistical descriptors. *IEEE Trans. on Systems Man and Cybernetics*, 37(2):248–257.

Porter & Neuringer (1984). Music discrimination by pigeons. *Journal of experimental psychology. Animal behavior processes*, 10(2):139–148.

Pye, D. (2000). Content-based methods for managing electronic music. Dins *Proc. ICASSP*.

Quinlan, R. (1986). Induction of decision trees. *Machine Learning Journal*, 1:81–106.

Rauber, A., Pampalk, E., & Merkl, D. (2002). Using psychoacoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarity. Dins *Proc. ISMIR*, ps. 71–80.

Redner, R. & Walker, H. (1984). Mixture densities, maximum likelihood and the em algorithm. *Society for Industrial and Applied Mathematics Review*, 26(2):195–239.

Rentfrow, P. & Gosling, S. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Pers. Soc. Psychology*, 84(6).

Resnick, P. & Varian, H. R. (1997). Recommender systems. *Communications of the ACM*, 40(3):56–58.

Rich, E. (1979). User modeling via stereotypes. *Cognitive Science: A Multidisciplinary Journal*, 3(4):329–354.

Rizo, D., Ponce, P., Pertusa, A., & Inesta, J. M. (2006a). Melodic track identification in midi files. Dins *Proc. of the 19th Int. FLAIRS Conference.*

Rizo, D., Ponce, P. J., PerezSancho, C., Pertusa, A., & Inesta, J. M. (2006b). A pattern recognition approach for melody track selection in midi files. Dins *Proc. ISMIR*, ps. 61–66.

Rosch, E. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605.

Ruppin, A. & Yeshurun, H. (2006). Midi music genre classification by invariant features. *Proc. ISMIR.*

Samson, J. (2007). Genre. Grove Music Online ed. L. Macy. Http://www.grovemusic.com.

Saunders, J. (1996). Real-time discrimination of broadcast speech/music. *Proc. ICASSP*, ps. 993–996.

Scaringella, N. & Mlynek, D. (2005). A mixture of support vector machines for audio. Dins *Proc. ISMIR - Mirex.*

Scaringella, N. & Zoia, G. (2005). On the modeling of time information for automatic genre recognition systems in audio signals. *Proc. ISMIR.*

Scaringella, N., Zoia, G., & Mlynek, D. (2006). Automatic genre classification of music content: a survey. *Signal Processing Magazine, IEEE*, 23(2):133–141.

Scheirer, E. (1998). Tempo and beat analysis of acoustical musical signals. *J. Acoust. Soc. Am*, 103(1):558–601.

Schroeder, M. R., Atal, B. S., & Hall, J. L. (1979). Optimizing digital speech coders by exploiting masking properties of the human ear. Dins *Journal of the Acoustic Society of America*, volum 66, ps. 1647–1652.

Scott, P. (2001). Music classification using neural networks. Report tècnic, Stanford University.

Shamma, S. (2008). On the role of space and time in auditory processing. *TRENDS in Cognitive Sciences*, 5(8):340–348.

Shardanand, U. & Maes, P. (1995). Social information filtering: Algorithms for automating word of mouth. *Proceedings ACM Conference on Human Factors in Computing Systems*.

Shlens, J. (2002). A tutorial on principal component analysis. Report tècnic, Systems Neurobiology Laboratory, Salk Insitute for Biological Studies.

Silverman, H. F., Yu, Y., Sachar, J. M., & Patterson III, W. R. (2005). Performance of real-time source-location estimators for a large-aperture microphone array. *IEEE Transactions of Speech and Audio Process*, 13(4):593–606.

Smiraglia, R. P. (2001). Musical works as information retrieval entities: Epistemological perspectives. Dins *Second International Symposium on Music Information Retrieval*, ps. 85–91. Bloomington, USA.

Smola, A. & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222.

Soares, V. (2005). Audio artist identification mirex 2005. Dins *Proc. ISMIR - Mirex*.

Soltau, H., Schultz, T., Westphal, M., & Waibel, A. (1998). Recognition of music types. *Proc. ICASSP*.

Sternberg, R. J., editor (1999). *The Nature of Cognition*. The MIT Press.

Stevens, S. S. (1956). Calculation of the loudness of complex noise. Dins *Journal of the Acoustic Society of America*, volum 28, ps. 807–832.

Stevens, S. S. (1962). Procedure for calculating loudness: Mark vi. Dins *Journal of the Acoustic Society of America*, volum 33, ps. 1577–1585.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *J. R. Stat. Soc.*, B(38):44–47.

Streich, S. (2007). *Music Complexity: a multi-faceted description of audio content*. Tesi Doctoral, Universitat Pompeu Fabra.

Streich, S. & Herrera, P. (2005). Detrended fluctuation analysis of music signals: Danceability estimation and further semantic characterization. Dins *Proceedings of the AES 118th Convention*.

Terhardt, E. (1979). Calculating virtual pitch. Dins *Hearing Research*, volum 1, ps. 155–182.

Tolonen, T. & Karjalainen, M. (2000). A computationally efficient multipitch analysis model. *IEEE Transactions on Speech and Audio Processing*, 8(6).

Toynbee, J. (2000). *Making Popular Music: Musicians, Creativity and Institutions.* London: Arnold.

Turk, M. & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86.

Tzanetakis, G. (2007). Marsyas sunmissions to mirex 2007. Dins *Proc. ISMIR - Mirex.*

Tzanetakis, G. & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5).

Tzanetakis, G., Essl, G., & Cook, P. (2001a). Audio analysis using the discrete wavelet transform. Dins *Proc. of WSES International Conference, Acoustics and Music: Theory and Applications (AMTA).*

Tzanetakis, G., Essl, G., & Cook, P. (2001b). Automatic musical genre classification of audio signals. Dins *Proc. ISMIR.*

Tzanetakis, G., Essl, G., & Cook, P. (2002). Human perception and computer extraction of musical beat strength. Dins *Proc. DAFx.*

Tzanetakis, G. & Murdoch, J. (2005). Fast genre classification and artist identification. Dins *Proc. ISMIR - Mirex.*

Uitdenbogerd, A. (2004). A study of automatic music classification methodology. *Proc. ISMIR.*

Uitdenbogerd, A. L. & Zobel, J. (1999). Matching techniques for large music databases. *Proceedings of the 7th ACM International Multimedia Conference*, ps. 57–66.

Vanden & Hubert, M. (2005). Robust classification in high dimensions based on the simca method. *Chemometrics and Intelligent Laboratory Systems*, 79:10–21.

Vapnik, V. (1972). *Estimation of dependences based on empirical data (in Russian).* Nauka, Moskow.

Vapnik, V. (1995). *The nature of statistical learning theory.* Springer-Verlag.

Vassilakis, P. (2001). *Perceptual and Physical Properties of Amplitude Fluctuation and their Musical Significance.* Tesi Doctoral, University of California, Los Angeles.

Vassilakis, P. (2005). Auditory roughness as a means of musical expression. *Selected Reports in Ethnomusicology 12 (Perspectives in Systematic Musicology)*, ps. 119–144.

Vickers, E. (2001). Automatic long-term loudness and dynamics matching. Dins *Proceedings of the AES 111th Convention.*

Vidakovic, B. & Müller, P. (1991). *Wavelets for Kids: a Tutorial Introduction.* Duke university.

Vinyes, M., Bonada, J., & Loscos, A. (2006). Demixing commercial music productions via human-assisted time-frequency masking. Dins *Proc. of AES 120th Convention*.

West, K. (2005). Mirex audio genre classification. Dins *Proc. ISMIR - Mirex*.

West, K. & Cox, S. (2005). Finding an optimal segmentation for audio genre classification. *Proc. ISMIR*.

Whitman, B. & Lawrence, S. (2002). Inferring descriptions and similarity for music from community metadata. *International Computer Music Conference*.

Whitman, B. & Smaragdis, P. (2002). Combining musical and cultural features for intelligent style detection. *International Conference on Music Information Retrieval*.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, second edició.

Wold, E., Blum, T., Keislar, D., & Wheaton, J. (1996). Content-based classification, search and retrieval of audio. *Journal IEEE Multimedia*, 3(3):27–36.

Wold, S. (1976). Pattern recognition by means of disjoint principal components models. *Pattern Recognition*, 8:127–139.

Xu, C. (2005). Automatic music classification and summarization. *IEEE Transactions on Speech and Audio Processing*, 13(3):441–450.

Zhang, T. & Kuo, C. (1999). Hierarchical classification of audio data for archiving and retrieving. Dins *Proc. International Conference on Acoustic, Speech, and Signal Processing*.

Zhao, W., Chellappa, R., & Krishnaswamy, A. (1998). Discriminant analysis of principal components for face recognition. Dins *Proc. of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, ps. 336–341.

Zhu, H. & Rohwer, R. (1996). No free lunch for cross-validation. *Neural Computation*, 8:1421–1426.

Ziv, J. & Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343.

Zwicker, E. & Fastl, H. (1990). *Psychoacoustics - Facts and Models*. Springer.

Zwicker, E. & Terhardt, E. (1980). Analytical expressions for critical-band rate and critical bandwidth as a function of frecuency. *J. Acoust. Soc. Am*, 68(5):1523–1525.

Appendix A: Publications by the author related to the dissertation research

Guaus, E., Herrera, P., Method for audio classification based on Soft Independent Modeling of Class Analogy. Patent in progress.

Sordo, M., Celma, O., Blech, M., Guaus, E. (2008). The Quest for Musical Genres: Do the Experts and the Wisdom of Crowds Agree?. *Proc. of the 9th International Conference on Music Information Retrieval.*

Guaus, E., Herrera, P. (2007). A basic system for music genre classification. *MIREX07.*

Guaus, E., Herrera, P. (2006). Music Genre Categorization in Humans and Machines. *Proc. of the 121th. AES Convention.*

Guaus, E., Herrera, P. (2006). Towards better automatic genre classifiers by means of understanding human decisions on genre discrimination. *International Conference on Music Perception and Cognition.*

Guaus, E., Gomez, E. (2005). Storage and Retrieval of relevant information from music master classes. *Annual Conference of the Association of Sound and Audiovisual Archives (IASA).*

Guaus, E., Herrera, P. (2005). The Rhythm Transform Towards A Generic Rhythm Description. *Proc. of the International Computer Music Conference.*

Cano, P., Koppenberger, M., Wack, N., G. Mahedero, J., Masip, J., Celma, O., Garcia, D., Gomez, E., Gouyon, F., Guaus, E., Herrera, P., Massaguer, J., Ong, B., Ramirez, M., Streich, S., Serra, X. (2005). An Industrial-Strength Content-based Music Recommendation System. *Proc. of the 28th Annual International ACM SIGIR Conference.*

Cano, P., Koppenberger, M., Wack, N., G. Mahedero, J., Aussenac, T., Marxer, R., Masip, J., Celma, O., Garcia, D., Gomez, E., Gouyon, F., Guaus, E., Herrera, P., Massaguer, J., Ong, B., Ramirez, M., Streich, S., Serra, X. (2005). *Content-based Music Audio Recommendation. Proc. of the ACM Multimedia.*

Guaus, E. (2004). New approaches for rhythmic description of audio signals. *Master Thesis.*

Guaus, E., Batlle, E. (2004). A non-linear rhythm-based style classifcation for Broadcast Speech-Music Discrimination. *Proc. of the 116th. AES Convention.*

Batlle, E., Masip, J., Guaus, E. (2004). Amadeus A Scalable HMM-based Audio Information Retrieval System. *Proc. of the 1st. International Symposium on Control, Communications and Signal Processing.*

Guaus, E., Batlle, E. (2003). Visualization of metre and other rhythm features. *Proc. of the 3rd. Internacional Symposium on Signal Processing and Information Technology.*

Batlle, E., Guaus, E., Masip, J. (2003). Open Position Multilingual Orchestra Conductor. Lifetime Opportunity. *Proc. of the 26th ACM/SIGIR International Symposium on Information Retrieval.*

Batlle, E., Masip, J., Guaus, E. (2002). Automatic Song Identification in Noisy Broadcast Audio. *Proc. of the Signal and Image Processing (SIP).*

Appendix B: Other publications by the author

Guaus, E., Bonada, J., Maestre, E., Perez, A., Blaauw, M. (2009). Calibration method to measure accurate bow force for real violin performances. *Proc. of the International Computer Music Conference.*

Perez, A., Bonada, J., Maestre, E., Guaus, E., Blaauw, M. (2008). Measuring Violin Sound Radiation for Sound Equalization . *The Journal of the Acoustical Society of America, vol. 123, issue 5, p. 3665.*

Perez, A., Bonada, J., Maestre, E., Guaus, E., Blaauw, M. (2008). Score Level Timbre Transformations of Violin Sounds. *Proc. of the International Conference on Digital Audio Effects.*

Guaus, E., Bonada, J., Perez, A., Maestre, E., Blaauw, M. (2007). Measuring the bow pressing force in a real violin performance. *International Symposium on Musical Acoustics.*

Maestre, E., Bonada, J., Blaauw, M., Perez, A., Guaus, E. (2007). Acquisition of violin instrumental gestures using a commercial EMF device. *Proc. of the International Computer Music Conference.*

Perez, A., Bonada, J., Maestre, E., Guaus, E., Blaauw, M. (2007). Combining Performance Actions with Spectral Models for Violin Sound Transformation. *Proc. of the Intarnational Conference on Acoustics.*

This thesis has been written using LaTeX