Technische Universität Berlin

Fachgebiet Audiokommunikation

Masterarbeit

Evaluation of Accent-Based Rhythmic Descriptors for Genre Classification of Musical Signals

Athanasios Lykartsis
[redacted]

Technische Universität Berlin

Fachgebiet Audiokommunikation

Masterarbeit

Evaluation of Accent-Based Rhythmic Descriptors for Genre Classification of Musical Signals

Vorgelegt von:	Athanasios Lykartsis
[redacted]	[redacted]
Erstgutachter:	Prof. Dr. Stefan Weinzierl
Zweitgutachter:	Dr. Alexander Lerch
Datum:	April 16, 2014

Hiermit erkläre ich an Eides statt gegenüber der Fakultät I der Technischen Universität Berlin, dass die vorliegende, dieser Erklärung angefügte Arbeit selbstständig und nur unter Zuhilfenahme der im Literaturverzeichnis genannten Quellen und Hilfsmittel angefertigt wurde. Alle Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind kenntlich gemacht. Ich reiche die Arbeit erstmals als Prüfungsleistung ein.

Berlin, den 16. April 2014 Athanasios Lykartsis

Acknowledgements

I would like to thank the following people who all helped me - each in their own way- towards finishing this Thesis:

Prof. Dr. Stefan Weinzierl for his constant help, motivation and trust in my abilities from the beginning of the masters programme until the present day.
Dr. Alexander Lerch for his extremely valuable time, expertise, help and advice, without which this work could not have been completed.
Andreas Pysiewicz for his support, his suggestions and our fruitful discussions, as well as for the last-minute proofreading.
Henrik von Coler for his helpful tips on feature selection and evaluation.
Marc Voigt for his IT-expertise and for making parallel processing with MATLAB available at the right time.
Mina Fallahi for giving me valuable time on the department supercomputer during her simulations.
Fabien Gouyon, Andreas Homburg, Klaus Seyerlehner, Giorgos Tzanetakis and the people behind ISMIR for providing the datasets (or making them freely available on the internet) and giving advice (where applicable).
My parents who supported me with wise words but also with the occasional wake-up call whenever necessary.
Last, but not least, **Marie** for tolerating and supporting me throughout the whole process, even if she did not always understand everything that was involved. This one's for you!

Abstract

In audio content analysis, there exists a scarcity of methods which can efficiently identify and retrieve musically similar audio content based solely on its rhythmic or temporal structure elements. This is mainly because rhythm and structure in sound are easily recognized by listeners but difficult to extract and represent efficiently in an automatic fashion. As rhythm is one of the physical and perceptual properties which plays a significant role in the characterization of music similarity, it is important to evaluate the relevance of adequate rhythmic content descriptors for the musical genre classification task, which is one of the most demanding in the music information retrieval literature. In the context of this thesis, a musical genre classification system based on accent-related rhythmic content descriptors is described, implemented and evaluated. Based on an musical accent model, novelty functions of audio features based on different relevance criteria are extracted. These are then used to create a rhythmic content representation of the acoustic signal, the beat histogram, which serves as a basis for the extraction of features for genre classification. Different implementations of features and their combinations are evaluated and tested. In order to assess the performance of the rhythm-based classification, other well-known descriptors are also extracted from audio and their performance for the classification task evaluated as a baseline. The evaluation takes place for five music genre datasets, in order to allow the comparability of the classification with other results published with respect to those datasets and to assess the suitability of the predictors for different kinds of musical genre hierarchies. For the classification part, two supervised methods were used: the kNN algorithm and the Support Vector Machines. An experimental setup is implemented and the performance of the algorithms are evaluated through their accuracy. Finally, feature selection methods are applied in order to identify the most relevant features. Results of the experiments show promising classification accuracy for the most datasets using the accent-based rhythmic descriptors. With respect to other audio descriptors, the rhythmic content ones show comparable results. Furthermore, the SVM algorithm shows better results for all datasets with respect to the kNN. Finally, feature selection methods allowed the identification of the best descriptors, which in their turn show comparable results to the full feature set. In all cases, the result are similar to those in other previously presented systems, which warrants the use and further evaluation of the proposed method in the future. Due to the generic character of their calculation, their perceptual relevance and their adequate description of the rhythmic content of an audio signal, the best descriptors are hoped to be of value in other related tasks, such as automatic language identification based on rhythmic cues.

Zusammenfassung

In Audioinhaltsanalyse, ein Mangel an Methoden, welche musikalisch ähnliches Audioinhalt auf Basis seiner rhythmischen oder zeitlich strukturellen Elementen in einer effizienter Art und Weise identifizieren und abrufen können, ist festzustellen. Das liegt hauptsächlich daran, dass Rhythmus und Struktur in Sound von Hörern leicht erkannt werden können, aber ihre Extraktion und effiziente Repräsentation in einer automatischer Weise ist eine schwierige Aufgabe. Da Rhythmus eine von den wichtigsten physikalischen und perzeptuellen Eigenschaften sind, die eine Rolle in der Charakterisierung von musikalischer Ähnlichkeit spielen, es ist wichtig, relevante Deskriptoren des rhythmischen Audioinhaltes für Nutzung in der anspruchsvollen Aufgabe der musikalischer Genreklassifizierung zu gestalten. Im Rahmen dieser Arbeit, ein Musikgenreklassifizierungssystem, das auf akzent-relevante rhythmische Deskriptoren basiert, ist implementiert und evaluiert. Mithilfe eines Models musikalischen Akzentes, Novitätsfunktionen von Audiofeatures auf Basis von verschiedenen Relevanzkriterien sind extrahiert. Sie sind dann verwendet um eine Repräsentation des rhythmischen Inhaltes eines akustischen Signals, das Beat-Histogramm, zu generieren. Letzteres dient als Basis um Features für die Genreklassifizierung. Verschiedene Implementierungen von Features und ihre Kombinationen sind getestet und evaluiert. Um die Leistung der rhythmusbasierten Klassifizierung zu beurteilen, andere bekannte Deskriptoren sind auch extrahiert und ihre Leistung wird als eine Baseline benutzt. Die Evaluation findet für fünf verschiedene Datensätze statt. Somit ist die Vergleichbarkeit der Ergebnisse der Klassifizierung mit diesen anderer Publikationen gewährt. Ausserdem, die Deskriptoren können dann für unterschiedlichen musikalischer Genrehierarchien evaluiert. In dem Klassifizierungsteil, zwei überwachte Klassifizierungsmethoden sind eingesetzt: Die kNN und SVM Algorithmen. Ein experimenteller Aufbau ist implementiert und die Algorithmen sind auf Basis ihrer Genauigkeit evaluiert. Schliesslich, Methoden zu Feature Selektion sind angewendet, um die relavanteste Deskriptoren zu identifizieren. Die Ergebnisse zeigen vielversprechenden Genauigkeit für die meisten Datensätze mit Nutzung der akzentbasierten rhythmischen Deskriptoren. Bezüglich der anderen Audiodeskriptoren, die Rhythmischen zeigen eine vergleichbare Leistung. Des weiteren, der SVM-Algorithmus zeigt bessere Ergebnisse für alle Datensätze im Vergleich zum kNN. Die Methoden zur Auswahl der Features erlauben die Identifizierung der besten Deskriptoren, die vergleichbare Resultaten zu denen des ganzen Deskriptorsatzes zeigen. In allen Fällen, die Ergebnisse sind ähnlich zu denjenigen von anderen präsentierten Systemen, was auf die weitere Evaluation und Nutzung der vorgeschlagenen Methoden hinweist. Wegen des generischen Charakters ihrer Berechnung, ihrer perzeptuellen Relevanz und ihrer Leistung für die Beschreibung des rhythmischen Inhalts eines akustischen Signals, es ist beabsichtigt, die beste Deskriptoren in verwandten Aufgaben zu verwenden, wie z.B. die automatische Sprachidentifizierung basierend auf rhythmischen Cues.

Contents

Part I.

Introduction

1. Problem Description and Previous Research

1.1. Problem Description

Music, widely defined as *organized sound* [93], has been a solid part of human culture since its beginning and bears great importance to humans as an acoustic medium, alongside with speech. In contrast to the latter, its purpose is not primarily to be used as a tool for efficient communication of facts and ideas, as its depth and openness to interpretation is quite phenomenal. Music is, among other things, a medium which serves the communication of feelings and emotions. It also serves as the motivation and companion for human movement or dance, and is widely regarded as a means of pleasure and enjoyment. It is for all these reasons that it continues to be a mainstay of human behavior and occupation, but also serves as an inexhaustible subject for discussion, research and analysis, both from a theoretical and from a technical perspective.

The richness encountered in music is a consequence of its importance: Music comes in countless forms and varieties, which traverse the boundaries of culture and historic period. Musical excerpts which share common elements are grouped under categorical labels known as *genres*. Those labels, albeit subjective in nature, help listeners to define in what way one musical excerpt differs from another, or to find similar excerpts to ones heard before based on specific acoustic, perceptual or cultural aspects. One very important dimension of music concerns its temporal structure - what is often summarized under the concept of *rhythm*. Together with harmony and melody, rhythm is one of the fundamental aspects of music - and, in fact, of any acoustic signal [69]. However, due to the semantic gap between the perceived rhythmicity and the manifest temporal structure of the audio signal, the definition, description and extraction of rhythm presents a challenging research subject, which is far from conclusion.

Researchers and scholars of music theory have analyzed music since ancient times, resulting in the emergence of numerous models of musical structure and content. Especially in the case of western, tonal music tradition, a stable knowledge framework has been produced and refined, remaining applicable for most of contemporary music. Likewise there has been much research in the areas of music cognition and psychology, mostly in the twentieth century, attempting to illuminate the ways listeners perceive and process musical signals, as well as which behavioral effects are related to the listening of music. One of the most interesting aspects combining these two views lies in the capability of listeners to easily and quickly extract abstract information from musical content (e.g., a song's rhythm or the genre to which it belongs [34]) with just a minimal amount of acoustic information available to them.

With the advent of the internet era, the automatic processing of audio signals became more relevant and, at cases, even necessary [31]. At the technical level, fully automatic

processing of music has not been possible until relatively recently, but advances in information technology in the last twenty years have allowed the emergence of various tools and applications. The interdisciplinary field which deals with this processing is *Music Information Retrieval* (henceforth MIR), and combines the research areas of computer science, engineering and signal processing with music theory and auditory perception and cognition [27, 66]. One of the most important subfields of MIR is *Audio Content Analysis* (henceforth ACA) [52], which focuses on the automatic analysis of digital audio signals and the extraction of useful information from them. This last area is also the focus of the thesis at hand.

One of the most important applications in ACA, *automatic musical genre classification* [74], addresses issues which have emerged due to the huge amount of digital audio material available to everyday users since the 1990s. With individuals and institutions having access to the equivalent of thousands of hours of sound material and few or incomplete metadata to accompany it, interesting questions arise: how can one organize, browse and analyze efficiently such a massive amount of information? Furthermore, how can this be performed in a fast and computationally efficient way, while at the same time retaining perceptual relevance of the information extracted? The general field addressing such subjects related to sound in general is called *audio signal classification*. Musical genre classification aims at solving the problem of automatically classifying a given musical excerpt to one or more genres, based on information extracted directly from the acoustic signal - its **content**. Given the complexity of music and the fuzziness of the definition of musical genre [34, 3]), the task of performing efficient and accurate musical genre classification emerges as non-trivial. Its relevance is however warranted, as it represents a broadly defined, very ambitious task with numerous applications [74].

ACA systems for automatic genre classification consist of a feature[1] extraction and a classification module [52]. While the choice of the classifier is relatively arbitrary and based mainly on performance issues, an important subject concerns the extraction of suitable audio descriptors for the considered application. With an almost endless amount of features and their combinations to extract ([52, 70, 67]), the design and choice of relevant descriptors is a difficult task. In the context of more specific applications, the features to be extracted are determined mostly based on the desired outcome, e.g., in beat-tracking, features must be found which allow an efficient and valid extraction of the dominant periodicity in the signal. However, in musical genre classification practically **all** categories of features may be relevant to the task [52, 74], which renders the search for appropriate features quite arduous. As such, it becomes evident that the design of more elaborate and, at the same time, perceptually meaningful features, or the reduction of the problem to a specific aspect of musical content is potentially a good strategy.

The design of descriptors for automatic musical genre classification is a much researched topic in audio content analysis the last years. Unfortunately, it has received far less attention then the subject of classification, since, in contrast to the latter, it is *domain specific*: knowledge about the domain of application has to be incorporated when attempting to produce novel, adequate descriptors. When dealing with sound, this prior knowledge

[1] The term *feature* will be used interchangeably with the term *descriptor* throughout the text. They both refer to low-level, measurable quantities which can be extracted directly from the audio signal or a transformation thereof.

concerns either perceptual matters, which help create features which try to imitate the way listeners perceive audio stimuli, such as perceptual models of loudness; or theoretical considerations, such as models of musical structure (e.g. pitch theory or harmony), which have been used up to date for the creation of relevant features. One of the subjects which has received somewhat less attention is rhythm-based genre classification, since this aspect of music is very difficult to quantify in a satisfactory manner which allows the extraction of numerical features. However, there is a number of publications which have dealt with the subject of automatic rhythm description. Furthermore, the related subjects of beat tracking and music similarity have provided a basis for the design of relevant rhythmic descriptors, albeit with a focus on singular aspects such as tempo. A more detailed discussion about such approaches will be given in section 1.2. It suffices here to point out an important shortcoming of previously applied methods: The descriptors used up to now give only moderate classification results with comparison to other features, since their scope is limited, i.e. they do not take into account the different levels of rhythm inherent in the audio signal. Since the design of new features based on mathematical considerations is relatively easy in comparison to a more conceptual approach, the current situation of rhythm-based genre classification shows an abundance of subfeatures for the classification task, but only a few methods for extracting perceptually relevant periodicities from the signal in a meaningful way. Especially, the number of studies attempting to connect musical theory with the feature extraction process are relatively few; to our knowledge, none of them has been applied in musical genre classification up to date.

In this context, this thesis is concerned with the problem of automatic genre classification of musical signals with the use of adequate rhythmic content descriptors, derived in part from a music theoretical approach concerning rhythm and its perceptually important constituents, *accents*. The parts of designing new features and their extraction, classification and the evaluation of the results, as well as individual areas which are involved in the task are described in detail. Those questions are linked with the matters of musical genre, rhythm and the features which can be extracted that describe the latter in a useful way, so that automatic musical genre classification can be conducted efficiently. Furthermore, the finding of suitable descriptors for the automatic classification task can help provide valuable insights regarding the way genre classification is performed from human listeners and help the improvement of music retrieval applications.

1.2. Previous Research

As Scheirer [75] and Tzanetakis [91] point out, the precursor of musical genre classification is found in the area of automatic speech recognition (ASR), where feature representations of the speech signal are used to distinguish phonemes in an audio stream or even at a higher level, for example in speaker recognition. Expanding this idea, the audio signal to be classified does not comprise only speech, but also music or other types of audio, and the categories to which it can be classified into can also be more diverse. Those considerations, along with the increasing demand for automatic indexing and browsing systems for the internet and music industry has spurred much research and led to the development of various musical genre classification systems the latest years, which will be discussed in more detail in the following.

In common musical genre classification approaches up to date, the acoustic material to be categorized is in the form of digital audio data (audio samples). Since the samples cannot be used directly for the classification (as their dimensionality is extremely high, the information in them very confounded and the gap to the abstract concepts used by listeners too big [74], there is a need to create reduced but relevant (in the sense of useful) representations of the audio data. There are several matters which come into consideration while attempting to design and construct a musical genre classification system [74, 3]:

- **Properties to be represented for genre classification** Which musical and/or perceptual properties represent musical genre and can (or must) be taken into consideration?
- **Relation of perceptual properties to features and feature design** How do these properties relate to the actual *features* (numerical values and quantities) to be extracted from the signal?
- **Classification methods** Which classifier should be used in a specific implementations and what are the advantages and disadvantages in each case?
- **Evaluation of genre classification** How can the performance of such a system be evaluated in a meaningful way and what do the results signify about the dataset and the features used?

All of those subjects are relevant for the thesis and will be discussed in some depth in the following chapters. It must be noted in advance that a broadly defined category such as genre can not be fully described through the variability explained through acoustic features alone [74, 3]. However, since the focus of most approaches lies on automatic processing, relevant studies have attempted to extract as much information as possible from the signal, in order to ensure a connection to all perceptual and musical aspects of the signal: timbral, tonal, dynamic, temporal (rhythmic), instrumentation-related, production-related and others [74, 52]. Such approaches have given encouraging results and could even be suitable for commercial applications, as they provide a very comprehensive representation of the signal at hand, with a number of publications which have explored the problem and became very influential in this aspect. We will give here a brief account of the most important musical genre classification studies in the last years. It must be noted that only publications conforming to the standard scheme of audio content analysis, i.e. feature extraction followed by classification, will be mentioned here, leaving aside others which depart from this model using either symbolic approaches or other schemes.

Scheirer and Slaney In one of the earlier works in automatic audio classification, Scheirer and Slaney [75] propose a system for the discrimination of speech and music signals. They extract features from the audio excerpts which pertain to different aspects of their temporal and timbral content. They proceed in using a Gaussian Mixture Model (GMM) and a k-Nearest-Neighbor (kNN) classifier for the multidimensional classification and achieve good discrimination results for speech and music in a broad dataset, which is however unfortunately not documented in detail.

Foote Foote [32] proposes a method for audio classification and retrieval which has parallels to the task of content-based image retrieval. He focuses on detecting similarity between different musical signals by applying an mel-frequency cepstral coefficient parameterization of the signal. He then uses a supervised vector quantization method to extract statistics about the musical signal, serving as "templates" for the classification, which is based on a distance metric between different templates.

Tzanetakis et al. In their seminal work [91], Tzanetakis et al. propose an automatic classification system for audio signals which operates on a simple hierarchy of ten musical genres with two sub-genres, although they also consider non-musical signals such as speech. They use three categories of frame-based features, referring to the timbral texture, pitch content, and rhythmic content of the audio excerpts. For classification, they employ also a Gaussian, a GMM and a kNN classifier. Their results show an overall classification accuracy of 61% and is one of the most pioneering in the area of musical genre classification. Conducting a listening experiment, they show that this classification rate is actually close to the one achieved by human subjects. In a related publication, Li and Tzanetakis [56] present a classification scheme which is based on the same feature set and dataset as in [91]. However, they use Linear Discriminant Analysis (LDA) and Support Vector Machines (SVMs) for the classification. This study can be seen as a continuation of the work in [91] and presents a deeper evaluation of the features used therein. The results are comparable to the ones in the previous study, but show the need for using more feature combinations and classifiers in musical genre classification studies.

Burred and Lerch In a work presented shorty after [91], Burred and Lerch [14] apply an hierarchical approach to the task of automatic musical genre classification. They also extract three categories of frame-based features (timbral, rhythmic and other, more technically oriented quantities) as well as MPEG7 descriptors to represent the content of an audio excerpt and use a Gaussian Mixture Model for the classification phase. However, they focus on performing the classification in an hierarchical scheme, since that provides a more accurate classification, and evaluate the features used in a systematic way. Their results are promising and will be taken into account in the present study.

Gouyon et al. In 2004, Gouyon et al. [39] proposed an automatic musical genre classification scheme which is based on rhythmic descriptors only, with help of a Nearest-Neighbor classifier. They focus on this aspect of the musical content because of the relevance of rhythm for musical genre classification and in order to create features for the classification which bear a close relationship to the cognitive patterns which are used from humans in order to perform the genre classification task. The features they used will be described more closely in part II, as they are of relevance for this work as well. One of the important elements of this study is that they also evaluate the descriptors in a systematic way, allowing to pinpoint those which provide a good classification performance.

Lidy and Rauber Lidy and Rauber [57] also focus on rhythmic content descriptors, but additionally examine the importance of psychoacoustic transformations for the calculation of the audio features. One of the novelties of the study is the use of multiple datasets and

multiple feature combinations for the classification, resulting in an increased count of experiments. They use SVMs for classification and calculate various performance measures, so as not to be binded only by the accuracy of the algorithms. Their results are promising and highlight the importance of both rhythmic content features and SVMs for automatic genre classification tasks.

Bergstra In his master's thesis, Bergstra [8] presents an automatic genre classification system which is based on a variation of a very often used features, the MFCCs. He achieves good classification accuracy on a small dataset, while at the same time examining the effect of different parameters on the genre classification and various machine learning methods. In two related publications [9, 10], he examines the subjects of the feature aggregation and the dataset used more closely.

West West [96] introduces a new classification scheme, concentrating on the problem of increasing accuracy while using well-known predictors which have already been tested extensively. He also focuses on the parameters of feature extraction in order to quantify their effect on classification accuracy. The features are then evaluated on a small dataset, while the study shows good results for several classifiers.

Mandel and Ellis Mandel and Ellis [61] use whole-song level features and SVMs for artist and excerpt classification. Their dataset is a subset of the *uspop2002* and the features used mainly MFCCs. Their contribution lies mainly in the use of support vector machines for classification, along with specific distance metrics and methods for parameterization.

Soltau In his diploma thesis [85] and a related publication [86], Soltau analyses a musical genre classification system in depth. He uses neural networks and HMMs as classifiers, and focuses also on the temporal structure of the music. To that end, he derives a transformation of the audio excerpt in abstract acoustic events, from which he extracts statistical features, and uses them for the recognition of the genres in a small dataset of modern music. His results are promising, although his model does not conform completely to the feature extraction and classification scheme used, for example, in [91, 14, 39].

Scaringella and Zoia Scaringella and Zoia [73] present a system which uses timbral and rhythmic features for a medium-sized dataset. The excerpts are then classified through the use of SVMs, Neural Networks (NN) and Hidden Markov Models (HMMs) with specific implementations. They report good results on their version of the classifiers, which warrants their further use.

Dixon et al. Dixon et al. in [25] work with the same dataset as in [39] and also extract rhythm related features, pertaining to the tempo and other periodicities in the signal. Their extracted representation is called a rhythmic pattern, which they then use to derive features and classify using a kNN classifier. Their results using the rhythmic patterns alone are not extremely good, but in combination with other statistical features they achieve a good accuracy on their dataset.

This list is by no means complete, as it focuses on the approaches which are relevant to the work at hand. The multitude of the above approaches shows that musical genre-based classification has been a crucial research topic with a steadily rising number of interesting results. However, two possible issues exist when implementing such approaches ([74]): First, the lack of parsimoniousness when selecting a descriptor set, which leads to the curse of dimensionality[2]; second, the lack of information about which aspects and for what reason exactly are important in defining genre. One solution to overcome both problems is to take into account only one perceptual quality of the music and try to build descriptors which are representative of this quality. To this end, we will choose to focus later in the thesis on those publications which focus on one specific aspect of music, namely its rhythm. Previous work done in this area includes the mentioned work of Gouyon [39, 40], who has examined in depth the evaluation of rhythmic descriptors alone for genre classification. However, he has also based his research on other findings ([14, 91, 57, 55]), which have also used and evaluated rhythmic content features. An important part in trying to extract such features concerns the definition of rhythm itself and its representation or description through automatic systems based on low-level features extracted from the audio signal. In general, the features extracted and the system used depend heavily on the application at hand. A comprehensive review of rhythm description systems can be found in [40]. In chapter 4, more information will be given on possible rhythm description strategies with a focus on the ones relevant for this thesis.

Before continuing to the following chapters, two important remarks have to be made with respect to the approach followed in the thesis. In this work, the system at hand has the classical form of an audio content analysis system [52], in which features (quantities corresponding to properties of the acoustic signal) are extracted directly from the signal, and then used as input for machine learning classification algorithms which allow their automatic classification. Thus, the discussion will be limited to methods conforming to this paradigm. An important distinction to be made here concerns the context of classification: audio recognition and classification can be performed either with knowledge of the categories in which the audio samples should be classified (one speaks of *supervised classification* in this case); or with a category of algorithms and statistical methods which do not need any prior information about the classes to which the audio belongs prior to classification and attempt to "cluster" the audio samples with respect to the statistical properties of their feature representation (*unsupervised classification*). Because of the much more interesting nature of the first category of problems and the mathematical and computational robustness of methods associated with it, we will consider only such approaches in the context of this thesis. Of course, such approaches bear the drawback of the need of manual classification of the samples prior to classification. However, since all the datasets considered here are already manually labeled, this does not represent a problem in the present work. We will give some more information about supervised and unsupervised methods for automatic genre classification in chapter 5.

[2]The term *curse of dimensionality* refers to the problem occurring with the use of a large number of possibly irrelevant or redundant features in classification problems, which can lead to poor classifier performance. More information about the problem will be given in chapter 5.

2. Thesis Aim and Applications

2.1. Thesis Aim

In this work, a similar approach to the ones described in section 1.2 is adopted. The aim, however, consists on focusing only on the rhythm (or the temporal structure) of the music and the rhythmic content features associated with it to perform musical genre classification. Thus, a differentiated view in the contribution of rhythm to the recognition and classification of musical genres can be given. A musical work, as every acoustic signal, evolves in and throughout time, and the evolution of the constituent parts is what drives the attention and helps to "follow the music". In the context of this thesis, the concept of *rhythm* encompasses **the temporal structure of the signal-inherent qualities**. Beginning from the acoustic surface of the signal, human listeners can perceptually derive many other abstract temporal representations, such as the *meter*, the *beat* or a specific and repeated rhythmical pattern of a musical quality, which then allow the calculation of similarity between the signal at hand and others or their belonging to a common class. It is those patterns which are to be represented through appropriate features in this thesis.

An important cue to extracting the aforementioned rhythmic patterns present in the signal and generalizing on their basis are *accents*, or points of perceptual prominence in the acoustic signal. These can be defined on the basis of a music theory approach, with the purpose of obtaining salient features, much as human listeners do when they try to classify music in genres ([74, 34]). This part is of great importance, because an appropriate feature design is the key to finding relevant features that can allow for the successful function of a classification algorithm. Based on those accents, novelty detection methods are used to quantify the amount of change pertaining to events associated with specific accentuations in the signal, which provide the ground for the creation of periodicity representations capturing the relevant rhythmic structure of parts of the audio excerpt. The features calculated on these representations which can eventually separate not only rhythmically similar pieces, but also those belonging to the same genre. As mentioned above, the task of musical genre classification is one of the most demanding and challenging in ACA, and by far not exhausted as a research area. Considering, however, that temporal (rhythmic) cues are sufficient for human subjects to group together genre similar musical excerpts [34, 58, 72], the finding of suitable features seems justified: One can think about the standard and recurring idiomatic expressions present in well-defined genres, such as the off-beat riffs and kick drum in most reggae music songs, the syncopated baseline typical for salsa, the articulation of the beat triplet in a waltz excerpt or the verbose and fine-grained beat/impulse sequences in techno music. However, to capture such precise constructs in more complex (although relatively well-defined) genres such as jazz or experimental music could be much more demanding - perhaps it is exactly the absence of repeated structures and the presence of great diversity which can help define those genres rhythmically. In this context, the thesis thus attempts to clarify the following questions:

- Is it possible to conduct a successful genre classification of musical pieces based only on rhythmic descriptors and if yes, to what extent?

- What are the features which allow for high classification accuracy and how can they be derived from a priori knowledge such as through an approach delivered by musical theory?

Following these research questions, the approach of this thesis is essentially an experimental one. After a description of established rhythmic description systems for musical genre classification, novel features are proposed, which are based on categories of defined accents and a correspondence between those accents and the features which can describe them. Those accent-based descriptors aim at explaining as much rhythm-related variance in the signal as possible, taking into account different levels of accentuation - not only referring to the signal envelope (loudness-related accentuation) but also to spectral changes. This is achieved by extracting novelty functions which then serve as input to create a periodicity representation of the signal. The subfeatures calculated on the basis of this representation provide feature vectors, which serve as a compact representation of the rhythmic content of the signal. Those are then used to train a supervised classification algorithm, allowing it to learn how to classify new signals with the use of the rhythmic features to a specific genre. This procedure is repeated for five datasets, two supervised classification methods and different parameter settings with the goal of evaluating the classification performance. As a comparison baseline, other frame features (which do not describe only rhythmic content but other aspects of the music, such as timbre, instrumentation and tonality) are also extracted and their performance evaluated, both alone and in combination with the rhythmic content features. Since the features are highly correlated to each other and, as such, perhaps irrelevant or redundant for the classification, feature selection methods are applied in order to pinpoint only those features which allow for good classification accuracy and are therefore, adequate rhythmic content descriptors.

Although the number of publications concerning musical genre classification and automatic rhythm description is relatively large, not many works exist which discuss the automatic recognition and use of accents in the musical signal. One attempt comes from Müllensiefen et al. [64]: They define an exhaustive list of binary accent rules, which pertain to all possible accentuation effects in the music and conduct listening experiments as well as clustering, in order to test their salience and usefulness. Phenomenal accents (or accents actually manifested in the signal) were used from Seppänen in his thesis [81], in order to find perceptually prominent points in a beat sequence, which could be candidates for metrically salient beat positions in the signal flow. He then uses the extracted metrical grid to create a real-time beat tracking system which is then evaluated. Those publications have shown promising results regarding the definition, extraction and use of accents in order to perform beat tracking as well as their perceptual relevance. To our knowledge, however, accent-based rhythmic features have not been explicitly used for musical genre classification yet. The next chapter presents the aim of the thesis with respect to this observation.

2.2. Applications

Answers to the questions posed in 2.1 can be of importance in three main areas:

1. The clarification of the relationship between perceived and automatically extracted rhythm
2. The adequacy of rhythmic content descriptors extracted from digital audio for musical genre classification or other related tasks.
3. The creation of successful and efficient musical genre classification systems based on rhythmic elements of the music.

Furthermore, results can be helpful in the design and implementation of automatic systems for rhythmic similarity, genre recognition based on rhythm and music recommendation systems. As such applications (e.g., LastFM and Pandora) become more and more prevalent, their profiting from the results seems a desirable goal.

The thesis is structured as follows:

In the second part, a brief account of the theory underlying the fundamental aspects of the thesis is given. First, an introduction to music theory and cognition, focusing on the concepts of rhythm in general and accent in particular, is given. Second, information regarding the feature extraction process is provided, with a focus on the automatic description and extraction of rhythm. Finally, an introduction to machine learning and the classification methods used in this work is presented.

In the third part, the method and implementation of the novel features describing rhythm is presented. Specifically, the design of the features which correspond to accents in music is laid down, together with the subfeatures resulting from them and their relevance to the perceived rhythm. Furthermore, the specifics of the feature extraction and the details of the classification process are presented and explained.

The fourth part describes the experimental setup used to test and evaluate the rhythmic content and other descriptors, as well as the datasets used in the thesis. Subsequently, the results of the experiments are presented in table form.

In the fifth and final part, the results and the approach are discussed, in order to pinpoint advantages and disadvantages in comparison to other methods and to gauge the possibility of using those descriptors in other similar task. Finally, an outlook is given as to which tasks are further conceivable for the improvement and use of the approach presented here.

As detailed explanations and mathematical foundations of the subjects presented here can also be found in well-known and acclaimed textbooks and publications, we will focus only on the most relevant aspects for this work and otherwise refer to the literature for further reading. More specific information about the features and the datasets employed here, as well as more detailed results of the evaluation can also be found in the appendices. We assume that the reader has some background concerning the subjects of digital signal processing, statistics and basic music theory.

Part II.

Background Theory

3. Rhythm

In order to properly analyze the rhythmic content descriptors which are presented and evaluated in this work, an introduction to the subject of rhythm and its related concepts is needed. In this chapter, definitions and explanations are given concerning rhythm in general and the important notions of beat and musical meter. Finally, the concept of accent and its relation to rhythm is outlined.

3.1. Definition of Rhythm

Rhythm is one of the fundamental dimensions of analysis and perception of music. Although difficult to define, it is a very familiar concept to both musicians and listeners. The term refers to temporal structure and is therefore primarily not music-specific ([69], p.96); it is used to generally designate a temporal structuring of events which are in close relationship to each other (possibly having the same cause), bear significance for attention (i.e., they are in some way *accented*) and contribute to the creation of perceived sound patterns through the alternation and repetition of different layers of similar elements. In other words, every arrangement or structuring in time of similar sound events (such as the onsets of notes, musical chords or the beats of a drum) can denote a rhythm, one of its key properties being that it describes an explicit, recurring pattern of sounds, phenomenally present in the acoustic signal [53]. The pattern can refer either to the sound events themselves or to the durations of the intervals between them. However, not all possible patterns of sound events are perceived as different rhythms, making clear that the acoustic realization of rhythm and its perception are two separate phenomena.

There have been numerous attempts to give an acceptable definition of rhythm. One of the first ones comes from Platon and Aristoxenos, who denote rhythm as ”measure of movement” and ”order of times” (i.e. durations) which is accessible to the senses [80]. From that point on and until modern times there have been many other definitions, which however do not deviate much from the original one. As this work concerns itself primarily with modern, western and tonal music, we will consider some later definitions which attempt to capture a more general essence of rhythm. Cooper and Meyer [18] define rhythm as the way in which accented and non-accented notes are grouped in a time unit (the measure). Joel Lester [54] gives a definition which considers the patterns of duration between musical events. This definition has the advantage that it takes into account events pertaining to various musical qualities, giving rise to the idea that more than one rhythms can be defined for a musical piece. One of the most interesting definitions comes from Lerdahl and Jackendoff, which consider rhythmic structure to be result of the interaction of individual rhythmic dimensions ([53], p.12), which mainly concern the perceptual grouping of similar elements and the inferred regular patterns of strong and weak beats, which they refer to as the meter. Fraisse denotes rhythm as ”...the ordered characteristic of succession”

[33] and stresses its close relationship with tempo, whereas London [59] defines rhythm as "the sequential pattern of durations relatively independent of meter or phrase structure".

More psychologically or cognitively motivated definitions link rhythm to the perceived patterns generated by recurring events and how they interact and are categorized by listeners ([90, 46, 49, 60]). Here, rhythm refers to what we perceive, when we are listening to a piece unfolding in time and the abstract representation we are able to create from it. It is, thus, very closely associated with expectation, reaction and the optimization of information processing [44]. Furthermore, rhythm is perceived *categorically* ([16, 17, 79]), i.e. a specific temporal pattern of musical events is identified as such, even when there are some deviations from the basic form of the pattern or when it is implemented with different events. This points towards a strong tendency to a extract a general pattern from the audio signal, something which is linked to the perceptual phenomenon of meter. There are many approaches as to how listeners extract those rhythmic patterns: One of the most well-known is the dynamic attending theory from from Jones et al. [46] which propose that perceived rhythm is the result of different attending modes (future-oriented and analytic) which involve anticipatory behaviors to coherent temporal events and their durational patterns. An interesting experimental approach has been undertaken by Desain and Honing [21], which could verify the categorical perception of rhythm proposed by Clarke ([16, 17]) and Schulze [79]. The main point of those studies can be summarized as follows: the percept of rhythm results from the ability of listeners to extract an abstract representation from similar, recurring temporal patterns of musical events.

To recapitulate: When focusing on the acoustic signal itself, rhythm can be broadly defined as the specific, repeated patterning of accentuated and non-accentuated events and durations which is phenomenally present in the signal, and which results through the interaction between patterns comprising groups of similar elements. As such, rhythm is closely related to periodicities inherent in the signal, and to the quantities (e.g., loudness or pitch) which give rise to them. This working definition will be used in the context of the thesis, as it provides a stable ground for the extraction of meaningful representations of rhythm. It must be noted that very often a confusion occurs between the *concept* of rhythm itself - as the general form of the temporal structure of musical events and durations - or the *rhythm* of a musical piece, which refers to the recurring patterns of notes and durations in this specific excerpt (such as for example, waltz or samba). Here the term rhythm will be used with its first meaning unless mentioned otherwise, whereas the respective terms denoting a specific rhythmic pattern will be used to refer to the second meaning. At this point it is important to look more closely at some concepts which are related to rhythm and paramount to its understanding: *meter, beat and accent*.

3.2. Beat and Meter

3.2.1. Beat

The term *Beat* refers to a perceived or real sequence of regularly spaced sound events, but which have no specific accentuation pattern associated with them. In short, the beat is a method of time measurement or division of a time span in intervals of the same duration, where none of the division points bears special significance for attention (or is actually

different in comparison to the others in some way). Another usually applied term for beat is *pulse*, as the regular construct of a beat represents that of a pulse train - for example, the stroke of a clock or the ticks of a metronome. This kind of basic temporal grid with which one can establish a time measurement unit for a musical piece also plays a big role in the definition of *tempo*, as the amount of beats in one minute gives us the tempo value of a song. As there can be more than one levels of beats in a piece (e.g. one main beat and its denominations, even if they do not appear explicitly in the signal), the term tempo refers to the speed estimate which is extracted from the main beat, also known as *tactus* ([53]). Once a beat has been established (mentally) or recognized (in reality), human listeners to stress or accentuate some of the onsets perceptually. That leads us to the notion of *meter*.

3.2.2. Meter

Meter is a much discussed subject in music theory. The most important property of meter is that it is a perceptual content (contrasted to rhythm), which is inducted from the phenomenally accented points of the music surface [60]. Furthermore, meter is hierarchical, regular and stable: Combining those qualities allows meter to serve as a kind of enhanced temporal grid, which helps shape our expectations about "what comes next" and thus be able to anticipate and predict events in time ([60, 44]). In this sense it reveals its role in helping to create the categorical perception of rhythms, as those patterns which bear a certain relationship to a specific temporal grid. Although meter is visualized as a train of pulses with different salience, or durationless events in time which conform to a highly regular pattern of hierarchical accentuation, the onsets of actual musical events does not have to coincide with the points in time where those pulses are situated. However, in order to establish a meter, some regularity *has* to be manifested in the acoustic signal in the first place: the meter is inferred exactly through the repetitions of similar events with well-defined durational patterns. Once a meter has been established, all other events are perceived with reference to this regular pattern. In light of new events which do not concur with the inferred pattern, a new meter will be extracted and established [60]. This points towards the character of meter as a cognitive phenomenon which primarily serves the listener's need to be able to impose *some* structure on very complex or even inconsistent events, so as to maximize the extracted information while minimizing the available attentional resources [60, 44].

The direct extraction of beat and meter from acoustic signals is a very interesting research subject which has attracted much attention the last years [81, 76, 23, 35, 94]. However, because of its complexity and the somewhat specific character of the results provided from it, and which are not necessarily of relevance for musical genre classification, we will focus on more abstract, statistical properties of the rhythmic content of a musical piece which can be manifested through features extracted directly from the audio data. In this and the previous section, the concept of *accent* has been used extensively to denote points which are in one or more ways emphasized for attention. In the next section we will take a close look at this important concept, as it is also of added relevance to this work.

3.3. Accent

In the context of this work, it is of great interest to investigate and attempt to define the concept of *accent*. Accents play a central role in music, serving as demarcation objects for new sequences, but also as carriers of meaning, denoting, for example, a point or part of high importance in a musical piece. Furthermore they function as tools to enhance interpretation and convey strong emotions to the listeners. Very often, accents are set willfully by the interpreter, for instance by playing a certain note substantially louder than the rest, i.e. *emphasizing* it; but accents can also emerge and be perceived as intrinsic to the musical structure surface, even if the score is not signifying them explicitly. Furthermore, accents are also perceived in non-musical signals, or even artificially produced sounds, where no interpreter is present [60].

To account for this complex character, a definition of accent is needed which is either abstract enough to encompass all possible cases, or which differentiates between types of accents and how they originate. This subject has occupied music theory and sound scientists for quite a while, and virtually all scholars which have examined the temporal structure of sound and music have made a contribution or at least a reference to the concept of accent. Here we will focus on certain treatments which attempt to explain and describe musical structure in general; in their context it is most possible to receive a good insight about accents which is not fragmental in nature, but rather set in the premises of a more general theory. It must be made clear again at this point that these theories mainly concern themselves with the discussion of western tonal music; however, the treatment of accents is more general and can be interpreted as valid for music and sound in total.

Cooper and Meyer In *The rhythmic structure of music* [18], Cooper and Meyer define an accentuated sound as "one, which is similar to its neighboring sounds, but differs in some aspect enough from them, in order to stand out for perception" [18]. This kind of definition is a good example for a very general view of the concept, which does not make any specific statement about the nature of deviation; this can pertain to any aspect of the sound, e.g. its duration, loudness, pitch, timbre etc. However, it sets a very basic premise, that can guide the search for accents in an acoustic signal, and which can be rephrased as follows: Every point (and the sounds associated with it) of perceived change or differentiation in an acoustic/musical signal is a candidate for bearing an accent. To give a very simple example: The onset of a new note or chord always denotes an accentuated point or event in time; while the sound is being held or slowly dissipates, no accent can be attributed to it. However, a question directly arises: Is really every sound sequence which is held steady over time a totally non-accentuated event? A positive answer would seem to contrast with our perception of meter (at least in musical signals), where a listener, under the guidance of his own "internal metronome" ([60]) assigns importance over regular points in time, even if no event is there which explicitly brings novelty to the structure. It seems, therefore, that a definition is warranted which takes into account different aspects of accentuation.

Lester In *The rhythms of tonal music* [54], Joel Lester gives a descriptive definition of accent, attempting to define its properties in full. Firstly, he stresses the character of accent

as an interpretation tool and its frequent identification with *stress*, or *emphasis*, that is, a willful intensification of a specific event. This misconception originates from the common use of the term "accent" in daily musical routine. Secondly, he makes the point that accent can not be reduced to just the metrical accent - i.e., the one which is given by the bar or measure of the piece or to similar abstract patterns, such as the poetic feet schemata [18], without risking the loss of many other accentual aspects or the misattribution of accents to time points and not specific events. His view can be summarized in the following points:

- Accent is always a point of emphasis, mostly occurring through the appearance of a new event or a group of events.
- Accent is relative, denoting the difference of a point not absolutely, but in comparison to the others close to it in time.
- The metrical accent represents a special case, as it is a perceptual phenomenon.

He then continues by listing various factors which can invoke accentuation. He includes long durations (for example after short notes), new events (e.g. a change in harmony, the appearance of a new instrument or motivic change) and changes in dynamic and articulation. The model of Lester is decisive because it gives a comprehensive listing of many factors and properties which play a role in defining accent. However, it lacks a good grouping of the accents based on their nature or character.

Lerdahl and Jackendoff The last model to be shown here is drawn from the work of Lerdahl and Jackendoff, *A generative theory of tonal music* (henceforth abbreviated as GTTM) [53]. Their work has been extremely influential concerning a unified approach to describing musical structure. In it, they deal with various different subjects, such as rhythmic structure in general and its constituents, grouping and metrical structure. They proceed by defining a rule system pertaining to specific preference and well-formedness principles, resulting in a model which can theoretically explain any kind of rhythmic structure in western tonal music. We will not go into more details about this model, as this would lie beyond the scope of this thesis. Instead, we will focus on their definition of accent, which they use to assess the salience of different music events in the grouping or metrical structure. In their judgment, it is important to differentiate between three kinds of accent: *metrical*, *structural* and *phenomenal*. They define those three categories in the following way [53]:

- **Phenomenal accents** pertain to "...any event at the musical surface that give emphasis or stress to a moment in the musical flow" ([53], p.17). This category includes "...attack-points of pitch events, local stresses such as sforzandi, sudden changes in dynamics or timbre, long notes, leaps to relatively high or low note, harmonic changes, and so forth" ([53], p.17). It is clear that all possible events in music which carry an added weight for attention are covered under this label.
- **Structural accents** refer to accents "...caused by the melodic/harmonic points of gravity in a phrase or section" ([53], p.17). Those accents are therefore indicative of changes which denote events that are *structurally* important in the harmonic context of a given musical excerpt.

- **Metrical accents** are carried by "...any beat which is relatively strong in its metrical context" ([53], p.17). The strength in a metrical context refers to the salience of a beat, i.e. its importance in the meter hierarchy - for example, the downbeat. They continue by mentioning that the perceptual basis for the metrical accent are the phenomenal accents, in the sense that listeners base the extraction of a metrical accent on evidence from phenomenal accents. In the case of absence of regularity of phenomenal accents, metrical accents become less salient [53].

The authors proceed by explaining the interaction of those accents in some depth. The interested reader might refer to [53] for more information. For the thesis at hand, we keep one important element: The categorization of accents in GTTM provides a valuable tool for their division on the basis of their causes, which can then be used to connect them to specific features. This endeavor will be carried out in section 6.3. We will then take into account the models described here and attempt to create a correspondence of accents to specific features, which will then be useful both for the feature extraction and the classification steps. Of course, the matter of accent is not exhausted through its consideration here, but the basis for the implementation is certainly set.

4. Feature Extraction

After the basic concepts regarding rhythm and accent have been defined we will proceed to the more technical matters of the thesis. In the following, an introduction is given regarding the audio descriptors, the methods used to extract them and the general architecture of a feature extraction system. Most of the audio features discussed here are very well known in the audio processing and the music information retrieval literature; therefore, special attention will be pointed towards features describing rhythmic content, which are of increased relevance for the thesis.

4.1. Feature Extraction Fundamentals

Feature extraction signifies the process of the creating a numerical representation called a *feature* out of the signal or its spectrum. The features can either denote simple, statistical properties of the audio signal, such as its mean or its standard deviation, or more complex properties such as the shape of the spectrum or the magnitude of a specific harmonic component. When dealing with more than one audio samples as basic processing units, the feature extraction process results in the generation of a feature vector v, which encodes information about the value and distribution of the desired quantity for all audio samples and can be used as input to a classifier method.

In audio content analysis, it is desirable that features pertain to some perceptual quality of the signal, as they can then be used as a proxy for the representation of that quality. However, the semantic gap between the abstract, perceived elements of audio (such as timbre, pitch and rhythm) and the low-level values of the feature can be quite large, which calls either for the extraction of many different features or of important, perceptually relevant features in order to perform tasks such as musical genre classification efficiently. In both cases, the goal is the same: To explain as much of the variability present in the audio signal as possible. Achieving this goal is not easy, but since this variability is also the basis for the perception and cognition of musical qualities from humans, its pursuit is in any case a fruitful task. To this end, a large number of features for all kinds of tasks in MIR has been designed up to date. A comprehensive list of most known features can be found in [52]. Before that, we will address three important subjects which lie at the core of almost every feature extraction system and also find application here: frame-based feature extraction, spectral representation and preprocessing.

4.1.1. Frame-Based Feature Extraction

The feature extraction process is essentially a digital signal processing task: All numerical values are derived from the samples of an audio signal through direct calculation of physical properties or by applying known mathematical operations and transformations.

However, the calculations can either take place on the signal as a whole, or on parts of it, called *blocks* or *frames*. In this case, the algorithm selects only a fraction of the total samples of the audio data, and performs all necessary calculations, optionally applying a windowing method prior to that. afterwards, the frame slides to a next chunk of the audio signal, by advancing at a certain amount of samples and continues the processing. At the end, instead of the calculation of one feature per whole audio file, a trajectory of the feature value with respect to time has been extracted. This fact alone is decisive, since the temporal trajectory of a feature value can be expected to hold much more information than one value for the whole audio excerpt. Apart from that, the new representation is itself a time-series of values, i.e. a signal: that means, that transformations can be applied to it and features can also be extracted from it, which greatly increases the scope of available information regarding the signal at hand. Frame-based processing has some other advantages (for a full list, see [52]), which usually have to do with the specific operational mode of computer systems or algorithms such as the Discrete Fourier Transform. The discussion of such topics does not lie in the scope of the thesis. Here, we will rather focus on the specific terminology concerning block-based feature extraction. There are three important parameters in frame-based processing:

- **Frame length**: The frame length denotes the amount of audio samples considered as a basic processing unit. It will be denoted in the following with N_{FR}.
- **Hop size**: The amount of samples at which the frame advances in order to consider a new set of signal values. In the following, the symbol N_{FRHOP} will be used.
- **Overlap factor**: The overlap between consecutive frames when $N_{HOP} \leq N_{FR}$. We will use here N_{OVLP} to refer to this.

We will meet this terms again in part III when discussing the implementation of feature extraction in this thesis. It is important to note at this point, that all values given here in samples can also be given in a time unit of seconds. The relation between samples N and seconds t in the context of digital is given by $t = \frac{N}{f_s}$, where f_s denotes the sampling rate, i.e. the amount of samples of an audio excerpt in one second.

4.1.2. Spectral Representation and STFT

The spectral representation of digital audio signals, i.e., their frequency content, is almost exclusively calculates by applying the Discrete Fourier Transform (DFT). This transformation essentially maps the discrete-time signal into a frequency representation, by taking the product of the samples of the signal with basic, complex exponential functions. Detailed descriptions of the DFT can be found in a multitude of signal processing textbooks. In frame-based processing, the spectral representation of the DFT refers to a short-time version of the signal, effectively acquiring the name Short Time Fourier Transform and having the following formula (n denotes the index of the current frame, k the DFT frequency bin with frequency $f(k) = \frac{f_s \cdot k}{N_{FR}}$, i the sample index and N_{FR} the frame length):

$$X(k,n) = \sum_{i=1}^{N_{FR}} x(i) e^{-jk(i-1)\frac{2\pi}{N_{FR}}} \tag{4.1}$$

The STFT is most often used to create a *spectrogram*. This refers to a mixed time-frequency representation of an audio signal. Its calculation is relatively straightforward: for each frame of overlapping audio signal data, the STFT is calculated. Then, consecutive frames are mapped on an x-axis, whereas the y-axis denotes the indexes of the frequency bins in increasing order (i.e., the frequency constituents of the spectral representation). The magnitude of each bin is then shown by a color or brightness value, resulting in a pseudo-3D image [52]. Apart from being very useful for the visualization of the spectral form of a signal over time, the STFT can be used as the basis for the calculation of spectral features over time 4.1.2, making it a powerful tool.

One last note concerns the size of the frame. In the creation of STFT, the frame size takes values between a hundred and several thousand samples (or between 10 and 180 ms for usual sampling rates such as $f_s = 44100$ samples), remaining small enough to ensure that the signal is stationary (i.e. its statistical properties do not change over time) in the given consideration area. In other applications, such as the rhythmic feature extraction presented in this thesis, the frame size can be in the range of one to several seconds. This increased size allows for the capture of periodicities in the rhythmic relevant area, which require much larger amount of samples to manifest themselves.

4.1.3. Preprocessing

Preprocessing is an integral part of each feature extraction process. The reasons for that are multiple: Avoiding noise or undesired tonal components in the audio signal, creating versions of the signal which can be processed more effectively, or retaining only the relevant parts of the spectral information for a specific application. We will present here shortly some methods which are commonplace for feature extraction algorithms, although not all of them are used in the proposed implementation.

Normalization Audio excerpts are very often converted to have a maximum amplitude of one. This is achieved through the division of all samples with the greatest absolute value present in the signal and serves the purpose of having standardized signals, ensuring that features extracted from each of them will also be in the same range and only relative differences between them will be relevant.

Conversion to Mono Converts a stereo signal to mono by taking the average of both channels or by retaining only one channel, as in most audio content analysis applications (excluding the ones having to do with spatial content), the use of two channels is not necessary.

Down-Sampling Reduces the amount of samples needed to represent an audio excerpt. This could lead to shorter computation times in problems where this is necessary, without dramatically changing the shape of the signal waveform. However, down-sampling must be applied with caution if problems of aliasing[1]. Therefore, most algorithms apply a low-

[1] aliasing refers to the phenomenon of appearance of unwanted frequency components in a digital signal when its sample rate is reduced below the double value of the highest frequency component in the signal are to be avoided

pass filter before downsampling, in order to ensure that there is no frequency content in the signal above the new Shannon frequency ($\frac{f_s}{2}$).

DC Removal Subtracting the mean from a signal. Since DC components have 0 Hz frequency, any analysis aiming at analyzing spectral content does not profit from them. Instead, DC is seen us unwanted noise in most problems, which has to be removed in order not to contaminate results of statistics and other measures extracted from the signal.

Windowing The application of a window function is very often a prerequisite in the context of frame-based processing. A window function is essentially a block of samples with the same length as the frame, which has a specific form, defining its spectral content, which is multiplied sample per sample with the frame in question. This procedure always results in a change in the spectral content of the signal frame itself, which has to be taken into account in the subsequent analysis. A multitude of window functions exist for specific tasks.

Filtering It is very often the case that some of the spectral content of the audio signal is irrelevant for a specific application. In this case, filtering can be applied in order to reduce the signal to only low, high or middle frequencies. Filtering can also be used to stress specific frequencies in an audio signal, for example in the extraction of the envelope, where it is relevant to keep only low frequencies.

4.2. Instantaneous Features

In the following sections, the features extracted in the main part of the thesis are presented. These fall in two broad categories: Features describing the rhythmic content of an excerpt and general instantaneous features. Another category is pitch-related features [52], but as this thesis is dealing mainly with polyphonic music, they will not be presented here. For all features, the discrete time audio signal is denoted with $x[n]$ with n the number of the current sample, the discrete complex spectrum of the signal is denoted with $X(k)$, whereas k refers to the specific bin of the FFT. The length of the FFT-Transform is considered to be N_{FFT}, and we will use the index n for the consecutive frames of an STFT as well. Furthermore, $i = i_s(n)$ shows the beginning index of the signal frame and $i_e(n)$ the end index of the signal frame. This nomenclature borrows heavily from Lerch ([52]).

4.2.1. Spectral Shape, Tonalness and Intensity Features

Instantaneous features ([52]) are numerical quantities which can be extracted from small samples of audio material. They are also called low-level features, as they represent quantities which are well-defined and physical, as opposed to high-level features, representing abstract and perceptual qualities. The amount of those features is quite large, therefore many attempts have been made at classifying them in different taxonomies. Lerch lists some of the possible categorization schemes [52]. A very approximate information retrieval based segregation could be time and spectral domain features, on account of the

type of information that they provide about the audio signal. Of course, other categorizations are possible, such as tonal, spectral shape, intensity and statistical properties. We will give here an account of all the features used in the present thesis, together with a short description of their meaning. The abbreviations given here for the features will be used throughout the rest of the thesis to refer to them.

Spectral Flux (SF): Measures the amount of change in spectral shape of the signal. It gives an approximation to the sensation of "roughness" in the sound, taking large values for abrupt changes in the spectral composition of the signal.

$$f_{SF}(n) = \frac{\sqrt{\sum_{k=0}^{N_{FFT}/2-1} \left(|X(k,n)| - |X(k,n-1)|\right)^2}}{\frac{N_{FFT}}{2}} \tag{4.2}$$

Spectral Centroid (SCD): Is a measure of the spectral center of gravity. Greater values signify more content in the higher frequencies, therefore brighter or sharper sound.

$$f_{SCD}(n) = \frac{\sum_{k=0}^{N_{FFT}/2-1} k \cdot |X(k,n)|^2}{\sum_{k=0}^{N_{FFT}/2-1} |X(k,n)|^2} \tag{4.3}$$

Mel-Frequency Cepstral Coefficients (MFCC): They have been used extensively in the area of speech recognition, as they provide a compact representation of the spectral shape of the signal. The details of their calculation can be found in [52]. The formula for their calculation is given here. The symbols with an apostrophe denote the mel-warped spectrum [97]. The transform here corresponds to the DCT of the the mel-frequency bands, and instead of the spectrum magnitude its logarithm is taken (resulting to the cepstrum). Those adjustments offer a number of improvements, which have made the MFCCs a valuable tool in many tasks. The index j denotes the order of the coefficient to be taken, in this thesis the 13 first coefficients are used.

$$f_{MFCC}(j,n) = \sum_{k'=1}^{N'_{FFT}} log(|X'(k',n)|) \cdot cos(j \cdot (k' - \frac{1}{2}) \frac{\pi}{N'_{FFT}}) \tag{4.4}$$

Spectral Flatness (SFL): It is a measure of the tonality of the signal based on the spectrum, as it represents the ratio of geometric and arithmetic mean of the magnitude spectrum.

$$f_{SFL}(n) = \frac{\sqrt[N_{FFT}/2]{\prod_{k=0}^{N_{FFT}/2-1} |X(k,n)|}}{(2/N_{FFT}) \cdot \sum_{k=0}^{N_{FFT}/2-1} |X(k,n)|} \tag{4.5}$$

Spectral Pitch Chroma (SPC): The pitch chroma is a twelve-dimensional vector where each dimension represents on pitch class. Each coefficient is calculated through the following formula (o denotes each octave):

$$f_{SPC}(j,n) = \sum_{o=o_1}^{o_u} \left(\frac{1}{k_u(o,j) - k_1(o,j) + 1} \sum_{k=k_1(o,j)}^{k_u(o,j)} |X(k,n)|\right) \tag{4.6}$$

Spectral Tonal Power Ratio (STPR): Another measure of the tonalness of the signal, the tonal power ratio is defined through the ratio of tonal power $E_T(n)$ of the signal to the overall power $E(n)$. For the calculation of the tonal power, the tonal components of the signal have to be identified through finding tonal peaks in the signal, usually through finding local maxima or peaks that lie above a given threshold ([52]).

$$f_{STPR}(n) = \frac{E_T(n)}{\sum_{k=0}^{N_{FFT}/2-1} |X(k,n)|^2} \tag{4.7}$$

Spectral Spread (SSP): The spectral spread describes how concentrated the spectrum of a signal is around the spectral centroid, essentially being a measure of the bandwidth of the signal.

$$f_{SSP}(n) = \sqrt{\frac{\sum_{k=0}^{N_{FFT}/2-1} (k - f_{SCD}(n))^2 |X(k,n)|^2}{\sum_{k=0}^{N_{FFT}/2-1} |X(k,n)|^2}} \tag{4.8}$$

Time Peak Envelope (PE): A peak envelope is extracted by taking the absolute maximum value of the signal amplitude in a block of samples.

$$f_{PE}(n) = \max_{i_s(n) \leqslant i \leqslant i_e(n)} |x(i)| \tag{4.9}$$

Time RMS (RMS): The RMS feature gives the root mean square of the signal amplitude in one block. It results from summing the squared magnitude spectrum for a processing block.

$$f_{RMS}(n) = \sqrt{\frac{1}{N_{FFT}/2} \sum_{i=i_s(n)}^{i_e(n)} x(i)^2} \tag{4.10}$$

Time Zero Crossing Rate (ZCR): The Zero Crossing Rate feature is a measure of tonalness of the signal. It is calculated as the amount of zero crossings of the signal in a given time period. It takes large values for noisy, high frequency signals, whereas small values for purely tonal signals.

$$f_{ZCR}(n) = \frac{1}{2 \cdot N_{FR}} \sum_{i=i_s(n)}^{i_e(n)} |sgn(x(i)) - sgn(x(i-1))| \tag{4.11}$$

A full list of the features described here along with their abbreviation is given in table 4.1.

Instantaneous features
Spectral Flux (SF)
Spectral Centroid (SCD)
Mel-Frequency Cepstral Coefficients (MFCC)
Peak Envelope (PE)
Root Mean Square (RMS)
Spectral Pitch Chroma (SPC)
Spectral Tonal Power Ratio (STPR)
Root Mean Square (RMS)
Spectral Flatness (SFL)
Spectral Spread (SSP)
Zero Crossing Rate (ZCR)

(4.12)

Table 4.1.: Instantaneous (spectral shape, tonal, intensity) features

4.2.2. Distribution Features

In this section, a special category of features pertaining to the distribution of audio signal values is given. It is important to stress that those features are applicable for any kind of signal, since they consider it as a distribution of random values from which statistics and other measures can be extracted. It follows that they can also be extracted from a spectral or periodicity representation without problems regarding their validity, although their conceptual meaning might be somewhat vague. Those features will be referred to from now on as subfeatures, since they will be used in the context of this thesis as features extracted on transformations, temporal trajectories or periodicity representations of other features. In the following, $x(i)$ denotes the current signal frame with index i for its samples, n stands for the index of the current frame, i denotes the index of the current sample, $i = i_s(n)$ shows the beginning index of the signal frame and $i_e(n)$ the end index of the signal frame.

Mean (ME): The subfeature measures the average of a signal or signal block. It is given by:

$$ME_x(n) = \frac{1}{N_{FR}} \sum_{i=i_s(n)}^{i_e(n)} x(i) \tag{4.13}$$

Geometric Mean (ME): This subfeature measures the average of a set of values ordered logarithmically:

$$GM_x(0,n) = \sqrt[N_{FR}]{\prod_{i=i_s(n)}^{i_e(n)} x(i)} \tag{4.14}$$

Standard Deviation (SD): Standard deviation measures the spread of the values of a signal around their arithmetic mean:

$$SD_x(n) = \sqrt{\frac{1}{N_{FR}} \sum_{i=i_s(n)}^{i_e(n)} (x(i) - ME_x(n))^2} \tag{4.15}$$

Centroid (CD): The centroid is a measure of the center of gravity of an input signal or distribution:

$$CD_x(n) = \frac{\sum_{i=i_s(n)}^{i_e(n)} (i - i_s(n)) \cdot x(i)}{\sum_{i=i_s(n)}^{i_e(n)} x(i)} \tag{4.16}$$

Skewness (SK): Known also as the centralized moment of third order, skewness measures the asymmetry of a probability distribution.

$$SK_x(n) = \frac{1}{\sigma_x^3(n) \cdot N_{FR}} \sum_{i=i_s(n)}^{i_e(n)} (x(i) - ME_x(n))^3 \tag{4.17}$$

Kurtosis (KU): Kurtosis measures how much the form of a distribution deviates from that of a normal gaussian.

$$KU_x(n) = \frac{1}{\sigma_x^4(n) \cdot N_{FR}} \sum_{i=i_s(n)}^{i_e(n)} (x(i) - ME_x(n))^4 - 3 \tag{4.18}$$

Flatness (KU): Flatness is a measure similar to the one of spectral flatness presented before and measures the ”peakiness” of a distribution.

$$FL_x(n) = \frac{\sqrt[N_{FR}]{\prod_{k=1}^{N_{FR}} |x(i)|}}{(N_{FR}) \cdot \sum_{k=1}^{N_{FR}} |x(i)|} \tag{4.19}$$

Entropy (EN): Entropy (for a distribution) measures the average information content stored in a probability density function. For signals, it gives an amount of ”novelty” in the signal values.

$$EN_x(n) = -\sum_{i=i_s(n)}^{i_e(n)} x(i) \cdot log_2(x(i)) \tag{4.20}$$

High Frequency Content (HFC): This subfeature measures is closely related to the centroid and measures the amount the values appearing at the higher end of a distribution. The implementation here takes into account the squared distribution and does not scale by the sum of all values in order to created a better scaled form of the feature.

$$HFC_x(n) = \sum_{i=i_s(n)}^{i_e(n)} (i - i_s(n)) \cdot x^2(i) \tag{4.21}$$

We will present a last category of subfeatures, the peak related features in the next section, since they are derived from the specific periodicity representation and can not be categorized with the features presented in this section. A full list of the features presented in this section can be seen in table 4.2. Here, we will also include two more features, the mean and standard deviation of the derivative, since they are easily calculated and therefore often used [14]: they denote the computation of the mean and standard deviation from the first derivative of the signal, in order to detect sharp changes in the signal or distribution. Since the taking of the first difference is equivalent to a high pass filtering for digital signals, the features are essentialy a measure of the mean and standard deviation of a signal version with less low-frequency content.

Distribution Features
Mean (ME)
Standard Deviation (SD)
Mean of Derivative (MD)
Standard Deviation of Derivative (SDD)
Skewness (SK)
Kurtosis (KU)
Entropy (EN)
Geometric Mean (GM)
Centroid (CD)
Flatness (FL)
High-Frequency Content (HFC)

(4.22)

Table 4.2.: Distribution features

4.3. Rhythmic Content Features

In the following, the discussion will focus on the features describing the rhythmic content of an audio excerpt. We will begin by giving a short account of onset detection and novelty functions, as they are important concepts for the ensuing discussion. Afterwards, we will analyze briefly the general scheme of periodicity representations and focus on the *beat histogram*, a method for creating a simple yet effective rhythm content representation, out of which features can be extracted. At the end of the section, some more subfeatures will be presented to complete the list in table 4.2, for which through the discussion of the beat histogram is necessary.

4.3.1. Onset Detection

A part of this work is concerned with the field of onset detection, more precisely with the concept of the novelty function, which will be described in more detail now. Onset detection is a wide field whose results are important for applications such as segmentation (see [95] for some uses), tempo detection or automatic transcription [52]. The goal of onset detection lies in the finding and tracking of onsets in the audio signal, defined as the beginnings of new sound events, such as for example, musical notes. When analyzing musical events, it becomes clear that the onsets are not "clear cut" points in time, as they have a specific duration which is not infinitesimally small. The onset event can be described by a short time period, the attack or rise time, in which the note or event makes its appearance and reaches its maximum amplitude. The signal in this period does not possess a steady-state character, but is in a phase of very quick evolution. The onset time is then taken to be one time point in this duration which denotes the event has reached a quasi-periodic state. As Lerch notes [52], the aforementioned terms are often used in an inconsistent manner and, to make it worse, the times extracted from the acoustic signal very often do not coincide exactly with the times perceived by listeners. Therefore, one distinguishes between note onset time, acoustic onset time and perceptual onset time [71]. As we deal with systems in the field of ACA here, we will choose to define an onset (and its respective time of occurrence), as the acoustic onset time, i.e. the point in the signal where a signal or acoustic event is theoretically measurable [52, 71]. As Lerch notes [52], the accuracy of automatic detection systems is usually too poor to allow to distinguish between an acoustic and a perceptual onset. Therefore, we can take as granted that the perceptual onset time of an acoustic event (which would be desirable to extract) coincides with the acoustic onset time, since the goal of the thesis is not perfectly accurate onset detection (as, e.g., in [95]), but accurate detection of the periodicities caused by them, which is guaranteed by the algorithms consistency in the way of detecting onsets in the signal.

4.3.2. Novelty Function

The novelty function (also known as novelty curve or detection function [6]) constitutes the first building block of an onset detection system [52]. The novelty function is essentially a trajectory denoting points in time where new events take place, i.e. the points where changes of the acoustic signal are very marked. There is no standardized way to extract a novelty function from an acoustic signal. However, most studies up to date take into account the form of the signal envelope, extracted through taking the first difference of the audio waveform, smoothing, and applying half- or full-wave rectification. The envelope analysis can either apply to the whole time signal or to specific bands in it. Schloss [77] was one of the first to present such a system as described above for onset detection, followed by Klapuri and Scheirer which used the envelope of the signal in several subbands and applied psychoacoustic transformations ([76, 47]), whereas newer publications compute the novelty function by STFT methods ([24, 6]). As Lerch states [52], the advantage of such methods is that not only changes in the amplitude (or its envelope) are taken into account, but also changes in the spectral domain. Such methods unfortunately come with the drawback of a lower time resolution, which is, however, mostly not relevant since the resolution of a block-based STFT with appropriate parameters is good enough for most applications.

Many onset detection studies ([24, 6]) have stressed the need to proceed in such a way. Most publications have then tested those automatically extracted onsets against manually labelled ones, in order to assess the precision of the algorithms [95, 24]. Others have tried to extract the exact tempo of given musical excerpts or make assertions about the metrical structure, attempting to denote for example the downbeat [36, 37, 35]. There are relatively few studies however on using those spectral onsets to describe rhythm and the perform genre classification based on the extracted features. The present work attempts exactly that, by addressing first the problem of finding perceptually relevant novelty functions. A detailed analysis of onset detection functions can be found in ([47, 24, 6, 4, 19, 5]). For the sake of completeness we give here the following categories, in which such functions fall:

- **Spectrum-based novelty detection:** Comprises novelty detection based on spectral flux, phase deviation, weighted phase deviation, complex domain (departure from steady-state behavior), rectified complex domain etc. [24, 6]
- **Statistics novelty detection:** Comprises methods which are based on entropy calculation or probability models [24, 6].
- **Time-signal-based novelty detection:** Performs novelty detection based on the envelope of the time signal. This can be performed either for amplitude or energy, and either directly on the waveform or in subbands of the signal.

Finally, we will give a short overview novelty detection methods applied directly in the spectral domain. The most widely used spectral novelty method is the spectral flux, described in section 4.2.1. Lerch [52] mentions three other methods from Laroche, Duxbury and Hainsworth ([50, 30, 42]), which all use different versions of the spectral flux (in complex form or with different distance measures) in order to calculate spectral novelty from the signal, whereas Bello [4] use the phase relations, as they can be indicative of transients (since their phase is not linear) and which most often appear in the case of onsets. Since the subject of novelty functions is a wide and upcoming field of research, we will not give more information about other methods at this point. However, it is important to stress that this work is also based on a similar approach: Spectrum-based methods for novelty detection up to date have not taken into account the trajectory of other quantities of the signal based on instantaneous features, such as spectral flatness, spectral centroid etc. (see 4.2.1), although the change of those quantities in time might provide information which cannot be found only in the spectral flux or its variations. We will see in part III how novelty functions can be extracted from those instantaneous features and how they can be used for rhythmic description. In the following section, a description of another important part of the rhythmic content analysis, specifically periodicity detection and the features involved in it, will be discussed in some depth.

4.3.3. Beat Histogram

Approaches for the extraction of features for the representation of rhythm and temporal structure belong to the field of automatic rhythm description. In an important publication from Gouyon [38] such those approaches are presented and discussed. The span of the applications concerning the subject is fairly wide: Rhythmic representations can be used

for tempo induction, beat tracking, time signature or swing determination, and, of course, applications of a wider scope such as musical genre classification. The common analytical frame work underlying the creation of such rhythmic content representations can be summarized as follows:

- An audio excerpt is preprocessed and subjected to an onset or novelty detection function, which creates a list of onset times or a novelty curve for some musical quality such as pitch or loudness.
- The resulting novelty curve is then used as a basis to extract a periodicity representation with the use of a suitable method.
- The periodicity representation is then used for further tasks: The extraction of features and statistics, the comparison with existing templates for similarity calculation or the determination of more abstract elements of an audio excerpt, such as meter or tempo.

The above steps can be performed with or without the use of preprocessing for the audio signal, and there are a number of different operations which can be applied in each part of the procedure, depending on the end goal. As the subject is extremely complex, both concerning the methods of the periodicity extraction and the possible tasks which follow it, we will focus here only on those approaches which use rhythmic content features for musical genre classification. The term "rhythmic content features" (originally from Tzanetakis et al. [91]) represents a special category of audio descriptors, which denote quantities related to the rhythmic structure of music. As stated in section 1.2, those have been already used in the context of musical genre classification, mainly by [91, 39, 14]. Such studies share a common approach for the (frame-based) processing scheme, which is very close to Scheirer's original approach [76], who was the first to extract periodicities in a completely methodical way:

- They create a reduced version of the time domain input signal by filtering (through normal filters or by applying, e.g., a Discrete Wavelet Transform) or applying other methods such as downsampling. Optionally, the signal is preprocessed with the methods shown in 4.1.3.
- The signal's envelope is extracted by half-wave rectification and smoothing, possibly taking the derivative (or, for discrete-time signals, the first difference) to stress changes in the signal. Here, the points of novelty can be extracted as singular onsets.
- From the signal envelope curve, which serves as a novelty function, the extract a periodicity representation in the low frequency area (between, e.g., 30 and 240 BPM (0.5 to 4 Hz)) using an autocorrelation function or a comb filterbank. Alternatively, if the onsets are present as discrete points on the time domain signal, an computation of an *Interonset-Interval histogram* (IOI-Histogram) is attempted, to derive the durations between all onsets.

After the periodicity representation has been created, features are extracted from it in the same way they are extracted from a signal or a distribution of values. As such, the

extraction of all the distribution features shown in table 4.2 and discussed in section 4.2.2 is applicable. However, even the spectral and other features discussed earlier can be considered, as the resulting periodicity representation is, in fact, in many accounts similar to a spectrum, but constrained in the very low frequencies. Gouyon [39] attempts that by extracting MFCCs from an IOI-Histogram and reports very good results on genre classification by using the first 15-MFCCs on a periodicity histogram. However, taking into account the definition of MFCCs (4.2.1), their perceptual meaning considering what they state about a periodicity representation is not so clear to us. Therefore, in order to keep the features explainable and their number as low as possible, we will refrain from extracting truly spectral subfeatures from the beat histogram. Burred and Lerch [14] extracted statistics of the beat histogram, together with a *rhythmic regularity* feature, denoting evenly spaced peaks in the beat histogram and signifying periodicities in integer relations to another. Tzanetakis, in [91], extracted another group of features, which refer to the most salient peaks in the beat histogram and its overall strength. We will focus here only on the ones from the last, as they are of the most important used in the literature and their meaning relatively straightforward:

Amplitudes of the most salient peaks (A0, A1, RA): The amplitudes of the two highest peaks in the beat histogram measure the beat strength (i.e., how pronounced they appear in the signal) of the predominant signal periodicities. apart from that, their ratio is also taken in order to account for more complex rhythmic structure.

Periodicities of the most salient peaks (P1, P2, P3): The frequencies of the two highest peaks in the beat histogram denote the BPM value for which the beat is the strongest. Most often, they coincide either with the tempo of the piece or with a related integer multiple or division of it. Here, we also consider a third periodicity (P3), which is given as the mean of the two others and measures the central tendency of the beat histogram.

Sum and Sum of Power (SU, SP): Those two subfeatures measure the overall beat histogram strength and are expected to be high for excerpts which have a rich or very pronounced rhythmic structure. The sum of power is calculated by taking the square of the beat histogram, i.e. its power.

We will refer to those descriptors as *peak features*, since they are related to the peaks in the beat histogram. A full count is given in table 4.3. Taking those features for the rhythmic representation plus the subfeatures mentioned earlier in subsection 4.2.2 gives in total 19 different subfeatures for the beat histogram. As mentioned before, the count of features rises very quickly with the addition of new possibilities, therefore we stopped at this number, considering it sufficient for the experiments to follow and having the surety of account for many sources of variance in the beat histogram representation. The most adequate features will be selected later through feature selection.

Peak Features
Amplitude of the most salient peak (A0)
Amplitude of the second most salient peak (A1)
Ratio of A0 to A1 (RA)
Periodicity of the most salient peak (P1)
Periodicity of the second most salient peak (P2)
Mean of P1 and P2 (P3)
Sum of beat histogram (SU)
Sum of power of the beat histogram (SP)

(4.23)

Table 4.3.: Peak features

5. Machine Learning

This chapter provides an account of machine learning fundamentals and the classification methods used in the musical genre classification task. In the context of thesis, two instance-based algorithms for supervised learning are used: The k-Nearest-Neighbor (*abbr.* ***kNN***) and the Support Vector Machine (*abbr.* ***SVM***). These methods possess a solid mathematical basis, they have been used extensively (also in the context of genre classification) and have proven efficiency. They are also already described in many machine learning textbooks [28, 12, 78, 20] and have numerous implementations in existing toolboxes, allowing their free and unproblematic use. As the SVM algorithm is a state-of-the-art classifier which has shown very good performance results in a variety of problems, it will receive a special focus in the following discussion. The algorithms in this work have been utilized as components of the MATLAB environment (kNN) or as toolboxes which can be used therein (libSVM [15]).

5.1. Machine Learning Fundamentals

Machine learning (*abbr.* ***ML***) is a field of techniques and methods in artificial intelligence, which concerns computer algorithms which can learn from data, i.e. to extract meaningful information out of them. This can then be used either for analysis or for prediction. The information derived is often referred to as a *pattern*, which shows the affinity of the field to pattern recognition. The basic function of an algorithm in ML deals with analyzing relevant data observations given in a numerical form performing a set mathematical operations on them, and creating a model which then can be used to generalize to new observations for a given task. As the term is somewhat vague, we will use it here to refer to those algorithms and systems which automatically infer the structure in the data by extracting this information from the data itself. ML algorithms come in two fundamental forms: **supervised** and **unsupervised**. The intermediate category of **reinforcement learning** is not relevant to the thesis and will therefore not be covered here.

Unsupervised learning This collection of problems concerns the finding of structure in data without possessing any prior information about the categories to which the samples belong to. It encompasses approaches ranging from clustering to principal component analysis, though the exact taxonomy of the methods is not very clear since they are used in the context of many disciplines for various problems. Although at the beginning this seems to be a difficult and somewhat ill-defined problem, Duda et al. [28] note several reasons as to why it is useful to employ unsupervised learning. The basic rationale behind those reasons is that although large amounts of data are nowadays very easy to come by, their manual annotation remains very costly and tedious, even if one knows how to perform it correctly. Furthermore, there are cases where no annotation at all is possible; in

this case, it is well-advised to probe the data to explain *something* about itself, essentially to see what it suggests. Another advantage of those methods is that they can help in dimensionality reduction: When dealing with data which is represented through a large amount of features, unsupervised learning can disclose hidden structure in the data which in turn may help to reduce the number of features greatly, by discarding or combining them. However, this category of algorithms has a considerable drawback: When faced with complex tasks, the output of the algorithm might not be very informative with respect to the specific task, as it essentially has no information about **what** should be achieved - it simply uncovers some information in the underlying data. In such cases, the use of supervised learning algorithms is of advantage.

Supervised learning In supervised learning, the discovery of structure in the data is performed through providing the algorithm with labeled instances, or instances whose class membership is already known. This setting has the advantage that the algorithm receives information about the correct output which has to be achieved. In this sense, the algorithm just has to learn a *mapping* from the input data to the output (or target) data. This mapping, encoded as a set of mathematical operations and values, is stored as a model which can subsequently be used for prediction of the labels of data not yet presented to the algorithm. It becomes apparent that the functionality of an ML algorithm in this setting bears a parallel to the procedure of inductive reasoning in humans: the algorithm derives a rule which connects the samples and the features, based on some observations at hand, and then used this rule to generalize to observations which have not been encountered earlier. An important implication of those procedures and the existence of labels for the sample data is that the learning performance of the algorithm can be evaluated in a much more meaningful way with respect to unsupervised learning, as the outcome labels of the algorithm and the "real" labels of the data (also known as *ground truth*) can be compared directly. Another important distinction concerns the methodology of the learning procedure. In supervised learning, there are two steps which have to be followed: first, a group of examples together with the labels are presented to the algorithm. Subsequently, the algorithm analyzes the data in a specific way and "learns" the way they relate to their labels based on the features representing the data. This step is called the *training phase*. Afterwards, a second group of unknown data is presented to the algorithm without their labels. The algorithm now has to predict their labels by applying the model it derived from the training data. This second step is called the *prediction phase*. After those two steps, the performance of the algorithm can be evaluated using common evaluation methods and its parameters tweaked accordingly.

Classification The term *classification* is used frequently in pattern recognition and statistics to refer to the problem of generally assigning a class label to an observation, based on a rule which is derived by analyzing a subgroup of the data. Considering automatic classification in the field of machine learning, the data is almost always represented through the numerical values of the features, which allows for the study of the problem of classification independently from the domain of application ([28]). Although classification is a very diverse problem which involves many procedures ([12, 28]), we will focus here in its application and use for supervised learning problems. In this context, the classification

algorithm has to learn a specific *decision function*, given the input and their class membership. The formulation of the problem is fairly simple: Given a vectorial representation of a set of samples in the feature space, the goal of classification is to assign a given input vector $\mathbf{x}$ to one of more discrete classes C_k , $k = \{1, 2, ..., K\}$ [12]. As it is apparent, the primary problem is of a binary nature: Every instance can either be classified as belonging to one class or not. Its generalization to multiple classes has to be dealt with separately and will be examined later on in this section. The algorithms which perform classification are called *classifiers* and the two methods used in this thesis fall under this category. In the following, we will discuss briefly the problem of linear classification, as it is fundamental for the explanation of the SVM method presented in 5.3

5.1.1. Linear Classification

The problem of linear classification is one of the most important in machine learning and has ties to linear discriminant analysis and linear programming. In linear classification, the decision surface (as we are referring to data in n-dimensional spaces) is a linear function of the input vectors [20], so that:

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b = \sum_{i=1}^{n} w_i x_i + b \tag{5.1}$$

From equation 5.1 it becomes apparent that the decision surface is actually a hyperplane (i.e. a plane in an n-dimensional space), separating the feature space in two disjoint regions. Observing its sign, a decision can be made about which class the instance belongs to: if $f(\mathbf{x}) > 0$, then the sample belongs to the positive class C_p, otherwise it belongs to the negative class C_n. The two parameters in the equation have the following significance: $\mathbf{w}$ is called the weight vector and is perpendicular to the hyperplane (thus determining its angle) and b is called the bias and distance of the hyperplane to the origin of the feature space. The formulation given here has been used in both statistics and neural processing, being known respectively as the *linear discriminant function* and the *perceptron* [20].

The category of problems which can be solved exactly by defining a linear hyperplane are called *linearly separable* [12]. Although a large category of problems exists for which such a solution is achievable, there are many problems which are non-separable: A linear decision surface which perfectly separates the instances does not exist, due to the form of the data distribution, so that any attempt to classify instances in this way will result to some misclassifications. At this point, it is important to refer to the concept of the *functional margin* of an instance $\mathbf{x}$ to the hyperplane $(\mathbf{w} \cdot \mathbf{x})$: It is defined as the quantity [20]:

$$\gamma_i = f(\mathbf{w} \cdot \mathbf{x}\rangle + b \tag{5.2}$$

This quantity is of great importance, as the greater the margin is for a training sample (or for the whole data set), the better separation a decision surface achieves [20]. The SVM is an example of a classifier which attempts to maximize this margin, leading to an optimum decision surface. The relevance of the margin will be examined closer in section 5.3.

5.1.2. Multiple Classes

The extension of the two class problem to multiple ($K > 2$) classes is unfortunately no trivial matter, as the classification procedure described in 5.1.1 is essentially a two-class problem. Several methods can be used to overcome this difficulty [12, 20]:

One-versus-all classification In this case, the multiclass problem is decomposed in $K-1$ subproblems, where each of the classifiers associated with it has to solve a binary classification problem, where the samples belonging to the class in question constitute the one case, whereas all the others are considered as belonging to the other case. The final solution results from combining (as a union) the solutions of each of the $K-1$ classifiers.

One-versus-one classification In one vs. one classification, the problem is stated as the sum of $K(K-1)/2$ binary classifiers, each of them separating one class from the other, for all pairs of given classes. An instance is then classified to a single class by a majority vote amongst the discriminant functions.

Direct multiclass discrimination The discriminant function here has the form: $\langle \mathbf{w_i} \cdot \mathbf{x} \rangle + b_i$, where w_i and b_i refer to the weight vector and bias of each of the discriminant functions. The decision function is then given through $c(\mathbf{x}) = arg\,max(\langle \mathbf{w_i} \cdot \mathbf{x} \rangle + b_i)$ over all K classes. Thus, the sample $\mathbf{x}$ is classified in the class whose hyperplane is furthest from it.

Referring to the first and second method, the combination of several binary linear discriminant functions for classification has the drawback of possibly creating regions in the feature space whose samples cannot be assigned to any class [28, 20, 12]. Furthermore, the partition of the problem in subproblems might create difficulties if the data is unbalanced, i.e. if many more samples are available for one class than for the other(s). The third method is relatively straightforward and has the advantage of creating K simply connected and convex decision regions in the feature space. Thus, its application is preferred whenever possible. However, this is not always the case - especially for the SVM as a binary linear classifier, very often one has to resort to the first or second method. However, as will be shown in section 5.3, there is a multiclass implementation of the non-linear SVM which functions similar to the direct method presented here. The kNN method does not suffer from such problems, as the decision boundaries result essentially from estimating the data distribution directly from the samples and without making any assumption for the form of the decision surface (which is also not directly a function of the data samples). Therefore, the existence of multiple classes is inherent in the "native" form of the kNN.

5.2. k-Nearest-Neighbor

The kNN algorithm is one of the most simple and yet efficient methods in the field of statistical pattern recognition. It is a non-parametric method for supervised learning, which can be used for the estimation of unknown probability distributions underlying the data, utilizing the nearest-neighbor rule [28]. When used for supervised classification, the input

consists of a matrix of m observations and n features or dimensions across which each of the observations is defined, along with a class label for some of the observations. During classification, an observation whose class is unknown is assigned a class label based on the majority of the classes of its k nearest neighbors - the proximity measured based on some distance metric such as euclidean, manhattan or mahalanobis [28]. The only algorithm parameters[1] which have to be defined beforehand are therefore k, the number of nearest neighbors on which the class label decision is based, and the distance metric for the proximity calculation. Essentially, the kNN algorithm lets the data alone define the decision boundary for classification, based on its intrinsic structure in the feature space.

The theoretical basis for the kNN algorithm can be found in the field of non-parametric density estimation of an arbitrary distribution and is described in depth in [28, 12, 2]. It is beyond of the scope of the thesis to provide details about the inner workings of the algorithm, but the interested reader can refer to the aforementioned literature. It is however, useful to provide a rationale for its use in this thesis. A first argument is that, exactly owing to its simplicity and power, the kNN algorithm has been - and still is - used for almost every supervised classification task, regardless of the data type involved. As such, its performance results can serve as a good comparison basis with those of other publications - especially in [91, 14, 39], where it has been used for genre classification as well. At the same time it allows us to perform a "reality check" and establish a baseline classification performance result which can then be compared to the performance the SVM.

Notwithstanding its widespread use, kNN is not always the best possible method which can be applied for classification, as it is sensitive to local structure present in the data when using tractable implementations ($k < \infty$), which can lead to poor performance. Furthermore, as a non-parametric method, kNN does not make any a priori assumptions about the statistical distribution or structure underlying the data. This property is one of its advantages but also a drawback: the absence of a need for previous analysis or assumption about the data allows its application to almost every problem, even (or especially) when little is known about the data model. However, it can lead to very suboptimal or misleading results when dealing with data which are unscaled, noisy, possess complex inner structure and high dimensionality if its parameters are not selected properly [11]. Regarding the values of these two algorithm parameters, the following considerations are of relevancet:

- **Number of nearest neighbors k:** The number of k to use for classification is not fixed. It can be theoretically shown that as k increases, the error rate of the algorithm approaches the theoretical optimum (Bayes rate) [28]. However, since it is impossible to use a very large value for k due to reasons of computational cost, most publications use the first few uneven numbers (e.g., $k = 1, 3, 5, ...$), since they usually provide satisfactory results. It is also advisable to choose an uneven value for k, as otherwise the vote of the nearest neighbors could theoretically result in a tie, which would render the assignment of a label to an unknown instance impossible. A plausible approach is to test for several values of k, until an optimum (which is problem-dependent) has been reached.

- **Distance metric:** The distance metric is also an open subject in the use of the kNN

[1] here, the term *parameters* refers to the hyperparameters of the algorithm, i.e. its settings, and not to the statistical parameters of a distribution.

in a given problem. The most often used distance metric is the euclidean norm (also known as L2-Norm). Although its suitability is not warranted for all kinds of problems (since it cannot deal with translation or rotation of the data, see [28]), when used with data scaled in the same way (as is the case here), it can be guaranteed to give similar results to other distance metrics, like manhattan or mahalanobis [28]. One possibility is to define a new distance metric with properties related to the specific problem. However, as the euclidean metric is easily interpretable (as the standard distance between two point in an n-dimensional space) and has been used in most publications dealing with genre classification, we use this distance metric in our experiments as well.

There exist some further enhancements for the kNN algorithm. For example, a distance-related weighting scheme can also be applied on the voting participants, so that neighbors lying further away from the observation to be classified have less influence on the decision [29]; or misclassification costs for specific classes (which, e.g., possess a small count of observations) can be defined in order to create adaptive versions of the algorithm [89]. For the sake of retaining the comparability with other experiments and because the focus of the work does not lie on the parameter selection of the kNN algorithm, we will refrain from applying any of those refinements. Their use in future experiments is, however, not ruled out.

5.3. Support Vector Machines

SVMs were first introduced from Vapnik in [92]. The evolution of the SVM based on the need to progress from linear programming algorithms with margins to algorithms which could achieve even more difficult partitions of the feature space, especially in defining non-linear, complicated separation surfaces.

SVMs were first introduced from Vapnik in [92]. They originated in the context of linear discriminant analysis, based on the need to progress from linear programming algorithms with margins to algorithms which could achieve even more difficult partitions of the feature space, especially in defining non-linear, complicated separation surfaces. Although the original formulation of the SVM is a linear one, the main property which helps the calculation of complex decision surfaces is the kernel trick: a transformation in the heart of the algorithm which projects the data in an higher-dimensional space (where they can be linearly separated) and then transforms it back to the original space, delivering a non-linear decision surface. It falls in a category of algorithms known as kernel methods, which essentially use linear or non-linear transformations (kernels) to ”map” the features in a space where classification can be performed more efficiently. Detailed information and the mathematical background about kernel learning and specifically SVMs can be found in [78, 20]. For the purposes of the thesis, the basic functionality will be presented here.

As it name gives away, the SVM is an algorithm which calculates a decision surface (a so-called hyperplane), its limits denoted by the support vectors (which are in fact samples that act as limits for the separation surface, see figure 5.3). This decision surface is calculated in a very high dimension (through the kernel) as an n-dimensional plane and the surface is then transformed back to the original space.

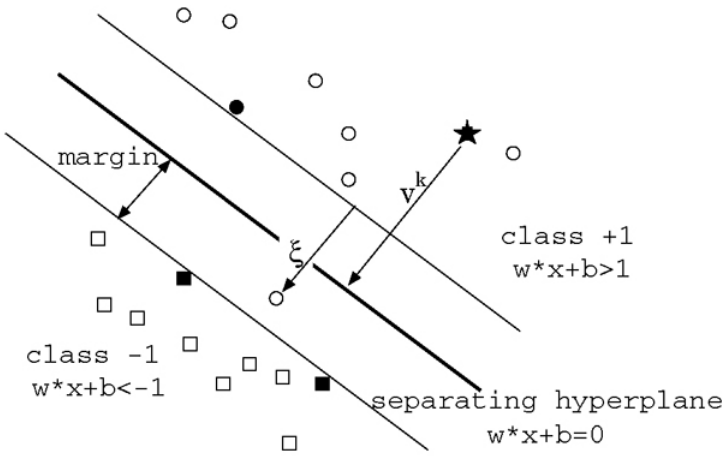

Figure 5.1.: Operating mode of SVMs, image originally from [51]

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \tag{5.3}$$

$$\mathbf{y}_i(\mathbf{w}^T(\mathbf{x}_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 \tag{5.4}$$

The equations shown give an insight in the algorithm parameters of the SVM. As the kNN method, SVM also does not make any assumption about the underlying structure of the data. However, in this case there is a clear training step, in which a part of the data with their classes are presented to the algorithm. The SVM then "learns" the weights of the vectors as well as the offsets in order to satisfy equation 5.3 for the given samples and classes, that is for the given separation surface. Afterwards in the validation/test step, new samples (feature vectors of the observations) are inputted without their labels, and the SVM uses the previous weights and offsets, essentially performing a regression, in order to separate the data into classes so that the structurally most efficient separation can take place. That means of course, that the SVM optimizes the separation based on the model generated before but on the new samples, which leads to a classification with a specific accuracy. Through the knowledge of the labels for the training test as well, the accuracy of the SVM can be computed and through cross validation the best model generated, which will be able to generalize for completely unknown data. The SVM can also be cast as a multiclass classifier: Mostly this happens by building a set of $m = \frac{n(n-1)}{2}$ classifiers for n classes, therefore decomposing the original n-class problem in m binary classification problems, i.e. "do the samples belong to this class or any other?". Of course in this case, problems might result with unbalanced data which can be dealt with either by separating the dataset accordingly or by applying a weighting scheme on the samples which is inversely proportional to their number. Such a weighting scheme has been also applied in the thesis and will be presented in section 7.2 .

There are two important aspects of the SVM which deserve some attention at this point. The first is the non-linear nature of the SVM. As stated before, the SVM in its initial form is a maximum margin linear classifier: it tried to find the hyperplane which best separates

the data, i.e. to find the support vectors which allow the finding of a decision surface which allows a maximum distance between classes. If the dot product of the original linear function is substituted by a non-linear kernel function, the finding of a non-linear surface becomes evident. There are many choices for the kernel function, like polynomial, hyperbolic tangent and gaussian radial basis function. The latter has been used extensively the last years, as it allows a very good non-linear separation of the data.

5.3.1. Kernel Methods

As stated, the linear classification problem described in 5.1.1 suffers from the problem of non-separability of the data in the feature space. In order to overcome this difficulty and apply the SVM to non-linearely separable problems, the so-called *kernel trick* is applied [20, 78, 28]. The basic consideration is very simple: In section 5.1.1 we saw the basic condition for the linear classification problem (also presented in equation 5.4). The kernel trick consists in applyind an (arbitrary) non-linear mapping to the vectors $\mathbf{x_i}$, so that equation 5.4 becomes

$$\mathbf{y}_i(\mathbf{w}^T(\phi(\mathbf{x}_i)) + b) \geq 1 - \xi_i, \xi_i \geq 0 \tag{5.5}$$

Through the application of the kernel trick, the SVM equation 5.3 (optimization problem) becomes:

$$\underset{\mathbf{w},b,\xi}{min}\frac{1}{2}\mathbf{w}^T K(\mathbf{x}_i, \mathbf{x}_j)\mathbf{w} + C\sum_{i=1}^{l}\Sigma_i \tag{5.6}$$

This procedure allows the SVM to map the original feature space of $N_{Features}$ dimensions to a higher-dimensional space (theoretically possibly infinite), where the samples are linearly separable. That way, the handling of very non-linear problems is facilitated, albeit at the cost of a greater computational time and resources. There are many possibilities for the kernel function, including linear, polynomial, sigmoid and radial basis function. A complete listing of those methods can be found in [20, 78, 28]. In our implementation, we used the radial basis kernel function, as it has proven its efficiency in a multitude of very non-linearly separable problems [28, 12]. Its formula is given by:

$$K(x_i, x_j) = e^{-\gamma||x_i - x_j||^2}, \quad \gamma \geq 0 \tag{5.7}$$

The hyperparameters usually take values which are powers of two, with γ in the range of $\gamma = 2^{\{-11,\ldots,1\}}$ and C in the range of $C = 2^{\{-2,\ldots,7\}}$ [20, 78, 28]. The hyperparameter γ denotes the distance to which one sample influences others in the features space during classification, similar to the number of neighbors in the kNN algorithm. It thus defines how much the distributions of samples are allowed to overlap in the feature space. Very high overlap conceptually means that the samples are very interrelated with each other, whereas low overlap results in the samples not being able to influence another in the feature space. The other hyperparameter of this algorithm is the cost parameter C (see equation 5.6), which denotes the misclassification cost for the algorithm, essentially controlling the "smoothness" of the decision surface: low values for C result in a very smooth decision surface (which may however lead to many misclassifications), whereas high values

for C result in a very "jagged" decision surface, which might allow for extremely good classification performance in terms of accuracy, but could be unable to generalize to new samples. This latter condition is called *overfitting* and is a problem that can be solved either by selecting a moderate value for C (the default in most implementations is $C = 1$)

This kernel function has been used extensively and has given good results in a variety of problems [65], including very difficult ones, as in the context of automatic text classification. Its advantage consists in the form of the kernel, which makes the decision surface in the feature space smooth enough to avoid overfitting, but also complicated enough so that complex relations in the data can be represented. Except from that, its hyperparameters are tunable and therefore present a good ground for experimentation. In this thesis, we exploit exactly those advantages for the classification procedure.

5.4. Classification Performance Metrics

The evaluation of classification performance is an important subject, as it helps to evaluate the different setting and features used for classification. In most supervised classification problems, the result of the classification procedure is called a *confusion* or *contingency* matrix. The rows of this matrix correspond to the ground truth labels, whereas are the columns correspond to the labels predicted by the classification algorithm. In that sense, each field of the confusion matrix shows the amount of samples classified by the algorithm as belonging to class C_p ad which actually bear the label of class C_k. In multiclass settings, and especially in the binary SVM used in this thesis, the confusion matrix results by taking together the predictions produced by each binary classifier. For the kNN algorithm this is not necessary, since every sample is assigned a class label separately: thus, the confusion matrix results by comparing the predicted sample label with the ground truth and summing up the results. All of those schemes are usually performed in a context of *n-fold cross validation*, meaning that the whole sample is randomly separated in n subsets (or folds), each of them bearing a count of samples proportional to the original class distribution. The classification procedure is then repeated n times, and the results summed together at the end.

It follows that error-free classification leads to a matrix with non-zero elements only in its diagonal: Every sample was assigned a class label which coincides with the label of the ground truth, whereas other class combination counts are all zero. This, however, is far by being the case in most problems; some misclassifications will occur using any classification algorithm, but the question is whether those errors are meaningful, or if the overall performance of the algorithm is not very low, i.e., an amount of correct classifications has been achieved. In order to quantify such considerations, we proceed to showing some classification performance metrics which are commonly used.

The performance metrics used in this context are most commonly *accuracy*, followed by other metrics which can be extracted from a confusion matrix: *precision*, *recall* and *specificity*. Those terms are being used interchangeably in statistics proper and machine learning contexts to refer to the same metrics and formulas, although with a different meaning depending on the field of application. We will explain their significance in the machine learning area here [84]. We will use the standard terminology for confusion matrices in a two-class problem (TP, TN, FP, FN) and consider a binary classification problem where

one class is positive whereas the other negative.

accuracy Accuracy (Correct classification rate) refers to the count of correct classifications of the algorithm with respect to the total count of classifications. It is given through

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} = \frac{Correct\ Classifications}{Total\ Count\ of\ Classifications}$$

precision Precision gives the amount of true positive classifications as a percentage of all *positive* classifications

$$Pre = \frac{TP}{TP + FP} = \frac{Correct\ Positive\ Classifications}{Total\ Count\ of\ Positive\ Classifications}$$

recall Recall (sensitivity) expresses the amount of true positive classifications as a percentage of true positive and false negatives, i.e. with respect to all correct decisions.

$$Rec = \frac{TP}{TP + FN} = \frac{Correct\ Positive\ Classifications}{Correct\ Classifications\ and\ Correct\ Rejections}$$

specificity

$$Spe = \frac{TN}{TN + FP} = \frac{Correct\ Rejections}{Total\ Count\ of\ False\ Classifications}$$

In the context of this thesis we will concentrate only on accuracy for four reasons:

1. It is easy to calculate and is usable for all kinds of classifiers.
2. Accuracy is the main performance metric which shows a total measure of the classification performance without taking into account other misclassifications.
3. It is the most commonly used classification performance metric, which allows good comparability with other results.
4. The other classification metrics refer basically to a binary and not a multiclass classification problem. Thus, their use in our case is not necessarily helpful. However, accuracy, as the total amount of correct classifications (i.e. the sum of the diagonal of the confusion matrix) gives a satisfactory idea of the algorithm performance.

5.5. Feature Selection

Feature Selection is an important procedure in the context of machine learning and pattern recognition, which is characterized by the involvement of methods for dimensionality reduction. The goal of feature selection algorithms is to determine a subset of the whole feature set, which gives better or at least comparable results in the classification task. Such methods can be applied both for supervised and unsupervised classification, but the goal in both cases is the reduction of the feature count while at the same time retaining features which have good discriminative power and account for as much of the variance in the data

as possible. In the context of the thesis, only methods which take into account the ground truth (i.e., the classification labels) are considered. There are three basic methods used in the context of feature selection: *filters*, *wrappers* and *embedded* ([41, 48, 45]). We will only discuss the first two here, as their application is sufficient for the goals of the thesis. A further method which is non-mathematical in nature but has been proposed by Duda et al. [28] as a good approach and will be discussed here is feature selection based on domain knowledge.

5.5.1. Filter Methods

Filter selection methods, as the name implies, have the goal of "filtering" out irrelevant features prior to the classification process. Irrelevance in this context signifies the property of features which provide little or no information concerning with regard to classification. As such, in the best case those features do not make any contribution to the performance of the algorithm, but only increase the overall computational load in the extraction and classification phases; in the worst case, they impede the performance of the classifier, essentially representing only noise in the data. As is the case in many applications, noise is unwanted and must be eliminated since it obscures the results. Especially in machine learning, irrelevant features hamper the performance of many an algorithm [28, 12, 41], such as the kNN and SVM which are used in this thesis. As various of the features which are extracted might be irrelevant due to their sheer number or their non-specificity for rhythmic content classification, the need to apply a filter method for feature selection is warranted.

There are various examples of filter methods in the literature [41]. Out of them, the two most important are correlation with target data and mutual information with target data. The first measures the correlation of the target data with the feature for a specific amount of samples by use of the correlation coefficient. However, correlation with target data can only detect linear relationships between data, making it a less useful feature in this thesis, since it is not at all expected, that the features and the target are linearly related. The second method, mutual information with target data [13, 41] is an information-theoretic tool, measuring the similarity of the distribution of the target data and a given feature. The mutual information is between two discrete vectorial variables x, y is defined as follows [41]:

$$I(x,y) = \sum P(x,y) log \frac{P(x,y)}{P(x)P(y)} \tag{5.8}$$

The probability P is used to denote the frequency distribution of the vectors x and y across all samples. In our case, x could represent the ground truth label, where y the current feature vector. The mutual information calculation then provides a measure of the relevance of the feature for the classification with respect to the ground truth. When conducted for all features, the result is a ranking of the features based on this relevance. A standard method applied afterwards is taking the m first features as the best or to use in the next feature selection step. Although this method does not ensure that the combination of those m best features also *indeed* denotes the best feature subset for classification, it is at least sure that those features are relevant for classification. In this thesis, an implementation of the mutual information filter method from the MI-Toolbox for MATLAB [13] was

used.

5.5.2. Wrapper Methods

The goal of wrapper methods consists in using the classification method itself in order to find the best features for classification. In this sense, they "wrap" around a classifier by taking individual feature subsets and testing their performance in classification. It is very easy to see that with a moderately high amount of features, the possible subsets are extremely many. Therefore, those methods are often much more computationally intensive and slower than the filter methods. The goal of those methods is to discard redundant features, i.e., features which do not offer more information than the ones already selected and whose exclusion from the feature subset does not. In this context, wrapper methods are very often applied after filter methods in feature selection processes, in order to further reduce the feature subset and achieve a good classification performance with only a fraction of the original features. It must be also mentioned that wrapper methods, since they are based on use of the classifier for performing the feature selection, must implement a criterium which allows them to evaluate the classification. This criterium can take various forms, but since we are using accuracy for the evaluation in this thesis, the misclassification rate ($1-accuracy$) is chose as the criterium for the wrapper method. A detailed review of wrapper methods can be found in [48, 41].

An important element of wrapper methods is the way they organize the search for the best features. Since it is theoretically (but not practically) advisable to evaluate all subsets, a strategy has to be devised that can be performed in relatively short computational time. There are two methods which take that consideration into account. Sequential backward selection, which starts by considering the whole feature subset and sequentially discards feature after feature while evaluating the classification performance, stopping when the change in performance in consecutive runs is below a threshold. This method, although it is sure to discard the worst features, has two significant disadvantages: Since the whole feature set is considered or marginally smaller versions of it, the selection time stays very high. Furthermore, the end feature set is most of the times almost as large as the original set, which leads to no significant gain with respect to which features are the "very best". For those reasons, another method, the sequential forward selection is advisable: This methods begins with very small feature subsets (essentially one feature) and incrementally adds more, until the change in performance with respect lies below a threshold. This way, faster selection is possible, while at the same time the end result is a very small feature subset which, however, guarantees good classification performance: the features selected can be considered to be the most valuable. For this reason, it is selected for the experiments conducted in this thesis. Concerning the implementation, the MATLAB internal forward feature selection method function *sequentialfs* is used with a misclassification criterium as mentioned before.

5.5.3. Domain Knowledge

Applying domain knowledge for the task of feature selection is a less mathematical method, which does not take into account any information theoretical or other criteria. Its goal consists in conceptually selecting features which might be valuable for classification. Of

course, this implies either deep knowledge of the domain of the problem which is submitted to classification, or a feature design step which allows for such a selection of a feature group. The second case can however be observed very often in many audio classification studies: For example, since the feature categories to be extracted are selected on basis of their perceptual relevance (comprising, for example, timbral, rhythmical or pitch feature groups [91, 14]) and pertaining to different qualities of the audio signal, the mere testing of each of those groups separately or their combinations denotes a simple feature selection method. Of course, in this case there is no aspiration that the selected features are actually the best; but their relevance to the classification task can surely be tested on a basic level. Such methods can also help when the original count of features is so high that other, more mathematical selection methods have a prohibitive computational cost [28]. In light of those observations, we will also attempt a "conceptual" feature selection in the following, by considering different groups of the accent-based descriptors.

Part III.

Method and Implementation

6. Method

6.1. Desired Goal and Strategy

As discussed in section 2.1, the main goal of this work lies in the design and assessment of useful rhythmic content descriptors for automatic musical genre classification. Automatic musical genre classification systems suffer mainly from three problems:

- The increased ambiguity of the classification task which comes as a consequence due to ambiguous ground truth labels.
- The difficulty of defining genre rigorously because of its dependence on many different aspects of music, which necessitates a laborious feature extraction phase.
- The semantic gap between the extracted acoustic features and the perceived qualities which denote a specific musical genre.

The first problem is inherent to the task and as such, it will be taken into consideration by extending the classification to different genre datasets, in order to evaluate the algorithm performance for numerous "user cases". The second problem is addressed through the use and evaluation of pure rhythmic content descriptors, which essentially constrains the problem only to this one aspect of sound. A solution to the third problem requires the design and evaluation of highly relevant yet easily interpretable descriptors. A small step in this direction is made by defining perceptually relevant descriptors in the following sections. Based on the music theoretical models presented in chapter 3, and especially on the ones described in [53], a new realization of possible rhythmic features to be used will be presented here. Those novel features are based on accents and their acoustic correlates in the signal, and are therefore hypothesized to contain valuable information about those components of the music which help listeners "navigate" themselves in a piece in time, anticipate events and eventually extract rhythmic patterns. The features can then be organized in groups and be tested separately to assess their importance. Finally, through the application of feature selection methods, a subset of the features which can give acceptable performance will be determined. The purpose is to allow for good classification accuracy while retaining few and relevant features. Results can be relevant for even more general tasks in the field of audio processing, such as automatic language identification based on rhythmic cues.

In the following sections, the rationale behind the accent-based features and their correspondence to the actual accents will be addressed.

6.2. Definition of Accents to Be Used

From the theoretical presentation of accents and their different forms given in section 3.3, it became clear that an inclusion of all possible accentuated quantities present in the signal is not expedient, due to their sheer number, multitude of types and their referring to very abstract musical qualities (such as a melodic arch or a very specific pattern of durations), which cannot be extracted in a compact numerical form. In the thesis at hand, we focus mainly on detecting greater accent categories which can then be matched to acoustic features. Therefore, we define four main accent groups which then serve as starting points for the feature extraction, mainly in accordance with [53], but attempting also to include in that way other types of accents defined in section 3.3:

1. **Phenomenal accents:** All accents which pertain to an occasional, situational change in the signal. This applies to loudness changes (stress), timbral changes related to instrumentation or musical texture, and changes in the signal contour or introduction of new musical motives.

2. **Structural accents:** Accents belonging to this group are related to changes in the "harmonic or melodic points of gravity of a musical piece" [53], most notably changes in the fundamental frequency (essentially caused by note onsets in monophonic music), chord changes (in polyphonic music), or musical key changes.

3. **Metrical accents:** This group of accents comprises the perceived accents which coincide with the points in a signal flow where stronger beats are perceived as such by listeners, after they have established the meter of a musical piece (3.2.2).

4. **Durational accents:** Those accents concern all notes or events which have a longer duration than others in a given time period and therefore carry an added perceptual weight in comparison with shorter events.

Those four groups capture, in our view, the greater part of accentuation effects perceived in any kind of musical context. Concerning their representability through audio features, however, several considerations have to be taken into account:

- From the four accent groups, the phenomenal ones can be *directly* extracted from the signal surface or its spectral representation in an automated processing scheme. Therefore the basic part of extractable accents relate to this group.

- Structural accents are also relatively straightforward to extract. However, for their extraction the acquisition of a fundamental frequency trajectory, measures of the signal tonality or chroma features is necessary. From a technical point of view, they are also "phenomenally" present in the signal and extractable as such (see 6.3). For the sake of keeping the conceptual separation intact, we will consider them as a special category.

- Metrical accents, resulting from a purely perceptual phenomenon as meter (3.2.2), are not directly present in the signal and therefore not extractable in a manner coherent to the other accents. The only possibility to conduct this would be through establishing a metrical grid (see [81] for an application of such a method), which allows to pinpoint their exact position in time and treat them as onsets. Nevertheless, the perceptual basis of their extraction are the phenomenal accents of the signal. For that reason, this group will not be comprise a special category in itself; it will be assumed that those accents are represented by the specific subfeatures of the phenomenal accent group, such as the period and salience of the most prevalent periodicity of the signal envelope, expressed through the root mean square feature.
- Durational accents are also not directly extractable from the signal by means of novelty functions and their subfeatures, as they pertain to the duration of specific sound events, which would require again the establishment of an onset grid. Furthermore, they do not relate to specific signal qualities but to all of them, therefore their grouping in a special category does not appear feasible. However, since similar events with longer duration relate to lower periodicities, their effect will be implemented through the application of a weighting scheme on the periodicity representation (i.e., the beat histogram). In this way, durational accents are accounted for in all possible changes manifesting themselves in the signal envelope or spectrum, but do not denote an explicit feature subset.

Based on those remarks, the correspondence between the two accent groups and their respective features as well as the realization of the durational accent weighting have to be clarified. The next section deals with this subject.

6.3. Relationship Between Accents and Features

In order to be able to provide the classification algorithms with input data in a numerical form, an equivalence between accents and the corresponding features must be established first. We will address the different stages of the beat histogram extraction system (see chapter 4) in order to determine how exactly the perceptual accents can relate to audio features.

6.3.1. Novelty Functions and Subfeatures

Novelty functions As described in chapter 4, the beat histogram shows periodicities inherent in the signal, which are components of its rhythmic structure. The sum of those periodicities comprises the total rhythmic feel of the excerpt (see chapter 3). In its basic form, the beat histogram contains periodicities from the envelope of the audio signal or a filtered version thereof [91, 76, 39, 14]. However, periodicities in the signal have different origins and therefore do not all necessarily manifest themselves in the envelope of the signal. The question then arises, in what way other periodicities in the signal can be quantified through appropriate novelty functions. The various novelty functions which come into question for this task were discussed in section 4.3.2. They refer to changes either in the envelope of the signal or its spectral content. In order to account for as much temporal

variability as possible in the signal, novelty functions of the following important instantaneous features (4.2) were chosen to be extracted from the acoustic signal, referring to the aspects of the acoustic quantity that they measure:

- **Tonal Components:** Spectral Flatness (SFL), Spectral Tonal Power Ratio (STPR), Spectral Pitch Chroma Coefficients (SPC)
- **Spectral Shape Components:** Spectral Flux (SF), Spectral Centroid (SCD), Mel-Frequency Cesptral Coefficients (MFCC)
- **Envelope Components:** Root Mean Square (RMS)

In total, 30 novelty functions are extracted here, taking into account that one novelty function has to be extracted for each MFCC and SPC coefficient. It is important to refer to the rationale behind taking different novelty functions for the creation of beat histograms. The following graphic 6.3.1 shows a representation of the different temporal trajectory (i.e., its novelty function) of an envelope (Root Mean Square, RMS) and a tonal (Spectral Tonal Power Ratio, STPR), together with their respective Beat Histograms (after extraction of the ACF) in the relevant area (30-240 BPM), extracted with the same set of parameters. The audio track is a disco excerpt from the GTZAN dataset (section 8.2 for more details on the datasets), and the features are extracted over the whole length of a texture frame (3 *s*):

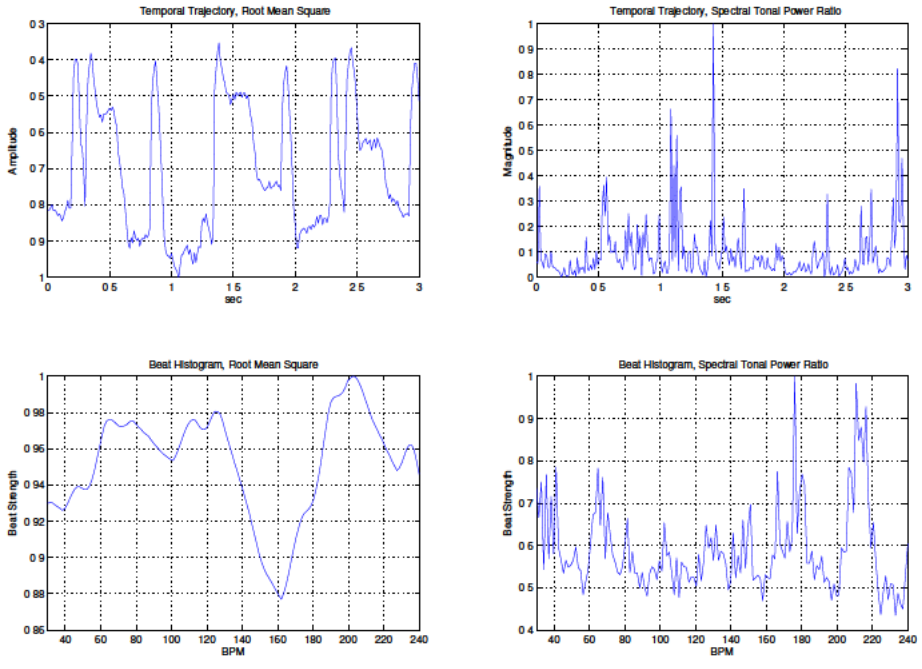

Figure 6.1.: Temporal Trajectory of Root Mean Square and Spectral Tonal Power Ratio over 3 seconds of a disco audio excerpt (upper row) and corresponding Beat Histograms from 30 to 240 BPM (lower row)

Based on this example, it becomes clear that since the changes tracked by both curves for the same audio excerpt are substantially different, the periodicity extraction leads to different novelty functions and beat histograms. Both the general form of the histogram and the strength and exact BPM value of the most prevalent periodicities can be seen to differ greatly in the two examples, giving the proof that the two novelty functions track, in fact, different rhythmic levels in the signal.

As the end amount of features (and therefore the computational time and cost for both feature extraction and classification) rises rapidly with the addition of more novelty functions (as subfeatures have to be extracted from the corresponding beat histogram) several of the novelty function detection methods and their respective novelty functions given in the literature [6, 24] had to be excluded:

- The performance of phase deviation and its derivatives on onset detection in polyphonic signals has been shown to be inferior to other methods [6]. As most of the audio tracks included in the datasets are polyphonic, their use did not seem to be warranted.
- Wavelet regularity detection methods [6] were excluded as they represent a special category of transformation which is not easily interpretable through the accent models described earlier. Apart from that, they tend to produce a very "spiky" novelty function which is not useful in creating beat histograms.
- Probability-model based novelty detection methods [6] also denote a specific processing scheme which was not conforming to feature extraction scheme of the rest of the features in our work. Therefore, despite their good performance in onset detection, they too were not taken into account.
- Other methods, such as complex domain, rectified complex domain [24] and novelty functions based on other spectral components such as, e.g., spectral-rolloff, spectral crest factor and spectral spread were also not included, as their results are expected to be the same with those of of the tonal and spectral shape components mentioned earlier and are therefore considered covered.

We acknowledge of course at this point that the amount and type of novelty functions not considered here potentially leaves much to be desired. However, taking into account the parsimoniousness principle in constructing features, as well as the added computational cost that the inclusion of those functions would bring, our decision seems to be justified. Furthermore, our work can be seen as a first step in the direction of using several novelty functions as basis for the beat histogram in musical genre classification, and therefore the potential usage and evaluation of the remaining methods can be based on the preliminary results shown here.

Subfeatures The rhythmic content features themselves in their end form (subfeatures) do not seem to offer much room for improvement. The reason is that they are basic statistical, distributional or peak-related features extracted from a beat histogram, in the same way that they can be extracted from, e.g., the magnitude spectrum of a signal. However, in order to ensure that no information contained in the beat histogram is neglected, it would

be advisable to extract as many subfeatures as possible. A complete list of the subfeatures used is given in the next chapter (7). One of the characteristics of those descriptors is that they exhibit a high degree of interrelation - e.g., the arithmetical and geometrical mean are highly correlated [52]. However, a difference in the exact form of the subfeature might have an influence on its relevance for the classification task (as, for example, some features are better scaled than others, even if they refer to the same base quantity). For that reason, the tendency is to begin with a full feature set and to eliminate the redundant or irrelevant ones through feature selection. This approach will also be followed in this thesis. However, we will refrain from using combinations or transformations of those subfeatures, for three reasons:

1. The main focus of this work lies on the novelty functions and not on the subfeatures themselves.
2. The total count of features is indeed already high enough, without having to resort to more combinations.
3. The more transformations applied before the final form of the feature, the less its interpretability at the end. As the features here result through a transformation in the first place (the beat histogram), the need to apply even more does not seem justified.

One subfeature class that was not included in our work were the MFCCs on the beat histogram, originally calculated on the periodicity histogram by Gouyon [39]. As described in section 4.2, the MFCCs are directly related to the mel-frequency transformed cepstrum of the signal. Notwithstanding the analogy of the beat histogram representation to the amplitude spectrum, it is not clear to us what the meaning of a cepstrum on the beat histogram signifies. Furthermore, the mel-frequency band division reflects psychoacoustical properties and refers to signal frequencies in the area of a 20 to 20000 Hz [97], and therefore its application on the beat histogram does not seem plausible. Although those features are reported in [39] to give very good accuracy on the BALLROOM dataset, for the reasons mentioned here we chose not to include them.

Durational accent The durational accent pertains to sound events which have a longer duration than others. It seems therefore plausible to assume, that it also refers to lower periodicities, as they are a direct indication of events with longer duration. A possibility could be the extraction of subfeatures which detect a greater low-frequency content in the beat histogram. Indeed, two subfeatures on the beat histogram, the high-frequency-content (HFC) and the centroid (CD) seem in our opinion to cover this aspect in a meaningful way. However, motivated through the application of a weighting scheme stressing the preferred tempo ([68]) on the periodicity histogram from Gouyon et al. [39], we will attempt a similar method: In our work, this type of accent is dealt with by applying a weighting scheme on the beat histogram with a logarithmic curve, ranging from the value of 2 for the lowest, to the value of 1 for the highest frequency in the beat histogram, which therefore emphasizes lower periodicities rather than higher ones. In that sense, a periodicity occurring at 30 BPM is two times as accented as one occurring at 240 BPM. Such a weighting scheme has not been applied before. However, its theoretical meaning is apparent and therefore its use seems warranted. Results of preliminary experiments have

shown that classification performance was, in total, better when using the weighted histograms than without. Therefore, we chose to include it in the end version. All further considerations will refer to beat histograms weighted as explained.

6.3.2. Correspondence Table

Having presented the range of novelty functions and subfeatures which can be implemented, a correspondence of perceptual accents and novelty functions 3.3 is given on table 6.1. The main idea behind this correspondence is that phenomenal accents can be represented through the extraction of novelty functions from instantaneous features which denote changes in the envelope and spectral shape of the signal, whereas structural accents can be represented through novelty functions resulting from instantaneous features measuring the tonality and specific pitch content of the signal.

Phenomenal	**Structural**
Spectral Flux	Spectral Pitch Chroma Coefficients (1-12)
Spectral Centroid	Spectral Tonal Power Ratio
MFCCs (1-13)	Spectral Flatness
Root Mean Square	

(6.1)

Table 6.1.: Correspondence of Accents and Novelty Functions of Signal Quantities

We are of the view that the subdivision of accents in two groups creates feature subsets which are theoretically "orthogonal" to each other on account of which and how much rhythm-related variability they explain, since they pertain to independent sources of change in the signal. Furthermore, using only one group of features greatly reduces their amount, which is a desirable condition in classification problems. The expectance is that those features can still explain a considerable amount of the rhythm content related variance in the signal and therefore produce good classification results. In oder to establish a baseline with which to compare the rhythmic content features, the instantaneous features presented in section 4.2 where also extracted in a frame-based scheme and used for classification both alone and in combination with the rhythmic content feature set.

In the next chapter, the specifics of the implementation of the musical genre classification system will be presented in some depth. It should be noted at this point, that the full implementation (feature extraction and classification), the experiments and the evaluation of the results was performed in MATLAB [63].

7. Implementation

In this chapter, an overview and the specifics of the automatic musical genre classification system implemented in the context of this thesis is given. The structure of the system follows that of a standard audio content analysis module for musical genre classification ([52]). It comprises of the following basic blocks:

- **Signal Processing:** Preprocessing of audio signal, feature extraction
- **Machine Learning:** Feature preprocessing, supervised classification, feature selection, evaluation of results

We will address each of those parts in detail, drawing on the theoretical considerations in chapters 4 and 5.

7.1. Feature Extraction Implementation

The feature extraction module is essentially a beat detection algorithm, or else a periodicity detection in the tempo area (30-240 BPM). Such systems have appeared in various flavors ([76, 47, 36, 22, 91]), involving differentiations in some processing steps or addition of new ones. The basic scheme remains the same: The signal is preprocessed (converted to mono, band-filtered, DC-freed etc.), then brought in a "reduced" form (e.g. through smoothing and undersampling) and then periodicities are detected from this version, using either comb filterbanks [76] or an autocorrelation function [91]. This transformation leads to the *beat histogram*, which gives an overview of the periodicity content of the signal in the very low frequencies (e.g., $40 - 200$ BPM) and its distribution in a specific time period, which is dictated by the texture frame length, i.e. the part of the signal which is considered as a basic unit of processing. Deviations from this scheme involve using different filterbanks, various preprocessing steps, taking the first difference of the envelope (novelty function) or even extracting more information in order to account for the phase, and not only the salience, of the periodicities [76].

The realization here is based on the system proposed in Tzanetakis [91], with some adjustments in order to account for the use of the novelty functions. Specifically, no band filtering of the signal is performed (which in [91] is applied through a Discrete Wavelet Transform), and instead of using the envelope of the signal itself, various instantaneous features extracted from a spectrogram representation are utilized (6.3.1). We will keep this terminology here and refer to the results of the analysis as beat histograms as well. The feature extraction module includes the following parts (performed for each excerpt in every dataset):

1. **Preprocessing of the audio excerpt:** Conversion to mono (by channel averaging), DC removal (through mean subtraction), normalization (so that the maximum amplitude is equal to 1), resampling to $f_s = 22050$ Hz, truncation to a maximum length of 30 seconds.

2. **Frame-based processing:** Separation of the signal in overlapping (with a factor of $N_{OVLP} = 75\%$), texture frames of an appropriate size for rhythmic feature extraction ($N_{TFR} = 3$ seconds)

3. **Spectrogram calculation:** For each frame, an STFT is performed with 75% overlap between consecutive frames, a length of $N_{AFR} = 1024$ samples for each frame (corresponding to a time duration of approximately 46 ms for the given sample rate), a hann window and an FFT resolution of $N_{FFT} = 2048$ samples.

4. **Novelty function extraction:** The temporal trajectory of the instantaneous features in table 6.1 is extracted, resulting in a novelty function for each one of them. The novelty function has a standardized length of $N_{NAFR} = N_{SPHOPS} + 1$, where N_{SPHOPS} denotes the count of hops for the spectrogram calculation.

5. **Beat histogram extraction:** Extraction of a beat histogram for each frame and for each one of the novelty functions through the calculation of an unbiased ACF, i.e. scaled with $\frac{1}{N_{NAFR}-CL}$, where $CL = \{1, 2, ..., N_{NF}\}$. This way the ACF does not exhibit its inherent tendency to decrease towards higher lags. Only half of the symmetric ACF is retained.

6. **Beat histogram averaging:** The beat histograms for all texture frames are summed and divided by their count, $N_{FRHOPS} + 1$, the latter denoting the amount of hops resulting from the texture frame-based processing.

7. **Subfeature extraction:** From the summary beat histogram of a track for a specific novelty function, the subfeatures listed in table 7.1 are extracted and stored as a partial feature vector. The concatenation of all subfeature groups for all novelty functions produces the final feature vector for an audio excerpt.

For the baseline classification system, i.e. the classification based on the feature set using instantaneous features only, the features where extracted in the same way with the scheme presented above, excluding steps 4 and 5. In order to ensure comparability in the way the end feature values for a single file are computed, step 6 is also implemented in the baseline system, with the difference that instead of the averaging of the beat histograms, the instantaneous feature values themselves are averaged over all frames.

The novelty functions used for the creation of the beat histogram and extraction of the rhythmic content features were presented in 6.1 in section 6.3.1. A list of all subfeatures for both the instantaneous and rhythmic content feature set is given in table 7.1. Concerning the scope of the instantaneous feature set, the feature extraction scheme conforms to that of the rhythmic content features. The subfeatures extracted can also be seen in table 7.1 and comprise all the distribution features, but not the peak ones since they are beat histogram related. The complete list of features for the baseline classification system can be found in table 4.1.

Distribution	Peak
mean (ME)	salience of strongest peak (A1)
standard deviation (SD)	salience of second stronger peak (A0)
mean of derivative (MD)	period of strongest peak (P1)
SD of derivative (SDD)	period of second stronger peak (P2)
skewness (SK)	period of peak centroid (P3)
kurtosis (KU)	ratio of A0 to A1 (RA)
entropy (EN)	sum (SU)
geometrical mean (GM)	sum of power (SP)
centroid (CD)	
flatness (FL)	
High Frequency Content (HFC)	

(7.1)

Table 7.1.: Subfeatures on Instantaneous Features (Distribution) and Beat Histograms of Novelty Functions (All)

After presenting the feature extraction scheme, we will now discuss four aspects of the beat histogram and their respective implementation or exclusion in this thesis:

Psychoacoustic transformation: Quite a few publications ([76, 47, 81]) have applied a psychoacoustic transformation or weighting on the sound signal before processing. This ranges from a transformation of the signal amplitude to represent loudness more closely, or applying weighting schemes. In just one publication ([39]), the psychoacoustic weighting is applied on the periodicity histogram itself, to model the effect of the preferred tempo by listeners. The advantage of such approaches is that they try to mimic human auditory perception, thus leading to possibly perceptually more relevant features. However, features derived from such transformations have not shown much better results than those extracted from "plain" versions of the signal. This leaves their use an open subject, which is perhaps advisable when the actual auditory perception has to be modeled. In our case, we follow the example of Gouyon et al. [39] and apply the low frequency weighting on the beat histogram (see 6.3.1), as this scheme is conform to the music theoretical considerations pertaining to durational accent.

Band related analysis: The separation of the signal in bands before processing is also a "classic" in the literature of feature extraction. The point is that by creating filtered versions of the signal, the spectral content of specific bands (low, middle or high) can be processed more effectively. Following that reasoning, band separation has been performed in many studies of rhythmic feature extraction ([76, 47, 91]). However, its application is not mandatory: Band separation is useful in the case of focusing on specific aspects of the signal (for example when attempting to analyze the spectral content in the low frequencies to perform instrument recognition), but brings with it some difficulties. One argument against its use in our work is that if by taking more bands of the signal as separate signals, the count of the features would be multiplied with the count of the bands. That would lead to

very a high feature count, which would be questionable with respect to their redundancy. A strategy to avoid this problem is to add the filtered envelopes and extract features from their result, as in [91]. However, since this would raise the computational cost considerably and does not bring any information which was not there in the signal to begin with, we chose to relay this aspect to future work.

Peak picking: Most beat or periodicity histogram algorithms employ one or other form of *peak picking*: the procedure refers to the automatic selection of peaks from a curve based on their relative amplitude to other points. For example, Tzanetakis takes the three strongest peaks from the beat histogram in [91] over each frame to extract the final periodicity representation. However, the application of peak picking presents some problems: Its suitable performance is very dependent on its parameters and it is often the case that peaks are not found by automatic algorithms, even if they are easily recognizable from human subjects. Therefore, we just take the beat histogram as a whole in our case, in order to avoid leaving out relevant peaks. However, for the peak-related subfeatures, some form of peak picking must be applied. For that, we used the MATLAB native peak picking function.

Periodicity extraction: One last point concerns the use of the autocorrelation function (ACF) instead of a comb filterbank for periodicity extraction. The great advantage of using the ACF to extract periodicities is its reduced complexity and very quick computation. Unfortunately, it provides information only about the periodicities and not about their phase. On the other side, the comb filters can give information about the phase of the different periodicities in the signal, which can be useful when trying to locate, e.g., the downbeat [76, 81]. Since in this work we focus on the periodicities of the novelty functions and their relationships or distribution, we adopted the ACF method.

7.2. Classification Implementation

Prior to classification, a preprocessing phase takes place. Preprocessing plays an essential role in every classification setting. The reason is that the data in their original form might be unsuitable for directs use in classification, when the value range of each feature is very different, resulting in feature spaces which are not scaled. Such unscaled features might hamper the performance of an algorithm by letting features which have a greater range or absolute value dominate the others ([28, 12]). In the case of machine learning, preprocessing methods are generally application-specific. In our case, only normalization of the features was necessary in order to ensure that they conform to the same distribution of their values. Usual schemes for that are normalization in the interval $[-1, 1]$ or $[0, 1]$ by dividing with the maximum absolute value or *z-score* normalization of a row of values, which subtracts the sample mean (therefore turning their mean to 0) and divides with the sample variance, therefore creating variance equal to 1. In this thesis, the z-score normalization is used.

At the end of this procedure, each audio excerpt has been transformed to a z-scored normalized feature vector, containing the features for each novelty function. Repeating this for all excerpts in a dataset yields a $N_{samples}$ times $N_{features}$ matrix for the classification.

The count of the samples is the total number of tracks in a dataset, whereas the count of the features is given through $N_{Features} = N_{NoveltyFunctions} \cdot N_{Subfeatures}$. The data are then submitted as input to the two supervised classification methods described in chapter 5, kNN and SVM. From this point on, the system follows the standard schema applied in classification problems:

- **Classification:** The data is partitioned in a train and a validation set and classification with a n-fold cross-validation procedure is performed. In order to ensure comparability with other publications ([39, 91, 14]) and since it is considered to be a plausible value for many reasons, we used $n_{CV} = 10$ in our implementation. The results from the 10 cross-validation runs are summed and stored as a confusion matrix. The procedure is repeated 10 times over and the end results are averaged. More information about cross validation is given in section 8.1. The labels for the classification are generated by observing the track distribution in the genres for a dataset and creating a vector of different integer values for each track in a dataset, as is the standard way for the multiclass kNN and SVM algorithm implementations in MATLAB.

- **Parameter selection:** This step involves inspecting the results and tuning the algorithm parameters (e.g., number of neighbors k for the kNN and C and γ parameters for the RBF-SVM). In total, four settings are used for the kNN and three for the SVM algorithm. The specifics of this step are given in section 8.1. An extra setting is calculated for the SVM classification for the unbalanced datasets, setting the SVM weights in such a way as to account for the sample distribution.

- **Feature selection:** Conducting feature selection, first with a filter method (Mutual Information with Target Data, using the maximum relevance CMIM metric from the MI-Toolbox [13]) and then with a sequential forward feature selection method (native in MATLAB). Additionally, the feature groups defined in table 6.1 are evaluated separately. After the feature selection, steps 1 and 2 are repeated in order to assess the performance using the reduced feature sets.

- **Evaluation:** The results of all experiments (confusion matrices) are evaluated, in order to assess the accuracy of each algorithm setting and feature group.

One last remark has to be made with respect to the handling of the unbalanced datasets. In order to overcome the problems caused by the unbalanced datasets, two strategies exist ([1]): Either to create a balanced form of each dataset, so that prior probabilities for each genre are equal; or to set the classification algorithm in such as a way as to compensate for the unbalanced samples size for each genre in a dataset. The first method, also known as random subsampling [1], has the advantages of not being specific to the classification algorithm and allowing for the creation of comparable datasets. However, it also exhibits serious drawbacks: Choosing only a part of the tracks for the genres with a strong presence in a dataset essentially means changing the dataset itself, which reduces comparability with other results. Furthermore, random subsampling can lead to datasets which produce invalid (overly optimistic or pessimistic) results. One could argue here that the genre distribution in a dataset was random to begin with (see section 8.2), so that a choice of some of the tracks could not create problems that were not already there to start with due

to the subjective nature of the dataset creation. The only way to counter this would be to create more random versions and average the results, which however would lead to an unacceptable increase in computational time. Therefore, we follow here only the second strategy: Setting up the classification algorithm in such a way as to allow for the use of unbalanced datasets. Unfortunately, this strategy is not applicable for the kNN algorithm, as its standard implementation cannot take into account unbalanced samples. For the SVM algorithm, a satisfactory strategy exists: When using the RBF kernel, it is possible to set the misclassification costs C for each class through a weighting scheme in such a way that too large or too small class sizes are compensated 5.3. This scheme essentially sets a greater misclassification cost C for the classes with few samples, so that it will be attempted to classify them correctly even at the cost of reduced generalization or a more "jagged" decision surface (see section 5.3); whereas classes with more samples are treated more leniently, allowing misclassifications in order to ensure that they do not "dominate" the whole feature set through their sheer presence. The question is at this point, what kind of weighting scheme could be applied. Although there are some examples in the literature, we follow here the method proposed in [1]: Since we have a one vs. one multiclass SVM, class weights are assigned inversely proportional to the class prior, scaled by the prior of the smallest class. That is, the smaller class receives the highest weight, equal to 1 (and therefore the highest c), whereas the largest class receives the smallest weight, equal to $\frac{prior(smallest\ class)}{prior(largest\ class)}$. Of course, this approach is also far from being perfect, as no universally acceptable weighting scheme exists.

Part IV.

Experimental Setup and Results

8. Experimental Setup

In the following sections, the experimental setup which was used for the evaluation of the rhythmic descriptors, as well as the datasets used in this work are presented and analyzed.

8.1. Setup Description

In order to test the distribution and peak-related subfeatures based on the beat histograms resulting from many different novelty functions described in subsection 4.3.2, an experimental setup has been proposed which has two goals: First, the testing of the accent-based descriptors for various datasets and parameter settings. This first step is performed in order to assess the generality of the descriptors, as the available datasets used in the task are very diverse and as the results of musical genre classification experiments are necessarily always dataset and parameter dependent. Second, the feature evaluation continues with a feature selection process which can identify which of the features are the most informative and relevant for classification. In this part, information theory based feature selection procedures (5.5) are applied in order to pinpoint the features which allow for the best performance, and conceptually selected feature subsets (6.2) are also evaluated.

Based on those considerations, the experimental setup for the genre classification task is structured as follows:

1. First, a baseline experiment is conducted, which aims at measuring the effectiveness of the instantaneous features described in section 4.2. This setting uses 231 descriptors in total, resulting from 21 instantaneous features and 11 subfeatures for each of them across a texture frame. The subfeature values are then averaged over all texture frames.

2. Second, an extended experiment is conducted with the rhythmic content descriptors and the results compared with those of the baseline experiment. This setting uses 570 descriptors in total, resulting from 30 novelty functions and 19 subfeatures for each of them, calculated on the beat histograms averaged over all texture frames.

3. Finally, a last experiment is conducted with all features pooled together, in order to assess the total classification accuracy using all available features. This experiment uses $570 + 231 = 801$ features in total.

All experiments take place with 10-fold cross-validation. n-fold cross validation randomly separates the dataset in a training and a validation part in a ratio of $n - 1$ to 1, retaining the track distribution relationships. It then repeats this procedure n times with, as much as possible, disjoint subsets of the dataset (i.e., datasets containing the minimum amount of common samples). The final results are summed over the validation folds. Furthermore, every experimental run is conducted ten times (runs), and the results over the

ten runs are averaged, so as to ensure the averaging out of chance effects as well as to provide an estimate about the statistical deviation of the experiment results across experiments. The *chance guess accuracy* [1] remains in all cases the inverse of the number of genres in the dataset (for a balanced dataset), or the genre presence in the dataset, i.e., the count of tracks belonging to one specific genre as a percentage of the total track count, also called the *prior*. In order to counter the effect of the unbalanced datasets, we used a weight vector which assigns weights for the cost parameter C in a manner inversely proportional to the prior of each genre class in the dataset. This way, the classes with very few samples have an advantage and an attempt is undertaken by the algorithm to classify them correctly, even at the cost of generalization, whereas classes with many samples are more easily misclassified. This selection is one of the many possibilities listed in relevant literature [1, 15], but we have refrained from using others to reduce the computational cost and because there exists no standard scheme for the selection of the weight vectors.

For the experiments, four parameterizations of the kNN algorithm are used ($k = 1, 3, 5, 7$, euclidean distance metric in all cases) and four parameterizations of the RBF-SVM algorithm: unoptimized ($C = 1, \gamma = 1/N_{Features}$), optimized for C and γ separately and optimized through Grid Search. The last case is performed with and without a weight vector for the cost parameter in case of the unbalanced datasets. *Grid search* is the name of the procedure of trying to find the best set of hyperparameters C and γ for the RBF-SVM algorithm. It is performed in a context also of n-fold cross validation and conducts an exhaustive search of the parameter space (i.e., the values of C and γ) in order to find the set which provides the best accuracy. This procedure is computationally costly, but ensures that the classification accuracy of the SVM is the best possible. In the experiments listed here, it has been performed once for each experimental setting where the dataset considered or the number of features changed, since the choice of best C and γ depends heavily on the dimensionality of the feature space and the specific data samples to be classified [78, 20]. The grid values used where: $\{2^{-1}, 2^{1}, ..., 2^{5}\}$ for C and $\{2^{-11}, 2^{-9}, ..., 2^{-1}\}$ for γ, which are values suggested by the researchers who implemented the libSVM algorithm [15] and results in a total of 24 cases. The goal of this first baseline setup is threefold:

- Determining the performance of the full rhythmic content descriptor set and the combined set in comparison to the baseline set of the instantaneous features.
- The assessment of the best mode for the classifier parameters and the comparison of the performance of the two classifiers involved.
- A first estimation of the dataset differences in performance and an assessment of performance in the case of unbalanced datasets.

After performing those basic experiments, feature selection methods are applied for reasons explained in section 5.5. The explicit goal of the thesis is to find a small set of descriptors which give comparable accuracy to the feature set, so as to determine which novelty functions and subfeatures are relevant for musical genre classification. Furthermore, conceptually selected descriptors, applying prior knowledge of the problem which led to their design are also evaluated. With those considerations, the classification setup with feature selection for the full rhythmic content feature set is structured in the following way:

[1] chance guess accuracy denotes the accuracy a classifier would exhibit by assigning samples a random label.

1. Feature selection based on the criterion of mutual information with target data is applied, resulting in a relevance ranking of all 570 rhythmic content features. Out of those, the best 20 are selected and classification performed with the best parameters for the kNN and SVM algorithms with them. The number of 20 features is selected as a very suitable value for fast classification, while still allowing the identification of a small set of relevant features. Preliminary tests showed that classification with more features from the mutual information ranking $(50 - 100)$ did not improve the results dramatically with respect to the 20 features set.

2. A sequential forward feature selection with the optimized RBF-SVM algorithm as wrapper is applied to the 20 rhythmic descriptors which result from the mutual information feature selection process. After that, classification takes place once more in order to evaluate the performance of the best, final features. The results are to be compared with all the others in order to assess the performance of the heavily reduced feature set in comparison with the full feature set.

3. The full rhythmic content descriptor set is separated in the phenomenal and the structural feature subset 6.2. Each of them is evaluated in order to assess the importance of the perceptual feature subsets.

In all cases, the measure reported is the mean average accuracy rate computed from the output of the multi-class classification in the form of confusion matrices. In multiclass classification, the overall accuracy is given by the trace of the confusion matrix divided by the total number of classifications. We refer to it here as mean average accuracy, since it is given over 10 validation folds and 10 experimental runs.

8.2. Dataset Description

In order to ensure comparability of the results with other publications and to properly evaluate the rhythmic content descriptors for different genre hierarchies and tracks, five datasets were used in this thesis. These are given here, together with their naming convention used fortwith:

- The Tzanetakis Dataset [91] (GTZAN)
- The Ballroom Dataset [39, 26] (BALLROOM)
- The Homburg Dataset [43] (HOMBURG)
- The ISMIR04 Dataset [7] (ISMIR04)
- The Unique Dataset [83] (UNIQUE)

Specific information on each dataset can be found in the cited publications. A brief but comprehensive description of their content (included genres, total number of tracks and distribution, length and file format) is given in Appendix B[2]. It suffices to say at this point,

[2]We will exclude information about the exact tracks in the datasets, as it requires great space and it is not the focus of the present work. However, detailed information or links to it can be found in the publication associated with each dataset.

that none of those datasets are "perfect", in the sense that there is no claim of completeness concerning the scope of musical material that they comprise, and that the ground truth label assignment corresponds to the available data and opinions of the researchers who collected and arranged them. However, since they have been used extensively up to date, we will also resort to using them as well, with the caveat that their suitability for the musical genre classification task is a matter of scientific discourse. There are, however, current attempts to create standardized datasets for specific tasks in MIR, and especially for musical genre classification, by the ISMIR community.

Another subject concerns the distribution of the tracks in the genres of the datasets. Four out of five datasets are unbalanced (i.e., they do not contain the same amount of excerpts for each genre), which can present a challenge for the classification algorithms [28, 12]. However, each one has been compiled by one or more audio experts (but no musicologists), so that there is relative safety in assuming that they bear at least ecological validity. Even more importantly, most of the datasets have already been tested and evaluated numerous times on a multitude of MIR tasks (including contests), which confirms their relevance for related research and contributes to the comparability of the results. We will give here a short description of the content of each dataset:

- **GTZAN:** A balanced dataset with a medium high track count and many different but well-known genres (100 tracks per genre, 10 genres, 1000 tracks in total). It is a standard in the area of musical genre classification, although its validity has been challenged [87, 88].
- **BALLROOM:** An unbalanced dataset, but almost uniformly distributed (circa 80 tracks per genre, 8 genres, 698 tracks in total). Included genres refer exclusively to dance music, with a very strong rhythmic content and pronounced beat, thus extremely helpful for rhythmic genre classification.
- **ISMIR04:** A relatively small, quasi-unbalanced dataset (circa 100 tracks per genre with some deviations (the *pop-rock* genre comprises almost half of the tracks, whereas *jazz* has too few), 6 genres, 729 tracks in total). Genres included are very basic or compound.
- **UNIQUE:** A very large, unbalanced dataset (ranging from 50 to 700 tracks per genre, 14 genres, 3115 tracks in total). Genres are very diverse (e.g., *world* music and *speech* are included, which is not the case by any other dataset), but many have too few examples relative to others.
- **HOMBURG:** A larger, but also unbalanced dataset (from 50 to 300 tracks per genre, 9 genres, 1886 tracks in total). Included genres are similar to those in the GTZAN dataset. This dataset has not been used extensively, but its selection of tracks appears to be ecologically valid.

Having described the experimental setup and the datasets used in them, the experiment results are presented in the next session.

9. Results

Results of the experiments are shown in tables 9.1-9.12. For each dataset and experimental setting a confusion matrix is the result of the classification procedure, which shows the correct classification examples and the misclassified ones. Due to considerations of space, the confusion matrices of only the best experimental setting (optimized SVM, referred to as SVM-Best) and the full rhythmic content feature set are presented in Appendix A. Concerning classification, we report here on the mean average classification accuracy (percentage of correct classifications over all genres with respect to all classifications in each experiment for every class) along with its standard deviation across the experimental runs and the weighted average chance guess accuracy of the classifiers. In the captions, the algorithm settings and the feature subset used are also provided. Finally, tables comprising the results of classification after the feature selection process are presented here. The results are discussed in depth in the next chapter. The best results in each table are marked by bold fonts. We give here again a brief account of the structure of the tables, in order to help the reader navigate more efficiently.

- The columns of the result tables correspond to the datasets, which bear the abbreviations introduced in 8.2.
- The rows of the result tables correspond to the experimental settings. $k-NN$ denotes the number of neighbors used in the kNN algorithm scheme. For the RBF-SVM, *Basic* denotes the setting with $C = 1, \gamma = 1/N_{Features}$, $Opt - C$ and $Opt - \gamma$ denote the separate optimization for the hyperparameters, where a value of $C = 4, \gamma = 1/N_{Features}$ has been used in the first case and a value of $C = 1, \gamma = 1/2^k$, whereas 2^k is the next closer value for γ as an inverse power of two that is smaller than $\gamma = 1/N_{Features}$. *Best* denote the setting where the values of C and γ have been optimized through grid search, and $B\&W$ denotes the optimized case with the application of the weight vector for the cost parameter C.
- The last row of each table gives the chance guess accuracy of the algorithm.
- Each field in a table contains the mean average accuracy as percentage of correct classifications to all classifications, followed by the standard deviation over the experiment runs in parentheses.
- Further information about the parameterization and feature group reported in each table is listed in the captions.

9.1. Classification Prior to Feature Selection

In the next tables, the results presented are produced for all basic feature groups (instantaneous (baseline), rhythmic content feature group and combined) with various settings for

the two algorithms.

S/D	GTZAN	BALLROOM	ISMIR04	UNIQUE	HOMBURG
1-NN	$53,7\%$ $(1,1)$	$40,7\%$ $(1,1)$	$55,1\%$ $(0,7)$	$47,3\%$ $(0,5)$	$30,0\%$ $(0,7)$
3-NN	$52,4\%$ $(1,0)$	$43,3\%$ $(1,4)$	$56,9\%$ $(0,8)$	$50,2\%$ $(0,4)$	$32,5\%$ $(0,8)$
5-NN	$55,5\%$ $(1,1)$	$46,3\%$ $(1,4)$	$57,5\%$ $(1,3)$	$54,4\%$ $(0,4)$	$36,2\%$ $(0,4)$
7-NN	$56,4\%$ $(1,0)$	$47,7\%$ $(0,6)$	$\mathbf{57,8\%}$ $(\mathbf{0,8})$	$55,9\%$ $(0,4)$	$38,2\%$ $(0,6)$
Chance	$10,0\%$	$12,5\%$	$16,7\%$	$7,1\%$	$11,1\%$

(9.1)

Table 9.1.: Classification results (mean average accuracy and standard deviation over 10 runs), all datasets, kNN, $k = 1/3/5/7$, euclidean distance, instantaneous features full dataset (S=Setting, D=Dataset)

S/D	GTZAN	BALLROOM	ISMIR04	UNIQUE	HOMBURG
1-NN	$41,2\%$ $(0,4)$	$36,2\%$ $(1,4)$	$61,1\%$ $(1,1)$	$52,5\%$ $(0,4)$	$32,8\%$ $(0,4)$
3-NN	$39,8\%$ $(0,9)$	$38,9\%$ $(1,5)$	$61,5\%$ $(0,8)$	$55,2\%$ $(0,6)$	$35,4\%$ $(0,5)$
5-NN	$42,6\%$ $(0,9)$	$44,3\%$ $(1,2)$	$61,2\%$ $(0,8)$	$58,5\%$ $(0,4)$	$38,7\%$ $(0,6)$
7-NN	$43,5\%$ $(0,8)$	$45,2\%$ $(1,1)$	$\mathbf{61,2\%}$ $(\mathbf{0,7})$	$60,1\%$ $(0,6)$	$40,3\%$ $(0,8)$
Chance	$10,0\%$	$12,5\%$	$16,7\%$	$7,1\%$	$11,1\%$

(9.2)

Table 9.2.: Classification results (mean average accuracy and standard deviation over 10 runs), all datasets, kNN, $k = 1/3/5/7$, euclidean distance, rhythmic content features full dataset (S=Setting, D=Dataset)

S/D	GTZAN	BALLROOM	ISMIR04	UNIQUE	HOMBURG
1-NN	$53,6\%$ $(0,6)$	$45,4\%$ $(0,9)$	$61,2\%$ $(1,10)$	$54,9\%$ $(0,2)$	$34,2\%$ $(0,6)$
3-NN	$53,3\%$ $(0,5)$	$47,0\%$ $(1,7)$	$61,5\%$ $(1,0)$	$57,8\%$ $(0,7)$	$36,1\%$ $(0,6)$
5-NN	$55,4\%$ $(0,7)$	$51,0\%$ $(1,7)$	$62,1\%$ $(0,6)$	$61,2\%$ $(0,3)$	$39,5\%$ $(0,4)$
7-NN	$56,1\%$ $(0,8)$	$53,3\%$ $(1,8)$	$61,5\%$ $(0,6)$	$\mathbf{62,3\%}$ $(\mathbf{0,3})$	$40,9\%$ $(0,4)$
Chance	$10,0\%$	$12,5\%$	$16,7\%$	$7,1\%$	$11,1\%$

(9.3)

Table 9.3.: Classification results (mean average accuracy and standard deviation over 10 runs), all datasets, kNN, $k = 1/3/5/7$, euclidean distance, all features combined (S=Setting, D=Dataset)

S/D	GTZAN	BALLROOM	ISMIR04	UNIQUE	HOMBURG
Basic	$65,4\%\ (0,6)$	$57,0\%\ (0,8)$	$66,9\%\ (0,9)$	$66,4\%\ (0,2)$	$48,5\%\ (0,4)$
Opt. c	$66,7\%\ (0,9)$	$57,5\%\ (0,7)$	$\mathbf{70,0\%\ (0,7)}$	$66,8\%\ (0,3)$	$47,8\%\ (0,6)$
Opt. γ	$61,5\%\ (0,7)$	$53,0\%\ (0,7)$	$61,1\%\ (0,4)$	$65,6\%\ (0,2)$	$47,6\%\ (0,2)$
Best	$66,6\%\ (0,7)$	$58,4\%\ (1,5)$	$69,8\%\ (0,7)$	$64,3\%\ (0,2)$	$49,3\%\ (0,5)$
B & W	N/A	$57,9\%\ (0,7)$	$68,0\%\ (0,6)$	$49,4\%\ (0,1)$	$41,2\%\ (0,5)$
Chance	$10,0\%$	$12,5\%$	$16,7\%$	$7,1\%$	$11,1\%$

(9.4)

Table 9.4.: Classification results (mean average accuracy and standard deviation over 10 runs), all datasets, RBF-SVM, different parameter settings, instantaneous features full dataset (S=Setting, D=Dataset)

S/D	GTZAN	BALLROOM	ISMIR04	UNIQUE	HOMBURG
Basic	$58,2\%\ (0,7)$	$59,1\%\ (0,8)$	$62,5\%\ (0,7)$	$64,1\%\ (0,3)$	$44,4\%\ (0,3)$
Opt. c	$58,7\%\ (1,0)$	$59,6\%\ (0,8)$	$\mathbf{66,2\%\ (0,8)}$	$65,6\%\ (0,3)$	$44,2\%\ (0,4)$
Opt. γ	$57,1\%\ (0,8)$	$59,7\%\ (0,7)$	$61,3\%\ (0,6)$	$64,3\%\ (0,4)$	$44,5\%\ (0,3)$
Best	$59,2\%\ (0,9)$	$60,4\%\ (0,9)$	$65,9\%\ (0,7)$	$64,6\%\ (0,2)$	$44,3\%\ (0,4)$
B & W	N/A	$57,9\%\ (0,7)$	$53,2\%\ (1,0)$	$51,5\%\ (0,8)$	$38,8\%\ (0,3)$
Chance	$10,0\%$	$12,5\%$	$16,7\%$	$7,1\%$	$11,1\%$

(9.5)

Table 9.5.: Classification results (mean average accuracy and standard deviation over 10 runs), all datasets, RBF-SVM, different parameter settings, rhythmic content features full dataset (S=Setting, D=Dataset)

S/D	GTZAN	BALLROOM	ISMIR04	UNIQUE	HOMBURG
Basic	$68,2\%\ (0,4)$	$64,0\%\ (1,0)$	$67,0\%\ (0,8)$	$68,7\%\ (0,1)$	$48,1\%\ (0,3)$
Opt. c	$69,3\%\ (0,6)$	$65,5\%\ (1,2)$	$69,2\%\ (0,7)$	$69,9\%\ (0,2)$	$48,9\%\ (0,4)$
Opt. γ	$68,1\%\ (0,5)$	$64,2\%\ (0,7)$	$65,5\%\ (0,6)$	$68,9\%\ (0,2)$	$48,2\%\ (0,5)$
Best	$59,2\%\ (0,9)$	$65,8\%\ (1,0)$	$68,6\%\ (0,5)$	$\mathbf{70,4\%\ (0,1)}$	$48,1\%\ (0,3)$
B & W	N/A	$69,4\%\ (0,8)$	$56,7\%\ (0,4)$	$58,1\%\ (0,3)$	$43,5\%\ (0,5)$
Chance	$10,0\%$	$12,5\%$	$16,7\%$	$7,1\%$	$11,1\%$

(9.6)

Table 9.6.: Classification results (mean average accuracy and standard deviation over 10 runs), all datasets, RBF-SVM, different parameter settings, all features combined (S=Setting, D=Dataset)

9.2. Classification After Feature Selection

In this section, results from the feature selection methods concerning the feature rankings as well as their classification performance are given. They are shown in three steps: Results after mutual information feature selection, after sequential forward selection and after the accent group based feature selection.

9.2.1. Classification After Mutual Information Feature Selection

In the next tables, the results presented are produced after the mutual information feature selection for the rhythmic content feature group with the basic and best performance settings for the two algorithms. For details on the count and final features produced in each case, see table 9.8.

S/D	GTZAN	BALLROOM	ISMIR04	UNIQUE	HOMBURG
1-NN	$40,0\%$ $(0,8)$	$45,3\%$ $(1,1)$	$54,5\%$ $(1,3)$	$48,9\%$ $(0,5)$	$28,1\%$ $(0,5)$
7-NN	$44,0\%$ $(0,9)$	$51,9\%$ $(1,0)$	$60,5\%$ $(0,8)$	$58,3\%$ $(0,3)$	$36,8\%$ $(0,6)$
SVM-Basic	$51,3\%$ $(0,5)$	$58,1\%$ $(0,7)$	$61,6\%$ $(0,7)$	$62,5\%$ $(0,3)$	$41,7\%$ $(0,2)$
SVM-Best	$51,8\%$ $(0,6)$	$57,9\%$ $(0,9)$	$\mathbf{62,8\%}$ $\mathbf{(0,6)}$	$62,5\%$ $(0,2)$	$42,5\%$ $(0,3)$
Chance	$10,0\%$	$12,5\%$	$16,7\%$	$7,1\%$	$11,1\%$

(9.7)

Table 9.7.: Classification results (mean average accuracy and standard deviation over 10 runs), all datasets, kNN, $k = 1/7$, euclidean distance, SVM basic and optimized, 20 best features from rhythmic content full dataset after CMIM ranking (S=Setting, D=Dataset)

9.2.2. Classification After Mutual Information and Sequential Forward Feature Selection

In the next tables, the results presented are produced after the whole feature selection with the SVM algorithm as wrapper and with the best performance settings. For details on the count and final features produced in each case, see table 9.10.

S/D	GTZAN	BALLROOM	ISMIR04	UNIQUE	HOMBURG
SVM-Best	$49,4\%$ $(0,7)$	$56,5\%$ $(0,8)$	$\mathbf{63,6\%}$ $\mathbf{(1,3)}$	$61,4\%$ $(0,2)$	$42,4\%$ $(0,42)$
Chance	$10,0\%$	$12,5\%$	$16,7\%$	$7,1\%$	$11,1\%$

(9.9)

Table 9.9.: Classification results (mean average accuracy and standard deviation over 10 runs), all datasets, optimized SVM, final best features from rhythmic content full dataset after CMIM ranking and sequential forward feature selection (S=Setting, D=Dataset)

R/D	**GTZAN**	**BALLROOM**	**ISMIR04**	**UNIQUE**	**HOMBURG**
1	MD_RMS	P1_SF	MD_MFCC2	SD_MFCC1	SD_RMS
2	FL_MFCC1	P2_RMS	CD_MFCC1	GM_SFL	SD_SPC3
3	GM_SFL	A0_SFL	P3_SPC8	MD_MFCC2	FL_SFL
4	SD_MFCC2	CD_SPC12	A0_SF	SP_SCD	SD_SPC1
5	EN_MFCC4	P2_SFL	P3_SPC12	SD_SPC7	GM_SCD
6	CD_SPC1	SD_SPC3	HFC_SPC7	SD_MFCC2	EN_MFCC2
7	HFC_MFCC11	P1_SFL	KU_MFCC1	P3_SPC2	EN_SPC7
8	SDD_RMS	P2_SF	P3_SPC11	FL_RMS	KU_SFL
9	SD_MFCC13	HFC_SCD	SK_SPC1	P3_STPR	SU_SFL
10	SD_SPC9	P3_SPC2	SD_SCD	P1_SF	SD_SPC7
11	SD_SPC12	P3_SPC7	SD_MFCC2	A0_SF	SD_MFCC2
12	KU_MFCC1	CD_SPC10	P3_RMS	KU_MFCC1	SD_SPC5
13	P3_SPC7	SD_MFCC2	A0_SPC6	CD_SPC11	GM_MFCC11
14	SD_SPC7	SD_MFCC1	P3_SFL	MD_RMS	SP_MFCC3
15	SK_SF	P1_STPR	SK_SPC9	P1_SFL	A0_RMS
16	SD_MFCC4	CD_SPC5	P3_SPC2	SD_SPC1	SD_MFCC3
17	P3_SPC4	SDD_RMS	P3_SPC4	HFC_SPC1	SU_MFCC4
18	CD_SPC12	RA_SF	SD_SCD	P3_SPC5	P3_STPR
19	FL_SPC8	P3_STPR	A0_SPC7	CD_SPC8	P3_SPC11
20	SD_MFCC11	SD_SPC9	P3_SPC1	SD_SPC12	CD_SPC10

(9.8)

Table 9.8.: Best 20 rhythmic content features per dataset after selection through mutual Information with target data (CMIM Method) (R=Ranking,D=Dataset)

R/D	GTZAN	BALLROOM	ISMIR04	UNIQUE	HOMBURG
1	MD_RMS	P1_SF	MD_MFCC2	SD_MFCC1	SD_RMS
2	FL_RMS	A0_SFL	CD_MFCC1	GM_SFL	SD_SPC3
3	GM_SFL	SD_SPC3	A0_SF	MD_MFCC2	FL_SFL
4	SD_MFCC2	HFC_SCD	P3_SPC12	SP_SCD	EN_SPC7
5	EN_MFCC4	P3_SPC2	HFC_SPC7	SD_SPC7	SD_SPC7
6	CD_SPC1	P3_SPC7	P3_SPC11	SD_MFCC2	GM_MFCC11
7	HFC_MFCC11	SD_MFCC4	SK_SPC1	FL_RMS	SP_MFCC3
8	SD_SPC9	P1_STPR	HFC_SCD	P3_STPR	SU_MFCC34
9	KU_MFCC1	SD_RMS	SD_RMS	P1_SF	P3_SPC11
10	P3_SPC4	P3_STPR	SK_SPC9	A0_SF	
11	FL_SPC8		P3_SPC4	KU_MFCC1	
12				CD_SPC11	
13				P1_ SFL	
14				SD_SPC1	
15				SD_SPC12	

(9.10)

Table 9.10.: Best final rhythmic content features per dataset after selection through sequential forward method with SVM wrapper on first 20 features after CMIM ranking (R=Ranking,D=Dataset)

9.2.3. Classification after Feature Selection by Accent Groups

In the next tables, the results presented are produced after the accent group feature selection from the rhythmic content feature set with the best settings for the kNN and RBF SVM algorithms. For details on the novelty functions for the features used in each case, see table 6.1 in section 6.3.2. The subfeatures used are all 19 shown in table 7.1 in section 7.2.

S/D	GTZAN	BALLROOM	ISMIR04	UNIQUE	HOMBURG
7-NN	$40,6\%$ $(0,8)$	$37,1\%$ $(0,9)$	$54,4\%$ $(1,2)$	$53,6\%$ $(0,4)$	$36,6\%$ $(0,8)$
SVM-Best	$52,7\%$ $(0,5)$	$46,4\%$ $(1,4)$	$60,8\%$ $(0,6)$	$\mathbf{61,3\%}$ $\mathbf{(0,4)}$	$41,8\%$ $(0,4)$
Chance	$10,0\%$	$12,5\%$	$16,7\%$	$7,1\%$	$11,1\%$

(9.11)

Table 9.11.: Classification results (mean average accuracy and standard deviation over 10 runs), all datasets, best kNN, $k = 7$, euclidean distance, optimized SVM, phenomenal accent feature group (S=Setting, D=Dataset)

S/D	**GTZAN**	**BALLROOM**	**ISMIR04**	**UNIQUE**	**HOMBURG**
7-NN	$31,2\%$ $(1,0)$	$35,7\%$ $(1,0)$	$51,5\%$ $(1,0)$	$50,2\%$ $(0,2)$	$33,9\%$ $(0,8)$
SVM-Best	$46,0\%$ $(0,8)$	$51,9\%$ $(0,9)$	$57,9\%$ $(0,4)$	$\mathbf{58,5\%}$ $(\mathbf{0,4})$	$41,6\%$ $(0,5)$
Chance	$10,0\%$	$12,5\%$	$16,7\%$	$7,1\%$	$11,1\%$

(9.12)

Table 9.12.: Classification results (mean average accuracy and standard deviation over 10 runs), all datasets, best kNN, $k = 7$, euclidean distance, optimized SVM, structural accent feature group (S=Setting, D=Dataset)

Part V.

Discussion and Outlook

10. Discussion

In this chapter, the results presented in the previous chapter will be interpreted and discussed in some depth and a conclusion is reached about their relevance.

The results in general present a mixed picture. It is clear that the rhythmic features alone cannot account for all the variance present in the datasets with respect to the musical genre classification task. It should be noted again here, however, that genre classification is not a well-defined problem in any case: The ground truth genre labels do not reflect (only) a physical reality, but rather a **subjective** categorization based on acoustic, musical and social criteria. Therefore, it cannot be expected that the extracted low-level features can separate the genres *perfectly*: it is much more an attempt to define up to which extent the track of a dataset are separable in given genres through such features. That is also the case for the baseline instantaneous features, which serve as a compact representation of other aspects pertaining to the audio sample, in the hope of being able to capture as much variance present in the audio samples as possible, which could be important for the widely defined task of musical genre classification. Given this as a fact, the objective shifts towards the determination of the performance of the feature subsets and the best features resulting from the feature selection process, as well as towards the study of the misclassified examples in order to draw conclusions about the relatedness (especially with respect to their rhythmic content) of specific genres.

Hereafter, specific aspects of the feature performance in the musical genre classification task will be discussed.

10.1. Performance of Basic Classification

The results of the baseline experiment with the full set of the instantaneous features, as well the rhythmic content features full set and the combined feature set experiments for both classification methods and all datasets are given in tables 9.1-9.6. We will refer to them in the following sections and compare the performance with respect to datasets, classification methods, parameterization of the algorithms. Finally a total comparison of the feature set performance will be given.

10.1.1. Baseline

In the baseline experiment, using the instantaneous feature set alone, the range of mean average accuracy achieved for each dataset is 30% for the HOMBURG dataset with a 1-NN classifier (table 9.1) and 70% for the ISMIR04 dataset with the C-optimized RBF-SVM classifier (table 9.4). Those results lie in any case above the average prior for each dataset and are indicative of the very acceptable classification performance that can be achieved by using the instantaneous feature set. We will compare the results regarding the three aspects mentioned in the beginning of this section.

Datasets For the baseline experiment, the best classification results are achieved for the ISMIR04 dataset, followed by the GTZAN and UNIQUE dataset, whereas the BALLROOM and HOMBURG datasets lie at the end of the scale. This tendency remains the same independent of classification parameters and method. This can be interpreted as an artifact pertaining to the track distribution in the datasets (which in case of the HOMBURG dataset is very unbalanced) and the specific selection of tracks therein, indicating either that the ground truth could be reexamined or that the collection is too diverse or onesided to allow good classification. In total, the results lie close to the ones reported in the corresponding publications ([91, 39, 43, 7, 83, 62]) but in most cases marginally lower. However, this can be interpreted as the result of using different features and classifiers for their experiments than the ones used here.

Classification methods The results with respect to the classification methods are unambiguous: the RBF-SVM has better performance than the kNN across all settings and datasets. It can be seen as a matter of fact, that the results of the basic RBF-SVM version are better than the best version of the kNN (7-NN), at a rate of ca. 10%. That is a clear indication that the application of SVMs for the musical genre classification problem gives better results than very basic methods such as the kNN, which was an expected outcome.

Parameterization Regarding the parameterization, the kNN algorithm shows clearly better results for the versions with a higher k for all datasets, with a steady increase in accuracy from 1-NN to 7-NN in the range of 2 to 8%. For the SVM, the same remark holds true for most cases, although the optimization for C alone gives generally the best results, comparable to the ones obtained from grid search, and the optimization for γ alone gives generally worse results than even the basic method. The use of a weight vector for C with the optimized version generally lowers the mean average accuracy, but brings a gain in the correct classification of the classes which have few samples.

10.1.2. Rhythmic Content Features

With use of the full rhythmic content feature set alone, a mean average classification accuracy ranging from $32,8\%$ for the HOMBURG dataset to $66,2\%$ for the UNIQUE dataset is achieved. As such, those results lie close to the ones given by the baseline experiment, albeit marginally lower. This difference lies in the area of 5 to 10% for the kNN versions, and between 3 to 8% for the various SVM parameterizations. This is a good indication that the rhythmic content features are indeed very informative and useful in classifying musical genres. The discrepancy between the baseline and the rhythmic feature set alone can be explained by the fact that the rhythmic descriptors capture only one aspect of genre dependency (i.e., its rhythmic content), whereas the instantaneous descriptors somehow encode information about the general content of the music, thus leading to better classification. This outcome, however, was expected; but the performance of the rhythmic content feature set remains very satisfactory. The partial aspects are explained in the following.

Datasets The best classification results for the rhythmic content feature set are also achieved for the ISMIR04 dataset, followed closely by UNIQUE. An important difference to the

baseline experiment can be observed here: The BALLROOM dataset achieves very comparable results to those of the GTZAN one in all the RBF-SVM and the best kNN settings, which is an indication that the rhythmic content features are very useful in the case of a highly rhythmically relevant dataset such as BALLROOM (8.2). As before, the HOMBURG dataset lies far behind the others in all cases, but still considerably above the prior. We will examine the performance results for the rhythmic content descriptors for each dataset here in more detail and compare the results to relevant publications:

GTZAN The results presented here are better than the ones given by Tzanetakis in [91] for the GTZAN dataset using only rhythmic content features and a single gaussian classifier (28%). They actually are indeed very close to the results using his whole feature set ($60,1\%$), at least when considering the RBF-SVM classifier. One must note, however, that he uses only 6 rhythmic descriptors on a beat histogram (corresponding to the first 6 of our peak-related subfeatures in this work, see table 4.3), resulting in a very frugal feature set in comparison to our experiments.

BALLROOM For the BALLROOM dataset, results using only rhythmic descriptors (as Gouyon in [39] did, achieving $52,8\%$ accuracy with the periodicity histogram descriptors and a 1-NN classifier) are comparable in the case of the RBF-SVM setting. However, we achieved only $36,2\%$ using a 1-NN classifier, a difference which could be traced back to the specific implementation of the periodicity histogram features (comparable to our beat histogram features) in his case. His best results using IOI-MFCC features and correct tempo ($90,1\%$) could not be reproduced here.

ISMIR04 Marques [62] uses a novel method for feature extraction based on a codebook representation and achieves $75,0\%$ accuracy using an SVM classifier with very high C and $83,0\%$ with a Markov model. Since the method is not truly comparable to ours, his result must be taken with care. However, our $55,1\%$ can be interpreted as a satisfactory result for rhythmic content feature extraction alone.

HOMBURG The accuracies achieved for the HOMBURG dataset for the RBF-SVM setting ($44,5\%$) are comparable to those achieved with the use of a Naive Bayes classifier in [43] ($43,7\%$), but lie lower to the ones attained with a kNN classifier with adaptive distance metric ($53,2\%$).

UNIQUE For the UNIQUE dataset, Seyerlehner [82] reports $75,9\%$ accuracy, though it is unclear how exactly this value is achieved. It lies, however close to our own value of $66,8\%$ using only rhythmic content features, which can be interpreted as an encouraging result.

In table 10.1, the best results from our rhythmic content feature set (for the RBF-SVM method) are listed together with those of relevant publications in order to allow a direct comparison. It should be noted, however, that the comparison must be performed with care: Not all related publications have used only rhythmic descriptors for the classification

task and classification methods vary widely. Therefore, in the last rows, information about the classification method and the feature set used is given as well.

	GTZAN	BALLROOM	ISMIR04	UNIQUE	HOMBURG
ARCF	$59,2\%$	$60,4\%$	$55,1\%$	$66,8\%$	$44,5\%$
Other	$28,0\%$ [91]	$52,8\%$ [39]	$75,0\%$ [62]	$75,9\%$ [82]	$53,2\%$ [43]
Features	RF	RF	OF	OF	OF
Classifier	GS	k-NN	SVM	SIM	k-NN
Chance	$10,0\%$	$12,5\%$	$16,7\%$	$7,1\%$	$11,1\%$

(10.1)

Table 10.1.: Best classification results comparison between results achieved here (ARCF) and other approaches (Other) (RF=Rhythmic Features, OF=Other Features, possibly comprising RF, GS=Gaussian Single Classifier, SIM=Similarity)

Classification methods The results with respect to the classification methods conform to those in the baseline experiment: the RBF-SVM method shows better performance than the kNN across all settings and datasets in this case as well. The difference between attained accuracies is comparable to the ones in the baseline experiment.

Parameterization The results of the parameterization also conform to those in the baseline experiment. The kNN algorithm shows better performance for higher k and the grid search results for C and γ give the parameterization which provides the best results for the SVM. However, the differences are smaller than for the instantaneous feature set.

10.1.3. Combined Feature Set

When using all features pooled together, the mean accuracy is elevated up to $4-10\%$ for the RBF-SVM method with respect to the the rhythmic content features and $1-5\%$ with respect to the instantaneous features. The corresponding values for the kNN are $1-12\%$ and $4-7\%$. Those values hold almost independent of the dataset under consideration. That is primarily expected, as the more information a good classifier receives, the best separation it can achieve, regardless of the high correlation of the features used. In total, an accuracy range of $34,2\%$ (HOMBURG) to $70,4\%$ (UNIQUE) is achieved. The classification results with respect to the parameterization and the datasets conform to the ones in the baseline experiment, therefore we will not discuss them further here. It is however clear that the combination of the two feature sets gives the best results, showing that there are relevant descriptors which can be further used in both of them.

In figure 10.1, the results for the optimized SVM method for instantaneous, rhythmic content and combined feature set for all datasets can be seen, providing an overview of the classification results discussed here. The general tendencies to be observed are three:

- The best results are given for the GTZAN and ISMIR04 datasets, whereas the HOMBURG dataset performs worse, as was also the case in most settings.

- For all datasets except the BALLROOM and UNIQUE the combined feature set has a performance which is very close to the instantaneous feature set. The discrepancy in the two datasets mentioned, where the combined feature set provides better results, can be due to the added importance of the rhythmic content features in their case.

- The rhythmic content features achieve results which are very close (ISMIR04) or slightly better (BALLROOM) than the instantaneous feature set results. In the case of the BALLROOM dataset, the explanation is once more the very dance-oriented, rhythmically relevant track selection which can be found in it. For the ISMIR04 dataset, taken into account that the other feature sets show almost the same classification accuracy, it is an indication that there is no specific feature set which is more relevant.

Before proceeding to the results after feature selection, two notes should be made: first, the number of features used in this case remains too high. Taking into account that one can construct even more of them in order to account for more variability, the goal can certainly not be to go in this direction, at least from a theoretical and computational parsimoniousness point of view. Therefore, it is important to search out which features give at least comparable classification accuracy, but limiting their number to a few so that interpretation can be meaningful and calculation can be efficient. Second, the datasets given here are, as already mentioned, imperfect: The ground truth in genres is, in the best case, subjective. In light of this fact, it is not meaningful to try and achieve perfectly accurate classification, but rather to try and find a small set of features that achieve acceptable classification performance which is comparable to that of the full feature set.

10.2. Performance of Classification after Feature Selection

In this section, a discussion of the classification results using the features resulting from the feature selection process will be given. Tables 9.8 and 9.10 show a full list of the best descriptors after each step of the selection process, whereas in tables 9.7 and 9.9 the corresponding achieved accuracies with the best classification algorithm settings for all datasets are given. Finally, tables 9.11 and 9.12 show the results of the classification after feature selection based on accent groups. As an introductory remark, one must note that the best features are necessarily somewhat different for each dataset, as they have to conform to the special distribution of tracks in genres therein. We will not discuss all of the best features separately here, as this lies beyond the scope of the thesis. Instead, we will focus on the most prevalent ones and the ones appearing at the top of the list and proceed to some remarks concerning their classification accuracy and interpretation.

10.2.1. Feature Selection with Mutual Information and Sequential Forward Methods

Best resulting features from feature selection As can be seen in table 9.8, the best features of the ranking based mutual information with the target data, the best features for all five datasets are the following:

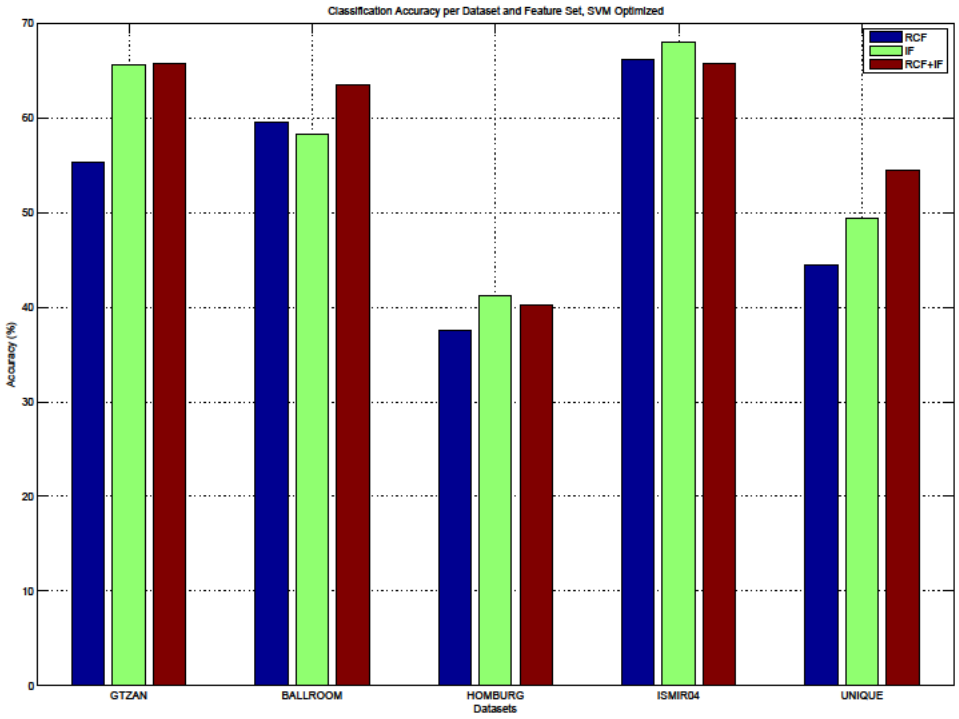

Figure 10.1.: Results of best settings for the RBF-SVM algorithm (SVM-Best), all datasets, prior to feature selection, rhythmic content features (RCF), instantaneous features (IF), combined (RCF+IF)

- **GTZAN:** The mean of the derivative of the beat histogram derived from the signal RMS (MD_RMS)
- **BALLROOM:** The most salient periodicity of the beat histogram derived from spectral flux ($P1\_SF$)
- **ISMIR04:** The mean of the derivative of the beat histogram derived from the second MFCC (MD_MFCC2)
- **UNIQUE:** The standard deviation of the beat histogram derived from the first MFCC (SD_MFCC1)
- **HOMBURG:** The standard deviation of the beat histogram derived from the signal RMS (SD_RMS)

Observing those results, we can surmise that the related subfeatures ($P1, MD$ and SD) as well as the instantaneous features whose novelty functions were used to create the respective beat histograms ($RMS, MFCC$ and SF) are amongst the most important in the sense that they explain the genre related variance in the best way. This was to be expected, since the perceptual accents to which they relate are the phenomenal ones, which are the most indicative of rhythmic (or, in fact, of any) changes in the signal.

Concerning the distribution of the best features in this list, preliminary analysis shows that with regard to the subfeatures, distribution features are more prevalent than peak features, with a tendency towards simple measures such as mean and standard deviation as opposed to skewness and kurtosis. However, the peak feature $P1$ not only figures high in the list, but also appears very often in the results. Since this descriptor is very close to the tempo of a piece (as it denotes the most salient periodicity), this result is not unexpected: Gouyon [39] has shown that only the tempo of the piece derived from the envelope is a feature with good discriminative power.

Concerning the novelty functions, the MFCCs, SPCs and RMS have the highest prevalence in this list. This demonstrates their relevance for extracting rhythmic content from an audio signal. To our understanding, this points towards two important considerations: The MFCC- and the RMS-based novelty functions, being essentially good descriptors of energy changes in the signal, signify the importance of the phenomenal accents for the description of rhythm. On the other hand, the features from beat histograms extracted through novelty functions of the SPCs are presumably very good descriptors of the structural (tonal) accents in a signal.

Table 9.10 shows the best features after sequential forward feature selection on the 20 first features from the ranking based mutual information with the target data. The results contain in the first rows the best features after the mutual information feature selection, which was to be expected as the most relevant features with respect to the ground truth will also be the less redundant in a wrapper feature selection context [41]. However, it is interesting to note that the distribution of the features across novelty functions and subfeatures remains roughly the same as before, indicating, that no novelty function or subfeature is actually unnecessary.

We will close the discussion about the best novelty functions and subfeatures from the mutual information feature selection process with another remark. Since the focus of the

thesis lay on classification, in-depth analysis of the best features was assigned to future research, which can be however performed on the basis of those preliminary results.

Classification with the resulting features from feature selection Using only the 20 first features of the mutual information ranking, a classification accuracy is achieved which is very close to the one using the whole rhythmic content feature set using the kNN and SVM with their best settings. This result is very important, since it shows that only **twenty** descriptors ($3,5\%$ of the full rhythmic content feature set) can produce very similar performance for genre classification, saving much time and allowing for better interpretation of the descriptors. The other tendencies observed here conform to the ones discussed in 10.1: the kNN shows worse performance than the SVM in all cases, and the performance of the individual datasets follows the same scheme (in decreasing order): ISMIR04, UNIQUE, BALLROOM, GTZAN and HOMBURG. It is interesting to note here that the performance of the BALLROOM dataset with fewer descriptors is actually better than with the full ones for the SVM setting. Similar considerations apply to the results after the sequential forward feature selection: The end resulting accuracy is very close to the one observed with the full rhythmic content feature set, which is an very promising result, as it shows that only very few features (between 9 and 15) can account for much of the rhythmic variance present in genres in the context of the musical genre classification task. Figure 10.2 shows a comparison of the performance of the best kNN and SVM settings, using the full rhythmic content dataset and the reduced versions.

10.2.2. Feature Selection by Accent Groups

The results of the evaluation of the feature subsets described in table 6.1 provides good conclusions concerning the usefulness of a conceptual categorization of features and its use for feature selection. As described in section 6.2, two groups of descriptors based on different accent types were used: phenomenal and structural. Using only the phenomenal features, a range for the mean average classification accuracy of $41,8\%$ for the HOMBURG to $61,3\%$ for the UNIQUE dataset was achieved for the RBF-SVM setting. Those results are very close to the ones achieved with use of the full rhythmic content dataset and the reduced versions with mutual information and sequential forward feature selection, showing that this categorization of features is potentially useful for the description of rhythm. It is interesting that the structural accent based features show generally slightly worse performance than the phenomenal ones, being possibly less important for audio description than the RMS or spectral flux feature when extracting a beat histogram from them. However, the difference between the results of the two accent groups is not extremely high, showing that *both* are needed in order to describe rhythm efficiently. Figure 10.3 shows a comparison of those results with the full rhythmic content feature set and the reduced versions after feature selection with the methods described above, for the RBF-SVM method with the best parameters.

It can be seen that the phenomenal and structural feature groups show only slightly smaller accuracy in comparison with the other settings. An important remark that can be made concerns the BALLROOM dataset: It is the only one for which the structural accent group features give better results than the phenomenal accent ones. This can be traced back to its very rhythmic character: it can be surmised that in view of the very pronounced beat

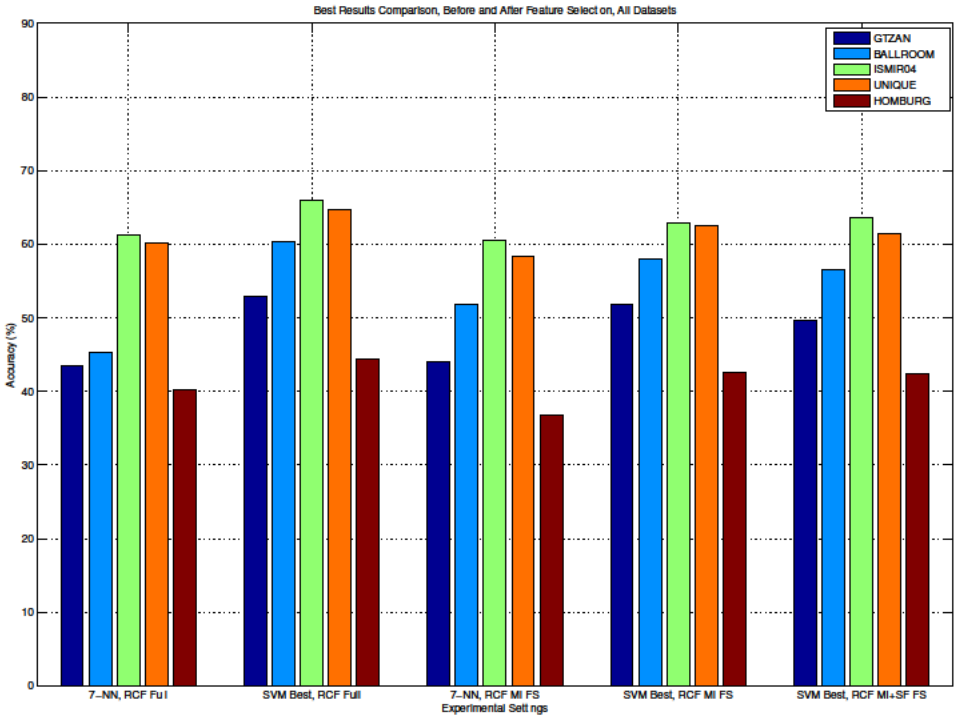

Figure 10.2.: Results of best settings (7-NN, SVM-Best), all datasets, prior to (RFC Full) and after Feature Selection (MI, MI+SF)

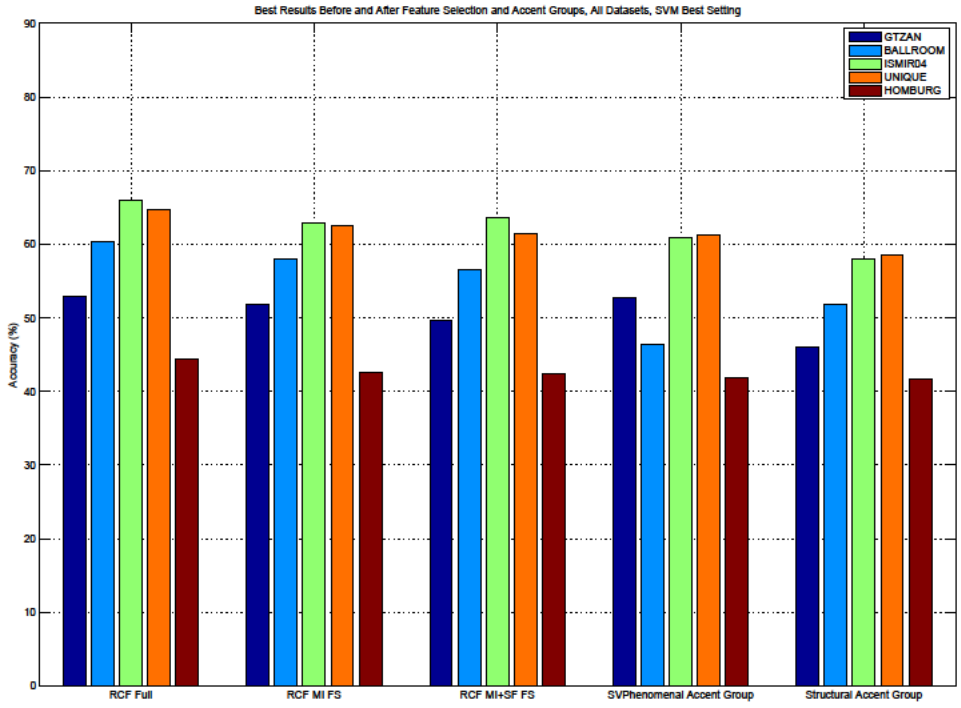

Figure 10.3.: Results of best settings (SVM-Best), all datasets, prior to (RFC Full), after Feature Selection (MI, MI+SF) and with Accent Groups

of most tracks in this dataset, the classification based on structural, i.e. tonal accents can allow the better separation of genres which have very similar phenomenal accent structure (e.g. *Waltz* and *Viennese Waltz* or *Jive* and *Quickstep*), but are different concerning the changes in the tonal levels of the signal. For the other datasets, the relative classification performance stays the same for all feature sets; only a slight fall can be observed from the full rhythmic feature set to the reduced ones with feature selection. However, that difference is very small, which encourages the use of very small rhythmic feature subsets for classification.

We close this section by making a remark on the analyses which were performed here. It is clear that an in-depth discussion of the whole feature selection experimental results is a copious undertaking, since there are many settings, feature sets and datasets to consider. However, a summary would read: For the further analysis of the features and subfeatures and experiments based on them, the reduced rhythmic content feature set after feature selection with RBF-SVM should be the first choice. Concerning the datasets, the HOMBURG dataset could definitely be excluded, whereas the BALLROOM and ISMIR datasets are promising choices for further research.

10.3. Interpretation of Misclassified examples

The discussion of the misclassified examples can provide good ideas concerning the "realistic" function of the algorithms and the relevance of the features. As can be seen in tables A.1-A.5, (see also the information about the datasets in tables B.1 to B.5), the kind of misclassification errors (referring only to the experiments using the rhythmic content features) in each dataset is very similar. Summing the results up and taking into account only the results using the best classification methods, the following tendencies can be observed:

- Given that some datasets are highly unbalanced, results tend to favor the classes which have the highest count of training examples. that results in the strongly present classes receiving a high classification rate, whereas small classes have often none correctly classified examples. This is an artifact of the classification algorithm which, as explained in chapter 5, can only be amended but not totally corrected, and as such must just taken for granted. The upshot of the procedure is that even in this case, a high mean average accuracy can be achieved, but this is only due to the fact that the classes with a large presence in the dataset reach very high scores, but samples of classes with smaller presence are inevitably assigned to the larger ones. However, in our view, this is similar to what a human listener would do: in the face of unbalanced genre preferences, one would expect a listener to classify unknown genres into something he or she knows better. With respect to the classification methods, we could see that the use of a weight vector for the cost parameter of the SVM can help improve the results slightly, but not mitigate the problem; this would require eventually the use of a balanced dataset. For the kNN, the unbalanced datasets present less of a problem; this is due to the simpler nature of the algorithm, which has the downside however of producing lower accuracy in all cases. The choice between the two is finally a matter of application or computational cost.

- For all datasets where it is applicable, the *rock*, *pop* and *alternative* genres tend to

be the worst classified (regarding their overall accuracy) and function as a "reservoir" for all other classes which are difficult to classify. This is a result of their very "general" musical character and rhythmic content, which does not allow for good separation. However, it is interesting to notice that the misclassification (false negative rate) takes place mostly between the similar classes, such as *country*, *metal* or *blues*, which share many common elements with *rock* and *pop*, especially concerning their rhythmic content.

- Varied or rhythmically rich genres, such as *classic* and *jazz* tend also to be mixed up with each other. For instance, in the GTZAN dataset, *classic* is only confused with *jazz* (and vice-versa), something which is also observable for the UNIQUE and ISMIR04 dataset. The same analysis holds for the genres which are "rock-affiliated" - *country* is often classified as *blues* - or conform to a similar production and consumption scheme, for example *dance* is often misclassified as *electronic* music. Those misclassifications can be explained through the use of similar rhythms in the context of those affiliated genres as a result of similar instrumentation (e.g., a drum kit with a steady beat) or their relevance for dancing (e.g., the use of a four-on-the-floor beat in most electronic tracks or of a specific rhythm in dance-specific music such as *salsa*, *samba* or *chachacha*).

- Rhythmically very strong (but not necessarily varied) genres, such as those of the BALLROOM dataset produce extremely good results in all cases regarding their misclassifications. That is a straight consequence of the relevance of very rhythmic descriptors for such genres. One such case is also the *metal* genre, which produces better results supposedly because of its very pronounced rhythmic character, as well as its fast tempo and high energy.

The above list of observations is far from being complete. However, a methodical analysis of misclassifications is a difficult matter which requires special attention, due to the large number of genres and relationships involved, let alone the facto of comparing different datasets with each other. Therefore, a strategy would be to focus on a specific dataset or the relationships between rhythmically very similar or different genres. Such a task lies beyond the scope of this thesis; however, an in depth analysis of the misclassifications in datasets which are more easy to analyze (such as the GTZAN and BALLROOM) is planned for future research.

10.4. Conclusion

From the previous discussion, the following conclusions can be drawn:

1. Musical genre classification remains a challenging task, even when using large sets of features, or features pertaining to specific sound aspects such as rhythm. However, when taking into account the difficulties in the definition of genre and its classification even from human listeners, high accuracy should not be expected in such tasks; a meaningful classification which achieves accuracies up to $60 - 70\%$ but makes the mistakes that human listeners would do (e.g., classifying *rock* as *pop*) is desirable and

attainable. In our results, such a classification accuracy could be achieved for the most datasets.

2. The datasets used are an important part in every musical genre classification setting, as the achieved accuracy depends highly on their size, track distribution and specific composition. It was established in the course of the thesis that datasets which are unbalanced or not thoroughly created can lead to suboptimal classification results. Therefore, attempts should be made to create standardized datasets from experts which are large, balanced and, as much as possible, musicologically valid.

3. Concerning the classification methods, the SVM algorithm has clearly shown the potential for good classification performance compared to the kNN. It is therefore proposed for further use in similar problems. Other methods can and should be used to compare the results of the SVMs to them.

4. The results achieved with the use of the accent-based rhythmic content features are promising and can be used as basis for further research in the area of automatic rhythm description or genre classification. Specifically, their use in other rhythm classification related problems such as beat tracking, tempo and meter induction or even language recognition based on rhythm is warranted.

5. From the results of the feature selection process with regard to the subfeatures, it follows that peak related features from the beat histogram are definitely among the best ones which can be used and therefore further research on rhythmic content should include them. However, basic statistics on the beat histogram such as standard deviation or the high frequency content also contribute to producing good classification results. Concerning the novelty functions, the results show that novelty functions derived from well-known and established features such as the root mean square, the mel-frequency cepstral and the spectral pitch chroma coefficients are amongst the most valuable for capturing the essence of the rhythmic content. Their use is therefore encouraged and further experiments will be conducted to assert their relevance. Finally, the features pertaining to phenomenal accents are much more useful in classification than the ones pertaining to structural ones. This was expected, since the phenomenal accent features track more "pronounced", actual changes in the signal and can therefore be considered as a good basis for the extraction of rhythm.

In general, it can be stated that the use of music theory approaches to designing new features for genre classification, in connection with a solid audio content analysis approach can give encouraging results. Further approaches conforming to this scheme or the application of the method to other related problems should, in our view, be undertaken.

11. Outlook

In the final chapter, an outlook will be given with respect to further experiments which could be undertaken in order to further evaluate the relevance of the proposed descriptors.

11.1. Improvement of Implementation

Through the course of this thesis, it became clear that the possible scope of experiments and parameterizations which can be undertaken is fairly wide. Therefore, we focused on conducting experiments taking into account many different settings, in order to gain a general idea of the suitability of our approach, but inevitably had to leave some subjects as material for future research. We give here a list of the possible examinations and experiments which can be undertaken to further increase the validity of the results, with respect to the subject area to which they apply.

Theoretical Approach The accent-based rhythmic content features described in this thesis have shown a good classification performance, even in the settings with the very few best features after the selection process. However, it is clear that the approach taken here for the categorization of accents and their correspondence to features is not unique. Further research could take into account other models of musical accent and attempt to encode them in features. A possible refinement would be the definition of more complex accents and their feature representation. Furthermore, musical accents are only one way of defining perceptually relevant points in the musical surface. Other methods could include the definition of other rhythmically relevant elements, such as the beat and or meter, and devise a classification based on features derived from them. Finally, the perceptual relevance of the accents used here was not tested directly with human listeners to assess their validity. A further goal would be the conduct of listening experiments in order to determine if and how listeners perceive accents in music, and base the feature design process on those results. This approach is one of great interest, as it would be very informative to compare the automatic classification results based on accent descriptors derived here, with the results produced by human subjects.

Feature Extraction The feature extraction procedure described in this thesis is one of the many possible implementations for periodicity extraction. There are several steps in the feature extraction procedure which could assume another approach. We list some of them here:

- Preprocessing of the features could include band-separation, transformations or other operations.

- The periodicity extraction process could be performed with other methods, such as the IOI-histogram or the application of a comb filter bank instead of an ACF.
- The instantaneous features selected to extract the novelty functions are but a fraction of the total descriptors available. Therefore, more could be included to test the importance of other signal quantities for classification. Concerning the subfeatures, the set used here is comprehensive, but other possibilities such as MFCCs on the beat histogram or derived features could be tested.
- Other transformations or steps in the feature extraction process, such as taking the first derivative of each novelty function or applying a different weighting scheme for the beat histogram than the one used here can be undertaken.

Classification The classification methods included here (kNN, SVM) where chosen due to their wide acceptance, ease of application, mathematical solidity and their good previous results in other experiments. However, a variety of other methods, such as decision trees, neural networks, gaussian mixture and hidden markov models can be used for classification. It remains open of those methods perform better than the kNN or the SVM, but their testing remains warranted. Furthermore, regarding the two methods employed in this thesis, other parameter settings, such as other distance metrics for the kNN and different parameterizations or kernel functions for the SVM could be used. The use of other settings for multiclass classification such as the one-vs-all method described in section 5.1.2 could help examine the problem of unbalanced datasets in more depth. Finally, in this thesis, unsupervised methods were not applied at all. It would be therefore important to apply methods from this area to the datasets with the descriptors used here in order to assess their performance in classification without given labels, and examine if the results are similar to those given here. This approach could help determine if the descriptors alone (without genre labels) can help explain the variance present in the datasets with respect to rhythm.

Datasets In our work, a greater number of datasets than in most publications was used. However, many others exist that can be evaluated, which gives a possibility for further evaluation of the descriptors proposed here. Concerning the datasets themselves, more in-depth evaluation such as classification in hierarchically organized datasets or subsets of them could be a possible research direction. Furthermore, an analysis of the datasets, either manually or by application of unsupervised learning methods, could be instrumental in determining if the ground truth is valid and whether specific tracks can be excluded to create more compact and consistent genre distributions.

Evaluation The evaluation of the classification results performed here is based mainly on classification accuracy. However, as it was described in section 5.4, there exist other possibilities of evaluating classification, with performance metrics which measure the precision, recall or specificity of the algorithm, and more compact measures such as the F-Score. In terms of comparability, this is note a problem since most publications report only on accuracy of the algorithm. However, in further settings other measures could be evaluated in order to assess the algorithm performance with regard to different "goodness" measures.

This step will be the next to be undertaken from the author as an expansion of the work performed in the thesis.

11.2. Further Research

A last word concerns the importance of the results for further research. Since it has been shown that the accent-based rhythmic features perform unexpectedly well in the task of musical genre classification with respect to the baseline features and results shown in other studies, it is in the view of the author that they can be of use for the description of rhythm *in general*: That is, their application in automatic rhythmic description problems such as the ones described in [38] could provide a valuable tool for the extraction of other rhythmic parameters, such as beat or meter, since their power in detecting rhythmic elements in music and discriminating music on their basis has been proved in the context of the thesis. Such use could also lead to the improvement of automatic musical genre classification systems for public or commercial use.

Furthermore, we believe that the descriptors shown here can be applied in non-musical signals, due to their generality and perceptual relevance. Therefore, the best ones will be evaluated further and used by the author in the task of automatic language identification, which bears many similarities to musical genre classification but focuses on speech signals of different languages. In this setting, the descriptors could be valuable for quantifying speech rhythm and testing its importance in distinguishing languages from one another. This evaluation will be performed in the context of the doctoral dissertation of the author, which concerns itself with the matter of automatic language identification based on speech rhythm and the descriptors that can be extracted to represent it.

Bibliography

[1] Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz. Applying support vector machines to imbalanced datasets. In *Machine Learning: ECML 2004*, pages 39–50. Springer, 2004.

[2] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[3] Jean-Julien Aucouturier and Francois Pachet. Representing musical genre: A state of the art. *Journal of New Music Research*, 32(1):83–93, 2003.

[4] Juan P Bello, Chris Duxbury, Mike Davies, and Mark Sandler. On the use of phase and energy for musical onset detection in the complex domain. *IEEE Signal Processing Letters*, 11(6):553–556, 2004.

[5] Juan Pablo Bello. Novelty detection, 2013. Electronic Presentation.

[6] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler. A tutorial on onset detection in music signals. *Speech and Audio Processing, IEEE Transactions on*, 13(5):1035–1047, 2005.

[7] Adam Berenzweig, Beth Logan, Daniel PW Ellis, and Brian Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, 28(2):63–76, 2004.

[8] James Bergstra. *Algorithms for classifying recorded music by genre*. PhD thesis, Citeseer, 2006.

[9] James Bergstra, Norman Casagrande, and Douglas Eck. Two algorithms for timbre and rhythm-based multiresolution audio classification. In *Proceedings of ISMIR*, 2005.

[10] James Bergstra, Norman Casagrande, Dumitru Erhan, Douglas Eck, and Balázs Kégl. Aggregate features and adaboost for music classification. *Machine Learning*, 65(2-3):473–484, 2006.

[11] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is nearest neighbor meaningful? In *Database Theory ICDT 99*, pages 217–235. Springer, 1999.

[12] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[13] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, 13:27–66, 2012.

[14] Juan José Burred and Alexander Lerch. A hierarchical approach to automatic musical genre classification. In *Proceedings of 6th International Conference on Digital Audio Effects '03*, 2003.

[15] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[16] Eric F Clarke. Categorical rhythm perception: an ecological perspective. *Action and perception in rhythm and music*, 55:19–33, 1987.

[17] Eric F Clarke. Rhythm and timing in music. *The psychology of music*, 2:473–500, 1999.

[18] Grosvenor Cooper. *The rhythmic structure of music*, volume 118. University of Chicago Press, 1963.

[19] Juan Pablo Bello Correa. *Towards the automated analysis of simple polyphonic music: A knowledge-based approach*. PhD thesis, University of London, 2003.

[20] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[21] Peter Desain and Henkjan Honing. The formation of rhythmic categories and metric priming. *PERCEPTION-LONDON-*, 32(3):341–366, 2003.

[22] Fabien Gouyon Dixon, Simon. Ismir 2006 tutorial: Computational rhythm description. Conference Tutorial, 2006.

[23] Simon Dixon. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30(1):39–58, 2001.

[24] Simon Dixon. Onset detection revisited. In *Proceedings of the 9th International Conference on Digital Audio Effects*, volume 120, pages 133–137. Citeseer, 2006.

[25] Simon Dixon, Fabien Gouyon, Gerhard Widmer, et al. Towards characterisation of music via rhythmic patterns. In *ISMIR*, 2004.

[26] Simon Dixon, Elias Pampalk, and Gerhard Widmer. Classification of dance music by periodicity patterns. In *ISMIR*, 2003.

[27] J Stephen Downie. Music information retrieval. *Annual review of information science and technology*, 37(1):295–340, 2003.

[28] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[29] Sahibsingh A Dudani. The distance-weighted k-nearest-neighbor rule. *Systems, Man and Cybernetics, IEEE Transactions on*, 4:325–327, 1976.

[30] Chris Duxbury, Juan Pablo Bello, Mike Davies, Mark Sandler, et al. Complex domain onset detection for musical signals. In *Proc. Digital Audio Effects Workshop (DAFx)*, pages 6–9, 2003.

[31] Daniel PW Ellis. Extracting information from music audio. *Communications of the ACM*, 49(8):32–37, 2006.

[32] Jonathan T Foote. Content-based retrieval of music and audio. In *Voice, Video, and Data Communications*, pages 138–147. International Society for Optics and Photonics, 1997.

[33] Paul Fraisse. Rhythm and tempo. In Diana Deutsch, editor, *The psychology of music*, Series in Cognition and Perception, chapter 6. Academic Press, 1982.

[34] Robert O Gjerdingen and David Perrott. Scanning the dial: The rapid recognition of music genres. *Journal of New Music Research*, 37(2):93–100, 2008.

[35] Masataka Goto. An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research*, 30(2):159–171, 2001.

[36] Masataka Goto and Yoichi Muraoka. A real-time beat tracking system for audio signals. In *Proceedings of the international computer music conference*, pages 171–174. San Francisco: International Computer Music Association, 1995.

[37] Masataka Goto and Yoichi Muraoka. Real-time rhythm tracking for drumless audio signals–chord change detection for musical decisions. In *Working Notes of the IJCAI-97 Workshop on Computational Auditory Scene Analysis*, pages 135–144, 1997.

[38] Fabien Gouyon and Simon Dixon. A review of automatic rhythm description systems. *Computer music journal*, 29(1):34–35, 2005.

[39] Fabien Gouyon, Simon Dixon, Elias Pampalk, and Gerhard Widmer. Evaluating rhythmic descriptors for musical genre classification. In *Proceedings of the AES 25th International Conference*, pages 196–204. Citeseer, 2004.

[40] Fabien Gouyon and Benoit Meudic. Towards rhythmic content processing of musical signals: Fostering complementary approaches. *Journal of New Music Research*, 32(1):41–64, 2003.

[41] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[42] Stephen Hainsworth and Malcolm Macleod. Onset detection in musical audio signals. In *Proc. Int. Computer Music Conference*, pages 163–6, 2003.

[43] Helge Homburg, Ingo Mierswa, Bülent Möller, Katharina Morik, and Michael Wurst. A benchmark dataset for audio classification and clustering. In *ISMIR*, volume 2005, pages 528–31, 2005.

[44] David Brian Huron. *Sweet anticipation: Music and the psychology of expectation*. MIT press, 2006.

[45] George H John, Ron Kohavi, Karl Pfleger, et al. Irrelevant features and the subset selection problem. In *ICML*, volume 94, pages 121–129, 1994.

[46] Mari R Jones and Marilyn Boltz. Dynamic attending and responses to time. *Psychological review*, 96(3):459, 1989.

[47] Anssi Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 6, pages 3089–3092. IEEE, 1999.

[48] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.

[49] Edward W Large and John F Kolen. Resonance and the perception of musical meter. *Connection science*, 6(2-3):177–208, 1994.

[50] Jean Laroche. Efficient tempo and beat tracking in audio recordings. *Journal of the Audio Engineering Society*, 51(4):226–233, 2003.

[51] Edda Leopold and Jörg Kindermann. Content classification of multimedia documents using partitions of low-level features. *Journal of Virtual Reality and Broadcasting*, 3(6), 2006.

[52] Alexander Lerch. *An introduction to audio content analysis: Applications in signal processing and music informatics*. John Wiley & Sons, 2012.

[53] Fred Lerdahl and Ray S Jackendoff. *A generative theory of tonal music*. MIT press, 1983.

[54] Joel Lester. *The rhythms of tonal music*. Pendragon Press, 1986.

[55] Tao Li, Mitsunori Ogihara, and Qi Li. A comparative study on content-based music genre classification. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 282–289. ACM, 2003.

[56] Tao Li and George Tzanetakis. Factors in automatic musical genre classification of audio signals. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 143–146. IEEE, 2003.

[57] Thomas Lidy and Andreas Rauber. Evaluation of feature extractors and psychoacoustic transformations for music genre classification. In *ISMIR*, pages 34–41, 2005.

[58] Stefaan Lippens, Jean-Pierre Martens, and Tom De Mulder. A comparison of human and automatic musical genre classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. (ICASSP'04)*, volume 4, pages iv–233. IEEE, 2004.

[59] Justin London. Rhythm. *The new Grove dictionary of music and musicians*, 21:277–309, 2001.

[60] Justin London. *Hearing in time*. Oxford University Press, 2012.

[61] Michael I Mandel and Daniel PW Ellis. Song-level features and support vector machines for music classification. In *ISMIR 2005: 6th International Conference on Music*

Information Retrieval: Proceedings: Variation 2: Queen Mary, University of London & Goldsmiths College, University of London, 11-15 September, 2005, pages 594–599. Queen Mary, University of London, 2005.

[62] Gonçalo Marques, Thibault Langlois, Fabien Gouyon, Miguel Lopes, and Mohamed Sordo. Short-term feature space and music genre classification. *Journal of New Music Research*, 40(2):127–137, 2011.

[63] MATLAB. *MATLAB:R2011a*. the MathWorks, Inc., Natick, Massachusetts, 2011.

[64] Daniel Müllensiefen, Martin Pfleiderer, and Klaus Frieler. The perception of accents in pop music melodies. *Journal of New Music Research*, 38(1):19–44, 2009.

[65] K Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, and Bernhard Scholkopf. An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on*, 12(2):181–201, 2001.

[66] Nicola Orio. *Music retrieval: A tutorial and review*, volume 1. now publishers Inc, 2006.

[67] François Pachet and Pierre Roy. Exploring billions of audio features. In *Content-Based Multimedia Indexing, 2007. CBMI'07. International Workshop on*, pages 227–235. IEEE, 2007.

[68] Richard Parncutt. A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, pages 409–464, 1994.

[69] Aniruddh D Patel. *Music, language, and the brain*. Oxford university press, 2010.

[70] G Peeters. A large set of audio features for sound description (similarity and description) in the cuidado project. *IRCAM, Paris, France*, 2004.

[71] Bruno H Repp. Probing the cognitive representation of musical time: Structural constraints on the perception of timing perturbations. *Cognition*, 44(3):241–281, 1992.

[72] Chris Sanden, Chad R Befus, and John Z Zhang. A perceptual study on music segmentation and genre classification. *Journal of New Music Research*, 41(3):277–293, 2012.

[73] Nicolas Scaringella and Giorgio Zoia. On the modeling of time information for automatic genre recognition systems in audio signals. In *Proceedings of the ISMIR 2005 6th International Conference on Music Information Retrieval, 12-15 September 2005*. IEEE, 2005.

[74] Nicolas Scaringella, Giorgio Zoia, and Daniel Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 23(2):133–141, March 2006.

[75] Eric Scheirer and Malcolm Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1331–1334. IEEE, 1997.

[76] Eric D Scheirer. Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*, 103(1):588–601, 1998.

[77] Walter Andrew Schloss. *On the automatic transcription of percussive music: from acoustic signal to high-level analysis*. PhD thesis, Stanford University, Center for Computer Research in Music and Acoustics (CCRMA), 1985.

[78] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[79] Hans-Henning Schulze. Categorical perception of rhythmic patterns. *Psychological Research*, 51(1):10–15, 1989.

[80] Wilhelm Seidel. Rhythmus/numerus. *Handwörterbuch der musikalischen Terminologie, hg. v. HH*, 1980.

[81] Jarno Seppänen. Computational models of musical meter recognition. Master's Thesis, Department of Information Technology, Tampere University of Technology, 2001.

[82] Klaus Seyerlehner, Markus Schedl, Peter Knees, and Reinhard Sonnleitner. Draft: A refined block-level feature set for classification, similarity and tag prediction. *Unpublished Paper (Draft)*, 2011.

[83] Klaus Seyerlehner, Gerhard Widmer, and Dominik Schnitzer. From rhythm patterns to perceived tempo. In *ISMIR*, pages 519–524, 2007.

[84] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[85] Hagen Soltau. Erkennung von musikstilen. *Master of Science Thesis*, 1997.

[86] Hagen Soltau, Tanja Schultz, Martin Westphal, and Alex Waibel. Recognition of music types. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 2, pages 1137–1140. IEEE, 1998.

[87] Bob L Sturm. An analysis of the gtzan music genre dataset. In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, pages 7–12. ACM, 2012.

[88] Bob L Sturm. The gtzan dataset: Its contents, its faults, their affects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*, 2013.

[89] Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378, 2007.

[90] David Temperley. *The cognition of basic musical structures*. MIT press, 2004.

[91] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, 10(5):293–302, 2002.

[92] Vladimir Vapnik. *The nature of statistical learning theory*. springer, 2000.

[93] Edgard Varèse and Chou Wen-chung. The liberation of sound. *Perspectives of New Music*, 5(1):11–19, Autumn-Winter 1966.

[94] Barry L Vercoe, William G Gardner, and Eric D Scheirer. Structured audio: Creation, transmission, and rendering of parametric sound representations. *Proceedings of the IEEE*, 86(5):922–940, 1998.

[95] Henrik von Coler. Blind note-level segmentation of monophonic music using hidden markov models. Master's thesis, Audio Communication Group, Technical University of Berlin, 2013.

[96] Kristopher West and Stephen Cox. Features and classifiers for the automatic classification of musical audio signals. In *ISMIR*, 2004.

[97] Eberhard Zwicker. *Psychoakustik*, volume 6. Springer, 1982.

List of Figures

List of Tables

Appendix

A. Confusion Matrices

In this chapter, the confusion matrices as result of the classification process are given. Due to space considerations and in order to focus on the important results, only the best results are presented here for each dataset, all rhythmic content feature sets before and after classification and with use of the optimized RBF-SVM method.

GTZAN	Predicted									
Ground Truth	blues	classic	country	disco	hiphop	jazz	metal	pop	reggae	rock
Blues	59.0	2.0	4.0	4.0	5.0	5.0	8.0	1.0	7.0	6.0
Classic	2.0	79.0	1.0	1.0	0	11.0	2.0	0	1.0	5.0
Country	11.0	3.0	56.0	4.0	2.0	7.0	3.0	1.0	1.0	13.0
Disco	3.0	0	4.0	56.0	6.0	3.0	7.0	8.0	6.0	10.0
Hiphop	3.0	0	0	8.0	64.0	2.0	2.0	5.0	13.0	3.0
Jazz	7.0	13.0	4.0	1.0	0	64.0	4.0	1.0	3.0	3.0
Metal	2.0	2.0	1.0	8.0	1.0	1.0	72.0	2.0	0	12.0
Pop	7.0	2.0	3.0	9.0	5.0	4.0	1.0	42.0	10.0	16.0
Rock	6.0	1.0	3.0	6.0	14.0	4.0	0	4.0	59.0	3.0
Reggae	7.0	4.0	8.0	4.0	4.0	4.0	15.0	9.0	4.0	42.0
Accuracy	59	79	56	56	64	64	72	42	52	42
Prior	10	10	10	10	10	10	10	10	10	10

(A.1)

Table A.1.: Genre confusion matrix for GTZAN dataset, full rhythmic content feature set

BALLROOM	Predicted							
Ground Truth	ChaChaCha	Jive	Quickstep	Rumba	Samba	Tango	Viennese Waltz	Waltz
ChaChaCha	85.0	1.0	5.0	5.0	10.0	5.0	1.0	0
Jive	8,0	29,0	8,0	2,0	10,0	3,0	1,0	0
Quickstep	4,0	4,0	46,0	8,0	9,0	6,0	4,0	2,0
Rumba	8,0	0	4,0	58,0	1,0	4,0	4,0	20,0
Samba	13,0	9,0	8,0	6,0	46,0	4,0	1,0	1,0
Tango	2,0	1,0	8,0	3,0	2,0	56,0	9,0	5,0
Viennese Waltz	1,0	0	10,0	12,0	0	6,0	21,0	15,0
Waltz	0	0	1,0	17,0	0	1,0	10,0	81,0
Accuracy	76,6	48,3	56,1	59,2	53,5	65,1	32,3	73,6
Prior	15,9	8,6	11,7	14,0	12,3	12,3	9,3	15,8

(A.2)

Table A.2.: Genre confusion matrix for BALLROOM dataset, full rhythmic content feature set

ISMIR04	Predicted					
Ground Truth	Classical	Electronic	Jazz	Metalpunk	Rockpop	World
Classical	298,0	7,0	0	1,0	2,0	13,0
Electronic	12,0	75,0	0	3,0	9,0	17,0
Jazz	13,0	3,0	0	0	7,0	2,0
Metalpunk	3,0	8,0	0	22,0	10,0	2,0
Rockpop	19,0	21,0	0	11,0	40,0	10,0
World	46,0	25,0	0	1,0	5,0	46,0
Accuracy	93,1	65,2	0	48,9	39,6	37,7
Prior	43,9	15,8	3,6	6,2	13,9	16,7

(A.3)

Table A.3.: Genre confusion matrix for ISMIR04 dataset, full rhythmic content feature set

UNIQUE	Predicted													
Ground Truth	Bl	Co	Da	El	Hi	Ja	Kl	Re	Ro	Sc	So	Vo	Wd	Wt
Blues	0	0	5,0	0	4,0	20,0	4,0	0	9,0	0	0	0	0	0
Country	0	0	8,0	0	3,0	17,0	8,0	1,0	20,0	1,0	0	0	1,0	0
Dance	0	0	661,0	1,0	18,0	21,0	16,0	0	47,0	0	0	0	3,0	0
Electronica	0	0	147,0	0	3,0	8,0	13,0	0	14,0	0	0	0	1,0	0
Hiphop	0	0	43,0	0	164,0	9,0	1,0	2,0	8,0	0	0	0	2,0	0
Jazz	0	0	41,0	0	16,0	140,0	78,0	0	34,0	0	0	0	1,0	0
Klassik	0	0	5,0	0	3,0	32,0	688,0	0	15,0	0	0	0	2,0	0
Reggae	0	0	23,0	0	34,0	13,0	1,0	1,0	3,0	0	0	0	0	0
Rock	0	0	23,0	0	5,0	15,0	8,0	1,0	345,0	0	0	0	1,0	0
Schlager	0	0	24,0	0	3,0	11,0	3,0	0	17,0	1,0	0	0	1,0	0
Soulrnb	0	0	7,0	0	9,0	14,0	4,0	0	5,0	0	0	0	1,0	0
Volksmusik	0	0	6,0	0	3,0	7,0	10,0	0	10,0	1,0	0	0	1,0	0
World	0	0	28,0	0	9,0	36,0	55,0	0	17,0	0	0	0	1,0	0
Wort	0	0	2,0	0	11,0	0	0	0	0	0	0	0	0	13,0
Accuracy	0	0	86,3	0	71,6	45,2	92,5	1,4	86,7	1,7	0	0	0	50
Prior	1,3	1,9	24,6	6	7,4	10	23,9	2,4	12,8	1,9	1,3	1,2	4,7	0,8

(A.4)

Table A.4.: Genre confusion matrix for UNIQUE dataset, full rhythmic content feature set

HOMBURG	Predicted								
Ground Truth	Alternative	Blues	Electronic	Folkcountry	Funksoulrnb	Jazz	Pop	Raphiphop	Rock
Alternative	6,0	0	5,0	13,0	0	19,0	3,0	8,0	91,0
Blues	0	5,0	1,0	12,0	0	27,0	0	12,0	63,0
Electronic	3,0	1,0	12,0	3,0	0	32,0	0	22,0	40,0
Folkcountry	7,0	3,0	2,0	44,0	0	61,0	2,0	11,0	93,0
Funksoulrnb	2,0	0	0	4,0	0	11,0	0	7,0	23,0
Jazz	4,0	3,0	8,0	29,0	0	149,0	1,0	33,0	91,0
Pop	2,0	0	2,0	14,0	0	28,0	3,0	11,0	56,0
Raphiphop	0	2,0	4,0	5,0	0	32,0	0	230,0	27,0
Rock	9,0	5,0	6,0	31,0	0	45,0	4,0	19,0	387,0
Accuracy	4,1	4,2	10,6	19,8	0	46,7	2,6	76,7	76,8
Prior	7,7	6,4	6	11,8	2,5	16,9	6,2	15,9	26,7

(A.5)

Table A.5.: Genre confusion matrix for HOMBURG dataset, full rhythmic content feature set

B. Dataset Description

In this section, the datasets used in the thesis are presented with respect their total track count, the genres they comprise, the distribution of tracks in genres, their prior (percentage of tracks of a genre in the dataset) and information about their length and audio data type. All of them can be found in the internet and are available at no charge for research purposes.

B.1. GTZAN

Total Track Count	1000	**Genres**	**Tracks per Genre**	**Prior**
Genre Count	10	Blues	100	10%
Balanced	yes	Classical	100	10%
Length (s)	30	Country	100	10%
Format	.au	Disco	100	10%
		Hiphop	100	10%
		Jazz	100	10%
		Metal	100	10%
		Pop	100	10%
		Reggae	100	10%
		Rock	100	10%

(B.1)

Table B.1.: GTZAN dataset

B.2. BALLROOM

Total Track Count	698	**Genres**	**Tracks per Genre**	**Prior**
Genre Count	8	ChaChaCha	111	$15,9\%$
Balanced	no	Jive	60	$8,6\%$
Length (s)	30	Quickstep	82	$11,7\%$
Format	.wav	Rumba	98	$14,0\%$
		Samba	86	$12,3\%$
		Tango	86	$12,3\%$
		Viennese Waltz	65	$9,3\%$
		Waltz	110	$15,8\%$

(B.2)

Table B.2.: BALLROOM dataset

B.3. ISMIR04

Total Track Count	729	**Genres**	**Tracks per Genre**	**Prior**
Genre Count	6	Classical	320	$43,9\%$
Balanced	no	Electronic	115	$15,8\%$
Length (s)	30	Jazz	26	$3,6\%$
Format	.mp3	Metalpunk	45	$6,2\%$
		Rockpop	101	$13,9\%$
		World	122	$16,7\%$

(B.3)

Table B.3.: ISMIR04 dataset

B.4. UNIQUE

Total Track Count	3115	**Genres**	**Tracks per Genre**	**Prior**
Genre Count	14	Blues	41	$1,3\%$
Balanced	no	Country	58	$1,9\%$
Length (s)	30	Dance	766	$24,6\%$
Format	.wma	Electronica	187	$6,6\%$
		Hiphop	229	$7,4\%$
		Jazz	310	$10,0\%$
		Klassik	744	$23,9\%$
		Reggae	74	$2,4\%$
		Rock	398	$12,8\%$
		Schlager	59	$1,9\%$
		Soulrnb	39	$1,3\%$
		Volksmusik	38	$1,2\%$
		World	146	$4,7\%$
		Wort	26	$0,8\%$

(B.4)

Table B.4.: UNIQUE dataset

B.5. HOMBURG

Total Track Count	1886	**Genres**	**Tracks per Genre**	**Prior**
Genre Count	9	Alternative	145	$7,7\%$
Balanced	no	Blues	120	$6,4\%$
Length (s)	30	Electronic	113	$6,0\%$
Format	.mp3	Folkcountry	222	$11,8\%$
		Funksoulrnb	47	$2,5\%$
		Jazz	319	$16,9\%$
		Pop	116	$6,2\%$
		Raphiphop	300	$15,9\%$
		Rock	100	$26,7\%$

(B.5)

Table B.5.: HOMBURG dataset