UNIVERSITY OF JYVÄSKYLÄ

# Testing a spectral-based feature set for audio genre classification

by

Martin Ariel Hartmann

Master's Thesis

in the
Faculty of Humanities
Department of Music

June 25, 2011

**JYVÄSKYLÄN YLIOPISTO**

| | |
|---|---|
| Tiedekunta – Faculty<br>**Humanities** | Laitos – Department<br>Music Department |
| Tekijä – Author<br>Martin Ariel Hartmann | |
| Työn nimi – Title<br>Testing a Spectral-based Feature Set for Audio Genre Classification | |
| Oppiaine – Subject<br>Music, Mind & Technology | Työn laji – Level<br>Master's Thesis |
| Aika – Month and year<br>June 2011 | Sivumäärä – Number of pages<br>88 |


Tiivistelmä – Abstract

Automatic musical genre classification is an important information retrieval task since it can be applied for practical purposes such as the organization of data collections in the digital music industry. However, this task remains an open question because the current state of the art shows far from satisfactory outcomes in terms of classification performance. Moreover, the most common algorithms that are used for this task are not designed for modelling music perception. This study suggests a framework for testing different musical features for use in music genre classification and evaluates the performance of this task based on two musical descriptors.

The focus of this study is on automatic classification of music into genres based on audio content. The performance of two sets of timbral descriptors, namely the sub-band fluxes and the mel-frequency cepstral coefficients, is compared. The choice of these particular descriptors is based on their ease or difficulty of interpretation from a perceptual point of view. Classification performance is determined by using a variety of music datasets, learning algorithms, feature selection approaches and combinatorial feature subsets yielded from these descriptors. The results were estimated upon overall classification accuracies, generalization capability, and relevance of these musical descriptors based on feature ranking.

According to the results, the sub-band fluxes, perceptually motivated descriptors of polyphonic timbre, performed better than the widely used mel-frequency cepstral coefficients. The former timbral descriptors showed better classification accuracies and lower tendency to overfit than the latter.

In a nutshell, this study gives support to using perceptually interpretable timbre desciptors for musical genre classification tasks and suggests the utilization of the sub-band flux set for further content-based tasks in the field of music information retrieval.



Asiasanat – Keywords
music information retrieval, music genre classification, polyphonic timbre, feature ranking

Säilytyspaikka – Depository

Muita tietoja – Additional information

# *Acknowledgements*

Many thanks to Petri Toiviainen for an outstanding supervision, day in, day out.
I would also like to thank Vinoo Alluri for the remarks at different stages of this process. Thanks to Olivier Lartillot for his suggestions, for the extraordinary MIRtoolbox and for his technical assistance. I am grateful to Pasi Saari for the mini tutorials and inspiring ideas. I want to thank Rafael Ferrer for his support and advices. Thanks to Tuomas Eerola for the tips and to Geoff Luck for sending me material. Thanks to Suvi Saarikallio for aiding me with my study plan. Thanks to Anemone Van Zijl, Birgitta Burger and Jonna Vuoskoski. Warm thanks to the Music Department at the University of Jyväskylä for offering a top-notch environment for studies.
Thanks to Valeri Tsatsishvili for the great explanations at a very early stage of the study. Alex Reed, Marc Thompson, Shriram Alluri, David Ellison, Sara Kolomainen, Anthony Prechtl: thanks for your support.
Special thanks to enthusiastic researchers in the field of Music Information Retrieval that kindly offered me help when needed: Elias Pampalk, Kris West, George Tzanetakis.
Thanks to Vasyl Pihur for offering enormous technical help.
Thanks to the Music Cognition researchers Silvia Español, Favio Shifres, Isabel Martínez and Pablo Hernán Rodríguez Zivic in Argentina for sharing ideas and for the encouraging comments. Thank you Ezequiel Grimson, thanks Marcelo Delgado, since you have been mentors for me.
Cheers to Pablo Traine, Tuukka Tervo, Pablo Butelman and Luciano Segura, great musical buddies.
Thanks to Eugenio Monjeau, Ángeles Piqué, Juan Martín Michelli, Pablo Vena, Matías Butelman, Ignacio Elguezábal, Gabriela Grinszpun, Julia Hacker, Diego Figueira, Boris Groisman, Julia Klein, María Angélica Cueto, Otto Kronqvist, Atte Hinkka, Maiju Ristkari, Li-Tang Tsai, Salla Kälkäinen and Mikko Norontaus for their presence. And to many others at Colegio Nacional de Buenos Aires, Universidad de Buenos Aires and Jyväskylän Yliopisto for their help in ways that I am not even aware of.
Thanks to Fanny Malinen, Hanna Hakalisto, Noora Mäenpää, Sonja Kekkonen and Waseem Rehmat, amazing work mates. Thanks Carolina Bollero, Merja Mälkki and Irma Hirsjärvi for the excellent work opportunities that were offered while I was writing this thesis. Thanks to Mercedes Krapovickas and Ignacio Maldovan for the hospitality; thanks to Tristana Ferreyra and Teemu Rantalaiho for the stimulus. Thanks to Auli and to her family, for many things. Thanks to my aunts Sofía Merajver (for the love and motivation) and Silvia Hartman (for predicting what I would love to investigate), thanks to Anna Ziff, to David Ziff and to Vera Kornylak. Last but not least, thanks to Alejandra Martín and Mariano Fonticelli...

# Contents

# List of Figures

# List of Tables

# Abbreviations

*Music Databases*

| | |
|---|---|
| **AF.RWC** | **A**rtist **F**iltered - **R**eal **W**orld **C**omputing |
| **GTZAN** | **G**eorge **Tzan**etakis |
| **RWC** | **R**eal **W**orld **C**omputing |

*Audio Descriptors*

| | |
|---|---|
| **MFCC** | **M**el-**F**requency **C**epstral **C**oefficients |

*Combinatorial Feature Subsets*

| | |
|---|---|
| `mfcc` | **m**el-**f**requency **c**epstral **c**oefficients (mean values) |
| `sbf` | **s**ub **b**and **f**lux (mean values) |
| `mfcc.sbf` | **m**el-**f**requency **c**epstral **c**oefficients (mean values) & **s**ub **b**and **f**lux (mean values) |
| `mfcc.std` | **m**el-**f**requency **c**epstral **c**oefficients (mean values and standard deviations) |
| `sbf.std` | **s**ub **b**and **f**lux (mean values and standard deviations) |
| `mfcc.sbf.std` | **m**el-**f**requency **c**epstral **c**oefficients (mean values and standard deviations) & **s**ub **b**and **f**lux (mean values and standard deviations) |

*Learning Algorithms*

| | |
|---|---|
| `bayes` | **n**aïve **b**ayes |
| `svm` | **s**upport **v**ector **m**achines |
| `knn` | **k**-**n**earest **n**eighbors |

*Feature Selection Approaches*

| | |
|---|---|
| **gainr** | **g**ain **r**atio |
| **wra.fs** | **w**rapper - **f**orward **s**election |
| **wra.be** | **w**rapper - **b**ackwards **e**limination |

*Dedicated to my beloved parents Tomás and Alicia, who have always been with me. To my dear siblings Alejandro and Irene, and to their delightful children: Ciro, Ema, Lena, and Violeta...*

# Chapter 1

# Introduction

In recent years, there has been a continuous increase of music available on the internet and a notable expansion of the digital music market. According to the International Federation of the Phonographic Industry (IFPI)[1], the number of tracks rose from 1 million in 2004 to 13 million in 2010. The revenue of the music industry from digital channels, including download stores, streaming services, internet radios, subscription models, online video channels and free-to-user sites, has grown in only 6 years from 2 % (US$ 420 million) to 29 % (US$ 4.6 billion). Moreover, in nine years, 297 million iPod units have been sold.

Due to the magnitude of the information that needs to be organized and offered for easy access, new computational tools were needed to satisfy the queries of the users and enrich their experience in a clear competition with illegal distributions. A popular example is the mobile-specific application *Shazam* [2], which has offered music retrieval tasks to over 125 million users according to the company.

The problem of automatic genre classification refers to the detection of one or many musical genres or styles that underly examples of music. This computational tool has gained interest in music research as well as in the worldwide digital music and media industry. In practice, automatic genre classification is advantageous wherever large collections of music need to be organized: libraries, internet, radio, databases, and so forth.

[1] http://www.ifpi.org/content/library/DMR2011.pdf
[2] http://www.shazam.com

In spite of the improvements during the last decade, automatic genre classification remains an open question. One of the reasons is that musical genres are typically defined by qualities that may not be evident in the audio signal but that require cultural knowledge. For example, an intuitive description of a musical genre could contain background information on the ethnic group where the music comes from, the location and period of time in which a music style was produced, the age of the listeners that are generally appealed by the music, the emotional aspects of the musical or lyrical content, other genres that hold relationship to it, and so on. In addition to this and from a musicological perspective, the common characteristics employed to group music into styles hold a relatively high level of abstraction, such as rhythm, tonality, and musical form qualities.

Another feature that stands out in genre descriptions is the instrumentation, orchestration, or musical *timbre*. Timbral features have been considered critical factors in the majority of the genre classification models so far suggested, thus indicating the role of timbre similarity as a strategy for class membership decisions. The spectral components of the audio signal have been widely investigated in the signal processing field and also within music perception and Music Information Retrieval (MIR). In automatic music genre classification, timbre-based approaches showed markedly better results when compared to rhythm- or harmony-based approaches. Arguably, timbral qualities of sound sources are crucial in prompting very fast sound recognition processes in humans –for example, familiar voices can be rapidly identified on the phone thanks to their particular timbres, among other sonic features. Further, the use of spectral features in music classification is motivated by the fact that humans are able to quickly identify known musical genres (Gjerdingen & Perrott, 2008), suggesting that low-level descriptions are important in this process.

However, there is no established method for modeling timbre, either over monophonic –understood as the way a single instrument sounds– or polyphonic signals. Furthermore, there is no general agreement on what timbre does exactly mean and the most common psychoacoustic definitions of timbre are inappropriate, since they generally focus on what timbre is not rather than in what it actually is. For example, the American Standards Association (1960) defines timbre as "that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar" (as cited in Toiviainen et al., 1998) and also notices its dependence on the spectrum. Timbre is thus defined as any perceptual description besides pitch and loudness, or as the perceptual representation of the spectrum. Neither the former definition –because it is a negative one– nor the latter positive definition –since the signal spectrum is a mathematical abstraction that also relates to loudness and pitch– are suitable to describe what timbre is.

Music perception aims to bridge the gap between lower and higher music processes, e.g., between timbre descriptions and knowledge of genres. The target is to find common characteristics between music examples based on the role of auditory perception and music cognition rather than on mathematical abstractions or purely acoustical relationships (Fuhrmann & Herrera, 2010). Since musical genres or styles are high-level representations of an inescapably social, emotional and embodied nature, the cognitive and perceptual aspects of music form an integral part in MIR research.

The development of perceptually motivated algorithms for the description of timbre qualities still imposes a challenge, particularly for polyphonic audio. It can be argued that the pace of the progress in many MIR tasks is partly set by the development in music perception. A recurrent problem for content-based approaches in automatic genre classification is that high level cognitive processing does not take any role in most of the models (Reed & Lee, 2006). This is probably one of the motives for MIR research to hardly be able to surpass a pinnacle or *glass ceiling* of performance. This problem was firstly observed by Aucouturier and Pachet (2004) for timbre similarity models that were solely based on the extraction of low-level features from the audio signal. It is clear that the glass ceiling in many MIR-related tasks is at least partly caused by a gap between the audio signals and its semantic labels such as styles or moods. Recent efforts to bear this problem by extracting symbolic representations and cultural features (McKay & Fujinaga, 2008; Silla, Koerich, & Kaestner, 2010) show that genre classification can still make much of its improvement through music perception and cognition.

The goal of the present study was to investigate the performance of two timbre-based descriptors for automatic music genre classification using a variety of approaches. We have compared a widely used set of descriptors, namely the *Mel-Frequency Cepstral Coefficients* (MFCC, Mermelstein, 1976), to the more recent *sub-band flux* set of features (Alluri & Toiviainen, 2010b). The former are of interest because these are prevalent multipurpose descriptors in MIR and speech recognition, while the latter were chosen due to their relevance from a perceptual viewpoint. To achieve our aim, we have compared the classification performance of these descriptors and also investigated their relative relevance trough a detailed analysis of optimal attributes for classification.

### Structure of the thesis

The remainder of this study is organized as follows: the next Chapter focuses on the state of the art in Music Genre Classification, including the most relevant descriptors and the function of music perception in this task; Chapter 3 presents the research design implemented in this study, as well as its data collection and posterior analysis of music

descriptors and classification models; Chapter 4 presents the accuracies of the obtained models, focusing on the music collections used and on the chosen subsets of timbral descriptors; the results of the study are discussed in Chapter 5; finally, three Appendices have been added to this study. Appendix A presents comprehensive plots containing the accuracies of all the models that have been computed in the study, comprising a variety of music databases, learning algorithms, feature selection approaches and feature subsets; Appendices B and C include the most pertinent code that has been utilized in this study, namely the extraction routine with *MIRtoolbox* in Matlab and the Borda Count method for rank aggregation.

# Chapter 2

# Background

Musicologists have strived for many years to understand the properties of timbre and to find its quantitative descriptions. These efforts find motivation in the fact that other music processes show clear acoustic correlates, such as frequency for pitch and amplitude for loudness. However, it remains unclear how humans perceptually organize the content in the audio signal to identify sound sources in the music.

Research on computer-based modeling of music perception could help us to comprehend how human music perception works, but in the case of genre classification the features in use appear to be inappropriate. In fact, a variety of spectral-based features have been investigated for automatic music classification, but there is little research on perceptually motivated features available in the literature. Further, these features either come from other domains or are not designed to model the overall timbre mixture or global timbral quality of musical pieces (Aucouturier, 2006). Great efforts to understand the correlates between timbral and spectral aspects on single instruments and sounds have been made since the 1970s using similarity measurements (see for example Grey, 1977; Grey and Gordon, 1978). In practice, timbre-based descriptors in music are particular in the sense that they often model the perception of timbral qualities of more than one instrument and sometimes concurrent notes of the same instrument. Monophonic timbre perception can be insufficient for modeling music similarity, and in a sense the descriptors that are normally borrowed from the speech domain are inadequate for music perception.

Etymologically, timbre was used in French language in order to refer both to *quality of a sound* and to *sound of a bell* [1]. It could have its origins in the greek word "tympanon", which means *kettledrum*. This rather ambiguous perceptual attribute is traditionally associated with the color or shape of any sound.

[1] See Harper, D. (n.d.) *Online Etymology Dictionary* (Retrieved 11.05.2010 from http://www.etymonline.com).

The purpose of the present chapter is to examine, compare and evaluate recent research in the MIR field with respect to genre classification and music perception that are relevant to our problem. This chapter is organized to cover these issues thematically. In section 2.1 the state of the art in the Music Information Retrieval scientific paradigm is presented. Section 2.2 offers a review on the general procedures for music genre classification based on musical features and also an introduction to feature selection. Next, Section 2.3 covers the most common musical descriptors in music genre classification, and includes the features that were used in this study. Finally, in Section 2.4 a chronological review of the *Music Information Retrieval eXchange* (MIREX) contest for classification of musical genre is presented.

## 2.1 General background in Music Information Retrieval

The rampant distribution of musical files on the internet through illegal sources starting in the late 1990s has been typically considered as a threat to the prevalent business models in the music industry. The digital music revolution has been accompanied by a fast development of lossy formats (.mp3, .wma) and applications for playback and compression (Winamp) or music sharing (Napster).

Based at the outset on text query and retrieval principles, the Music Information Retrieval (MIR) paradigm became a prominent viewpoint for studying the connection between computer sciences, digital signal processing, and human perception (Tzanetakis, 2002). This interdisciplinary research approach basically aims to harvest relevant information from musical data using a variety of methods (Fingerhut, 2004) and includes a number of applications that enable scaffoldings between music listeners and musical files. For example, it would be possible to automatically sort musical files according to music genre, mood, artists or composers based on labels that are automatically assigned to the audio. As opposed to manual labeling, these systems would provide rapid and effortless tagging of music in a large scale.

The international Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) aims to create "tera-scale" musical collections using audio, symbolic and metadata material that can be accessible as digital libraries (Downie & Futrelle, 2005)

### The MIREX Contest

In 2004, a community for discussion on MIR-related problems called Music information Retrieval eXchange (MIREX) was created, promoting a framework for an annual contest that takes place within the International Society for Music Information Retrieval

(ISMIR). Inspired by the Text Information Retrieval Conference (TREC), these MIREX tasks are evaluated by the IMIRSEL project (Downie, 2008). As an example from the 2009 MIREX contest, 17 tasks focusing on both low and high levels of music processing were evaluated, including Audio Classical Composer Identification, Audio Music Mood Classification, Audio Cover Song Identification, Music Structure Segmentation and Query-by-Tapping [2].

## 2.2 Feature-based music genre classification

Even for humans, associating musical excerpts to certain genres can bring difficulties, mainly due to the blurry boundaries of musical genres. These common qualities could at least be partly explained by a musical *attribute*, i.e., a quantity that describes an example (Kohavi & Provost, 1998). Given a set of musical examples, automatic music genre classification can be thus understood as an associative learning between attributes and class decisions.

Panagakis and Kotropoulos (2010) discuss the motivation on music genre classification studies and they refer to Aucouturier and Pachet (2003), who pointed out that probably genre is the most popular variable for organizing music. Scaringella, Zoia, and Mlynek (2006) have reviewed the music genre classification problem and stated relevant questions on this topic.

### Musical genres

In spite of the highly subjective ground of genre classification, we somewhat find a consensus among individuals belonging to a same culture on how to group popular pieces of music into genres. They typically organize music into genres according to their similarities with more "prototypical" examples. Individuals can classify music into genres, in other words, they can associate musical pieces with common characteristics and assign categorical labels as a way of grouping them (Tzanetakis & Cook, 2002).

However, some ambiguities and lack of consensus in this assignment are sometimes found. McKay and Fujinaga (2006) inquires into the the argument that the benefits of automatic genre classification are restrained since musical genres are unclear and subjective. Scaringella et al. (2006) suggest to loosen the classification paradigm by adding more than one genre class to a file in order to attain realistic systems.

[2] Downie, J.S. and West, K. (2009). *MIREX2009 Results* (Retrieved from http://www.music-ir.org/mirex/2009/index.php/MIREX2009).

**Chance level scores** It is clear that the complexity of a classification task increases with the number of classes into which the examples must be grouped. Therefore, given the same performance, there is more chance involved in settings where the examples belong to a lower number of classes.

### Music databases

There are two major problems that affect the music databases used for Music Genre Classification. Both issues relate to the data preparation for this task and have been discussed for different MIR problems.

**Ground truth** A ground truth refers to a description or set of descriptions used for validation of measurements or techniques. For example, the ground truth of a context-dependent descriptor –such as the style of a song– can be used for evaluation and validation of a technique that automatically assigns these descriptors to the data. In Music Information Retrieval, the ground truth can be collected from many sources, such as music catalogues, user ratings or tags, musicological studies, and so on.

**Artist and album filtering** These methods assure that all the music files that belong to the same artist or album will be either part of the training list or of the test set. It is important to consider that the validity of a classification model could be questioned without the use of artist and album filtering (Downie, 2008). A significant drop in the accuracy was noticed when the use of filtered and unfiltered data was compared in parallel evaluations by Pampalk, Flexer, Widmer, et al. (2005) (Downie, 2008). Remarkably, West (2005a) puts into consideration a future implementation of artist filtering in AGC tasks, but the only submission that explicitly includes the filter in the 2005 evaluation is Pampalk (2005) [3].

### Feature extraction

It is possible to extract a variety of features from the audio signal, such as brightness and tempo, that can be useful for automatically classifying music. The process of feature extraction aims to reduce the data into measurable properties (Duda, Hart, & Stork, 2001).

[3] A recent study on music semantic similarity for polyphonic instrument recognition by Fuhrmann and Herrera (2010) has also taken care that artist and album effects would not affect the results.

The canonical method for genre classification based on audio content –which is particularly noisy and of considerable computational size– is to extract features instead of working directly with the data. This procedure is used to reduce the data dimensionality and to obtain information that is considered most important for genre classification. The audio is analyzed as windows using audio frame decomposition for each feature in order to obtain a compact representation of an attribute for each window. As a partial result of a multiple feature extraction, each of the frames is represented by a vector of feature values and each music excerpt as a sequence of feature frames. Finally, the statistics (means and standard deviations) for each feature over the whole example are usually computed. In this section we will describe the pre-processing step for musical feature extraction and briefly present musical features that are commonly used in genre classification.

**Pre-processing** Prior to the implementation of the feature extraction methods below, the audio signal is divided into short overlapping or non-overlapping frames. In feature extraction for MIR tasks, the signal information is typically truncated into short frames that could correspond to the times and frequency resolutions of human hearing and cognitive processing (Aucouturier, 2008; J. O. Smith, 2010). In order to better capture the temporal evolution of the audio signal, the frames are of the order of only 20 milliseconds long, which is a standard resolution for the purposes of classification based on low-level features.

**Extraction of feature values** For each of the frames, a feature vector of a length equal to the number of features is computed. Next, the usual method implies the estimation of feature statistics, namely the arithmetic mean and the standard deviation of a feature over each of the music examples. The calculation of the standard deviations can serve as a compromise, for a given feature, between the details of each frame and the central tendency of the whole example.

### Feature selection and content-based classification

Classification is one example of pattern recognition, and it is pivotal in Music Information Retrieval. Feature selection is often implemented as a previous step to classification for dimensionality reduction purposes. Lessening the dimensionality of the feature space has a way to simplify the classification stage. As to this study, the latter is based on the actual content of the source instead on *metadata* such as tags or keywords, thus it is called *content-based* classification. A general introduction to the problem of

content-based classification and a rationale for feature extraction are mentioned in this subsection.

Pattern recognition or discriminant analysis roughly refer to methods for classification in machine learning. These approaches can learn to recognize statistical regularities and build models that can be applied to map data to suitable categories (McKay, 2010). Pattern recognition systems can be optimized for computational ease and decent performance. Once the model is deployed, it is possible to handle a significant amount of data for rapid evaluation. Another advantage is that pattern recognition can lift the level of complexity of many problems by concurrently embracing multiple characteristics or attributes. Moreover, these systems offer consistent upshots and therefore are reusable with new information and parameters.

A typical routine in Audio Genre Classification is to train a classifier with labeled data. This method involves two fundamental steps. In the first place, a teacher provides a training dataset, *i.e.*, information where the category labels are known. In the second step, a test set of new data without labels is inputed and the task of the algorithm is to predict its classes. This method is called supervised learning since it needs to be taught with a training dataset (Duda et al., 2001). Unlike regression and its continuous output, classification refers to supervised learning with a discrete output –a genre class for each song is predicted. For automatic genre classification tasks, supervised learning can be used in order to train the classification algorithm with the ground truth –a collection of factual data with the correct labels. In contrast, unsupervised learning is used in automatic timbre similarity tasks since the aim is to group musical pieces according to their overall sound (Aucouturier & Pachet, 2003).

#### Three paradigms for classification

Three main automatic classification paradigms are mentioned in the literature, namely expert systems, supervised learning and unsupervised learning (McKay, 2010; Scaringella et al., 2006).

**Expert systems** These systems explicitly compute sets of rules from high-level features in order to define classes. In other words, with very detailed and objective descriptions of the classes, expert systems could perform a classification based on a serious understanding of the examples (Scaringella et al., 2006).

**Unsupervised Learning** The second paradigm is called unsupervised learning or clustering and refers to a grouping of unlabeled instances. In other words, in clustering

there is no prior specification of a dependent attribute. Since class attributes are not provided, these techniques have to rely on the similarity between instances that is yielded by the extracted descriptors. Typical algorithms are $k$-means, agglomerative hierarchical clustering and Self-Organizing Maps (SOM).

In unsupervised learning, the system is not taught and therefore performs a grouping or clustering of the data according to given parameters Duda et al., 2001. Due to the absence of an explicit teacher, the classification of an example is based on their associated feature vectors. For this purpose, a distance measure between vectors can be computed by the clustering algorithm. There are many different distance measures that can be applied, we can mention Euclidean and Cosine distances. Another option is to build statistical models of the distribution of the features, like Gaussian Models and Gaussian Mixture Models. (Scaringella et al., 2006)

**Supervised Learning** In supervised classification techniques, which are utilized in this study, the learning algorithm maps instances into a category label that is given beforehand by a teacher. In comparison with expert systems, the categories are not explicitly described in supervised learning, since the learning algorithm forms relationships between the training set attributes and their classes.

### Data post-processing

Once the classifier is trained and is tested for classification, it is possible to measure its performance. This is done for example by calculating how fast, how accurate in prediction, how extensible towards other scenarios and how simple is the created system (Liu & Motoda, 1998). In order to evaluate the classification accuracy, the classifier is tested with different information with respect to the data used for the training process.

**Learning algorithms** A variety of machine learning algorithms for supervised learning exist in the literature. There are two general types of classifiers, namely non-parametric and parametric. Another recent distinction groups classifiers based on their generative or discriminative properties.

1. The parametric versus non-parametric opposition (Duda et al., 2001) refers to the knowledge of the forms of the probability distribution of the training data. In other words, parametric classifiers are based on assumptions on the underlying data distribution and the classifier training is constrained to estimated parameters

of these distributions. Unlike parametric classifiers, non-parametric classifiers do not require any assumption on the probability densities.

Among common non-parametric classifiers we find Fisher's linear discriminant function, Support Vector Machines (SVMs), and decision trees. These are considered as non-parametric in the sense that they do not depend on assumptions about the distribution of the data. Parametric classifiers like Naïve Bayes adopt assumptions of the statistical distribution of the training set.

The aforementioned distinction between classifiers can be confusing because for both groupings the classifiers estimate parameters during the training phase in order to apply them in the testing phase (see for example Kuncheva and Rodríguez, 2010). A different way to group them (Ng & Jordan, 2001) approximately maintains the same division, but based on different criteria.

2. According to this second distinction there are two to three main types of classifiers, namely discriminative, generative and probabilistic. The main difference between these approaches stems from the modeling method (Rattray, 2009). In discriminative classifiers, either the classification rules are directly generated from the training set or a direct function from the input instances to the labels is learned. Similarly to the parametric group of classifiers, generative classifiers model the joint probability distribution for the examples and the class labels. Both probabilistic classifiers and generative classifiers set a probabilistic model of the data within each of the classes. In contrast, in discriminative classifiers, the discriminant rule is modeled directly from all the classes during the training process.

**Generalization ability and overfitting** The notion of generalization (von Luxburg & Schölkopf, 2011) is useful to understand the relationship between the classification results for the training test and test sets. We can define a training set as as a set of points $X$ in a space $\mathcal{X}$, where each point has a known label denoted by $Y$. Using a learning algorithm and the training set $(X_1, Y_1), \ldots, (X_2, Y_2)$ we come up with a classifier $f_n$. Without a test set we cannot know how many unlabeled points are misclassified, i.e, the true underlying risk $R(f_n)$ of the classifier. Nevertheless, prior to the test classification we can calculate the empirical risk or training error based on the number of mistakes $(\sum \ell)$ that the classifier makes over the training set:

$$R_{emp}(f) := \frac{1}{n} \sum_{i=1}^{n} \ell(X_1, Y_1, f(X_i)).$$

The generalization performance of the classifier $R(f_n)$ is relatively small when the learning algorithm is able to explain most of the data in the training set $(X_1, Y_1), \ldots, (X_2, Y_2)$.

A good generalization ability is indicated by a small absolute difference between $R(f_n)$ and $R_{emp}(f_n)$, that is $|R(f_n) - R_{emp}(f_n)|$.

Arguably, the generalization ability can be simply obtained by calculating the absolute difference between the average train accuracy and the average test accuracy of a classification model. Further, a high generalization ability can correspond to low tendency of *overfitting* the data.

This problem occurs when the training set feature values are fitted too well into the learning algorithm. If the training data has noise or anomalies that are not part of the overall distribution of the data, the resulting trained model may not fit appropriately the testing examples unless it ignores the noisy training data. In other words, the model fits closely the training examples but cannot generalize well to other data. Overfitting occurs when not only the salient characteristics, but also the noise of the data is modeled. Since there are more parameters than those that are necessary for creating a statistical model of the data, it yields substandard accuracies. Noteworthy, feature dimensionality leads to overfitting models because it adds noise to the training data that may not be present in the testing data. (Sherrod, 2003)

A possible way of dealing with overfitting would be to use a large primary set in order to refine the classifier learning process. Another solution is a method called cross-indexing, which loops the selection process for a given learning method and a feature selection search (Saari, 2009).

#### Attribute selection

Prior to classification, the aim of this task is to obtain, when possible, an optimal feature subset for building a straightforward model. There are two general approaches to this process, namely *filter* selection and *wrapper* selection. The former does not interact with any learning algorithm, while the latter uses a meta classifier in order to provide results based on classification accuracy (Silla, Koerich, & Kaestner, 2008; Saari, 2009).

Attribute selection aims to lower the effects of three important issues that can affect the results of a classification (Saari, 2009). These problems are called *curse of dimensionality*, *feature relevance* and *feature redundancy*.

**Dimensionality** It is not recommended to cram features into pattern recognition systems (Reunanen, 2003). Dimensionality occurs when too many feature vectors are fed into the model, producing a highly dimensional feature space. The curse of dimensionality can be computationally demanding and is most likely to produce a constraint in the model interpretability.

**Feature relevance** This problem is often observed when more features than those that are needed are extracted from the audio signal. A feature is considered irrelevant for the classification if the same classification accuracy is obtained when the values of that single feature are absent from the model (John, Kohavi, & Pfleger, 1994).

**Feature redundancy** The third issue is called feature redundancy and occurs when two features are providing similar or even identical information about the studied data. Undesired results can occur if the classifier exaggerates the effect of the phenomenon that both of the features are measuring.

**Wrappers and filters** Feature selection makes possible to obtain a smaller set of dimensions, similarly as with the Principal Component Analysis (PCA) method. However, unlike the latter, which is an unsupervised method, feature selection can make full use of information about the target concept. In other words, feature selection techniques like wrapper selection can exploit the learning algorithm in order to find the best feature subset.

While the filter approach provides information about feature *redundancy* –i.e., what features are most correlated–, the wrapper approach can provide a ranking of the most relevant features therefore suggesting an optimal feature subset. However, the wrapper approach is more prone to produce overfitting than the filter approach (Saari, 2009). Sánchez-Maroño, Alonso-Betanzos, and Tombilla-Sanromán (2007) give account of an approach that combines wrappers and filters for feature selection. In addition, they compare the drawbacks of both approaches, in terms of computational demand, goodness of fit and number of attributes selected.

## 2.3 Musical descriptors

In this section we will present the most influential descriptors in music genre classification, placing a special emphasis on timbre-based features. Firstly, a rough introduction to digital signal processing is given in order to briefly mention the analog to digital conversion and the calculation of the signal spectrum.

### Overview of digital signals

Our ears respond to air vibrations that are called sounds. The "shape" of the vibrations is represented as the air pressure variations as a function of time, and this shape can be more or less regular. Sounds with higher periodicity could be heard as tones with a firm

pitch, while sounds with more fluctuations in its shape could be considered more noisy. (Moore, 1985)

**Frequency** The *pitch* and the *frequency* are two descriptions of the regular period of sounds. Sounds can be heard by humans if this period ranges between $\frac{1}{20}$ and $\frac{1}{20000}$ of a second (Moore, 1985). While frequency is a *physical* quantity that does not include any reference to the ear for its calculation, pitch is a *subjective* evaluation of the frequency that depends on other descriptions such as duration and loudness (Backus, 1977). Roughly speaking, pitch is the perceived "tonal height", in other words a logarithmic scaling of frequency that is perceptually meaningful for humans.

**Amplitude** Given an oscillation, its *amplitude* measures the strength of the pressure deviation with respect to the mean atmospheric pressure. The waveform shape is a result of amplitude as a function of time; this general shape of the vibrations could suggest a given tone quality or *timbre*. (Moore, 1985)

**Phase** Finally, the initial *phase* angle refers to the relative starting position of the oscillation. Given a period $T$ (measured as $T = \frac{1}{frequency}$), an amplitude $A$ and an initial phase angle $\phi$, it is possible to define a sine function $y$ at any given moment $t$ (Tempelaars, 1996):

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

### Analog to digital conversion

An audio signal can be recorded using analog or digital representations. The analog signal wave is a voltage wave that is analogous to the pressure wave. The digital signal is a representation based on a discontinuous range of values. In order to transport analog signals into a digital format, an analog-to-digital conversion is performed. The signal is sampled and quantized for data stream encoding. Next, the resultant discrete-time signal can be processed or analyzed using digital signal processing (DSP) algorithms.

### Time to frequency transformation

The 18th century French mathematician and physicist Jean Baptiste Joseph Fourier studied how periodic functions can be explained as a summation of a possibly infinite

number of trigonometric functions, each with a particular amplitude and phase (Moore, 1985).

Fourier's theorem associates sinusoidal vibrations and non-sinusoidal (or arbitrary) vibrations. It is possible to illustrate this by supposing that a tuning fork could produce only one sinusoid. In such a case, this tuning system would be a *harmonic oscillator*: unlike everyday sounds, the tuning fork would be everlasting and undamped. Ideally, the sound of a violin could be decomposed into a possibly infinite number of harmonic oscillators. (Tempelaars, 1996)

This reduction is called *Fourier analysis* or *Fourier transform* and it is applied in digital processing units for the calculation of the signal spectrum in audio and image processing (Rockmore, 2000).

**Spectrum** The frequency spectrum is a representation of the frequency and phase values of the sinusoidal components of the signal. The spectrum describes the amplitude over frequency bins for a given instant or time lapse. This representation is particularly relevant since it could resemble the perception of pitch in the human auditory system (Serra, 1989). It is possible that a similar filtering process in the cochlea detects frequencies from incoming sounds.

Analogously to light, the sum of all the frequencies at the same time is called white noise. When some frequencies are more prevalent than others, the sound obtains more color. When there are more harmonic multiples of the fundamental frequency, the sound becomes more pitched.

**Fourier series** The Fourier series is a possibly infinite number of sinusoidal components of the periodic signal (Tempelaars, 1996). These components include harmonics, in other words frequencies that are multiples of the fundamental $F_0$. The decomposition into a number of "basic blocks" is somewhat analogous to the calculation of all the prime factors of a positive integer (Moore, 1985).

**Fourier transform** The aforementioned splitting of an arbitrary signal into its components is referred to as a conversion from the *time domain* of the waveform to the *frequency domain* or signal spectrum. The Fourier transform can be defined in the following way (Serra, 1989):

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

where "$\triangleq$" means *equals by definition*, $x(t)$ is the signal in the time domain and $X(\omega)$ is the signal in the frequency domain. This formula can be applied to continuous waveforms, and each resultant $\omega$ is a frequency index measured in radians per second.

The frequency spectrum is a complex function, i.e. a non-sinusoidal function that contains more than one component (see Tempelaars, 1996). This function is decomposed into two real functions, namely the spectrum amplitude and the spectrum phase (S. Smith, 1997). The continuous frequency indexes are complex numbers that specify the frequency and the phase of each sinusoidal component, and are usually represented as $X(\omega)$, standing for the whole frequency spectrum. (Serra, 1989)

The Fourier synthesis is the inverse of the Fourier transform and makes it possible to reconstruct the arbitrary function from the sinusoidal components.

**Short-Time Fourier Transform** The Short-Time Fourier Transform (STFT) is used for signals that vary in time, such as musical signals. The STFT is obtained by dividing the signal into successive audio frames and performing a succession of Fourier transform calculations. The result of the STFT representation of the amplitude distribution as a function of frequency is a set of spectra called *spectrogram* (J. O. Smith, 2010).

**Discrete-Time Fourier Transform** The Discrete-Time Fourier Transform (DTFT) is roughly the FT for time-varying and discrete (i.e. sampled) waveforms such as digital signals. In other words, it is used for signals that are both discrete in frequency and in time.

**Discrete Fourier Transform** The Discrete Fourier Transform (DFT) is roughly the Fourier transform for periodic (i.e, they do not vary over time) and discrete waveforms. The DFT is calculated in signals that are continuous in time but discrete in frequency, such as digital periodic signals. The Fast Fourier Transform (FFT) is a fast algorithm to implement the DFT, and since it calculates the frequency spectrum of a discrete time-domain signal of finite duration, it is especially useful in Digital Signal Processing applications.

The DFT of a signal $X$ can be defined as

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

where

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

$t_n \triangleq nT = n$th sampling instant (seconds), $n$ is an integer $\geq 0$

$X(\omega_\kappa) \triangleq$ *spectrum* of $x$ (complex valued), at frequency $\omega_\kappa$

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

$N =$ number of time samples = number of frequency samples (integer number).

The FFT computation of the frequency spectrum for each audio frame of a signal is the STFT.

### Features for music content description

At this point, the features that have been used for this study are presented. Both features require a Fourier transform for their calculation and either divide or warp the frequency content following the idea that both speech and music processing in the auditory system could occur in separate frequency bands (Allen, 1994).

#### Mel-Frequency Cepstral Coefficients

Widely used for speech recognition, the MFCCs describe the spectral shape of the signal. Its computation involves five main steps including the conversion of signal frame into a mel scale representation (Logan et al., 2000) in order to emphasize the middle frequency bands (J. T. Foote, 1997). The MFCC transformation discards pitch information from the audio signal and it has been proved useful for computing music similarity (J. Foote, 1999).

There are other descriptors that have been recently developed and are similar to the MFCCs. The MPEG-7 standard *AudioSpectrumEnvelope* has been considered a more direct way of describing the signal spectral shape (J. B. L. Smith, 2010). This descriptor is obtained by creating a partition of the spectrum into bands –most of them logarithmic– and estimating the relative power for each of the bands.

A variation of the MFCCs using an octave-based scale has been suggested for discrimination between instrumental sections and other sections that contain both instruments and singing (Maddage, Xu, Kankanhalli, & Shao, 2004). The octave-scaled cepstrum coefficients (OSCCs) aim to characterize the spectrum frequencies within the singing range ( $250 - 1000$ Hz) for the purposes of music structure analysis. This change of the frequency scale yielded to a more robust algorithm for this discrimination task. (J. B. L. Smith, 2010)

#### Sub-band flux

The sub-band flux set of features is a representation of frequency and amplitude fluctuations as a function of time in ten octave-scaled spectrum bands (Alluri & Toiviainen, 2010b). It is as a matter of fact a spectro-temporal feature based on the spectral flux (see below), which is calculated for each sub-band of the filter bank, resulting in a set of 10 sub-band spectral fluxes.

This feature set was suggested in a recent study by Alluri and Toiviainen (2010b) that focused on finding correlations between perceptual dimensions and the acoustic ground of polyphonic timbre. In one of the experiments, 35 participants rated 100 musical excerpts of 1.5 seconds using 8 bipolar timbre semantic scales (Strong-Weak, Empty-Full, etc.). Using factor analysis, the results were grouped into three perceptual dimensions (Brightness, Activity and Fullness). A regression analysis was performed in order to find out which acoustic features could better explain these dimensions and it showed that the sub-band flux was a suitable descriptor for modeling of polyphonic timbre perception. Similar results were later found in a cross-cultural setting (Alluri & Toiviainen, 2010a).

Following, a list of the most frequently used descriptors in music genre classification is included. Some of these descriptors require to calculate one or more Fourier transforms for their computation.

- Energy descriptors

  **High energy-Low energy ratio** Ratio of frames showing energy below and above 1500 Hz. (Alluri & Toiviainen, 2010b)

- Spectral descriptors

**Delta Mel-Frequency Cepstral Coefficients** The $\Delta$ MFCC and $\Delta-\Delta$ MFCC (Furui, 1986) are calculated from the derivatives of the MFCCs and are considered dynamic features when compared to the "static" spectral description of the MFCCs. Like the latter, these descriptors also are originally from speech recognition but have been applied in MIR.

**Spectral Centroid** Statistical mean of the spectral distribution, gravity center of the magnitude spectrum (Tzanetakis & Cook, 2002; Saari, 2009).

**Relative Shannon entropy** Peakiness of spectral distribution (Toh, Togneri, & Nordholm, 2005).

**Roughness** Estimation of sensory dissonance. Average of dissonance curves (Sethares, 1998).

**Spectral Flux** Measures how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame. Some of the derivates of this descriptor are the sub-band flux and the perceptual spectral flux –a spectral flux computed from a compensated spectrogram with respect to Equi-loudness curves (Couvreur et al., 2008).

**Spectral Roll-off** Also termed as spectral extent (Theimer, Vatolkin, & Eronen, 2008), it describes the frequency below which 85% of the sound energy in the spectrum is concentrated (Tzanetakis & Cook, 2002).

**Spectral Slope** Description of how quickly there is a decrease in amplitude of successive partials as their frequency gets higher (Theimer et al., 2008).

- Perceptual descriptors

**Chromagram** The chromagram is arguably the most used harmonic descriptor in music perception. Also called Pitch Class Profile, the chromagram usually represents the likelihood of a pitch class in the audio, but can also be computed as the spectral energy collected in frequency bands of a pitch class (Bartsch & Wakefield, 2005). The perception of pitch with respect to a musical context can be graphically represented with a continually cyclic helix that has two dimensions, *chroma* (or pitch class) and *height*. Chroma refers to the position of a musical pitch within one cycle of the helix as if it was seen from directly above. The pitch height is the position in the vertical axis of the cycle and is related with the octaves for a same pitch class. (Goto, 2006)

**Loudness** According to Fletcher and Munson (1933), the loudness of a tone is the magnitude of an auditory sensation. Each of the equal-loudness contours is an averaged estimation of the human sensitivity to different frequencies at a given amplitude sound pressure level.

- Temporal Features

  **Zero crossing rates** Signal noisiness, measured by number of times the signal changes sign (Tzanetakis & Cook, 2002).

This introduction to musical descriptors covered many of the most used algorithms for feature extraction in music genre classification. In the next section, a review of recent approaches to genre classification, there is more detailed information on the general procedures for extraction of these features from the musical signal.

## 2.4 Approaches to automatic genre classification - Review of MIREX 2005-2009 submissions

In recent years, a variety of models for automatic classification of music into genres or styles has been suggested after first studies in MIDI by Dannenberg, Thom, and Watson (1997) and in audio signal by Tzanetakis and Cook (2002). In 2004, the MIREX Community started the Audio Genre Classification (AGC) Contest as an evaluation platform for this task. The overall goal in the AGC contest is to obtain a good performance in genre classification of music. For this purpose, the suggested routine would be to extract relevant musical features from the audio signal, train and finally test the classifier with the music datasets provided for the contest (see Figure 3.1 below).

The AGC MIREX Contest is organized within the ISMIR (International Society for Music Information Retrieval) and concurs with the ISMIR Conference. This contest has been run on six occasions: 2004, 2005, 2007, 2008, 2009 and 2010. In the present review the results of all the contests, except for 2004 and 2010 will be commented. There is only little material available on the 2004 AGC contest, and the 2010 contest results were not included in this review due to time constraints. The 2005 contest will be commented with more detail to illustrate the general properties of the classification models for this kind of contest.

### 2005 MIREX Evaluation

Two datasets were used for the 2005 Audio Genre Classification contest. The aim was to explore genre classification performance using both hierarchical –i.e, relationships of dependence between genres are established– and single level taxonomies. Either 3- or 5-

fold cross validation, depending on the computational times, was used for the evaluation and 13 participants out of 15 completed the task within the established time of 24 hours.

The datasets were composed of 1005 songs for the hierarchical set "Magnatune" and 940 songs for the single level set "USPOP". Approximately 66% of each set was intended for training the classifiers. The participants could choose to work with either mono (22.05 KHz) or stereo (44.1 KHz) PCM polyphonic music audio. However, all the candidates that have reported on this matter performed the task in mono.

Prior to the classification stage, this task normally requires the extraction of musical features from the audio signal. In the majority of the submissions for AGC contest 2005, timbre-related features were extracted. Other features, such as rhythmic and energy-based were also extracted, for example in the submission by Scaringella and Mlynek (2005).

The mel-frequency cepstral coefficients (MFCCs) were among the most frequently extracted attributes (Pampalk, 2005; Mandel & Ellis, 2005). Other features include spectral centroid (Tzanetakis & Murdoch, 2005; Burred, 2005), spectral roll-off (Tzanetakis & Murdoch, 2005; Burred, 2005; Bergstra, Casagrande, & Eck, 2005), spectral flux (Tzanetakis & Murdoch, 2005; Burred, 2005), zero crossing rates (Tzanetakis & Murdoch, 2005; Burred, 2005; Bergstra et al., 2005), low energy (Tzanetakis & Murdoch, 2005; Burred, 2005; Scaringella & Mlynek, 2005) and loudness (Burred, 2005; Lidy & Rauber, 2005).

There was no established song length for the analysis of the audio files in the task. Thus, the candidates could choose to extract information from the whole song or a part thereof. At least five of the participants have analyzed snippets instead of whole songs. Three of these submissions (Ahrendt and Meng, 2005; Burred, 2005 and Scaringella and Mlynek, 2005) operated with snippets of 30 seconds. However, the three best performing submissions (Bergstra et al., 2005; Mandel & Ellis, 2005; West, 2005a) have instead worked with whole songs.

A prevalent method for feature extraction is to decompose the signal into a number of frames using a window function for the analysis. The purpose of this is to capture more precise information than if the features were computed over the whole duration of the excerpt. This procedure is illustrated in this contest by Ahrendt and Meng (2005), Bergstra et al. (2005), Lidy and Rauber (2005), Scaringella and Mlynek (2005) and Tzanetakis and Murdoch (2005). The length of the windows usually varied between 20 and 23 milliseconds, nevertheless the submission of Bergstra et al. (2005) obtained the best classification accuracy in MIREX 2005 using a window of 47 milliseconds.

Frequently, the analysis windows are overlapped in order to attenuate the loss at the window edges. By overlapping, the analysis time origin between frames is advanced.

The hop size or step size refers to the number of samples by which each of the successive windows is advanced (J. O. Smith, 2010). Three participants have chosen an overlapping of 50%, with hop sizes varying between 10 milliseconds (Scaringella & Mlynek, 2005) and 400 milliseconds (Ahrendt & Meng, 2005). It is worthy of notice that Bergstra et al. (2005) and Mandel and Ellis (2005) did not use overlapping in their models, which were highly accurate in terms of classification performance.

Prior to the classification, the feature values can be aggregated into a window of a larger length than the analysis window. From this "texture window" (Tzanetakis & Cook, 2002) or segment that is shorter than the whole song length, statistics can be computed for building the classification model. In this contest, algorithms that included longer segmentations outperformed models in which shorter segmentations were used. For example, the submissions by Bergstra, Casagrande and Eck (13.9 seconds) and West (7 seconds segmentation for rhythmic features) have obtained better classification performance than Ahrendt and Meng (2005) and Tzanetakis and Cook (2002), who used 1.2 seconds and 3 seconds for computing statistics, respectively.

For the final step of the feature extraction, a multidimensional feature vector is obtained. Many differences were found in the submitted models with respect to the number of dimensions for the classifier input. These range from 18 in Scaringella and Mlynek (2005) to a total of 1668 dimensions in the submission by Lidy and Rauber (2005).

Usually, multi-class learning and classification is performed either with predictive density estimation methods like the Gaussian Mixture Models (GMM) or using discriminative exemplar-based classifiers like the Support Vector Machines (SVM) (Aucouturier, 2006; Mandel & Ellis, 2005). Of the ten participants that completed the tasks, six have used either SVM or GMM for their models. Other classifiers, such as simple Gaussian Distributions and ADABOOST are found in the three best ranked models (Bergstra et al., 2005; Mandel & Ellis, 2005; West, 2005a).

The classification accuracies of the MIREX 2005 AGC contest were as follows: Bergstra, Casagrande and Eck 82.34%; Mandel and Ellis 78.81%; West 75.29%; Lidy and Rauber 75.27%, 74.78% and 74.58%; Pampalk 75.14%; Scaringella and Mlynek 73.11%; Ahrendt 71.55%; Burred 62.63%; Soares 60.98%; Tzanetakis 60.72% [4]. A total overall accuracy of 72.84% with a standard deviation of $\sigma = 7.21$ was obtained in the 2005 contest.

[4]See Downie (2005). *2005 MIREX Contest Results - Audio Genre Classification (Contest wiki)*. Retrieved from http://www.music-ir.org/evaluation/mirex-results/audio-genre/index.html

### 2007 MIREX Evaluation

Two years after MIREX 2005, the genre classification contest was repeated –there was a MIREX contest but no AGC Contest in 2006. In this occasion, only one 10-genre dataset of 22.05 kHz mono music snippets of 30 seconds each was used for the contest. The major change for the evaluation of this year was the requirement of an artist filter for the train/test splitting in order to assure both validity and reliability in the submissions. As a result, there was an expected drop in the performance when compared to the results of the AGC 2005. The results of this evaluation have been analyzed in the context of the glass ceiling effect and the use of artist filtering in (Downie, 2008).

In the 2007 submission, the feature extraction techniques were similar to those in AGC 2005, because spectral and temporal features were extracted in the majority of the algorithms. However, a novel approach to feature extraction using a symbolic-acoustic combination was suggested by Lidy, Rauber, Pertusa, and Iñesta (2007). This method yielded an extraction of information about pitches, durations, inter-onset intervals (IOI), and other symbolic descriptions extracted from the audio signal.

While some submissions have based the feature extraction process on subjective measurements, the majority did not seem to focus on perceptually interpretable models. One of the exceptions was the submission of Mandel and Ellis (2007), who have given account of dissonance, loudness sensations and masking effects for feature extraction. In contrast, Guaus and Herrera (2005) investigated the extraction of variants of the MFCCs, such as the delta MFCCs and delta square MFCCs.

Compared to AGC 2005, the results in 2007 were less variant and have shown an overall lower performance (see Downie, 2008). While the performance of the 2005 contest had a mean of 72.84%, the average results dropped in AGC 2007 to 64.31%. Remarkably, the standard deviation of the average raw accuracies in the 2007 contest was of $\sigma = 4.48$ against more variant results in 2005 at $\sigma = 7.21$. Such a change in the performance of the overall evaluation results has been acknowledged as a glass ceiling and may have been partly caused by the implementation of the artist filter for the 2007 evaluation.

### 2008 MIREX Evaluation

Notably, only 6 candidates submitted accepted algorithms to the AGC contest in 2008, 4 of which had already participated in the 2007 contest. A new latin music dataset (Silla, Kaestner, & Koerich, 2007) was used in this contest in addition to the dataset used in 2007. Therefore, the classification task was enriched by increasing the diversity of analyzed musical genres. A summary of the styles evaluated in MIREX Audio Genre

| | **2005** Magnatune | **2005** USPOP | **2007** Mixed | **2008** Mixed | **2008** Latin | **2009** Mixed | **2009** Latin |
|---|---|---|---|---|---|---|---|
| **Genre classes** | blues<br>classical<br>electronic<br>ethnic<br>folk<br>jazz<br>newage<br>punk<br>rock | electronica/dance<br>newage<br>rap/hiphop<br>reggae<br>rock | blues<br>classical<br>country<br>edance<br>jazz<br>metal<br>raphiphop<br>rockroll<br>romantic | blues<br>classical<br>country<br>edance<br>jazz<br>metal<br>raphiphop<br>rockroll<br>romantic | bachata<br>bolero<br>forro<br>gaucha<br>merengue<br>pagode<br>salsa<br>sertaneja<br>tango | blues<br>classical<br>country<br>edance<br>jazz<br>metal<br>raphiphop<br>rockroll<br>romantic | bachata<br>bolero<br>forro<br>gaucha<br>merengue<br>pagode<br>salsa<br>sertaneja<br>tango |
| **Total** | 10 | 6 | 10 | 10 | 10 | 10 | 10 |
| **Songs** | 1515 | 1414 | 7000 | 7000 | 3160 | 7000 | 3227 |
| **Length** | whole | whole | 30s | 30s | unknown | 30s | unknown |
| **Format** | .mp3 | .mp3 128 | 22.05khz mono .wav | 22.05khz mono .wav | .mp3 | 22.05khz mono .wav | .mp3 |

TABLE 2.1: Datasets used in the MIREX AGC contests until 2009 (Bergstra, Casagrande, Erhan, Eck, & Kégl, 2006).

Classification is presented in Table 2.1 [5].

There is another addition that is worthy of mention for the feature extraction process and that was not present in past AGC evaluations. The submission by Peeters (2008) includes the extraction of Chroma / Pitch Class Profiles (PCP) coefficients (Peeters, 2008) in order to incorporate basic harmonic content from the signal.

Notably, the 2008 AGC Contest overall accuracy was lower and less variant than a year before. The overall accuracy in 2008 was of 63.89% for the mixed set, while the latin set reached 58.18%. In contrast, in AGC 2007 the total average was of 64.31% ($\sigma = 4.48$). The standard deviations of the 2008 contest are of $\sigma = 1.93$ (mixed set) and $\sigma = 9.35$ (Latin set), showing little differences in the evaluation results, especially for the mixed set AGC evaluation.

### 2009 MIREX Evaluation

For the evaluation of 2009, the Latin music dataset and a mixed collection were used, the latter consisting of 7000 30-second clips equally distributed into 10 genres (see Downie and West, 2009. The Latin set, which comprised 3227 audio files and was also divided into 10 music genres, encouraged the extraction of rhythmic features (Downie & West, 2009). The evaluation framework IMIRSEL asked the participants to submit algorithms that supported mono WAV sound files with a sample rate of 22 kHz 16 bit. We will focus on the three best ranked submissions of both sets.

[5] Information about the datasets can be found in http://www.music-ir.org/mirex/wiki/MainPage.

Based on the articles that were attached in the submissions, the studied participants submitted the same algorithm for both the Latin and the mixed dataset. However for the Latin set, which was comprised of whole songs, Seyerlehner and Schedl (2009) decided to decode up to the first 4 minutes of the musical files and analyze the central 2 minutes of the decoded signal.

With respect to the type of features that have been extracted from the audio signal, the most common were spectral, rhythmic and temporal, though some of the models did not include precise information about the feature extraction process. In addition to these features, Burred and Peeters (2009) and Grecu, Lidy, and Rauber (2009) extracted perceptual features from the audio signal. The windows used for analysis have differed as well, for example Seyerlehner and Schedl (2009) chose a Hanning window of 93 milliseconds with a hop size of 23 milliseconds while Burred and Peeters (2009) selected a Blackmann window of 60 milliseconds using 20 milliseconds of overlapping hop size.

We can find interesting variations in the number of dimensions for the feature vectors, and therefore in the complexity of the submitted models. For the Latin set, Grecu et al. (2009) reached an overall accuracy of 58.64% with a feature vector of 7320 dimensions, while the less complex submission by Seyerlehner and Schedl (2009) obtained a performance of 62.23% using a feature vector of 97 dimensions.

As regards the classification model accuracies, the first three places [6] for both the Latin set and the mixed set were all obtained using either an SVM classifier or its variants (SVM in Seyerlehner and Schedl, 2009, GSV-SVM in Cao and Li, 2009, C-SVM in Burred and Peeters, 2009, multi-class SVM ensemble using one-against-one principle in Grecu et al., 2009). It could be taken into account that both Cao and Li (2009) and Seyerlehner and Schedl (2009) included classifiers that were already pre-trained using other song datasets (i.e., different than those that were offered for the contest).

The overall average results for AGC 2009 were lower than in 2008, however the variance is higher than in previous contests. A total mean classification accuracy of 55.62% and 62.07% was obtained for the Latin and mixed sets, respectively. The standard deviation of the Latin set was of $\sigma = 9.35$ while for the mixed set it reached a value of $\sigma = 12.20$. Such growth in the variance compared with previous evaluations could be explained in part by an increase in the amount of submissions coming from new participants in the 2009 contest.

[6] AGC 2009 Classification results. http://www.music-ir.org/mirex/results/2009/MIREX2009ResultsPoster2.pdf

To sum up, between 2005 and 2009 some interesting differences were found between the models submitted to the MIREX contest. One of the main developments has arguably been the mandatory addition of an artist filter in 2007, which reduced the classification accuracy results but aimed to increase the validity of the models across datasets. Another interesting development consisted in the additions to the number of songs for evaluation in the contest, reaching a peak in 2008 and 2009 with 7000 songs. This could be an attempt to add complexity to the genre characterization. We can also mention the extraction of features that intend to resemble cognitive processes by taking into account loudness, localization, dissonance and harmony.

It could be interesting to assess these features against behavioral measurements, looking for their perceptual relevance. Another fundamental step for some of the models would be to look after a reduction of dimensionality as a way to attain validity and a better interpretation of the performance results. Undoubtedly, there is still a lot of room for improvement in automatic Music Genre Classification. Moreover, after the suggested models surpass the glass ceiling by far, it may be required to find plausible explanations from a perceptual viewpoint in order to satisfy both engineering and perceptual questions.

# Chapter 3

# Methodology

The methodological part of the present is divided into three major steps. The first stage consists in pre-processing the data by performing feature extraction for three music databases and preparing six combinatorial feature subsets. Next, three approaches to feature selection were used in order to obtain the five best attributes of each combinatorial subset. In the final stage, both the "Top 5" subsets and the original combinatorial subsets were classified using three learning algorithms. Generally, the suggested approach follows the layout for automatic music classification that has been illustrated by West (2005b) for the MIREX contest (see Figure 3.1).

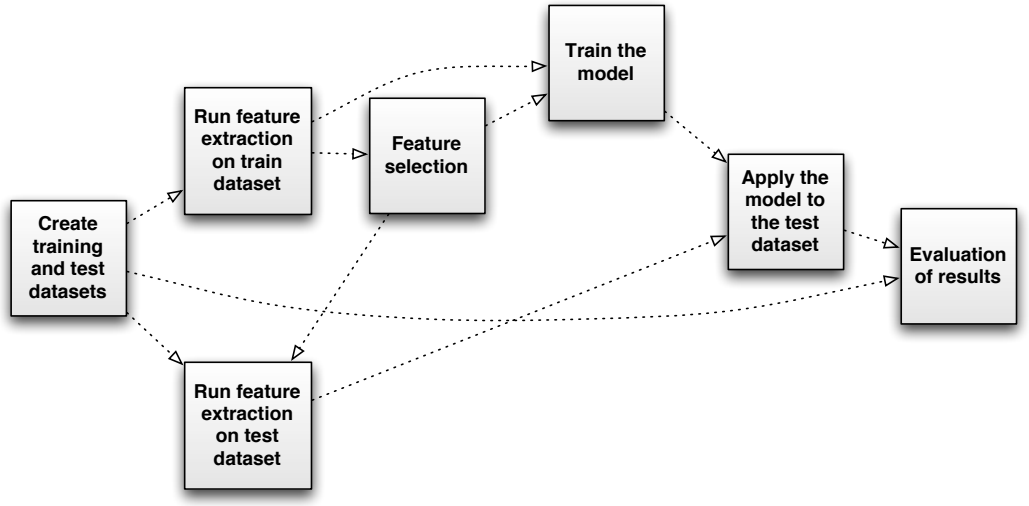


FIGURE 3.1: Simplified routine for Audio Genre Classification tasks (West, 2005) The train and test sets are subjected to a feature extraction process. A model can be trained with the aid of feature selection in order to select relevant features for the subset. The model is finally applied to the test set and the results of the implementation are evaluated using cross-validation.

| **factors→** levels ↓ | **learning algorithm** | **combinatorial subset** | **music dataset** |
|---|---|---|---|
| **1** | *K-NN* | `mfcc` | *GTZAN* |
| **2** | *NaiveBayes* | `sbf` | *RWC* |
| **3** | *Vector* | `mfcc.sbf` | *AF-RWC* |
| **4** | | `mfcc.std` | |
| **5** | | `sbf.std` | |
| **6** | | `mfcc.sbf.std` | |

TABLE 3.1: Full-factorial Design I: All attributes.

| **factors→** levels ↓ | **learning algorithm** | **attribute selection method** | **combinatorial subset** | **music dataset** |
|---|---|---|---|---|
| **1** | *K-NN* | *GainRatio* | `mfcc` | *GTZAN* |
| **2** | *NaiveBayes* | *WrapperBE* | `sbf` | *RWC* |
| **3** | *Vector* | *WrapperFS* | `mfcc.sbf` | *AF-RWC* |
| **4** | | | `mfcc.std` | |
| **5** | | | `sbf.std` | |
| **6** | | | `mfcc.sbf.std` | |

TABLE 3.2: Full-factorial Design II: Top 5 attributes.

**Full-factorial designs** The performance of the MFCCs and the sub-band flux set of features for music genre classification has been tested by implementing a variety of strategies based on four main factors, namely 1) the music datasets used, 2) the features extracted from the audio signal, 3) the learning algorithms used for classification 4) and the feature selection methods for dimensionality reduction. Since our aim was to focus on the performance of the sub-band flux feature set, we have investigated which of the descriptors were relevant for this purpose and which models yielded best accuracies.

Tables 3.1 and 3.2 present the factorial designs that were implemented in this study. The factor levels are described below in the present chapter; it can be noticed that the only difference between both designs is that the Design I does not include the feature selection stage. This design is called "All" design because the datasets do not undergo feature selection, as opposed to the "Top" design, were feature selection is included and the amount of attributes per dataset is reduced to a number of 5 "Top" ranked features. The data used for both designs is kept identical until the feature selection stage, which is skipped in the "All" design.

## 3.1 Data pre-processing

In this section, an in-depth description of the characteristics of the music databases that were utilized in the study is given. Furthermore, the data collection stage, that consisted in the frame-decomposed extraction of musical features from the song excerpts, is pointed up below.

### Music databases

In order to evaluate the classification using different scenarios, two main databases have been obtained and used as primary sets. One of the sets is called GTZAN, introduced in Tzanetakis and Cook (2002) and widely used, e.g. in Kotropoulos, Arce, and Panagakis (2010). The other music database is called RWC (Real World Computing) database and was used in 2004 for the MIREX Audio Description Contest of that year. Both databases are available on the internet [1].

These databases are publicly available and widely used in research on Audio Genre Classification. Unlike the Latin set that has been used in the most recent MIREX contests, only the most popular -and thus easily discernible- musical styles can be found in these music datasets. However, the datasets are fairly adequate regarding the general purposes of this study. The GTZAN, RWC and AF.RWC (a variant of the RWC) databases are presented below in Table 3.3.

**GTZAN Genre Collection** The dataset was created by Tzanetakis and used for the first time in Tzanetakis and Cook (2002). The files are untitled and have been recorded in different conditions such as cd and radio quality. It contains 1000 audio files of 30 seconds comprising 10 genres (blues, classical, country, disco, hip hop, jazz, metal, reggae, rock), each genre being represented by 100 tracks. The sample rate of the recordings is 22 050 Hz, and all the files are in Mono 16-bit .au format. Unlike the RWC, it is a *balanced* dataset, since each genre is represented by the same amount of songs (100).

**RWC database** The five Real World Computing (RWC) databases have been released in 2001 by the Real World Computing Partnership of Japan. The songs in the database were commissioned for RWC and they were not released for public consumption but are instead aimed to be used for research purposes.

[1] RWC: http://www.music-ir.org/mirex/2005/index.php/Audio_Genre_Classification.
GTZAN: http://marsyas.info/download/data_sets

| | **GTZAN** | **RWC** | **AF.RWC** |
|---|---|---|---|
| **Genre classes** | blues (100)<br>classical (100)<br>country (100)<br>disco (100)<br>hiphop (100)<br>jazz (100)<br>metal (100)<br>pop (100)<br>reggae (100)<br>rock (100) | classical (320)<br>electronic (115)<br>jazz (26)<br>metal (29)<br>pop (6)<br>punk (16)<br>rock (95)<br>world (122) | classical (40)<br>jazz (5)<br>metal (6)<br>pop (2)<br>punk (2)<br>rock (24)<br>world (19) |
| **Number of classes** | 10 | 8 | 10 |
| **Number of files** | 1000 | 729 | 98 |
| **Average Song length** | 30s<br>(song openings) | Whole | Whole |
| **Format** | .au | .mp3 | .mp3 |

TABLE 3.3: Music databases used for data collection. The music databases used in the present study are quite different from each other. Besides the differences in genre classes, both balanced (GTZAN) and unbalanced databases as well as an artist-filtered database (AF.RWC) were used.

The RWC-Magnatune dataset was used for the first time in the MIREX 2004 Audio Description Contest. It is comprised of 10 popular genres (ambient, blues, classical, electronic, ethnic, folk, jazz, new-age, punk, rock) that are hierarchically organized. The dataset contains 1515 instances (1005 training files and 510 testing files) that are prepared for Audio Genre Classification. This public collection holds entire songs in mp3 format. However, only 8 classes comprising 729 files of this database could be accessed, therefore ambient and blues were not included in our evaluations.

**Artist Filtered RWC database** The artist effect is an issue that particularly affects Genre Classification. In this context, it refers to a biasing in genre classification towards an unsuitable artist classification. For that purpose, one of the three datasets that are used for evaluation is an artist filtered version of the RWC music database (Pampalk, 2005). In this database, there is only one musical excerpt per artist and therefore each artist has been represented with only one song in order to avoid that the training set and the testing set contain excerpts from the same artist. The number of songs has been drastically reduced from 729 in RWC to 98 excerpts in Artist Filtered RWC (AF.RWC). Only 13,44% of the RWC set is used for its Artist Filtered version.

### Feature extraction

The musical feature extraction was performed using in *MIRtoolbox* 1.3, a Matlab toolbox for feature extraction and information retrieval of music (Lartillot & Toiviainen, 2007) that is released as a free software under the GPLv2 public license.

The audio files in the GTZAN database were resampled from 22500 Hz to 44100 Hz in order to use an identical sampling rate for all the databases. After this, 50 seconds of audio were trimmed from the middle of each audio file in the GTZAN database. For the other two databases, the whole snippets of 30 seconds were used. For all the databases, the snippets were labeled based on their corresponding musical genres. For the extraction of the attributes, frame decomposition was used using standard procedures in genre classification, namely 25 milliseconds frames and 50 % of overlapping between frames. A diagram of the data collection is presented in Figure 3.2.

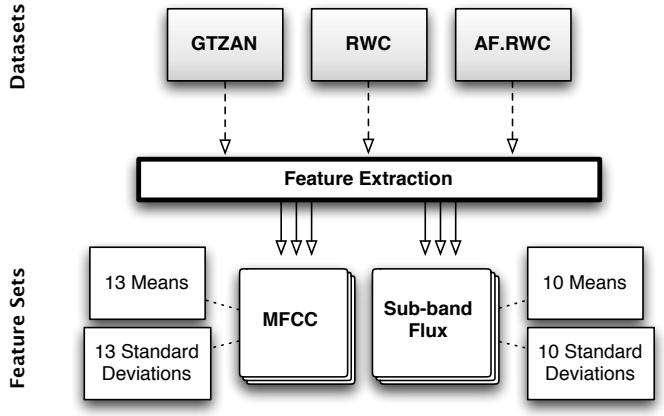


FIGURE 3.2: Data collection diagram. A total of 46 descriptors is obtained in each feature extraction. The extraction of sub-band fluxes and MFCCs is repeated for each music database – GTZAN, RWC and AF.RWC.

**Analysis window** The short segments were extracted from the audio signals using window functions in order to taper the audio data and thus avoid spectral leakage, which is a common effect of the DFT. Spectral leakage is produced when the signal contains frequency components that do not comprise complete cycles of the sine wave for a given sampling rate. The Hanning, Hamming and Blackmann windows are commonly used window functions that remove the tails produced by frequencies that cannot be represented by a single sample. There is no ideal analysis window that can remove the spectral leakage at a high resolution (J. O. Smith, 2010), but the Hamming window is

| **Category** | **Feature** | **Description** |
|---|---|---|
| **Spectral** | Mel-Frequency Cepstral Coefficients (13 features) | Used for speech recognition, they can describe the spectral shape of the sound. Its computation involves five main steps including the conversion of signal frame into a mel scale representation. (Logan et al., 2000; Alluri & Toiviainen, 2010b) |
| **Spectro-temporal** | Sub-band flux (10 features) | Fluctuations of frequency and amplitude in ten octave-scaled subbands of the spectrum. The spectral flux is calculated –taking the Euclidean distance of each successive spectrogram frame– for each of the filtered audio channels, obtaining 10 sub-band fluxes. (Logan et al., 2000) |

TABLE 3.4: Features extracted in the present study. For each feature the mean value plus its standard deviations was extracted.

considered most favorable for Fourier Transform (Lartillot, 2010) and it was therefore applied in this study.

### Spectral features

Two out of the eight features of the acoustic subset investigated by Alluri and Toiviainen (2010b) were extracted, namely the MFCC coefficients and the sub-band flux. However, unlike the aforementioned investigation, in this study both the means and standard deviations of each of the first 13 MFCC coefficients and the 10 sub-band flux frequency bands were extracted from the musical excerpts. Thus, a total of 46 attributes were extracted from the audio files. In Table 3.4, the features that were used in the study are presented.

The idea behind adding the standard deviations of these features was to make a compromise between detail and central tendency for the classification stage and to offer an estimate of the relative impact of the standard deviations over the classification performance.

**Sub-band flux set of features** The spectral flux of ten sub-bands was extracted using frame analysis and later the mean and standard deviation for each sub-band flux

was obtained. The extraction method and parameters suggested by Alluri and Toiviainen (2010b) were kept intact for this study.

The sub-band flux was calculated using two main steps, namely filter-bank decomposition and spectral flux. The first step performs a decomposition of the signal using a bank of filters. The second step computes a spectrogram for each frequency channel and calculates the distance between successive frames of the spectrograms. As a result, a spectral flux for each frequency channel is obtained. (Alluri & Toiviainen, 2010b)

In order to perform the filter-bank decomposition, which models a process of the auditory system by which the cochlea performs a frequency analysis of vibrations, the authors of the original study have chosen a method suggested by Scheirer (1998). The aim of this method was to perform tempo analysis and beat extraction of polyphonic audio signals with arbitrary timbral information. This filter-bank includes a set of low-pass, band-pass and high-pass elliptic filters that are non-overlapping. (Scheirer, 1998; Lartillot, 2010)

The chosen elliptic filters for the sub-band flux were of second-order, as against the original model by Scheirer (1998) using sixth-order filters. Arguably, in Alluri and Toiviainen (2010b) the attenuation of higher frequencies has been done gradually in order to model human hearing –in the manner of the equal-loudness contours, which do not have steep attenuations. The first elliptic filter used for the sub-band flux set is low-pass and the last one is high-pass, while the remaining eight are band-pass.

The ten sub-bands cover a full frequency range for a sampling rate of 44.1 kHz and are octave-scaled. The boundaries of the sub-bands are at around half-sharp G and the ranges are as follows for each sub-band: 0-50 Hz, 50-100 Hz, 100-200 Hz, 200-400 Hz, 400-800 Hz, 800-1600 Hz, 1600-3200 Hz, 3200-6400 Hz, 6400-12800 Hz, and 12800-22050 Hz (Alluri & Toiviainen, 2010b). The octaves covered by each sub-band are shown in Table 3.5 [2].

For the second step of the computation, which obtains the spectral flux for each of the sub bands, a spectrogram of each sub-band was calculated using STFT. Each spectrogram is a frame-decomposition of the audio signal energy for each of the frequency channels that were obtained in the first step. (Lartillot, 2010)

In order to calculate the spectrogram, a computation of the spectrum for each frame was used in Alluri and Toiviainen (2010b) using 20 milliseconds frames. The method used for calculating the spectrum of a signal is the FFT, or the computationally efficient form of the DFT (Cooley & Tukey, 1965; Lartillot, 2010).

[2] As a matter of fact, the octaves in Table 3.5 are slightly shifted from G. The exact pitch chroma for each frequency is G plus 35 cents (¢), or $\frac{7}{20}$ of a semitone. One octave is equal to 1200 cents and in equal temperament one semitone is equal to 100 cents.

| | **Frequency Range** | **Octave Scale** |
|---|---|---|
| Sub-band No. 1 | 0 - 50 Hz | – $G_1(+35)$ |
| Sub-band No. 2 | 50 - 100 Hz | $G_1(+35)$ – $G_2(+35)$ |
| Sub-band No. 3 | 100 - 200 Hz | $G_2(+35)$ – $G_3(+35)$ |
| Sub-band No. 4 | 200 - 400 Hz | $G_3(+35)$ – $G_4(+35)$ |
| Sub-band No. 5 | 400 - 800 Hz | $G_4(+35)$ – $G_5(+35)$ |
| Sub-band No. 6 | 800 - 1600 Hz | $G_5(+35)$ – $G_6(+35)$ |
| Sub-band No. 7 | 1600 - 3200 Hz | $G_6(+35)$ – $G_7(+35)$ |
| Sub-band No. 8 | 3200 - 6400 Hz | $G_7(+35)$ – $G_8(+35)$ |
| Sub-band No. 9 | 6400 - 12800 Hz | $G_8(+35)$ – $G_9(+35)$ |
| Sub-band No. 10 | 12800 - 22050 Hz | $G_9(+35)$ – |

TABLE 3.5: Sub-band flux frequency ranges and pitch intervals

Finally, ten sub-band flux values are obtained by calculating the spectral flux of each spectrogram. The spectral flux is a measure of distances between the magnitude spectra of successive frames, thus giving account of their temporal evolution. A common distance measure, which is the Euclidean distance $d$, was suggested by Alluri and Toiviainen (2010b) using the following formula:

$$d = \sqrt{\sum_{n=1}^{N} \left(A_t\left[n\right] - A_{t-1}\left[n\right]\right)^2}$$

where the audio frames at times t and t-1 are normalized to have Euclidean norm unit:

$$\Sigma A[n]^2 = 1$$

The optimal window size of the sub-band flux has been previously studied by correlating perceptual similarity ratings with the results of the feature extraction using different window lengths. Since no significant changes where found in the correlation values between the different analysis sizes, a 25 milliseconds window with an overlap of 50% was used, following previous literature in Music Information Retrieval (Alluri & Toiviainen, 2010b).

**Mel-Frequency Cepstral Coefficients** As often as not, the MFCCs 1-13 are extracted for MIR tasks. In this study, these standards were complied using frame decomposition (25 milliseconds window and half overlapping) and subsequently the mean and standard deviation of each coefficient was obtained. The first MFCC coefficient, usually called MFCC №0 is discarded since it correlates with the signal log energy.

These descriptors come from the area of speech recognition and from the aim to suggest a perceptually plausible representation of the human speech signal (Mermelstein, 1976).

The spectrum is logarithmically scaled into a mel-spectrum in order to obtain frequencies that are evenly spaced according to human perception. A common function (Deemagarn & Kawtrakul, 2004) for calculating a mel pitch from the frequency in Hz is

$$mel(f) = 2595 \times log_{10}(1 + \frac{f}{100}) \ .$$

For the computation of the MFCCs, firstly the short-time Discrete Fourier Transform is obtained on each analysis window in order to obtain the spectrum. The next step is to map the result onto mel banks of bandpass filters. The logarithmic scaling of the frequencies is then computed to "shrink" frequency ranges above 1 kHz and to stand out the range of frequencies under 1 kHz since this range is perceptually better for distinguishing frequency differences between notes. Next or simultaneously to the prior step, the signal is generally reduced into 40 mel bands. The logarithmic square of the obtained mel-spectrum is transformed using a method called Discrete Cosine Transform (DCT). The DCT is similar to the Discrete Fourier Transform but unlike the latter it is obtained only with the real part of the input. Each of the amplitudes of the obtained *cepstrum* is an MFC coefficient.

As aforementioned, the first coefficient (MFCC №0) is normally discarded in the literature because it correlates highly with the signal power, and only the next 13 periodicities of the log spectrum are calculated. It has been studied that very high MFCCs correlate highly with pitch, so they are not considered timbral descriptors and therefore are usually not calculated (J. B. L. Smith, 2010)

The MFCCs perform a *cepstral* analysis on the spectrum shape by computing, from the spectrum itself, a signal with periodicities (J. B. L. Smith, 2010). The signal cepstrum roughly involves a Fourier analysis on the observed spectrum. The real cepstrum of a signal is obtained as the inverse Fourier transform of the logarithm of the signal power spectrum (Yeh, 2008).

## Data processing

After the feature extraction, the primary sets were exported into *Weka* (Waikato Environment for Knowledge Analysis), a suite of machine learning algorithms that is released as free software under the GNU General Public License (Witten & Frank, 2005).

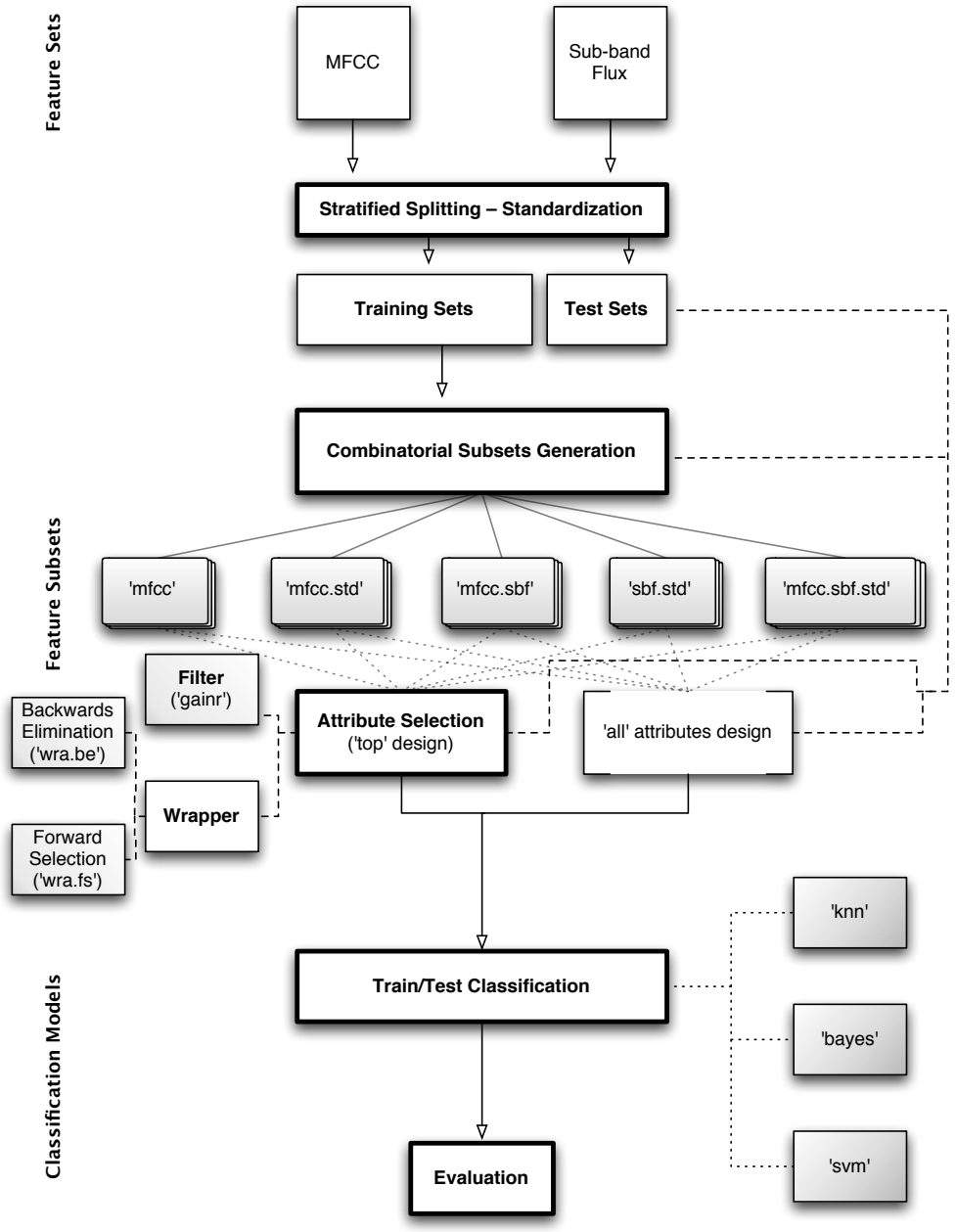


FIGURE 3.3: Analysis diagram presenting the main stages of the study. This process is repeated for each music database (GTZAN, RWC and AF.RWC). The factor levels, except for the databases, appear as shadowed boxes. For the evaluation, 10-fold cross-validation has been applied.

The collected data was processed using standard procedures for splitting into training and testing sets. Prior to the classification, the datasets were organized into combinatorial subsets. In the factorial design called "Top 5", these subsets were reduced using feature selection. Figure 3.3 illustrates in detail the analysis stage of the study.

### Stratified splitting

In order to ensure that both the training and test set had an equal distribution of songs belonging to a same genre, the primary sets that contained the feature vectors were split into training and testing datasets in a random and stratified fashion. This method, called *stratified splitting*, was implemented in order to reduce the chances of obtaining invalid classification performance results. (Saari, 2009)

### Standardization

The purpose of performing a *dataset standardization* is to ensure that initially each feature is equally relevant for the feature selection algorithm (Saari, 2009).

Each attribute value $x$ of the train set was subtracted by the mean value $\mu$ for that attribute across all instances and the result divided by the standard deviation $\sigma$ of that mean value. The z-values

$$z = \frac{x - \mu}{\sigma}$$

have been computed prior to the feature selection process. This method scales the train dataset in such a way that each of the extracted features has $\mu = 0$ and $\sigma = 1$. Also the test set feature values were scaled according to the standardization information of the train set. In order to prepare the test set with the same parameters as the train set, the batch filtering option has been used for standardization in Weka.

### Combinatorial subset generation

In order to analyze the relative relevance of the descriptors under study, an analysis of the dataset into subsets was implemented. These combinatorial subsets pointed to offer better comparisons between the musical features, by separating or combining them for the purpose of classification. Besides analyzing the interactions between features in the classification, the aim was to evaluate the importance of the standard deviations for this task. Of the chosen musical descriptors, a total of six combinatorial subsets were

created prior to the feature selection stage.The following combinatorial feature subsets have been generated:

`mfcc` Mel-frequency Cepstral Coefficients. Mean feature values (feature vector size = 13).

`sbf` Sub-band Fluxes. Mean feature values (feature vector size = 10).

`mfcc.sbf` Mel-frequency Cepstral Coefficients and Sub-band Fluxes. Mean feature values (feature vector size = 23).

`mfcc.std` Mel-frequency Cepstral Coefficients. Mean and standard deviation feature values (feature vector size = 26).

`sbf.std` Sub-band Fluxes. Mean and standard deviation feature values (feature vector size = 20).

`mfcc.sbf.std` Mel-frequency Cepstral Coefficients and Sub-band Fluxes. Mean and standard deviation feature values (feature vector size = 46).

## 3.2 Feature selection

For the "Top" full-factorial design of this study, an important factor was added that makes it distinguishable from the "All" design. This factor, namely attribute selection, consisted in three levels or approaches to feature dimensionality reduction. The aims were to reduce the dimensionality of the combinatorial subsets and secondly to achieve a proper estimation of the relevance of individual attributes within the sub-band flux and MFCC feature sets. Using an attribute ranking procedure, a total of five best ranked attributes were selected for each combinatorial subset of the "Top" design. The reduced combinatorial subsets were utilized both for *rank aggregation* (see Section 3.3 below) and for classification.

Two methods have been used in this study for ranked feature selection in order to select relevant features, avoid correlated features and reduce the dimensionality down to five attributes. These two popular approaches are wrapper selection and filter selection.

Wrappers were used in this study by implementing a *Greedy Stepwise* heuristic search method and both *backwards elimination* and *forward selection* approaches. Besides these two approaches, filter selection was used to obtain classifier independent results. The chosen method for filter selection was *GainRatio*, an optimized version of the Information Gain. This kind of filter selection is based on the Kullback-Leibler divergence.

For the wrapper selection, a total of three learning algorithms has been used as meta-classifiers, namely `bayes` (Naïve Bayes), `svm` (Support Vector Machines) and `knn` ($k$-Nearest Neighbors). The reduced combinatorial subsets based on wrappers utilized the same learning algorithm as meta-classifier and later for the classification stage. For example, `wra.be` (wrapper selection with backwards elimination search method) using `svm` (Support Vector Machines) meta-classifier for feature selection also maintained `svm` for train/test classification.

The aim of the Greedy Stepwise subset evaluator is to find the feature set with the highest merit based on its classification accuracy. In forward selection mode, the starting point is an empty set of features. Each attribute is firstly evaluated individually, so $n$ subsets are created, and each subset contains only one of the $n$ attributes. For this step, the best of the $n$ subsets is the one with the best single attribute –based on the accuracy provided using a meta-classifier. The "winner" subset is kept and the next step is to expand its size by adding another attribute. From the remaining $n - 1$ attributes, the attribute that improves the merit the most when it is added to the subset is selected and included in the subset, which is now expanded to two attributes. In the "ranking" mode of the Greedy Stepwise operation, the search continues expanding the subset until the whole search space is covered, even if the overall merit is reduced by the feature expansion. Since the aim is to create a ranking of attributes, the search is forced to the far side of the search space. The operation adds to the subset, at each step, a single best attribute. In backwards elimination mode, the process is similar but in inverse direction, so the starting point is a subset with $n$ attributes and these are removed one at a time. (Saari, 2009)

In this study, only the best five attributes from the ranking feature selection were retained in the combinatorial subsets. Therefore, each of the classification models of the "Top" design were based on final feature vectors of 5 dimensions. The ranking of attributes was obtained based on an arbitrary number of final dimensions that were set *a priori*. Different approaches like, for example, a search of the smallest feature set with the highest performance could have been more suitable at this point. However, the focus was more on the individual attributes and their relative weight in the classification than on obtaining the most efficient models.

## 3.3 Classification and evaluation

The last section of the present chapter focuses on the methods that were utilized for the culmination of the data analysis. The steps followed for the collection of classification metrics based on key factor levels are described, including explanations of the learning

algorithms that were implemented for train and test classification of the data. Moreover, the applied method for obtaining a relative valuation of attributes from the reduced combinatorial feature subsets is presented.

## Classification

Finally, the train classification models were built based on different learning algorithms. Train classification accuracies and test classification results were obtained from the datasets. The models were validated using 10-fold cross-validation, requiring an averaging step in order to summarize the results of the cross-validation folds.

### Learning algorithms

The learning algorithms used for performance assessment with the wrappers and for train/test classification were the $k$-Nearest Neighbors ($k = 5$), Sequential Minimal Optimization (an optimization of the Support Vector Machines) and a Naïve Bayes classifier.

**Support Vector Machines** These widely used algorithms belong to the family of binary linear discriminative classifiers, and operate by building a separating hyperplane between the data points of each class. In order to avoid overfitting, the data is separated into two classes by finding the separating hyperplane that maximizes the geometric margin between class members and non-class members (West, 2008). The training examples that are not close to the decision boundary receive zero weights. If the weighted examples are removed, the position of the separating hyperplane would change. Therefore, these the training examples that lie close from the hyperplane and thus take part in its specification are called supportive patterns or *support vectors* (Boser, Guyon, & Vapnik, 1992).

**$K$-Nearest Neighbors** Instance-based classifiers create a model of the data at the classification runtime. This goes in contrast with other classifiers such as `svm`, which attempt to create a model and the class division of the data before the classification. Therefore, the discriminative knn learning algorithms are commonly known as *lazy* classifiers, and calculate posterior probabilities by finding $k$ nearest neighbors of the new instances inside the training data. The parameter $k$ represents the neighborhood magnitude used for the classification of each instance. A distance function, like for example the Euclidean distances, can be used in order to compute the distance between instances. Finally, a posterior probability of class membership P ( $x$ — $y$ ) is estimated (West, 2008).

**Naïve Bayes** These popular probabilistic classifiers depend on three main assumptions. The first assumption is that no latent or hidden features would be influential in the classification. Another assumption for these classifiers is that there are no correlations between the features. The third is an assumption of normal distribution of the feature values within each class. Despite the second assumption, which is called class-conditional independence and that it is usually not met, this classifier provides fairly good results especially in combination with feature selection methods. The probabilistic model of the data is obtained by calculating a class-conditional probability density function for each feature vector using means and standard deviations (Witten & Frank, 2005).

A *Bayes' rule* can be used to form a classifier for each new feature vector by determining a posterior distribution for each feature vector, or the probability of a feature vector $x$ to belong to a class $y$ (Barber, 2010). This is possible to achieve because both the prior probability P ( $x$ ) – the probability of class membership before any observation is made about the classes– and the likelihood –probability of the class $y$ being the value $x$ present– have been estimated during the training process (Duda et al., 2001).

### Cross-validation

A prevalent procedure for evaluating the classifier performance by averaging out the error in the modeling of the data is called *cross validation*. This method partitions the data into training and test samples and repeats the whole classification routine a number of times using a given learning algorithm. Cross validation builds and evaluates multiple models, one for each cross-validation fold.

In this study, each time a new random stratified splitting of the primary dataset was performed. In other words, first the primary dataset was randomly split $n$ times into training and testing sets in order to obtain $n$ classification results. Finally, an overall classification accuracy was obtained by calculating the average of the cross-validated classification accuracies (Saari, 2009).

In a review on different validation methods by Refaeilzadeh, Tang, and Liu (2009), $k$-fold cross-validation is mentioned as the most widely used validation method for data mining. The drawback of this method is that the $k$ training sets can overlap to each other, however the $k$ test sets remain independent.

The most used cross-validation approach in data mining is 10-fold cross-validation, offering a compromise between generalizable predictions on one hand, and less overlapping training sets and relatively larger testing sets on the other hand. Multiple run $k$-fold

cross-validation (i.e., a repetition of the cross-validation and posterior averaging) produces a large amount of performance measurements and it has been recommended for reliable estimates (Witten & Frank, 2005), e.g. in binary data (Bouckaert, 2003), but it is not commonly used in data mining.

For the present study, one run of stratified 10-fold cross-validation has been performed in order to evaluate the extent to which the models could be generalized. Cross-validation is recommended in approaches where the amount of data is scarce; in studies where the available data was sufficient, this step was avoided (Downie, Ehmann, & Tcheng, 2005).

**Averaging** The training and test cross-validated models were subsequently averaged, obtaining 54 test classification accuracies for the "All" design and 162 test classification accuracies for the "Top" design.

### Evaluation

For the evaluation of the results, the performance measure *accuracy*

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

was used, and for each cross-validated model an average accuracy was calculated.

### Relevance estimates of individual attributes

The *rank aggregation* problem roughly refers to the search of an optimal ranking given two or more ordered lists. It is a combination of various ranking lists into one optimal rank ordering (Pihur, Datta, & Datta, 2007). In order to find an optimal ranking, a simple methodology called Borda count finds the average ranks of each element across all lists. Rank aggregation was performed using *R*, a programming language and software environment available under the GNU GPL [3]. An example of the code implemented for the Borda count can be found in Appendix C.

A total of 16 "optimal" lists of five attributes each were calculated based on the explored factors: six combinatorial feature subsets, seven attribute selection methods, and three music databases. For the attribute selection method we have calculated seven rank lists including the filter selection approach and six wrapper selection approaches, since we have explored three meta-classifiers for each of the greedy selection algorithms.

[3] http://www.r-project.org/

Moreover, a list combining all the wrapper selection methods was obtained. For the combinatorial subsets, six lists were created across all music databases. For the music databases, three lists summarize the ranks. These lists are presented in the second section of the next chapter.

# Chapter 4

# Results

The general purpose of the experiment was to compare the accuracies between the sub-band flux- and the MFCC-based models and to roughly estimate which of the attributes within these feature sets would be of higher relevance for music genre classification tasks. In the present chapter, the outcomes of our approach with respect to classification accuracies and relative relevance of attributes are examined in detail.

## 4.1 Performance

In order to obtain measures of performance for this study, the ten folds of each cross-validated model were firstly averaged. The results have been organized by averaging the results for each of the music databases, since the accuracies are substantially dependent on the data. Table 4.1 shows the average accuracies for both designs, each of these divided into the training and testing sets to offer an estimation of model overfitting.

For all the designs, the GTZAN model appears as the music database that yields the lowest accuracies. The table shows mostly better results for the "All" design when compared to the "Top" design. The "All" design contains models that were not subjected to feature selection.

The highest of the average test accuracies is been found for the "All" RWC music database at 62.46%, while the "Top" test models –i.e., those to which feature ranking and selection has been applied– for the GTZAN database show the lowest test accuracies at 36.24%. In spite of the design, the test accuracies showed lower accuracies than the train accuracies, following the expected tendency. Typically, lower results are found for the test accuracies, since the classification of instances that were not used for building the classification model is more challenging than the classification of the training set,

which has already been used for building the model. We also observe that for each of the databases, the higher accuracies are found for the "All" training models, while the "Top" test models yielded lower results.

| **Design and set** | **GTZAN** | **RWC** | **AF.RWC** |
|---|---|---|---|
| all – train | 58.67 | 67.66 | 68.31 |
| all – test | 50.66 | 62.46 | 55.35 |
| top – train | 44.03 | 62.07 | 62.91 |
| top – test | 36.24 | 57.41 | 52.57 |

TABLE 4.1: Mean accuracies per music database for both factorial designs.

Table 4.2 shows the results across music databases in order to offer an estimate of the mean performance provided by the combination of the spectral descriptors. It is clear, based on this information, that the feature ranking and selection leads to poorer classification results. If the results between the train and test sets are simply subtracted, a difference of 8.45 is found for the models that were not subjected to attribute selection and a difference of 7.41 for the "Top" designs.

| **Design and set** | **Accuracy** |
|---|---|
| all – train | 64.86 |
| all – test | 56.41 |
| top – train | 56.31 |
| top – test | 48.90 |

TABLE 4.2: Average accuracy for both factorial designs across music databases.

Table 4.3 shows results across music databases for each of the combinatorial subsets and each of the designs divided into train and test sets. The `mfcc.sbf.std` obtained the highest results for each category except for the "Top" test design accuracies, that where higher in average for the `sbf.std`. The `mfcc` combinatorial subset showed the lowest accuracies for each category. It is also apparent that the average accuracies were higher for the `sbf` than for the `mfcc`. When the test results between the "All" and the "Top" are compared, the highest amplitude for the `mfcc.sbf.std` is observed at 12.62, and

the lowest for the `sbf` means at 3.5, showing the expected tendency of obtaining better results when the combinatorial subsets have more attributes. A reduction from 10 to 5 attributes (`sbf`) yields to a much smaller difference in classification accuracy than a reduction from 46 to 5 attributes (`mfcc.sbf.std`). In addition, and for all the designs and sets, there is an increase in the classification accuracy for the models with standard deviations in the combinatorial subsets (`sbf.std` and `mfcc.std`) when compared to the models in which the subsets contained only mean values (`sbf` and `mfcc` respectively).

| **Design and set** | `mfcc` | `sbf` | `mfcc.sbf` | `mfcc.std` | `sbf.std` | `mfcc.sbf.std` |
|---|---|---|---|---|---|---|
| all – train | 59.79 | 60.81 | 66.26 | 64.59 | 64.53 | 73.31 |
| all – test | 48.79 | 54.15 | 58.40 | 54.74 | 56.88 | 63.98 |
| top – train | 51.42 | 57.85 | 57.99 | 52.94 | 58.74 | 59.08 |
| top – test | 42.06 | 51.1 | 50.98 | 45.17 | 51.79 | 51.36 |

TABLE 4.3: Accuracy across music databases for each combinatorial subset and for both factorial designs

Table 4.4 shows, for both experimental designs, the subtraction between the train and test set accuracies for each combinatorial subset. The data has been presented separately for each of the music databases. The sub-band flux mean subsets (`sbf`) yield the lowest absolute differences in all the cases, followed by the sub-band flux mean and standard deviation subsets (`sbf.std`). It can also be seen that the artist filtered music database (AF.RWC) shows the highest differences between training and test models, regardless of the combinatorial subset. The lowest differences are given for the RWC database. Also for all the music databases, the differences for the `mfcc.std` models are lower than the differences for the `sbf.std`, but they are higher when comparing `sbf.std` and `sbf`.

| **music database** | `mfcc` | `sbf` | `mfcc.sbf` | `mfcc.std` | `sbf.std` | `mfcc.sbf.std` |
|---|---|---|---|---|---|---|
| gtzan | 10.36 | 5.88 | 7.53 | 9.23 | 6.63 | 8.42 |
| rwc | 5.19 | 4.28 | 5.18 | 5.17 | 4.86 | 6.54 |
| af.rwc | 17.45 | 9.81 | 10.88 | 15.15 | 11.47 | 13.02 |

TABLE 4.4: Absolute differences in average accuracy between train sets and test sets for each combinatorial subset and music dataset

Figures 4.1 and 4.2 provide similar information to the results shown in Table 4.3. These two box plots, one for the models containing all the attributes and another for those only keeping the "Top" models, show descriptive statistics for the accuracies of each combinatorial subset. Two whiskers are shown for each subset; the black and white whiskers contain information about the training models, while the red and light blue whiskers refer to the test set models. For each whisker the thick horizontal line is the classification median, the lower and upper sides of the box are the first and third quartiles, the short horizontal lines are the sample minima and maxima (accuracies at $\pm 3\sigma$ from the mean) and the dotted circles are outliers.

These figures show clear differences between the combinatorial subsets, especially between the medians in Figure 4.1. Also for the same graph, the difference between training and test set medians in the combinatorial subset `sbf` is smaller when compared to other combinatorial subsets. In Figure 4.2 some differences can be found between models where sub-band flux means are or could be among the five chosen attributes; the difference between the medians is larger for `mfcc` and `mfcc.std` when compared to the rest of the combinatorial subsets.

It is also visible that Figure 4.1 presents less overlapping between the training and test set distributions than Figure 4.2. To illustrate this, the distance between the first and third quartiles and between sample minima and maxima for both training and test models is greater than in the models that were subjected to feature selection (presented in Figure 4.2). This suggests that the dimensionality reduction that has been implemented has increased the variability of the results.

Another difference between both designs refers to the outliers, which can occur due to chance, measurement errors or a tendency to leptokurtosis (high peakiness around the mean and fatter tails) in the distributions. The "All" design (Figure 4.1) presents less outliers than the "Top" design (Figure 4.2), and most of the outliers are for the training data –actually there are no outliers present in the testing data for the "Top" design. It can also be seen in both figures that the `mfcc` combinatorial subsets do not show any outlier models, and that the only combinatorial subset in which there are outliers at $+3\sigma$ instead of $-3\sigma$ is `mfcc.std` for the "All" design and the training set.

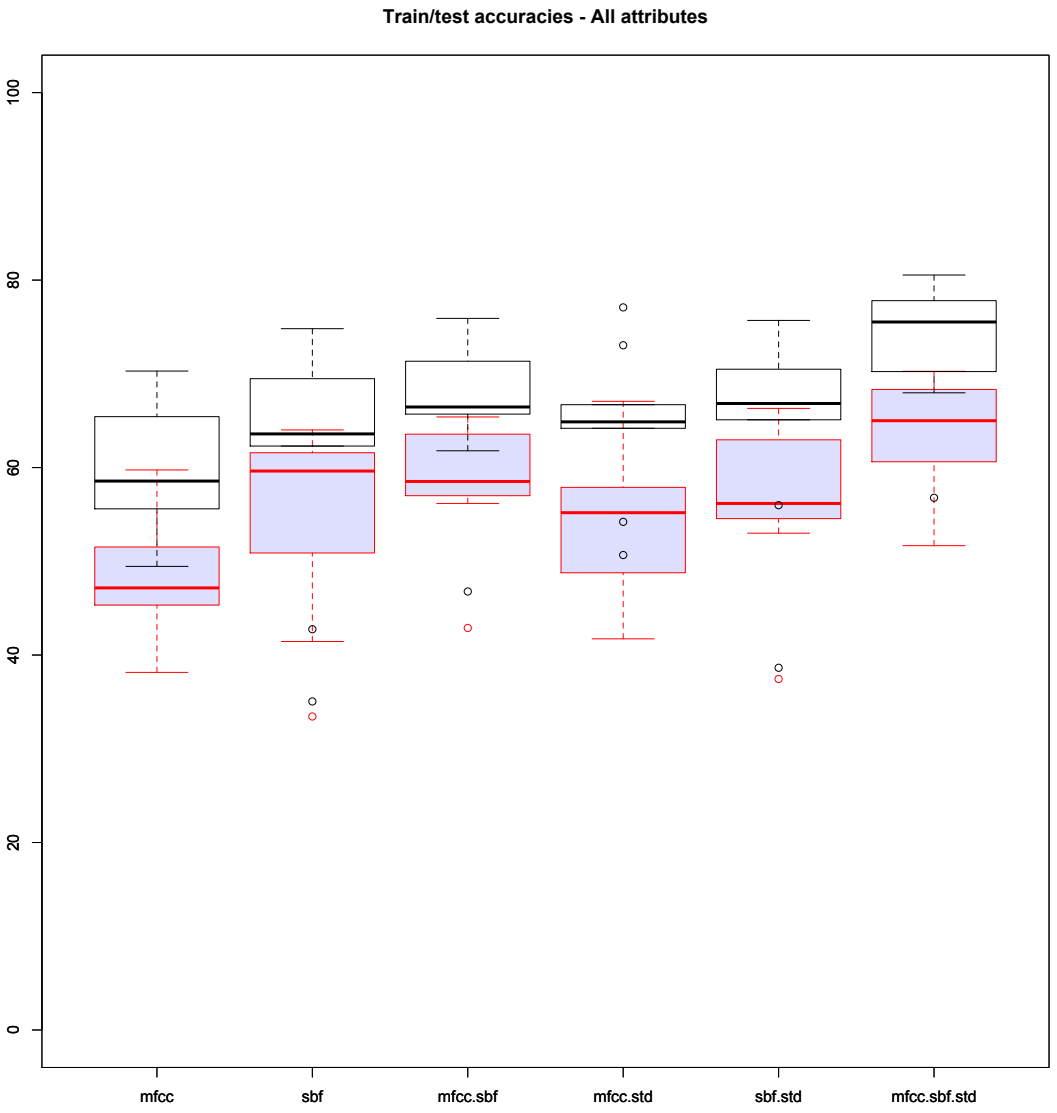


FIGURE 4.1: Accuracies for each combinatorial subset - "All" design. The white whiskers represent the training set descriptions while the blue whiskers depict statistical summaries for the test set.

Figures 4.3, 4.4 and 4.5 visualize, for each music database, classification results within each combinatorial subset. The results for each design and set are connected with lines for a better comparison. The red lines illustrate the combinatorial subsets with all the attributes and the brown lines represent the combinatorial subsets that were subjected to feature selection. Dashed lines and continuous lines were used to indicate training sets and test sets, respectively.

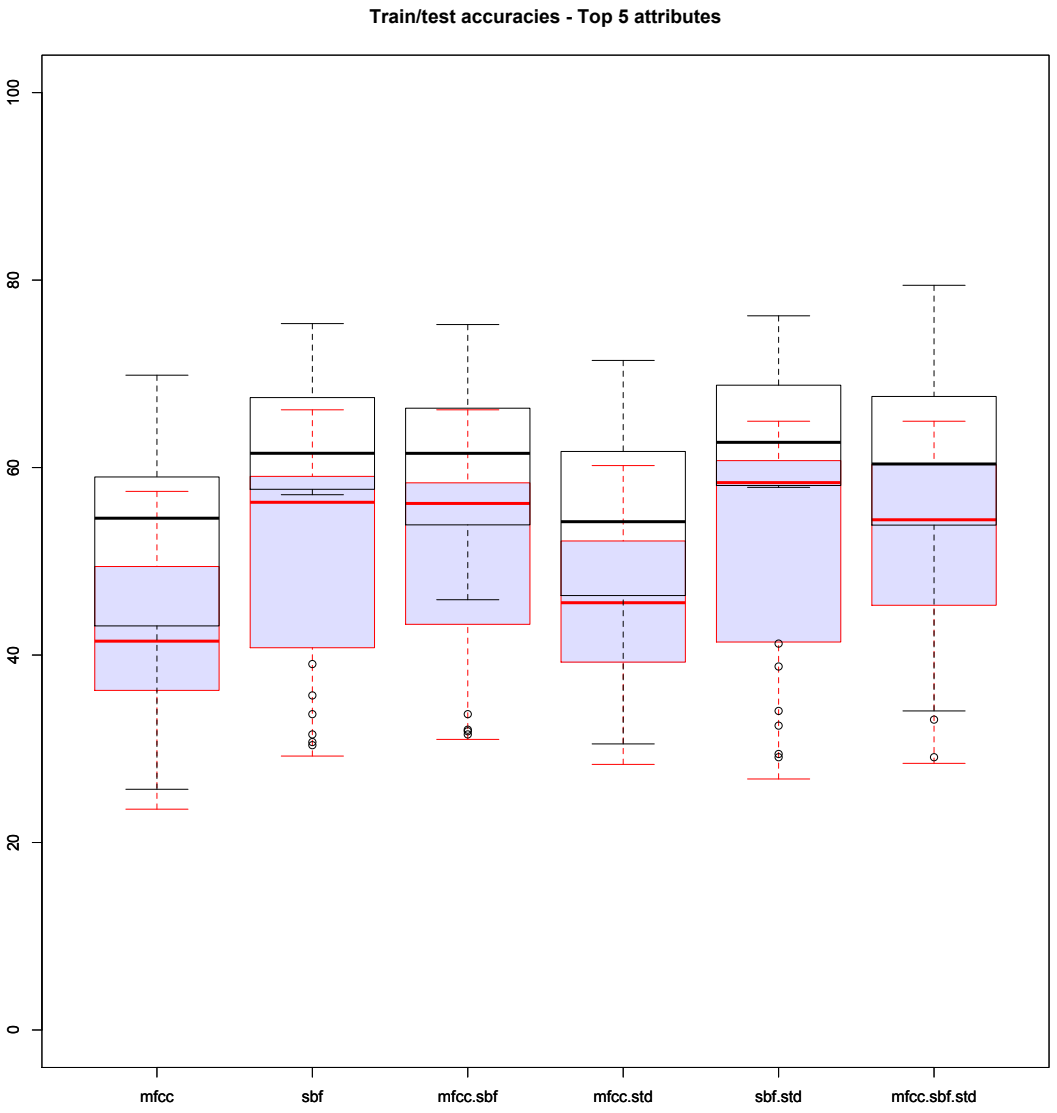


FIGURE 4.2: Accuracies for each combinatorial subset - "Top 5" design

In Figure 4.3 the accuracy profiles for the GTZAN music database are presented. For the `sbf` and `sbf.std` models, it can be seen that the attribute ranking and selection stage does not affect much the accuracies when compared to the "All" design accuracies. The most notable differences between designs appear in the `mfcc.sbf.std` models, perhaps either showing overfitting or an impractical attribute selection. Concerning the models with feature selection, their accuracy profiles are almost horizontally straight –as if the results were independent of the combinatorial subset.

In contrast with Figure 4.3, Figure 4.4 shows in overall higher accuracies for the RWC database. The accuracy profiles appear much more concentrated and there are clearer differences between each combinatorial subset. It can be noticed also in this figure

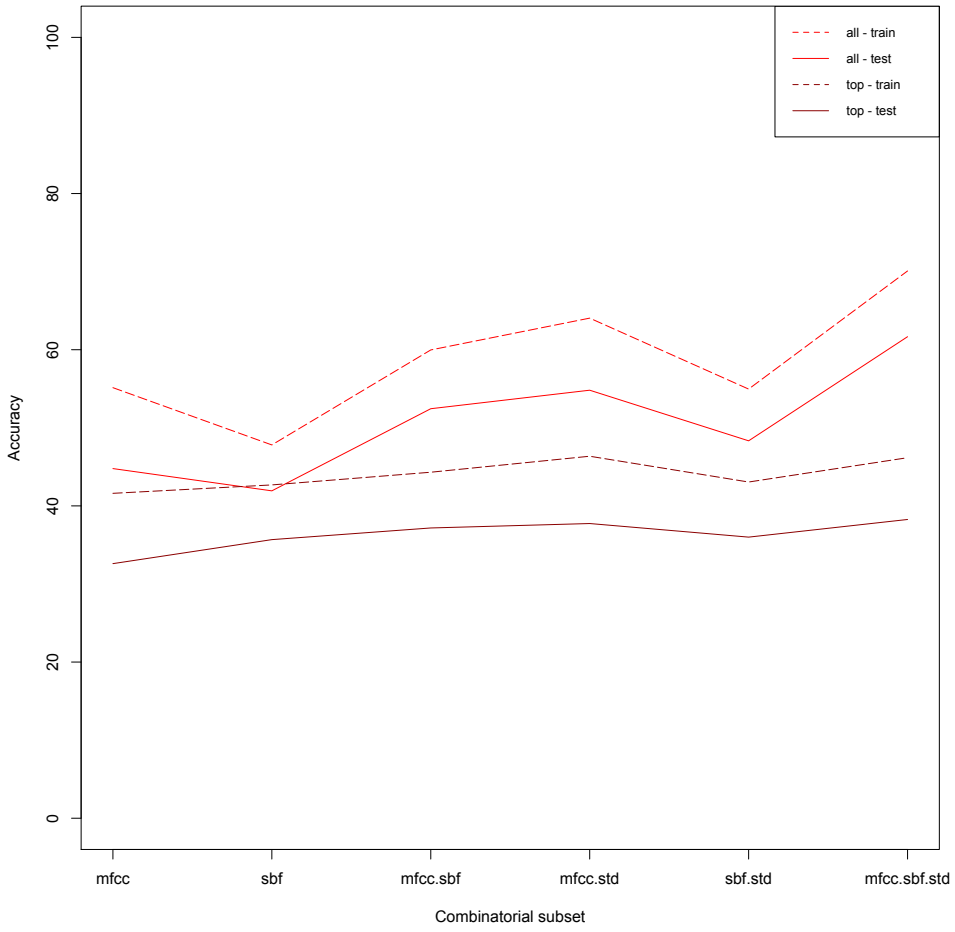


FIGURE 4.3: Accuracy profiles for each combinatorial subset and design - GTZAN music database.

that the `sbf` and `sbf.std` models do not decrease much in accuracy if dimensionality reduction is applied, as opposed to other models.

Figure 4.5 shows that the artist-filtered RWC music database yielded lower accuracies when compared with the unfiltered RWC database in Figure 4.4. These differences could be due to the reduction of both the artist and album effects and also to a lessening of the number of examples for some of the genre classes. In addition, it can be seen in Figure 4.5 for the artist-filtered RWC database that similar accuracies were obtained for "Top" and "All" design accuracies of a given combinatorial subset and set. It seems that the results were not sufficiently affected by the attribute reduction, and once again, this was especially the case for the `sbf` and `sbf.std` combinatorial subsets.

Considering the three music databases, and based upon the differences between the highest and the lowest accuracies that are present in the graphs, it can be noticed

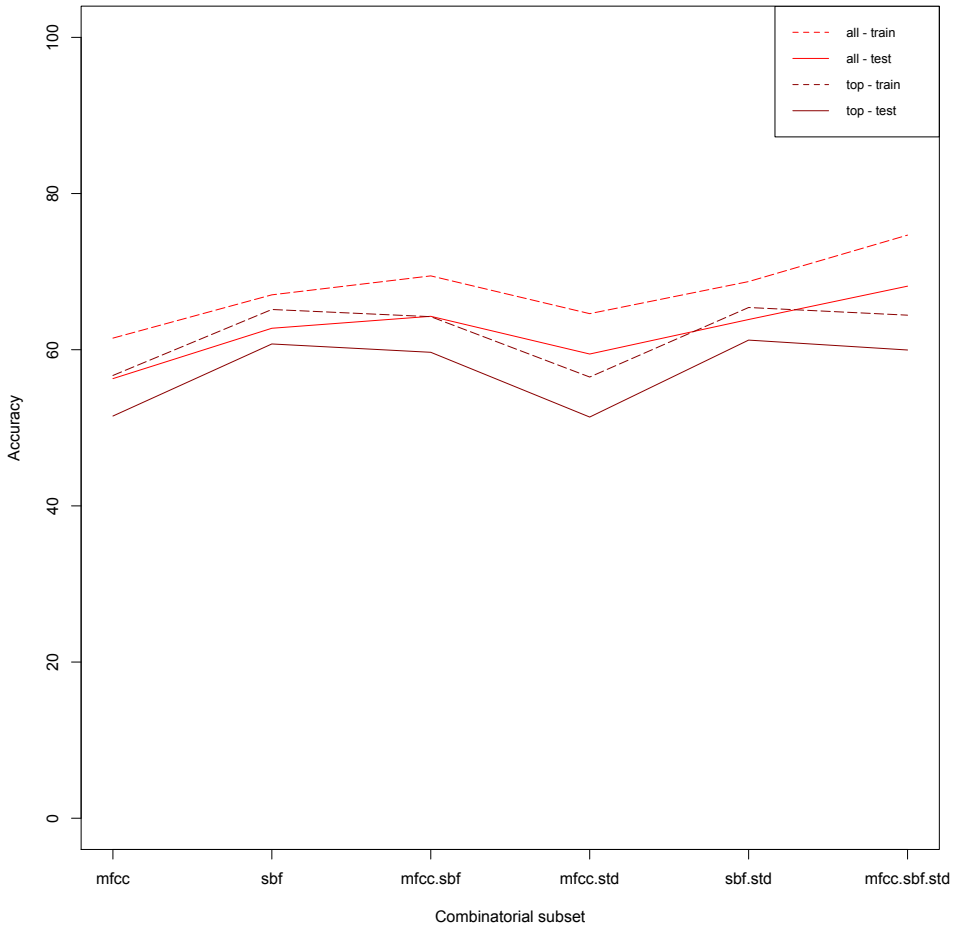


FIGURE 4.4: Accuracy profiles for each combinatorial subset and design - RWC music database

that the GTZAN database shows high variance in the accuracies, followed by AF.RWC. In comparison with the aforementioned databases, the RWC database models present relatively even results regardless of the design, set or combinatorial subset – except for `mfcc.sbf.std` and `mfcc.std` that yielded great differences between designs. Moreover, across combinatory subsets, RWC appears to be the database that leads to less over-fitted results due to small differences between training and test sets. On the contrary, the AF.RWC database models showed clear differences between the profiles of each design.

Overall it is observed that the `sbf` subset shows higher accuracies when compared to `mfcc` subset, although in the GTZAN database where the profiles are less regular and greater differences can be observed between the designs. However, based on the three figures

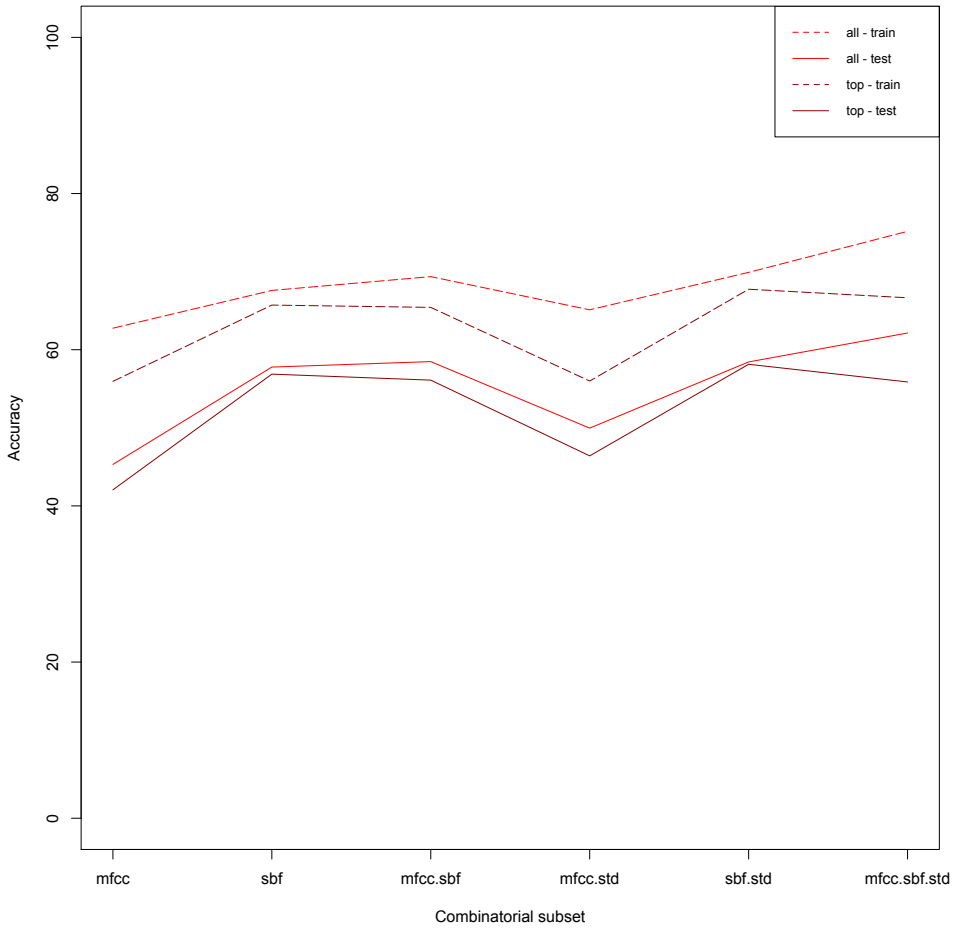


FIGURE 4.5: Accuracy profiles for each combinatorial subset and design - AF.RWC music database

it can be deduced that the dimensionality reduction affects more the `mfcc` accuracies than those of the `sbf` combinatorial subsets. This could suggest that there are more features in the sub-band flux set showing a high correlation than in the MFCC set and therefore the former presents more redundant features than the latter for an AGC task, since their inclusion does not significantly improve the results.

For the accuracies based on modeling with all the features of a subset, the combination of MFCC and sub-band fluxes (`mfcc.sbf`) can outperform the accuracies for `sbf` either mildly or highly (see GTZAN "All" design in Figure 4.3). However, the difference in accuracy between both "All" and "Top" designs is higher for `mfcc.sbf` with respect to `sbf`. In addition to this, the "Top" `mfcc.sbf` accuracies are higher than those of the "Top" `mfcc` subset accuracies, showing that for subsets of the same size, the addition of sub-band flux descriptors can improve the overall accuracy.

It can also be added that the combination of MFCC and sub-band flux (`mfcc.sbf`) yields quite similar results, for all music databases, designs and sets, to the combination of sub-band flux means plus its standard deviations (`sbf.std`). Finally, the "Top" design results between `mfcc.sbf` and `mfcc.sbf.std` are fairly similar to each other. We will analyze this information below based on the rank aggregation results.

## 4.2 Relative attribute relevance

As previously mentioned in Section 3.3, rank aggregation refers to a search for optimal ranked lists that offer a summarization of more than two rankings. This method was utilized in the "Top" design, since it consists in ranked datasets of five elements each. The aim was to estimate the relevance of the sub-band flux and MFCC individual attributes across these datasets. The results were organized into 16 "optimal" lists that are presented in Tables 4.5, 4.6 and 4.7. Notably, a total of 53 occurrences in the lists for the sub-band flux attributes, from which 39 corresponded to the mean values and 14 to the standard deviations. For the MFCC attributes 27 occurrences were found, from which 23 were mean values and 4 corresponded to their standard deviations.

The listed sub-band flux means, ranked by number of occurrences are №2 (9 occurrences), №10 (6 occurrences), №3 (4 occurrences), №4 (4 occurrences), №6 (4 occurrences), №8 (4 occurrences), №9 (4 occurrences), №7 (2 occurrences), №1 (1 occurrence) and №5 (1 occurrence). The sub-band flux standard deviations are №2 (6 occurrences), №10 (3 occurrences), №8 (2 occurrences), №9 (9 occurrences) and №3 (1 occurrence).

The MFCC mean that appears the most in the rankings was the MFCC №1 (9 occurrences). It was followed by MFCC means №3 (3 ocurrences), №13 (3 ocurrences) and №4 (2 ocurrences). The mean of the coefficients №5, №6, №8, №10, №11 and №12 were listed once. The MFCC standard deviations that were listed in the aggregated ranks are №1, №3, №8 and №11, with one appearance each.

In addition to the mentioned rankings, we have combined all the top 5 attribute rankings from each fold, except for the models using filter selection, in order to obtain a general optimal ranking across all wrapper methods and music databases.

In Table 4.5, each ranking combines a total of 18 datasets. Based on the results after grouping the datasets according to seven approaches for dimensionality reduction, it appears that the sub-band flux means №2 and №6, as well as the MFCC mean №1

are relatively relevant. It can be taken into account at this point that the wrapper method with backwards elimination (`wra.be`) for feature selection yielded relatively low accuracies, as shown below in Appendix A.

| **Ranks** (1 is best) | **gainr** | **wra.be knn** | **wra.be bayes** | **wra.be svm** |
|---|---|---|---|---|
| 1 | sbf mean 2 | sbf mean 6 | sbf mean 6 | mfcc mean 13 |
| 2 | sbf mean 10 | sbf mean 4 | sbf mean 8 | mfcc mean 10 |
| 3 | mfcc mean 1 | mfcc mean 8 | sbf mean 9 | sbf mean 5 |
| 4 | sbf std 2 | mfcc mean 13 | mfcc mean 11 | mfcc mean 5 |
| 5 | sbf mean 3 | mfcc mean 12 | sbf mean 7 | sbf mean 6 |
| | | **wra.fs knn** | **wra.fs bayes** | **wra.fs svm** |
| 1 | | mfcc mean 1 | mfcc mean 1 | mfcc mean 1 |
| 2 | | sbf mean 2 | sbf mean 2 | sbf mean 2 |
| 3 | | sbf mean 10 | mfcc mean 3 | sbf mean 4 |
| 4 | | sbf mean 8 | mfcc mean 4 | mfcc mean 3 |
| 5 | | sbf std 2 | sbf mean 10 | sbf mean 3 |

TABLE 4.5: Aggregated rankings of top attributes for seven feature selection approaches

In Table 4.6 there are six rankings for the combinatorial subsets; each list groups together 21 datasets belonging to a combinatorial subset. Sub-band flux means and standard deviations №2 and №10, as well as MFCC mean №1 appear more than once in the lists. It is noteworthy that the sub-band flux №5 is not ranked in this list nor in the following, since it clearly represents a part of the human audible range (between 400-800 Hz).

As we have mentioned above, the results between `mfcc.sbf` and `mfcc.sbf.std` are for this design quite similar, but the results for the `mfcc.sbf.std` aggregated rank do not show information that is consistent with those results. It would be expected that most of the aggregated ranks for this subset had been related with the mfcc.sbf subset, that is, either MFCC means or sub-band flux means. However, most of the ranks listed correspond to standard deviations. In any case, finding similar results does not imply that similar methods were used.

In Table 4.7 each of the columns refers to the rank aggregation of 42 datasets. The aggregated rankings for each music database show a general tendency towards descriptors

| **Ranks** (1 is best) | `mfcc` | `sbf` | `mfcc.sbf` |
|---|---|---|---|
| 1 | mfcc mean 1 | sbf mean 2 | sbf mean 2 |
| 2 | mfcc mean 4 | sbf mean 10 | sbf mean 10 |
| 3 | mfcc mean 6 | sbf mean 9 | sbf mean 8 |
| 4 | mfcc mean 3 | sbf mean 4 | sbf mean 3 |
| 5 | mfcc mean 13 | sbf mean 6 | sbf mean 9 |
| | `mfcc.std` | `sbf.std` | `mfcc.sbf.std` |
| 1 | mfcc mean 1 | sbf std 2 | sbf std 2 |
| 2 | mfcc std 11 | sbf std 10 | sbf std 10 |
| 3 | mfcc std 1 | sbf std 8 | sbf std 8 |
| 4 | mfcc std 3 | sbf std 3 | sbf mean 2 |
| 5 | mfcc std 8 | sbf std 9 | sbf std 9 |

TABLE 4.6: Aggregated rankings of top attributes for six combinatorial subsets

of low value, especially №1 and №2. However, there is a tendency towards high values in the aggregated rank for the GTZAN music database. One descriptor, namely the MFCC mean №1 appears ranked for the three databases.

| **Ranks** (1 is best) | **gtzan** | **rwc** | **af.rwc** |
|---|---|---|---|
| 1 | mfcc mean 1 | sbf mean 2 | sbf mean 2 |
| 2 | sbf mean 10 | mfcc mean 1 | mfcc mean 1 |
| 3 | sbf mean 9 | sbf mean 1 | sbf mean 3 |
| 4 | sbf mean 8 | sbf mean 7 | sbf mean 4 |
| 5 | sbf std 10 | sbf std 2 | sbf std 2 |

TABLE 4.7: Aggregated rankings of top attributes for three music databases

Finally, Table 4.8 summarizes the rankings of 108 datasets using a wrapper approach for feature selection. Based on this table, it appears that the sub-band flux means are relatively more relevant than its standard deviations and than the MFCC descriptors.

Overall, both the performance measures and the aggregated ranks offer clear tendencies that suggest the suitability of the sub-band flux feature set for the purpose of music

| **Ranks** (1 is best) | **Wrapper feature selection** |
|---|---|
| 1 | mfcc mean 1 |
| 2 | sbf mean 2 |
| 3 | sbf mean 7 |
| 4 | sbf mean 4 |
| 5 | sbf mean 8 |

TABLE 4.8: Aggregated rank of top attributes combining all the "wrapper" feature selection methods

genre classification, particularly against the MFCCs. The next chapter puts forward general considerations on the results described above.

# Chapter 5

# Discussion

In the present chapter, an exploration on the implications of the obtained results for music genre classification and Music Information Retrieval in general is presented. The focus of the analysis is set on model overfitting, classification performance and relative attribute relevance. Moreover, significant limitations of this study and suggestions for further research are discussed.

## 5.1 Overfitting models

The in-depth analysis presented in the last chapter suggests that the designs in which attribute selection was applied yielded results that might over-fit less, as it is illustrated in Figures 4.1 and 4.2, e.g. by comparing the overlapping between whiskers. In contrast, and for the design that yielded less likeliness of model overfitting, the MFCC means –and perhaps the standard deviations of these features as well– showed higher overfitting tendencies, as it can be inferred from the difference between medians in 4.2 and as suggested from Table 4.4.

Comparing the accuracies between the train and the test sets (Table 4.4) we find larger differences in the MFCC subsets than in the case of the sub-band flux. This finding could indicate a greater tendency to over-fit of both `mfcc` and `mfcc.std` when compared to `sbf` and `sbf.std`. A smaller tendency of overfitting in sub-band flux models -or in models where the sub-band flux could have been chosen as an attribute in the feature selection stage- can be visualized in Figures 4.1 and 4.2 when comparing the difference between medians for the training and test sets. This difference between the train and test accuracy can be referred to as generalization ability. As seen in Table 4.4, a higher generalization ability is found in the models where the only descriptors were sub-band

fluxes. Within small combinatorial subsets, the models that only contained sub-band flux means yielded the lowest tendency to over-fit.

Another finding that can be inferred from the accuracy profiles is that a reduced combination of sub-band flux means and MFCC means ("Top" `mfcc.sbf` models) performs on average better than a subset of top MFCC descriptors of the same number of dimensions ("Top" `mfcc`), while showing similar difference between train and test sets and thus overfitting risks of comparable magnitude.

## 5.2 Accuracies

Based on the averaged classification accuracies, a number of critical observations can be made about the studied descriptors. First of all, Tables 4.2 and 4.3 highlight the expected tendency of large combinatorial subsets to increase the classification rates. Other interesting remarks on the chance levels of the databases, on the musical features under study and on the dimensionality of the models are described below.

### Chance level

The chance level was lower for the classification models using the GTZAN database. This makes sense because this database has more classes than the other databases that were investigated. The lower chance level could partly explain the relatively low accuracies yielded by this database. Comparing the accuracy profiles of Figure 4.3 with the profiles for the databases RWR and AF.RWC, great differences are found for both designs as well as a poorer overall classification performance. Probably the differences between the GTZAN accuracies and the other models are better explained by their chance levels.

### Descriptors

About the performance that were found for the investigated descriptors, better results have been yielded by the sub-band flux than the MFCCs. Table 4.3 shows a better overall performance for the sub-band flux when compared to the MFCCs, and this difference appears to be more important when the number of attributes is reduced.

With respect to the accuracy profiles presented in Section 4.1 of the previous Chapter, it is possible to do some remarks on the sub-band flux means in comparison to the MFCC means. The combinatorial subsets that consisted only in sub-band flux means (`sbf`) showed less tendency to over-fit than the MFCC means subsets, suggesting a higher

ability to generalize to other scenarios. Moreover, the former yielded higher accuracies or little decrease in test accuracies (see Figure 4.3) with respect to the latter. It remains noteworthy that being the sub-band flux a set with less attributes than the MFCCs, the sub-band flux descriptors showed less decrease in performance when the number of dimensions was reduced to five.

The described tendency is similar when comparing the versions of these combinatorial subsets that included their standard deviations (`sbf.std` and `mfcc.std`, respectively). However, mixed results are seen after a comparison of mean accuracies between the sub-band flux combinatorial subset (`sbf`) and the sub-band flux plus the standard deviations subset (`sbf.std`) and equivalently between the MFCC means and its subset that included the deviations. Based on Tables 4.3 and 4.4, it can be suggested that the "addition" of the standard deviations to the sub-band flux means and to the MFCC means subsets could help in raising the classification accuracy. However, this addition apparently decreases the model overfitting in the case of the MFCCs, and increases overfitting of the sub-band flux models.

Based on these findings it seems that an optimal way to extract the sub-band flux would be by only including mean values; and that the best way to extract MFCCs from the signal would be including both means and standard deviations.

As we mentioned in Section 4.1, it is also noteworthy that the MFCC and sub-band flux sets combined subset (`mfcc.sbf`) is comparable to the combination of the sub-band flux means and standard deviations (`sbf.std`). This would suggest that the MFCC means perform similarly to the standard deviations of the sub-band flux, unless these similar results only occur when these descriptors are combined with the sub-band flux means.

### Dimensionality

Table 4.1 shows that the models that were subjected to feature ranking and selection did not perform as accurately as the models for which no reduction of dimensionality was made. However, it is worth of notice that in the designs with dimensionality reduction, the number of attributes was at least halved, so it was expected to find drops in the classification rates.

The accuracy profiles have shown that, for the models based on the RWC and AF.RWC music databases, the combinatorial subsets that contained only sub-band flux attributes (`sbf` and `sbf.std`) have yielded very similar results for the reduced design and for the design that was not subjected to dimensionality reduction. Yet the GTZAN reduced

model accuracies for these subsets were relatively similar to the full subsets, when compared to models corresponding to the other music databases. This relative independence of feature selection for the sub-band fluxes has at least one possible explanation, namely, that there could be dispensable attributes in the sub-band flux set as opposed to the MFCCs for a genre classification task.

## 5.3 Relevance of attributes

In the last chapter, compelling results have been presented on the lists of attributes that could be considered as relatively relevant for the purpose of music genre classification. From the analysis of the aggregated rankings for the best five attributes of each dataset, it has been discovered that the sub-band fluxes have been more times among the best attributes than the MFCC coefficients and that the sub-band flux set yielded more unique relevant attributes than the MFCC set.

The total number of sub-band flux attributes that were listed in the aggregated ranks has almost doubled the amount of MFCC attributes. From these amounts, the number of standard deviations was more than 17% for the MFCC attributes versus almost 36% for the sub-band flux descriptors. This means that the relative relevance of the standard deviations over the means is higher for the sub-band flux than for the MFCCs, as if it made more sense to incorporate the sub-band flux standard deviations to the sub-band flux means than to do the same with the MFCC standard deviations – however, this idea is in clear contradiction with the findings presented in Section 5.2.

According to the aggregated "Top 5" ranks, all the sub-band flux means are relatively relevant since the bands №1-10 have been listed in the aggregated ranks. From the MFCC means, all the coefficients except №2, №7 and №9 could be considered as relatively relevant. Out of 13 standard deviations for the MFCCs, only 4 attributes were listed, while half of the 10 sub-band flux standard deviations have been computed within the aggregated ranks.

Overall, the most frequent attributes in the lists were the sub-band flux mean №2 and the MFCC mean №1. For the standard deviations, the most frequently listed attribute was the sub-band flux deviation №2; the MFCC standard deviations were multimodal, with one occurrence each.

The relative relevance of MFCC mean №1 suggested by the rank aggregation can be explained by the fact that this coefficient correlates highly with the spectral slope, a representation of the decrease in amplitude as a function of frequency (Theimer et al., 2008). Further, strong negative correlations have been found between this second MFCC

coefficient and the spectral rolloff (Pedersen & Diederich, 2007). The spectral rolloff gives account of the brightness of the signal and yielded moderately high positive correlations with the dimension of Activity in the study on polyphonic timbre perception by Alluri and Toiviainen (2010a).

The relative relevance of the sub-band flux mean and standard deviation №2 can be linked to the moderate positive correlation of the mean №2 with the perceptual dimension of Fullness found by Alluri and Toiviainen (2010b), and also to a moderate correlation with the perceptual dimension of Activity in another study (Alluri & Toiviainen, 2010a). Another pair of attributes that were also found many times as relevant, the sub-band flux mean and standard deviation №10, has an analogous relationship with the moderate positive correlations between the sub-band flux mean №10 and the Activity perceptual dimension, found in the latter study.

## 5.4 Limitations

This study has a number of technical and design limitations that may have affected the results and should be taken into consideration. These barriers concern the general approach that was implemented, the music databases that were used, the chosen feature extraction parameters, the feature selection and rank aggregation stages, and the problem of overfitting.

### Bag-of-frames approach

One of the biggest limitations of the current "bag-of-frames" approaches in genre classification is the fact that merely the statistics of the feature values are used for building the classification model. Because of this, the classification results might remain the same if the content of the audio signal was a random rearrangement in time (i.e. "spliced and scrambled" version) of a musical signal (Aucouturier, 2008). To exemplify this, a randomly reorganized version of a classical music excerpt may be classified as experimental music by a human but a machine learning algorithm would still assign it to the classical group. An important outcome of the "bag-of-frames" approach is that any musical feature will offer a description of the average evolution of the signal based on short fragments. This is averse to spectrotemporal features, compelling these to offer a poorer description, if any, of the long-term development of the musical signal over time.

### Databases

A number of limitations have been found with respect to the databases that are originated by the lack of open data available for genre classification tasks. Besides that, other issues regarding the preprocessing and analysis of this study must be mentioned. First, the preprocessing of the AF.RWC database did not include the Electronic class, due to an error in the processing. This makes it more difficult to compare the results of RWC and its artist-filtered version AF.RWC. Second, a better analysis of the feature sets could have been done if the classification results were analyzed for each genre class. Furthermore, other limitations must be mentioned concerning the music databases:

**No validation set** It would have been possible to holdout, for each of the music databases, a percentage of the training data for validation in order to provide a more correct analysis of the classification performance, although it would have been necessary to reduce the training set size and therefore the number of examples used in order to build the model for each cross-validation fold.

**Improper investigation of artist and album effect** The effect of adding artist and album filtering could not be analyzed, mainly due to the lack of proper databases for investigating these effects. The number of examples contained in the AF.RWC database is fairly small, thus the results yielded by the use of this database could not be contrasted with those found for the RWC database.

**Duplicates in used databases** Both of the original music databases that were used in this study (GTZAN and RWC) contain a number of duplicated songs. For example, 7 duplicates were found in the GTZAN database. The classification accuracies should be underestimated since the same song could have been assigned to both training and test sets for some of the models.

**Unbalanced data and lack of examples per class** The databases RWC and especially AF.RWC are highly unbalanced, meaning that the number of songs per class is considerably variant. Further, some of the classes are significantly underrepresented due to an insufficient amount of songs. This might have operated as a confounding variable, thus affecting the classification accuracies.

### Feature extraction

For the feature extraction stage, the implemented design aimed to trim a total of 50 seconds from the middle of each file. However, the RWC and AF.RWC databases contained only 30 seconds excerpts. As a result, the extraction of these two databases was made with shorter audio when compared to the GTZAN dataset. Thus, the results for each database are harder to compare.

### Feature selection and rank aggregation

Some of the choices for the design of the feature selection and feature ranking stage where done in an arbitrary way. Since the generated combinatorial subsets `-mfcc`, `mfcc.sbf`, and so forth- do not contain the same subset size, it could be considered a mistake to select the "top" five attributes of each subset. To solve this problem, an option would have been to reduce the dimensionality using a ratio of the combinatorial subsets. This could have been done, for example, by retaining the best ranked $\frac{1}{4}$ ratio of attributes for each dataset. However, this option is not available in Weka, the software used for feature selection, and it would have also added challenges to the design of the rank aggregation stage.

About rank aggregation, the relative relevance of the individual attributes that were ranked is a matter not yet decided. The weights for the attribute ranking could not be collected, therefore the Rank Aggregation was done using an unweighted method. Consequently, it could be the case that all the ranked attributes had the first place in the aggregated ranks. This means that the results for the "Top 5" lists could have been randomly obtained.

### Overfitting

The problem of overfitting should be taken seriously, especially for datasets with a reduced number of instances (Kohavi & Sommerfield, 1995). In this study, the analysis included models that were likely to be highly over-fitted due to their factor level combination in the design. For example in the AF.RWC models, the `mfcc.sbf.std` subsets have too many attributes and only few instances. The final overall accuracies could show a bias in favor of these models. It would have been optimal to limit the analysis by excluding the results of the smaller datasets and of the bigger combinatorial subsets to avoid highly over-fitted outcomes.

Again, it must be considered that there is no validation set in the design. The difference between train and test accuracy is not enough as an estimate of overfitting and generalization ability. Therefore, the reported remarks on overfitting for this study should be taken with reservation.

## 5.5 Conclusions

In a nutshell, we found a total of four fundamental outcomes of this study. Firstly, the results show a better overall performance in accuracy for the sub-band flux than for the MFCC. These results were clearer when the number of sub-band flux bands and MFCC coefficients was reduced to five. Second, overfitting risk might not increased but instead decreased after the dimensionality reduction –an opposite scenario was expected for the case of the models subjected to wrapper feature selection. Third, the results suggest higher tendency to model overfitting for the MFCC models than for the sub-band flux models. It can be thus conjectured that the sub-band flux yields a higher generalization ability. Finally, the relative feature relevancy of sub-band flux bands on both ends of the frequency spectrum gives support to previous findings on the suitability of this feature set for polyphonic music similarity.

Future research can focus on whether or not the sub-band flux is convenient for other machine learning scenarios in Music Information Retrieval such as mood, artist and audio tag classification, since arguably spectrotemporal descriptions of polyphonic music are important in this tasks. Moreover, other challenges such as music structure analysis can be benefited from these outcomes. Clearly, these designs shall be more cautious in order to provide a more solid validation of the results, for example by incorporating a validation set. Finally, other musical contexts, such as latin or non-western music genres can deliver interesting insights on the potential of the sub-band flux feature set for music genre classification.

The models for music genre classification that have been tested in this study might result very simple when compared to the processes of human classification. Music conveys timbral information that probably serves as a crucial component for musical “class membership” decisions in humans, as supported by perceptual studies that found accurate musical genre recognition times of about 250 milliseconds (Gjerdingen & Perrott, 2008). However, the timbral information that is contained in the music is surely not enough as

a description for genre class decisions. Music genre classification probably depends on subjective and active processes that are developed early in life (Piaget, 1986).

Nevertheless, this study tried to shed light on the importance of perceptually interpretable features for music information retrieval purposes. The results have supported the evidence that these descriptors offer better accuracies and could be generalized better to other musical data than descriptors of unproved perceptual motivation, while lowering model complexity. Further research is needed to evaluate the scope of these findings and possibly extend it to other problems in Music Information Retrieval and computational music analysis.

# Appendix A

# Level plots for factorial designs and sets

In the present Appendix the level plots for the "All" and "Top" factorial designs are presented. These are of special interest for future reference, because they provide mean accuracies for the 396 classification models that were obtained based on all the levels of the full-factorial designs (see Tables 3.1 and 3.2).

The first figure (Figure A.1) presents the train set accuracies for the design that did not include dimensionality reduction ("All" design). The second figure (Figure A.2) shows the classification models for the test set accuracies of this design. The third (Figure A.3) and fourth (Figure A.4) figures refer to the design that included a feature ranking and selection stage ("Top" design) resulting in five attributes per model; these figures contain the classification accuracies for the training and test set models, respectively.

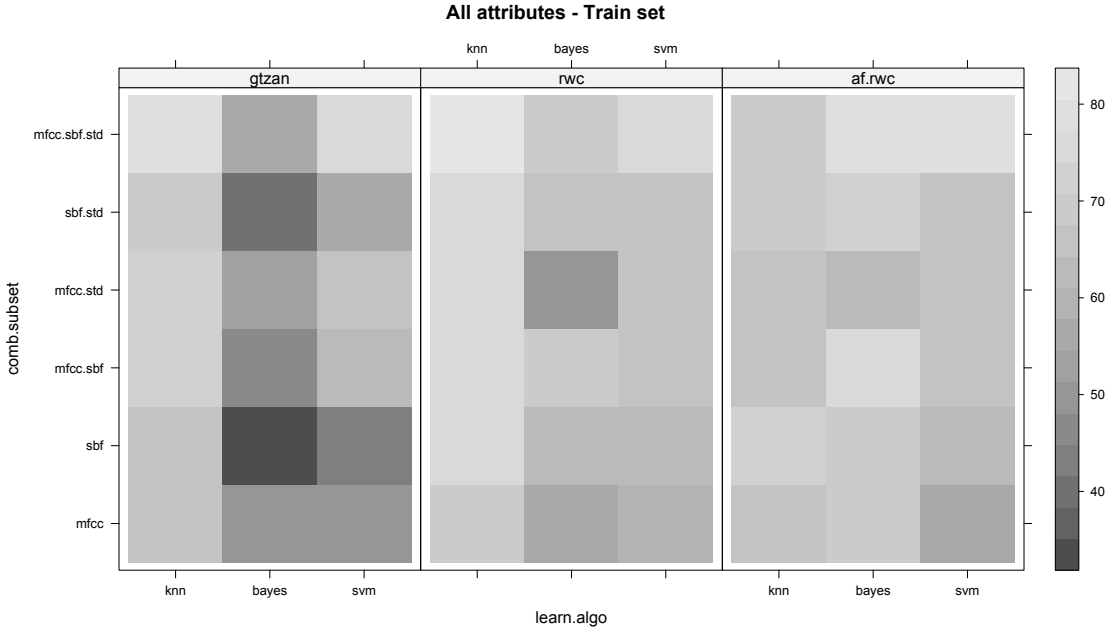


FIGURE A.1: Train set accuracies for the "All" design. A clear drop of accuracy can be observed for the GTZAN database models builded with the Naïve Bayes ("bayes") classifier. The k-nearest neighbor ("knn") classifier has yielded high train accuracies.

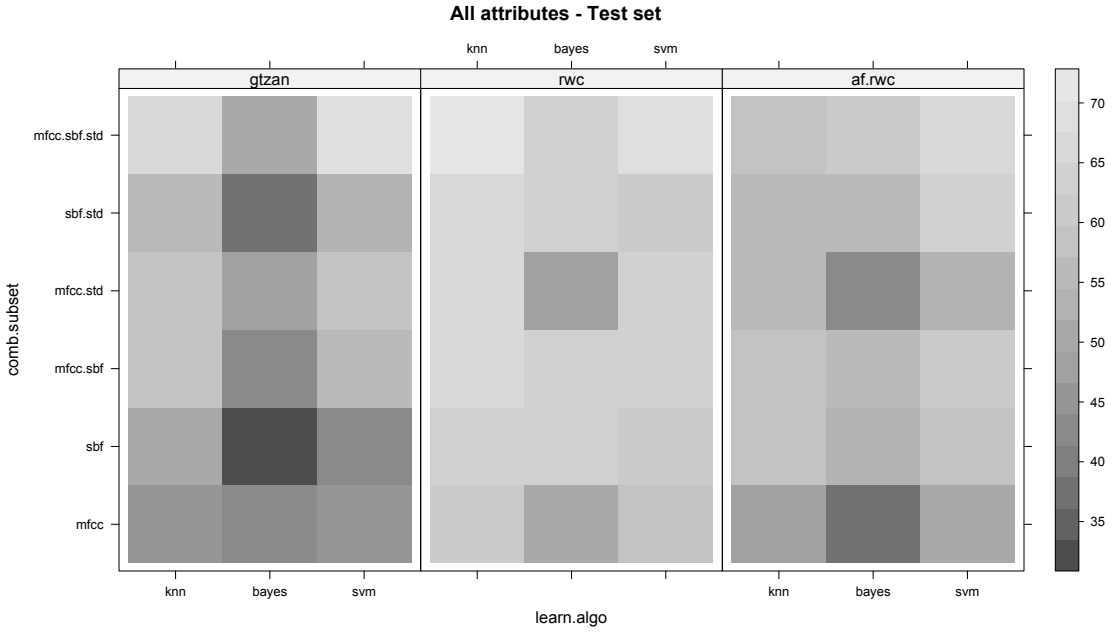


FIGURE A.2: Test set accuracies for the "All" design. The results follow a similar pattern with those of the previous figure, although the accuracies are lower (notice that the color scale has changed).

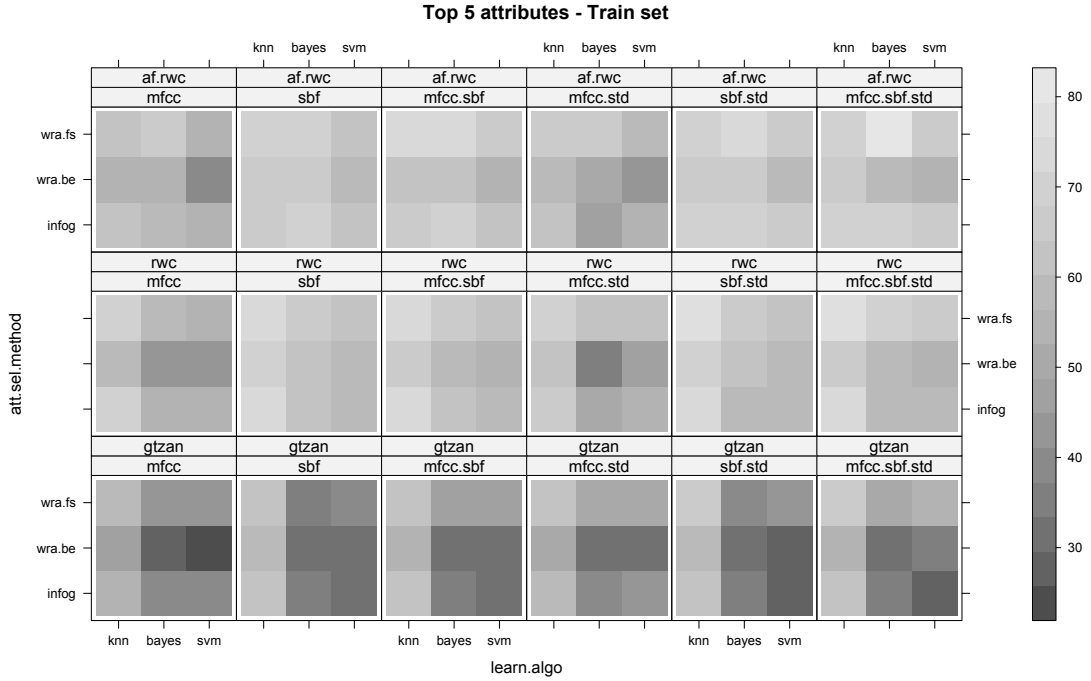


FIGURE A.3: Train set accuracies for the "Top" design. The wrapper feature selection method using a backwards elimination approach ("wra.be") yields relatively low results for the GTZAN database models. Also good overall results are obtained with the k-nearest neighbor ("knn") classifier.

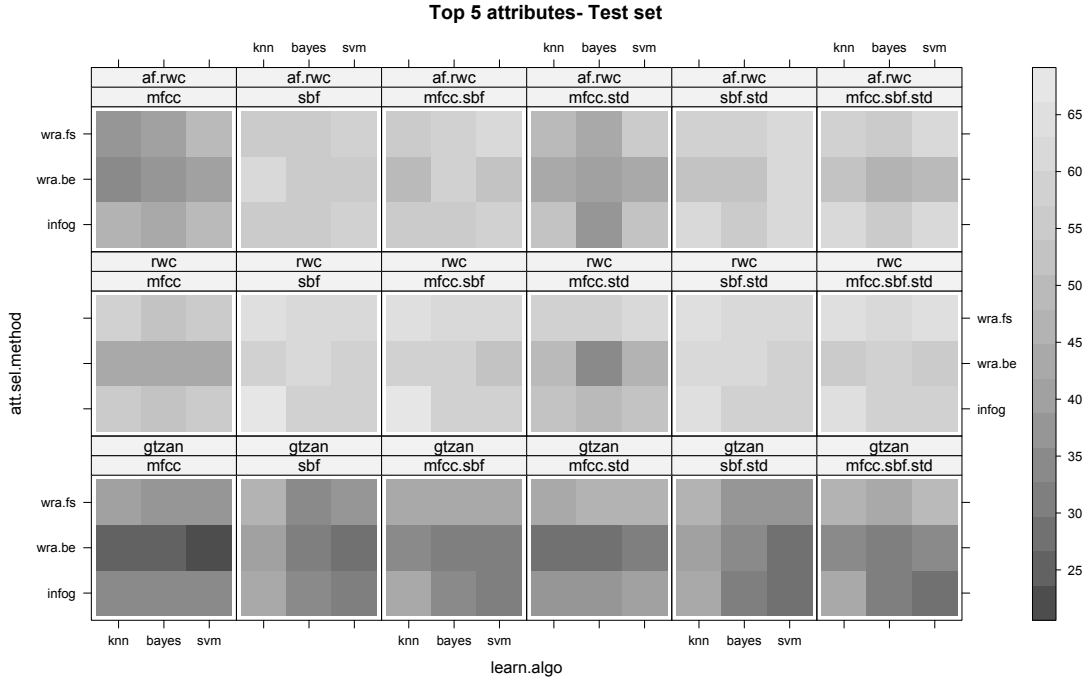


FIGURE A.4: The pattern of classification accuracies is similar to the previous figure, but showing lower accuracies (the color scale has changed in this figure with respect to the previous one).

# Appendix B

# Feature extraction in MIRtoolbox

The complete code for the feature extraction that was implemented for this study is presented below. For the following design in MIRtoolbox for Matlab we counted with the precious help of Olivier Lartillot, Vinoo Alluri and Pasi Saari.

```
1
3
% Flowchart design for feature extraction
5
a = miraudio ('Design', 'Extract', -25, +25, 'Middle',
7 'Sampling', 44100, 'Label', [1 2 3]);

9 % Mel-Frequency Cepstral Coefficients

11 myflow.mfcc = mirmfcc(a, 'Frame', .025, 's', 50, '%');

13
% Sub-Band Flux
15
myflow.sbflux = mirflux(mirfilterbank(a, 'Manual',
17 [-Inf 50*2.^(0:1:8) Inf],'Order',2), 'Frame', .025, 50, '%');

19 % Average and standard deviation along frames

21 myflow = mirstat (myflow);

23 % Application of the object to the audio files contained in folders (each
    folder name corresponds to the musical genre of the included excerpts)

25 output = mireval (myflow, 'Folders');
```

```
27 % Exportation of the statistical information into an attribute-relation
       file

29 mirexport ('export.arff', output);
```

# Appendix C

# Rank aggregation - Borda count method

The code for the Borda count method that was utilized for generating aggregated ranks of attributes is based on the the following example and function written in R. The author of the code, repeated verbatim, is Vasyl Pihur, developer of the R *RankAggreg* package.

```
2
x <- matrix(0, 10, 5)
4 for(i in 1:10)
  x[i,] <- sample(1:13, 5)
6
borda <- function(mat){
8   uniq <- sort(unique(as.vector(x)))
  ranks <- matrix(c(uniq, rep(0, length(uniq))), ncol=2)
10
  ranki <- list()
12  for(i in 1:nrow(ranks)){
    ranki[[i]] <- unlist(apply(x, 1, function(z) which(z == ranks[i, 1])))
14    ranks[i, 2] <- mean(c(ranki[[i]], rep(ncol(x)+1, nrow(x) - length(ranki
    [[i]]))))  # give rank ncol(x) + 1 when not appearing in the list
  }
16  names(ranki) <- 1:nrow(ranks)
  list(ranks[order(ranks[,2]),], ranki[order(ranks[,2])])
18 }

20 borda(x)
```

# Bibliography


Ahrendt, P., & Meng, A. (2005). Music genre classification using the multivariate AR feature integration model. Retrieved 2005, from http://www.music-ir.org/evaluation/mirex-results/articles/audio_genre/ahrendt.pdf

Allen, J. (1994). How do humans process and recognize speech? *Speech and Audio Processing, IEEE Transactions on*, *2* (4), 567–577.

Alluri, V., & Toiviainen, P. (2010a). *Cross-cultural similarities in polyphonic timbre perception*. Proceedings of the 11th International International Conference of Music Perception and Cognition.

Alluri, V., & Toiviainen, P. (2010b). Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Perception*, *27* (3), 223–241.

Aucouturier, J.-J. (2006). *Dix expériences sur la modélisation du timbre polyphonique [ten experiments on the modelling of polyphonic timbre]*. (Doctoral dissertation, University of Paris 6, France.).

Aucouturier, J.-J. (2008). Splicing: a fair comparison between machine and human on a music similarity task. *Department of System Studies, University of Tokio. Unpublished.*

Aucouturier, J.-J., & Pachet, F. (2003). Representing musical genre: a state of the art. *Journal of New Music Research*, *32* (1), 83–93.

Aucouturier, J.-J., & Pachet, F. (2004). Improving timbre similarity: how high's the sky? *Journal of Negative Results in Speech and Audio Sciences*, *1*.

Backus, J. (1977). The acoustical foundations of music. (Chap. 7). New York: W. W. Norton & company.

Barber, D. (2010). *Bayesian reasoning and machine learning*. Cambridge University Press (in press). Retrieved from http://web4.cs.ucl.ac.uk/staff/D.Barber/pmwiki/pmwiki.php?n=Main.Textbook . Accessed 14.05.2010.

Bartsch, M. A., & Wakefield, G. H. (2005). Audio thumbnailing of popular music using chroma-based representations. In *IEEE Transactions on multimedia* (Vol. 7, *1*, pp. 96–104).

Bergstra, J., Casagrande, N., & Eck, D. (2005). *Two algorithms for timbre- and rhythm-based multi-resolution audio classification*. Retrieved from http://www.music-ir.org/evaluation/mirex-results/articles/audio_genre/bergstra.pdf.

Bergstra, J., Casagrande, N., Erhan, D., Eck, D., & Kégl, B. (2006). Aggregate features and adaboost for music classification. *Machine Learning*, *65* (2-3.), 473–484.

Boser, B., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 144–152). ACM.

Bouckaert, R. R. (2003). Choosing between two learning algorithms based on calibrated tests. In T. Fawcett & N. Mishra (Eds.), *Proceedings of 20th international conference on machine learning* (pp. 51–58). Menlo Park, California: AAAI Press.

Burred, J. J. (2005). *A hierarchical music genre classifier based on user-defined taxonomies*. Retrieved from http://www.music-ir.org/evaluation/mirex-results/articles/audio_genre/burred.pdf.

Burred, J. J., & Peeters, G. (2009). *An adaptive system for music classification and tagging (MIREX 2009 submission)*. Retrieved from http://www.music-ir.org/mirex/2009/results/abs/BP_train_tag.pdf.

Cao, C., & Li, M. (2009). *Thinkit submissions for MIREX2009 audio music classification and similarity tasks*. Retrieved from: http://www.music-ir.org/mirex/2009/results/abs/CL.pdf.

Cooley, J., & Tukey, J. (1965). An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, *19* (297–301).

Couvreur, L., Bettens, F., Drugman, T., Dubuisson, T., Dupong, S., Frisson, C., . . . Mancas, M. (2008). *Audio thumbnailing* (QPSR Vol. I No. 2).

Dannenberg, R. B., Thom, B., & Watson, D. (1997). A machine learning approach to musical style recognition. In *Proc. international computer music conference* (pp. 344–347).

Deemagarn, A., & Kawtrakul, A. (2004). Thai connected digit speech recognition using hidden markov models. In *Proceedings of the international conference on speech and computer* (pp. 731–735). Citeseer. St. Petersburg.

Downie, J. S. (2005). *2005 MIREX contest results - audio genre classification (contest wiki)*. Retrieved from http://www.music-ir.org/evaluation/mirex-results/audio-genre/index.html.

Downie, J. S. (2008). The music information retrieval evaluation exchange (2005-2007): a window into music information retrieval research. *Acoustical Science and Technology*, *29* (4), 247–255. Availableat:http://dx.doi.org/10.1250/ast.29.247.

Downie, J. S., Ehmann, A. F., & Tcheng, D. (2005). Real-time genre classification for music digital libraries. In *Proceedings of joint conference on digital libraries*. Colorado.

Downie, J. S., & Futrelle, J. (2005). Terascale music mining. In *Proceedings of the 2005 acm/ieee sc 05 conference (sc'05)* (p. 71). IEEE.

Downie, J. S., & West, K. (2009). *Mirex2009 results*. Retrieved from http://www.music-ir.org/mirex/2009/index.php/MIREX2009_Results.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. Citeseer.

Fingerhut, M. (2004). *Music information retrieval, or how to search for (and maybe find) music and do away with incipts*. IAML - IASA 2004 Congress, Oslo. Retrieved from http://mediatheque.ircam.fr/articles/textes/Fingerhut04b.

Fletcher, H., & Munson, W. (1933). Loudness, its definition, measurement and calculation. *Journal of the Acoustical Society of America*.

Foote, J. T. (1997). Content-based retrieval of music and audio. In C.-C. J. K. et al. (Ed.), *Multimedia storage and archiving systems II, proceedings of SPIE* (Vol. 3229, pp. 138–147). Content-based Retrieval of Music and Audio.

Foote, J. (1999). Visualizing music and audio using self-similarity. In *Proceedings of 7th ACM international conference on multimedia (part 1)* (pp. 77–80).

Fuhrmann, F., & Herrera, P. (2010). Polyphonic instrument recognition for exploring semantic similarities in music. In *Proceedings of the 13th international conference on digital audio effects (DAFx-10)*. Graz, Austria: Citeseer.

Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, *34* (1), 52–59.

Gjerdingen, R., & Perrott, D. (2008). Scanning the dial: the rapid recognition of music genres. *Journal of New Music Research. Vol. 37* (2).

Goto, M. (2006). A chorus section detection method for musical audio signals and its application to a music listening station. *Transactions on Audio, Speech and Language Processing*, *14* (5), 1783–1794.

Grecu, A., Lidy, T., & Rauber, A. (2009). *MIREX 2009 enhancing audio classification with template features and postprocessing existing audio descriptors*. Retrieved from http://www.music-ir.org/mirex/2009/results/abs/GLR.pdf.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America. 2* (61), 1270–1277.

Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America*, *63*, 1493–1500.

Guaus, E., & Herrera, P. (2005). *A basic system for music genre classification*. Retrieved from http://www.music-ir.org/mirex/abstracts/2007/GC_guaus.pdf.

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In W. W. C. . H. Hirsh (Ed.), *Machine learning: proceedings of the eleventh international conference* (pp. 121–129). San Francisco, California: Morgan Kaufmann.

Kohavi, R., & Provost, F. (1998). Glossary of terms. *Machine Learning*, *30*, 271–274.

Kohavi, R., & Sommerfield, D. (1995). Feature subset selection using the wrapper method: overfitting and dynamic search space topology. In *Kdd-95*.

Kotropoulos, C., Arce, G., & Panagakis, Y. (2010). Ensemble discriminant sparse projections applied to music genre classification. In *2010 international conference on pattern recognition*.

Kuncheva, L., & Rodríguez, J. (2010). Classifier ensembles for fMRI data analysis: an experiment. *Magnetic resonance imaging*, *28* (4), 583–593.

Lartillot, O. (2010). *MIR Toolbox 1.3 user's manual*. Retrieved from https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox. Accessed 20.4.2010.

Lartillot, O., & Toiviainen, P. (2007). A Matlab toolbox for musical feature extraction from audio. In *International conference on digital audio effects*. Bordeaux.

Lidy, T., & Rauber, A. (2005). *Combined fluctuation features for music genre classification*. Retrieved from http://www.music-ir.org/evaluation/mirex-results/articles/audio_genre/lidy.pdf.

Lidy, T., Rauber, A., Pertusa, A., & Iñesta, J. (2007). *Combining audio and symbolic descriptors for music classification from audio*. Austrian Computer Society OCG. Retrieved from http://music-ir.org/mirex/2007/abs/AI_CC_GC_MC_AS_lidy.pdf.

Liu, H., & Motoda, H. (1998). *Feature selection for knowledge discovery and data mining*. Kluwer Academic Publishers, Boston, USA.

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *International symposium on music information retrieval* (Vol. 28, p. 5). Citeseer.

Maddage, N. C., Xu, C., Kankanhalli, M. S., & Shao, X. (2004). Content-based music structure analysis with applications to music semantics understanding. In *Proceedings of the 12th annual ACM international conference on Multimedia* (pp. 112–119). MULTIMEDIA '04. New York, NY, USA: ACM.

Mandel, M., & Ellis, D. (2005). *Song-level features and SVMs for music classification*. Retrieved from http://www.music-ir.org/evaluation/mirex-results/articles/audio_genre/mandel.pdf.

Mandel, M., & Ellis, D. (2007). *Labrosa's audio music similarity and classification submissions*. Retrieved from http://www.music-ir.org/mirex/2007/abs/AI_CC_GC_MC_AS_mandel.pdf.

McKay, C., & Fujinaga, I. (2006). Musical genre classification: is it worth pursuing and how can it be improved? In *Proceedings of the 7th international conference on music information retrieval* (pp. 101–106).

McKay, C. (2010). *Automatic music classification with jMIR*. (Doctoral dissertation, McGill University).

McKay, C., & Fujinaga, I. (2008). Combining features extracted from audio, symbolic and cultural sources. *Proc. of the Int. Society for Music Information Retrieval Conference*, 597–602. Combining features extracted from audio, symbolic and cultural sources. Proceedings of ISMIR 2008, p. 597-602.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence*, *116*.

Moore. (1985). *Digital audio signal processing: an anthology*. (J. Strawn, Ed.). William Kaufmann, INC.

Ng, A., & Jordan, M. (2001). On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes. *Advances in neural information processing systems (NIPS)* (841-848).

Pampalk, E., Flexer, A., Widmer, G. (2005). Improvements of audio-based music similarity and genre classification. In *Proc. of the Int. Society for Music Information Retrieval Conference* (Vol. 5). Citeseer.

Pampalk, E. (2005). *Speeding up music similarity*. Retrieved from http://www.music-ir.org/evaluation/mirex-results/articles/audio_genre/pampalk.pdf.

Panagakis, Y., & Kotropoulos, C. (2010). Music genre classification via topology preserving non-negative tensor factorization and sparse representations. In *Acoustics speech and signal processing (ICASSP), 2010 IEEE international conference on* (pp. 249–252). IEEE.

Pedersen, C., & Diederich, J. (2007). Accent classification using support vector machines.

Peeters, G. (2008). *A generic training and classification system for MIREX08 classification tasks: audio music mood, audio genre, audio artist and audio tag*. Retrieved from http://www.music-ir.org/mirex/abstracts/2008/Peeters_2008_ISMIR_MIREX.pdf.

Piaget, J. (1986). *The construction of reality in the child*. Ballantine.

Pihur, V., Datta, S., & Datta, S. (2007). Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach — bioinformatics. *Bioinformatics*, *23* (13), 1607–1615.

Rattray, M. (2009). *Probabilistic classifiers*. Retrieved from http://intranet.cs.man.ac.uk/ai//COMP60431/lectures/probabilistic_classifiers.pdf.

Reed, J., & Lee, C.-H. (2006). A study on music genre classification based on universal acoustic models. *Proc. of the Int. Society for Music Information Retrieval Conference*, 89–94.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Encyclopedia of database systems. (Chap. Cross Validation). Springer.

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *The Journal of Machine Learning Research*, *3*, 1371–1382.

Rockmore, D. N. (2000). The FFT - an algorithm the whole family can use. *Computing in Science  Engineering*, *2* (1), 60 –64.

Saari, P. (2009). *Feature selection for classification of music according to expressed emotion*. (Master's thesis, University of Jyväskylä, Finland).

Sánchez-Maroño, N., Alonso-Betanzos, A., & Tombilla-Sanromán, M. (2007). Filter methods for feature selection – a comparative study. In H. Yin, P. Tino, E. Corchado, W. Byrne & X. Yao (Eds.), *Intelligent data engineering and automated learning - IDEAL 2007* (Vol. 4881, pp. 178–187). Lecture Notes in Computer Science. Berlin / Heidelberg: Springer. Retrieved from http://dx.doi.org/10.1007/978-3-540-77226-2_19

Scaringella, N., & Mlynek, D. (2005). A mixture of support vector machines for audio classification. Retrieved from http://www.music-ir.org/evaluation/mirex-results/articles/audio_genre/scaringella.pdf.

Scaringella, N., Zoia, G., & Mlynek, D. (2006). Automatic genre classification of music content (a survey). *IEEE Signal Processing Magazine*, *23* (2), 133–141.

Scheirer, E. (1998). Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustical Society of America*, *103* (1), 588–601.

Serra, X. (1989). *A system for sound analysis/transformation/syntesis based on a deterministic plus stochastic decomposition*. (Doctoral dissertation, Stanford University).

Sethares, W. (1998). *Tuning, timbre, spectrum, scale*. New York: Springer.

Seyerlehner, K., & Schedl, M. (2009). *Block-level audio features for music genre classification*. Retrieved from http://music-ir.org/mirex/2009/results/abs/SS.pdf.

Sherrod, P. (2003). Predictive modeling software. Retrieved from http://www.dtreg.com/DTREG.pdf.

Silla, C., Kaestner, C., & Koerich, A. (2007). *The latin music database: uma base de dados para a classificação automática de gêneros musicais [the latin music database: a database for automatic classification of musical genres]*. 11th Brazilian Symposium on Computer Music (SBCM). Retrieved from http://www.ppgia.pucpr.br/~silla/publications/2007-SBCM-SillaKaestnerKoerich.pdf.

Silla, C., Koerich, A., & Kaestner, C. (2008). Feature selection in automatic music genre classification. *Proceedings of the Tenth IEEE International Symposium on Multimedia*, 39–44.

Silla, C., Koerich, A., & Kaestner, C. (2010). Improving automatic music genre classification with hybrid content-based feature vectors. In *ACM SAC 2010 - information access and retrieval track* (pp. 1702–1707).

Smith, J. O. (2010). *Spectral audio signal processing*. Retrieved from https://ccrma.stanford.edu/~jos/sasp/. California Technical Publishing.

Smith, J. B. L. (2010). *An evaluation and comparison of approaches to the automatic formal analysis of musical audio*. (Master's thesis, McGill University, Montréal, QC, Canada).

Smith, S. (1997). *The scientist and engineer's guide to digital signal processing*. California Technical Publishing. Retrieved 1997, from http://www.dspguide.com/

Tempelaars, S. (1996). Signal processing, speech and music. In M. Leman & P. Berg (Eds.). (Chap. 2 and 4). Swets & Zeitlinger.

Theimer, W., Vatolkin, I., & Eronen, A. (2008). Definitions of audio features for music content description. *Algorithm Engineering Report TR08-2-001, Technische Universitt Dortmund*.

Toh, A., Togneri, R., & Nordholm, S. (2005). Spectral entropy as speech features for speech recognition. In *Proceedings of PEECS* (pp. 22–25).

Toiviainen, P., Tervaniemi, M., Louhivuori, J., Saher, M., Huotilainen, M., & Näätänen, R. (1998). Timbre similarity: convergence of neural, behavioral, and computational approaches. *Music Perception*, *16* (2), 223–241.

Tzanetakis, G. (2002). *ISMIR 2002 tutorial: music information retrieval for audio signals*. Retrieved from http://webhome.cs.uvic.ca/~gtzan/work/talks/ismir/gtzanISMIRtut.pdf.

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10* (5), 293–302.

Tzanetakis, G., & Murdoch, J. (2005). *Fast genre classification and artist identification*. Retrieved from http://www.music-ir.org/evaluation/mirex-results/articles/audio_genre/tzanetakis.pdf.

von Luxburg, U., & Schölkopf, B. (2011). Statistical learning theory: models, concepts, and results. In S. H. D. Gabbay & J. Woods (Eds.), *Handbook for the history of logic, vol. 10: inductive logic*. Elsevier.

West, K. (2005a). *MIREX audio genre classification*. Retrieved from http://www.music-ir.org/evaluation/mirex-results/articles/audio_genre/west.pdf.

West, K. (2005b). *MIREX FR*. Retrieved 7.11.10 from http://www.music-ir.org/mirex/wiki/2005:MIREX_FR.

West, K. (2008). *Novel techniques for audio music classification and search*. (Doctoral dissertation, University of East Anglia).

Witten, I., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques*. Elsevier.

Yeh, C. (2008). *Multiple fundamental frequency estimation of polyphonic recordings*. (Doctoral dissertation, University Paris VI).