Automatic Characterization of Music Complexity: a multi-faceted approach

by

Sebastian Streich

Submitted in partial fulfilment of the requirements for
the degree of Diploma of Advanced Studies
Doctorate in Computer Science and Digital Communication
Department of Technology

Tutor: Dr. Xavier Serra

Universitat Pompeu Fabra

Barcelona, July 2005

“I think the next century will be the century of complexity.”
Stephen Hawking (January 2000)

Abstract

The aim of this work is to present the concept of a multi-faceted music complexity descriptor set. The complexity of music is one of the less intensively researched areas in music information retrieval. Especially an automated estimation based on the audio material itself has not been addressed by many researchers. However, it is not only a very interesting and challenging topic, it also allows for very practical and relevant applications in music information retrieval.

After giving some background information about the motivation for this research, we will discuss examples of practical applications, such as collection visualization, playlist generation, and music recommendation. In a review of former work we will see existing models for melodic, rhythmic, and harmonic complexity facets. Since these are not directly applicable for our needs, we will then give a set of operational definitions for the author's approach to the problem. Preliminary results for rhythmic and acoustic complexity are reported and discussed. From these we sketch the steps for future work within and beyond the scope of a PhD thesis.

Acknowledgements

At the first place I would like to thank Dr. Xavier Serra, my supervisor, for giving me the opportunity to work on this very interesting topic in his team. Also, and specially, I want to thank Perfecto Herrera for providing countless suggestions and constant support for my work in this research project.

Further thanks go to my colleagues from Office 316, Beesuan Ong, Emilia Gómez, Fabien Gouyon, and Enric Guaus for many fruitful discussions and important feedback. I also want to thank the other people – directly or indirectly – involved in the SIMAC research project for their appreciated cooperation.

This research was funded by a scholarship from Universitat Pompeu Fabra and by the EU-FP6-IST-507142 project SIMAC. I am thankful for this support, which was a crucial economic basis for my research.

A special thank goes to the nice people that are my family whose support and encouragement was and is always very important for me.

Barcelona Sebastian Streich
July 13, 2005

Contents

Chapter 1

Introduction

This chapter contains some remarks on the background, context, and motivation of the research presented and the researcher presenting in this document. Also, some clarifications about the terminology are given, while a concise demarcation of the topic under research will be done in chapter 2.

1.1 Background and Context

The author obtained the German academic degree of Diplom-Ingenieur (graduate engineer) in Media Technology from Ilmenau Technical University in 2002. Already during the last years of his studies he specialized on the field of digital audio signal processing. In his student research project in 2000 the author worked in a group that ported an audio watermark decoder to a DSP platform. Later, during his internship at NOKIA Research Center in Tampere, Finland, in 2001 he had the opportunity to get in touch with audio codec technology [VSV$^+$03, WS02]. He graduated with a thesis on the design of a speaker identification system for the Electronic Media Technology Team of Fraunhofer Gesellschaft.

After graduation, the author was employed by Fraunhofer Gesellschaft as a research engineer. His work was centered around metadata extraction from digital audio with a focus on melody and chord estimation. In September 2003 he joined the Music Technology Group[1] (MTG) of Pompeu Fabra University in Barcelona, Spain, where he is enrolled in the PhD program in Computer Science and Digital Communication and holds a university scholarship. The MTG

[1] http://www.iua.upf.es/mtg/

consists of around 40 researchers, mostly PhD candidates, who are working on music signal processing, music information retrieval, musician-computer interaction, and music software engineering tasks. Dr. Xavier Serra, the head and founder of the group, has also supervised the research presented here.

Another very important part of the research context is the EU-FP6-IST-507142 project SIMAC (Semantic Interaction with Music Audio Contents). The project was initiated by the MTG and started in January 2004. As the project leader, MTG is heavily involved in administration and research. The author is one of six researchers inside MTG who, lead by the research manager Perfecto Herrera, are working full-time for SIMAC. The project's main goal is the development of prototypes for the automatic generation of semantic descriptors and the development of prototypes for exploration, recommendation, and retrieval of music. Further information on the project can be found in the internet under the URL *http://www.semanticaudio.org*.

The author is also a hobby musician with a classical piano education. Apart from classical music for solo piano, he is also enjoying to play ensemble music from Baroque to Bossanova Jazz. In music listening generally speaking the author has no particular preference for certain genres or artists, but likes to select individual tracks from a wide range of styles depending on the given mood and situation.

1.2 Today's Music Consumption Behaviour

Music, in the western cultural world at least, is present in everybody's life. It can be found not only at its "traditional" places like opera houses or discotheques, but appears in kitchens and nurseries, in cars and airports, bars and restaurants, parks and sport centers and in many other places. With the increased mobility of music reproduction devices nowadays everybody can bring his own personal music along and listen to it almost anywhere. But this enhanced availability of music also means that concentrated and active listening has become rare. Often music serves as a background while the listener is doing something else, it is used as an acoustical "gap filler", or as an emotional trigger [NH97c]. New ways of dissemination for music have been arising and never before has it been as easy as today to get access to such a huge amount of music (more than a million titles are available from a single music portal). Yet, it is not always easy to find what one is looking for.

1.2.1 Music Tracks as Data Files

With a broader public becoming aware of efficient audio coding techniques during the last decade, the amount of music being stored and collected in digital formats increased rapidly. While in former times a music collection consisted of a conglomeration of physical media in form of vinyl discs, analogue or digital tapes, and later also compact discs, a new kind of collections emerged due to the new formats. No longer is the music in these collections attached to a physical medium, but exists merely as a string of bits that can, very easily and without degradation in quality, be moved and copied from any digital storing device to any other one. This property, apart from being very convenient, again reinforced the rapid spread of the new formats, by making it as easy as never before to share and exchange music with virtually anybody on the planet (at least with the roughly 13% of our world's population who have access to the internet).

Nowadays, the music industry managed to hinder the wild exchange of music by filing lawsuits against operators and private users of such content exchange networks. But digital music collections already exist in large numbers and the concept of storing music tracks as digital files with all its advantages is an established fact for many music consumers. Lately, legal options to buy digital music online and receive it by file transfer over the World Wide Web are arising, often combined with some sort of digital rights management that is supposed to prevent consumers from "shamelessly" copying and sharing their new acquisitions. Last but not least, alternative models of copyrighting and licensing musical content are emerging (e. g. creative commons[2]) and contribute a small but growing part to the music that can be found and obtained from the internet. David Kusek and Gerd Leonhard, two music futurists, even coin the term of "music as water" [KL05] in order to describe the change of attitude towards music, music consumption, and music ownership that is in the process of establishing itself:

> "Imagine a world where music flows all around us, like water, or like electricity, and where access to music becomes a kind of 'utility'. Not for free, per se, but certainly for what feels like free."

[2] http://creativecommons.org/

1.2.2 Problems with Digital Collections

Despite the many advantages that came with this “digital revolution” we can observe some negative effects, too. The flexibility that music tracks in form of individual files provide goes along with a demand for a proper storage organization and labelling. While in a physical collection an item might be declared lost when it cannot be found, in a large virtual collection this can be said already when there are no means to search for it. The freedom to create personalized play lists and to listen to music tracks independently from concepts like albums or compilations requires on the other hand a good knowledge of the collection at hand in order to arrive at satisfying results. When typing errors in filenames, a lack of systematics, and inconsistency in the manually assigned labels occur in such a collection, navigation and utilization become difficult and limited.

Another aspect of bit strings representing music is their volatility and reproducibility. With the move of a finger they can be erased and they are gone without leaving any trace. Or they can be replicated with very little effort basically infinite times at no cost other than the disk space to store them. Especially the latter has an impact on the value that is assigned to such an item. According to the theory of supply and demand a commodity with finite demand but infinite supply has a price of zero (of course this cannot be applied directly to music). As an illustration: If I go into a record shop to buy an album and the salesman gives me two copies for the price of one, I will consider this a true benefit, because I can give one copy to a friend or resell it. If I buy the same album from an online shop and they offer me to download it twice by paying only once, I would just find it silly. This gives a good hint for understanding why people are so “generous” to offer their complete collection of music files for anybody to download without any charge. And it also helps understanding why people do not feel shy to download or copy music from others without paying and without feeling they do something unjust. Especially in the times of free peer-to-peer music sharing networks this configuration boosted the dissemination and the enforcement of digital music formats, which are a matter of fact nowadays. On the other hand, this excess of digital tracks decreased the appreciation for the individual item. So some private collections have been extended for the sake of collecting rather than because of a particular interest in the material.

Summarizing we can say that to date probably many hard disks exist which contain a large collection of digital music, but the owner is ignorant about the full potential of it. Even if he or she is not, it means a lot of effort to fully exploit it and rather sooner than later the currently available tools reach their limits. With the new way of music dissemination through online portals like for example iTunes[3] or Y!music[4] the problem of searching, browsing, and navigating a (unfamiliar) music collection is brought to an even larger scale.

1.2.3 Semantic Descriptors as a Perspective

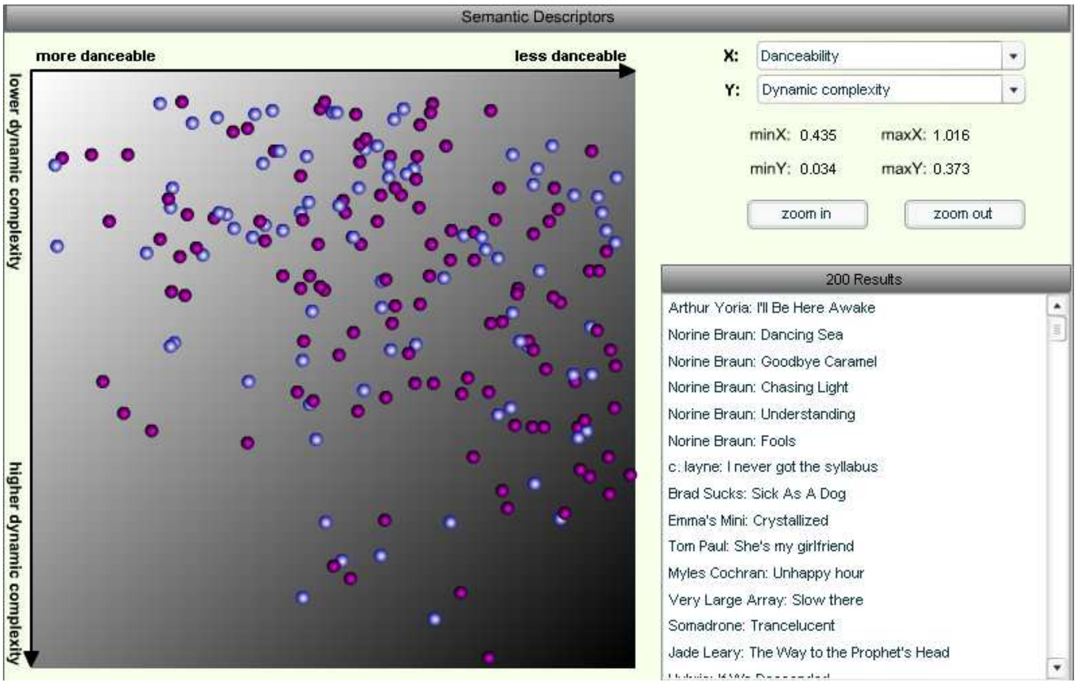
Slowly, tools and technologies are starting to spread that intent to enhance and facilitate the interaction with digital music collections. The key component of such tools is the assignment of semantic descriptors to the tracks. These descriptors capture certain properties and attributes of the music content, for example the tempo, the instrumentation, or the lead singer's gender. These properties usually reveal themselves automatically to a human who is listening to a music track and hence are potentially relevant information. Therefore a link between the pure digital audio data and the semantic concepts describing the content offers much more natural ways of searching in music collections than it is currently possible. Instead of being limited to titles, artists, and genres as the means of a query even very subtle or abstract aspects of music could be used.

Apart from the enhanced querying, also an extended interaction becomes possible. Browsing through a collection, be it one's own or a foreign one, where different musical aspects are visualized (see figure 1.1) is not only amusing, but also may lead to a better understanding of the music. Similarities between different styles or artists can be discovered, the evolution of a band can be tracked, extreme examples can be identified. This playful and educational side effect could then again lead to a more attentive way of music listening, increasing pleasure and appreciation.

But not only for the human do semantic descriptors bring additional opportunities. If they are organized in a machine readable way, there is an access for information processing devices to the properties of music which are relevant for a human listener. Basically, this allows computers to mimic a human's music perception behaviour. So instead of the human browsing the collection, the

[3] http://www.apple.com/itunes/

[4] http://launch.yahoo.com/



X: Danceability

Y: Dynamic complexity

minX: 0.435 maxX: 1.016

minY: 0.034 maxY: 0.373

Arthur Yoria: I'll Be Here Awake
Norine Braun: Dancing Sea
Norine Braun: Goodbye Caramel
Norine Braun: Chasing Light
Norine Braun: Understanding
Norine Braun: Fools
c. layne: I never got the syllabus
Brad Sucks: Sick As A Dog
Emma's Mini: Crystallized
Tom Paul: She's my girlfriend
Myles Cochran: Unhappy hour
Very Large Array: Slow there
Somadrone: Trancelucent
Jade Leary: The Way to the Prophet's Head

Figure 1.1: Screenshot from a music browser interface displaying part of a song collection organized by danceability and dynamic complexity.

computer could do this automatically and identify for example clusters of similar tracks. Thus, the computer is enabled to generate play lists according to given criteria or to recommend to the user similar tracks to a selected example track. So we see that providing machine readable semantic descriptors for the tracks in a collection opens the door for a large variety of interesting methods of sorting, searching, filtering, clustering, classifying, and visualizing. But how can we arrive there? Different ways exist to assign semantic descriptors to music. It is possible (although expensive) to have a group of music professionals annotate them. This is of course usually not an option for private collections. Still, the metadata - once annotated - could be stored in a public database and would then be associated through a fingerprinting service with the files in a private collection. To some extent the descriptors can also be assigned by a whole community of listeners by majority vote or by using collaborative filtering techniques. This involves quite some coordinative efforts and furthermore involves a delay until reliable data for a new track becomes available. The most practical and versatile option however is the automatic computation of descriptors based on the audio file itself. This way, an objective, consistent, inexpensive, and detailed annotation can be accomplished. The research presented here is about this third option of descriptor extraction.

1.3 Thesis Outline

After this introduction a very compact description of the research procedure will be given in chapter 2 defining the scope of this research. In the following chapter 3 we will then look at different views on complexity and discuss the details of the underlying concept for music complexity as we want to approach it. Afterwards, a review of related research is presented in chapter 4. We will see that several models capturing aspects of music complexity were already proposed in the literature, but that their application to the task we have in mind is problematic or at least not straightforward. In chapter 5 we first consider in more detail how the different facets of music complexity can be approached from the computational perspective. A description of two original contributions of the author follows. The implementations of acoustic and rhythmic complexity models (danceability) are explained, and first results are discussed. In chapter 6 the roadmap for the PhD thesis and some ideas reaching beyond will be rolled out. The document concludes with a bottomline of the research presented here.

Chapter 2

Working Thesis and Procedure

It is the goal of this research work to provide operational models for the automated computation of music complexity as it is perceived[1] by human listeners. We regard the complexity of music as a high-level, intuitive attribute, which can be experienced directly or indirectly by the active listener, so it could be estimated by empirical methods. In particular we define the complexity as that property of a musical unit which determines how much effort the listener has to put into following and understanding it (see also section 3.1).

The proposed models are intended to provide a compact description of the musical content, which can be utilized to facilitate the user interaction with music databases. Since music has different and partly independent facets, we will address these individually with separate models (see also section 5.1). We consider the following musical facets, that are also illustrated in figure 2.1:

- Melody
- Harmony
- Rhythm
- Timbre
- Acoustics (spatial/dynamic)
- Structure

[1] Throughout this document we will use the term "perception" although "cognition" might be considered more appropriate at some places. However, we want to omit this distinction here and consecutively use "perceptual" as the contrasting term to "technical".

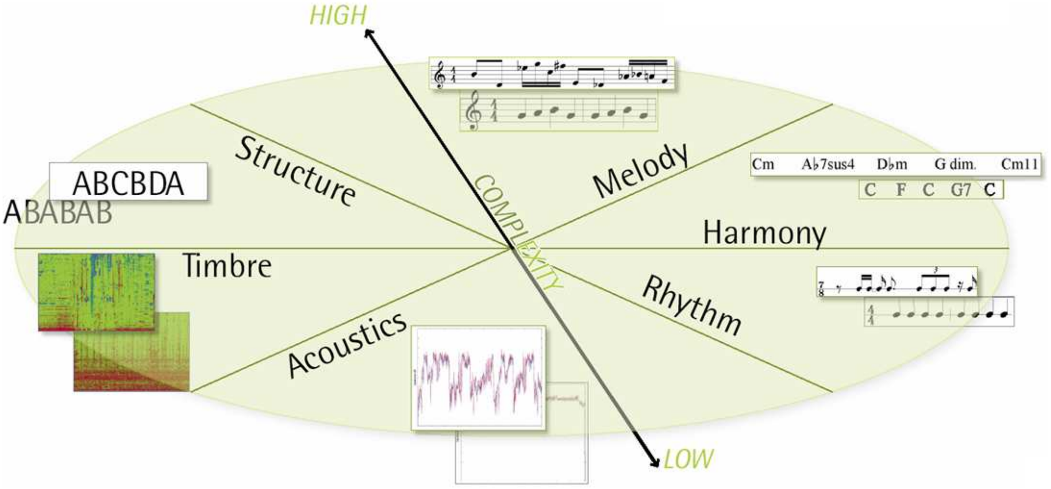


Figure 2.1: Circle of music complexity facets.

The descriptors are going to be designed for the track level. That means a complete track of music is considered the unit on which the models are working. The segmentation into tracks is assumed to be given already, so the application to a stream (as in broadcast monitoring) is not addressed.

Neither will we address the individual characteristics of users' complexity perceptions. Instead it is the goal to provide models for the "common sense" in music complexity rating (see section 3.1.3). We want to focus rather on a naïve complexity estimation without taking expert knowledge about style-specific musical peculiarities into account. That means we are interested in a classification on a large, but rougher scale rather than a precise ranking taking finest nuances into account. For the latter we would have to focus in detail on a particular user's musical background and listening habits. The former can be achieved by assuming a common background for a certain group of users and then restricting the validity of the provided models only to this group. We chose the group of non-expert music listeners with western cultural background as the intended users for our models, since it forms a large fraction of the users dealing with the above mentioned music databases.

Chapter 3

Music Complexity as a Semantic Descriptor

Semantic descriptors for music content can come in many forms, because music is rich in aspects and levels of abstraction. This said we immediately arrive at a new problem. Which are useful semantic descriptors to be used? There is no obvious solution to this problem and most likely there is more than just one answer depending on the application and maybe even on the content itself. Obviously the extracted descriptors should capture features of the music that are relevant to human listeners. Furthermore, it would be considered positively, if the descriptors are compact in their representation, and their meaning is easy to be comprehended. Of course there are a lot of descriptors fulfilling these demands. In the following we will mainly consider one small fraction of them: the multiple facets of music complexity.

3.1 What is Complexity?

The term complexity is understood differently in different contexts. It is a term that everybody seems to know, but still a unified definition is difficult to find. Furthermore, it is a fancy term to use in order to raise interest, which Edmonds in [Edm95] describes with the following metaphor: “The label of ‘complexity’ often performs much the same role as that of the name of a desirable residential area in Estate Agent’s advertisements. It is applied to many items beyond the original area but which are still somewhere in the vicinity.” So, despite a compact working definition for the use in this research work has already been given in chapter 2, we dedicate some space in this section to cover the scientific background of the term, and to clarify our understanding in the context of music.

3.1.1 Complexity in Information Theory

In information theory there exist different measures that can be associated with complexity. The most “traditional” one is probably the concept of source entropy introduced in 1948 by Shannon [Sha48]. Entropy measures the randomness of the output from an information source, by specifying the average amount of information for each symbol s_i emitted by the source according to the following formula:

$$H_0 = -\sum_{i=1}^{N} p(s_i) \cdot \log_2 p(s_i), \tag{3.1}$$

where N is the number of distinct symbols and $p(s_i)$ is the probability that symbol s_i is emitted. High entropy values would be associated with higher complexity, since the information rate is higher. Equation 3.1 assumes a memoryless source, like the flipping of a coin or the rolling of a dice, where the preceding symbol in the sequence has no influence at all on the symbols to follow. This assumption is not justified in many practical applications. For structured data as texts for example, the probability of a symbol depends strongly on its predecessors. In information theory the terms *Markovian source* and *Markov chain* are used in these cases. Equation 3.1 can be adapted for a Markov chain by taking the corresponding conditional probabilities into account. If only the direct predecessor is relevant (as in a first order Markov chain), the equation takes the following form:

$$H_1 = -\sum_{i=1}^{N} p(s_i) \sum_{j=1}^{N} p(s_j|s_i) \cdot \log_2 p(s_j|s_i), \tag{3.2}$$

where $p(s_j|s_i)$ denotes the conditional probability of symbol s_j appearing after symbols s_i in the sequence. H_0 is the upper bound for H_1 and any higher order entropy of a given source, since the structural regularities captured by the conditional probabilities can only decrease randomness.

The most widespread definition of complexity in the information theory field originates from the theory of algorithmic information. Referring to one of the main scientists behind it this definition is usually called *Kolmogorov complexity* (see [GV99] for a short introduction). In contrast to entropy it addresses the absolute information content of an object, which is the amount of data that needs to be transmitted in the absence of any other a priori knowledge. It is defined as the length of the shortest program p (in bits), that can generate the sequence s. Formally we can write:

$$K(s) = \min(|p| : T(p) = s), \tag{3.3}$$

where $T(p)$ refers to the output of a universal Turing machine executing p. Thanks to its general nature Kolmogorov complexity has been applied also in the digital audio domain (see [Sch01]) as a way to prove mathematically the advantages of structured audio in generalized audio coding.

The striking quality of both concepts is the objectivity of the measures. Kolmogorov complexity and Shannon entropy both exist as exclusive properties of the object and are completely independent from an observer. But when considering musical audio signals we face problems. First, the entropy measure needs a finite set of discrete symbols being emitted from the source, but this is not what we have. The signal would need to be converted somehow into symbols like the letters in a written text, which is not very straightforward. Despite that music can be generated with a keyboard it is usually much more than the output of a "typewriter"[1].

Secondly, both measures are rather technical focusing on an efficient encoding, while we are interested in the reception by humans. A practical example illustrating this difference very well is given by Standish in [Sta01]. He points out that a random sequence of numbers will always yield a maximum complexity although it does not contain any meaningful information for a human. Comparing a random string of characters with the manuscript of a theater play, the random sequence would yield the higher Kolmogorov complexity value, because there is no program shorter than the sequence itself. However, for a human the order would be exactly reversed, because the random string is perceived as meaningless noise, while the theater play is recognized as a sophisticated, multi-layered setting (at least if it is a good one). The apparent objectivity of the measure makes it meaningless when context has to be considered.

Standish suggests the use of equivalence classes to overcome this. An equiv-

[1] This can be experienced easily by comparing a monophonic mobile phone ringing tone with a live performance of the same piece.

alence class for him is the set of all mutations of a sequence that are equivalent in a given context. So for example different random sequences could hardly be distinguished by a human observer and would therefore form a large equivalence class. On the other hand, for a written text carrying a meaning, only relatively few mutations exist that would be judged as equivalent. This judgement depends not only on the data itself but also the context in which it is observed. If the equivalence classes are considered in the complexity computation, we arrive at more meaningful results from a human's point of view. But still we do not have a practical solution. The decision of what is equivalent and what is not does not seem to be very straightforward, especially when it comes to music. Furthermore there is no closed mathematical solution for computing the Kolmogorov complexity and it remains a rather theoretical measure.

3.1.2 A Philosophical Point of View

Edmonds, who is reviewing the matter of complexity from the philosophical point of view [Edm95], gives us the following definition: Complexity is "[t]hat property of a language expression which makes it difficult to formulate its overall behaviour, even when given almost complete information about its atomic components and their inter-relations."

This is a very general statement leaving (purposely) a lot of room for interpretation depending on the given context where it is to be applied. Although it is not explicit in the statement, Edmonds considers a subjective element of complexity, because what "makes it difficult" can vary with the context and the observer. We can agree on this definition also for the case of musical complexity, yet, it does not bring us much closer to an implementable model. How do we identify the aspects of music that make the *formulation of the overall behaviour* difficult? What are the *atomic components* we should consider in this context? However, we can take out at least two ideas from this statement. First, we can think of complexity as being a property that varies along an axis between *easiness* and *difficulty*. Second, it is a kind of *meta-descriptor*, because it describes a certain property of lower-level descriptors (atomic components) when they are arranged to form a bigger unit. So by nature a complexity descriptor is compact even though it might contain information about a wide-stretched temporal process.

In the case of music it is not unambiguous to speak of a unit. Music can be structured on different levels. For the applications we have in mind it is advantageous to consider the music tracks that can be found in a collection as the units whose complexity is to be described. This way, we can avoid additional problems of segmentation (still an unsolved problem for musical audio [OH04]) and simply rely on the assumption that one digital file corresponds to one musical unit.

It is worthwhile mentioning that it could also make sense to go below the track level and provide complexity estimates for smaller fractions. In fact, an instant computation of complexity estimates in a sliding window fashion itself could be helpful in segment identification or other tasks related to very high-level aspects of music. However, we save these considerations for the chapter on Future Work.

3.1.3 Perceptual Foundations of Music Complexity

If we move our focus more from the general towards the specific, we can find the following remark on music complexity by Finnäs. In [Fin89] he states that "unusual harmonies and timbres, irregular tempi and rhythms, unexpected tone sequences and variations in volume" raise the level of perceived complexity. This statement is neither exhaustive nor precise. However it brings some aspects into play, which are special for music and therefore interesting for us.

We can see that Finnäs mentions very different facets of music (harmony, rhythm, volume, . . .) that are at least partly independent from each other. Nevertheless, they all are or can be of relevance for the level of musical complexity. For example one music track can contain very sophisticated rhythm patterns, but no melodic voice at all. Another one might have unexpected changes in volume and timbre, but very straightforward melody and chord sequences. Which one should be rated more complex? Looking at the problem this way calls for a multi-faceted approach with an individual complexity estimation for each facet, as we already proposed in chapter 2. A joint measure of global complexity does not necessarily make sense under these circumstances.

The terms "unusual", "irregular", and "unexpected" again lead to context dependency and subjectivity. Whether an event is expected or unexpected is not completely immanent in the event itself. In extreme cases both could be true, depending on who we ask. This is of course not very encouraging for somebody who is supposed to develop a computer model that gets only the event description (in our case obtained from the audio data) as an input. It is obvious that not everybody will share the same opinion about the complexity of a particular piece of music, depending on his or her individual musical knowledge and the familiarity with special musical styles. However, if we restrict ourselves – as proposed in chapter 2 – to only deal with individuals from a certain cultural heritage (i. e. eliminating influences from different tuning or scale systems, etc.) we can still identify something like a musical "common sense". It can be understood for example as the ability to judge whether a singer is singing in tune or not, to clap on the beat, or to distinguish consonant from dissonant chords. Neurological experiments for example by Barbara Tillmann et al. proved the existence of such effects (e. g. in [TBB00]). Tillmann refers to the cause as *implicit learning of music by mere exposure*. So simply by frequent exposure to music in their everyday life humans unconsciously learn certain immanent regularities or rules of the music. Violation of these rules then increases the effort a listener has to put into "decoding" or processing the music and thus increases the perceived complexity according to our definition.

At this point there is also a link to Gestalt theory, which assumes the existence of certain universal principles like proximity, continuation, closure, etc. that are supposed to be "hardwired" in our perception (see e. g. [Bre90]). These principles have very similar effects as the implicitly learned rules, but they are given to us already by birth and hence do not have to be induced through frequent exposure to stimuli of a certain type. While learned rules are stored in long-term memory, the gestalt principles are operational directly in short-term memory due to the organization and design of our "operating system". This implies that the Gestalt principles do not even depend on cultural background and environment, but form – in the widest sense of the word – a "common sense" for the perception of stimuli. For example the implication-realization model for melodic complexity by Narmour [Nar90], which will be reviewed in section 4.2, is built on these principles.

We want to consider this "common sense" or implicit knowledge as the basis for our complexity estimations. So rather than adapting to individual peculiarities of single music listeners we want to provide a general model, which is able to make predictions on a large scale that everybody can agree on with a small amount of error tolerance. It is however very interesting to follow the thought of a learning complexity agent system adapting to its owner's preferences and advancing in its estimations along with him. We will get back to this idea in chapter 6.

3.2 Applicability of Music Complexity

We now explained what we refer to with the term "music complexity". After these, up to this point, quite academic considerations we will now turn to the practical motivation of the described concept. Why should musical complexity descriptors be particularly interesting for the normal music consumer?

The answer has two parts. First, it is not too far fetched that certain facets of complexity might be directly relevant for the listener. For example if I am interested in finding danceable music for a party, the rhythmic complexity already provides a useful parameter for my search. Or if I am looking for "easy listening" music, I might restrict my search to tracks at the lower end of the complexity scale on one or on several dimensions.

For the second part of the answer we have to go back to the year 1971 where we find a publication by Daniel Berlyne [Ber71]. In this publication he states that an individual's preference for a certain piece of music is related to the amount of activity it produces in the listener's brain, to which he refers as the *arousal potential*. According to this theory there is an optimal arousal potential that causes the maximum liking, while a too low as well as a too high arousal potential result in a decrease of liking. He illustrates this behaviour by an inverted U-shaped curve (see figure 3.1) which was originally introduced in the 19th century already by Wundt [Wun74] to display the interrelation between pleasure and stimulus intensity.

Berlyne identifies three different categories of variables affecting arousal (see [Ber71] for details). As the most significant he regards the *collative variables*, containing among others complexity, novelty/familiarity, and surprise effect of the stimulus. Since we are intending to model exactly these aspects of music with our descriptor, it is supposed to be very well suited for reflecting the potential liking of a certain piece of music. We will return to this topic and review several related experiments in section 4.1.

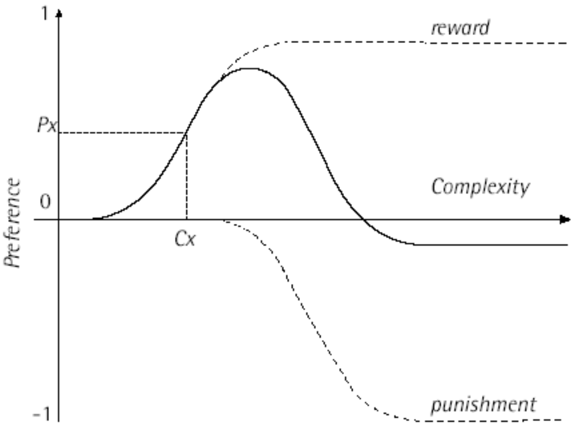


Figure 3.1: The Wundt curve for the relation between music complexity and preference.

This said we can now flesh out the motivations for the use of semantic descriptors given in section 1.2.3. When a user is interacting with a music database or a music collection three major tasks can be identified:

1. Providing an appropriate interface for navigation and exploration.
2. The generation of a program (playlist) based on the user's input.
3. The retrieval of songs that match the user's desires.

We can identify applications of complexity descriptors in all three tasks, which is discussed in the three following sections.

3.2.1 Enhanced Visualization and Browsing

For smooth navigation and exploration of databases a well-designed visualization of the contents is a crucial condition. This is a very difficult task when it comes to large amounts of complex data like music tracks. One example for such visualization is the Islands of Music application, developed by Elias Pampalk [Pam01]. This application uses the metaphor of islands and sea to display the similarity of songs in a collection. Similar songs are grouped together and represented by an island, while dissimilar ones are separated from them through “the sea”. The application uses features that are motivated from psychoacoustic insights, and processes them through a self-organizing map (SOM). In order to compute similarity between songs the sequence of instantaneous descriptor values extracted from each song has to be shrunken down to one number. Pampalk does this by taking the median. He reports satisfying results, but at the same time states that the median is not a good representation for songs with changing properties (e. g. bimodal feature distributions).

Here, the complexity descriptors have a clear advantage, because by default they consist only of one number which represents the whole track and thus do not need to be further reduced by basic statistical measures, like the mean or the median. Obviously, complexity represents a self-contained concept and is not intended to from an alternative to the use of such measures. As pointed out in the beginning of this section, the different complexity descriptors reflect specific characteristics of the music that are potentially of direct relevance for the listener. The descriptors are therefore very well suited to facilitate the visualization of musical properties the user might want to explore. It is straightforward to plot the whole collection in a plane showing for example rhythmic complexity versus loudness complexity without the need for specifying a similarity metric (see also figure 1.1 on page 6).

Another aspect is the possibility of a more “musicological” way of interaction with a music collection. By providing the link between the actual audio and the musical content description a user might increase his knowledge about the music in his own or a different collection. Common properties of music from different artists or different genres might be discovered. The changes in musical characteristics over time for a particular band can be made visible. Also here the complexity descriptors form an interesting addition, which opens new opportunities that are still to be explored.

3.2.2 Playlist Generation

A playlist is a list of titles to be played like a musical program. A user interacting with a database might ask for the automated generation of such a list. As Pachet, Roy, and Cazaly point out in [PRC00] the creation of such a list has to be taken serious, since “[t]he craft of music programming is precisely to build coherent sequences, rather than just select individual titles.” A first step towards coherence is to set certain criteria the songs have to fulfill in order to be grouped into one playlist. The user could be asked to provide a seed song for the playlist and the computer would try to find tracks from the database which have similar descriptor values. Pachet, Roy, and Cazaly go further and look at an even more advanced way of playlist generation capturing the two contradictory aspects of repetition and surprise. Listeners have a desire for both, as they state, since constant repetition of already known songs will cause boredom, but permanent surprise by unknown songs will probably cause stress. In their experiments Pachet, Roy, and Cazaly use a hand edited database containing, among others, attributes like type of melody or music setup. We can see a correspondence here to melodic and timbral complexity, that encourages the utilization of complexity descriptors for playlist generation.

An alternative way of playlist generation, which gives more control to the user, is that of using user specified high-level concepts as for example *party-music* or *music for workout*. Inside the SIMAC project methods are explored to arrive at such a functionality. A playlist could then be easily compiled by selecting tracks with the according label. The bottleneck here is the labelling of the tracks, which might be a lot of work in a big collection. Since the labels are personalized and may only have validity for the user who invented them, there is no way to obtain them from a centralized metadata service. Instead the user can try to train the system to automatically classify his tracks and assigning the personalized labels. For this process semantic descriptors are needed that help in distinguishing whether a track should be assigned a certain label or not. It depends of course very much on the nature of the label to identify descriptors that are significant for this distinction. In any case, the complexity descriptors certainly have a potential to be useful here, as can be seen from the examples at the beginning of this section.

3.2.3 Song Retrieval

For song retrieval there are different possibilities in a music database. The most obvious one is the direct specification of parameters by the user. Since the complexity descriptors consist of only one value per track, they can be used very easily in queries. The user can specify constraints only for those facets he is interested in and narrow down the set of results. This way it is very straightforward to find music that, for example, does not change much in loudness level over time, or contains sophisticated chord patterns.

A second way of querying is the so called *query-by-example* approach. The user presents one or several songs to the database and wants to find similar ones. So, as explained for the visualization using similarity measures, here the complexity descriptors can easily be integrated into the computation again. The weighting and/or the tolerance for the different descriptors could be specified by the user directly, extracted from the provided example, or taken from a pre-computed user profile. Such a user profile would be established by monitoring the user's listening habits (i. e. songs he/she has in his/her collection; songs he/she listens to very frequently, etc.) as in the *recommender* application developed in the SIMAC project.

Finally, we can think of a *music recommender* that does not even need an example song. If the user's common listening behaviour is known it should be possible to establish a "complexity profile". This can be understood for example as a histogram for the complexity values where either the number of tracks or the listening frequency is monitored. From such a histogram it should be possible to identify the user's optimal complexity level in Berlyne's sense (see figure 3.1 on page 17). Tracks matching this level could then be selected as recommendations for the user imitating the recommendations of friends with a similar musical taste. It should be stated that complexity is of course not the only criteria that should be used here, but according to Berlyne and others plays an important role for potential preference. Some experiments on this relationship are reviewed in section 4.1.

Chapter 4

Review of Former Work

After the subject of our research has been explained and specified, and the intended applications have been described we will now have a closer look at the work that has already been done in this area. We will first review several experimental findings in the context of Berlyne's theory of arousal potential. Then we address different complexity facets where former work is already available. These are melodic, rhythmic, and harmonic complexities. Finally, we take brief look at two alternative concepts of complexity.

4.1 The Preferred Level of Complexity

We pointed out in section 3.2 that according to Berlyne's theory of arousal potential [Ber71] the level of perceived complexity of a piece of music can be associated with the preference for it. Our research is not intending to provide evidence for the validity of this theory. Rather than exploiting the psychological aspects between perceived complexity and preference, we focus on the development of a computational model of complexity for music. However, the former aspect plays an important role in the motivation for our research. Hence, we want to give some room here to report about studies that have been published on the matter.

4.1.1 Psychological Experiments

Berlyne himself conducted several studies during the 1960s and 1970s ([Ber60], [Ber71], and [Ber74]) on the connection between arousal potential and hedonic value (liking) of artistic stimuli. He has considered not only auditory, but also visual and combined stimuli in his experiments. He has found strong evidence for the existence of the inverted-U relationship as depicted in figure 3.1 on page 17 which led to his theory of arousal potential. As mentioned above he identified different types of variables to contribute to arousal potential ([Ber60] pp. 170–179):

- Intensive Variables (e. g. pitch, color, energy)
- Affective Variables (e. g. anxiety, pain, gratification)
- Collative Variables (e. g. novelty, surprisingness, complexity)

Berlyne considered the collative variables the most important ones of the three types.

With respect to music complexity in particular, Heyduk in [Hey75] reports about an experiment with four custom-built piano compositions of different complexity levels. Complexity was varied in two facets: chord structure and syncopation. He found strong evidence for Berlyne's theory by analyzing the ratings for preference and complexity given by a group of 120 subjects. While Berlyne used the term *arousal potential* for the combination of different factors determining preference, Heyduk follows the terminology of Walker [Wal73] and talks about *psychological complexity*. It is important to note the distinction to a pure stimulus complexity here. The latter would be fixed and objective, very much like the Kolmogorov complexity mentioned in section 3.1.1. The former includes the latter, but is also determined by attributes like novelty, uncertainty and arousal properties. It is a subjective property, because it only becomes manifest in the encounter of an individual with the stimulus.

Steck and Machotka [SM75] conducted an experiment using random sequences of sinusoidal tones as stimuli. All tones were equally loud and had equal duration within one "composition" of approximately 10s length. 16 different levels of complexity were constructed by varying the tone duration (and thus the tone rate) from 2s down to 60ms. The analysis of preference ratings from 60 subjects revealed a clear inverted-U relationship on the objective complexity levels. However, when presenting only subsets of the test samples which were taken from adjacent complexity levels, again an inverted-U relationship was found within each subset. Even more interesting is their observation that the relative position of maximal preference inside each subset was fixed. That means the preferred level of complexity was not absolute, but relative for the presented stimuli.

Hargreaves and North point to two potential problems with respect to Berlyne's theory: the influence of the listener's mood and intention when selecting music, and the dependence on the appropriateness of the music for the listening situation [NH97a]. It seems obvious that the same listener will prefer different types of music whether he is driving a car, relaxing on the sofa, or dancing in the club. So when they asked subjects to rate the preference for the music in a cafeteria, they indeed found an inverted-U shaped relation between complexity and preference. However, the effects were mediated by musical style. Organ music of the same complexity level as New Age music was liked less in this listening situation. In another study [NH97b] they show that preference and pleasantness of music have to be distinguished. Preference and arousal potential were again found to relate through the inverted-U curve in this study. But North and Hargreaves argue that a subject with an antipathy against brass music for instance will not be pleased by this music, whether it matches the optimal complexity level or not. In this sense, optimal complexity only provides the potential for a maximum of pleasure a subject might experience, but does not determine it completely. Conversely, a subject might find pleasure in listening to music that does not possess the optimal level of complexity. However, pleasure would reach a maximum when the right level of complexity and a general liking coincide in a piece of music.

A quite recent study by Orr and Ohlsson [OO05] addressed the dependency of musical expertise on preference. They used natural stimuli, bluegrass and jazz solos, at different complexity levels, which were purposely performed and recorded for their experiments. Four different groups of subjects were asked to rate perceived complexity and liking for the stimuli. One group consisted of subjects with no musical training, a second one was composed of subjects with moderate training in music, the third and fourth group consisted of professional jazz and bluegrass musicians. The results reported by Orr and Ohlsson indicate two things. First, it seems that the significance of complexity for preference decreases with increasing musical expertise, unless complexity itself is learned as an aesthetic criterion. This can be seen from the fact that for the group of professional jazz musicians an inverted-U relationship could not be identified. For the professional bluegrass musicians the inverted-U relationship only appeared for their ratings of bluegrass music. The authors interpret this effect insofar as complexity might represent an explicit aesthetic criterion in bluegrass music, which has been learned by the professionals. The group of untrained subjects was the one where the inverted-U became apparent the most for either musical style. The moderately trained group revealed this effect only in case of the bluegrass samples. So secondly, the importance of optimal complexity seems also to depend on the music style. We have reviewed so far only psychological studies using selected stimuli where the complexity was purposely varied in a controlled test environment. It is also interesting to consider studies that where conceived the other way around, that means observing preference indicators of real music and then relating these with complexity estimations.

4.1.2 Musicological Studies

Eerola and North [EN00] report about their analysis of 128 Beatles songs, all written by and for the Beatles in the years between 1962 and 1970. From MIDI-transcriptions of the songs they extracted the melodies and analyzed the melodic complexity with their expectancy-based model (see section 4.2). A highly significant increasing trend was found for melodic complexity over the time period. Secondly, the authors compared the complexity values with indicators of commercial success of the songs and albums. The chart position and the time in the charts were both negatively correlated with melodic complexity. So the higher the complexity, the less popular were the songs. Although this is a clear sign of relevance between complexity and popularity, the authors point out that other factors of social, cultural, or commercial kind certainly have an influence as well. Since they were not considered in the study, care has to be taken in drawing conclusions. However, they make reference to Simonton [Sim94a], who also found a clear connection between melodic complexity and popularity. His results stem from an extensive study of 15,618 melodic themes of classical music.

A study by Parry [Par04] comes to similar conclusions. He analyzed the melodic and rhythmic complexity of 10 songs that were listed in the Billboard Modern Rock Top 40 within the period from January to June of 1996. The complexity was estimated using MIDI transcriptions of the songs and is based on a very basic self-similarity measure. As indicators for the chart performance the number of weeks in the charts, average weekly change in rank, peak ranking, and debut ranking were considered. Parry found the number of weeks in the charts being positively correlated with both, rhythmic and melodic complexity. The former was also positively correlated with the peak ranking. For the average change in rank a negative correlation was found with melodic complexity, indicating that higher melodic complexity inhibited rapid changes. The debut ranking revealed no statistically significant correlation with the two complexity measures.

4.1.3 Conclusion

As a conclusion we can say that a certain relationship between preference and complexity of music cannot be denied. Based on the rather scarce evidence the correlation, however, seems to be not as simple as we could guess from Berlyne's theory. Other factors, as mentioned, also show effects and might under certain circumstances completely overrule the influence of complexity on preference. It should therefore not be expected to have found the holy grail of music recommendation with the usage of complexity descriptors. On the other hand, the reported findings clearly prove the relevance of complexity in music listening, especially for non-expert listeners. Providing complexity descriptors for music therefore should be able to enhance human interaction with music collections, which is the goal of the research presented here.

4.2 Existing Models for Melodic Complexity

4.2.1 The Implication-Realization Model

Already back in 1990 Eugene Narmour proposed a model for melodic complexity. This *Implication-Realization* model, as he calls it, is extensively described in [Nar90]. The model hierarchically builds up larger structures from smaller elements and thus possesses different levels. The predictions are most clearly specified on the lowest level, which is the tone-to-tone level. Any melodic interval that is perceived as being "open" (incomplete sounding) is said to create an *implication* on the listener's side. That means it raises certain expectations in the listener about how the sequence is supposed to continue. Figure 4.1 illus-

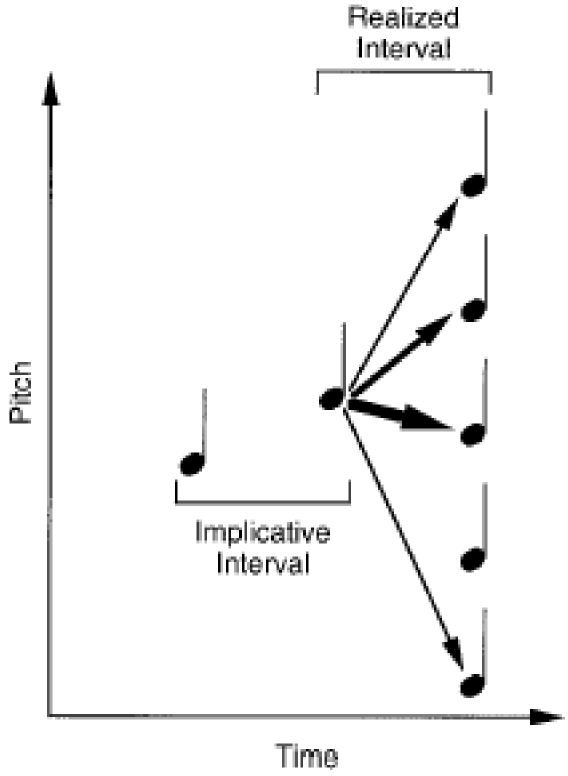


Figure 4.1: Illustration of the Implication-Realization Process (from [SAPM02]).

trates this process for the tone-to-tone level. The first interval implies a second one to follow, while the relative thickness of the arrows indicates that different continuations can be implied with different strength. Factors that contribute to closure (i. e. the opposite of openness) at the single-interval (tone-to-tone) level are [SAPM02]

- a longer duration of the second tone than the first one (e. g. eighth note followed by quarter note)
- a higher stability of the second tone in the established musical key (e. g. *ti* followed by *do*)
- a stronger metrical emphasis of the second tone (e. g. at the first beat of a measure).

Frequent realization of these expectations reveals a low level of complexity; frequent disappointment reveals a high level of complexity, because the structure of the melody is hard to decode. It has to be pointed out that the development of this model was guided by gestalt principles only, implying validity completely independent from the listener's cultural or musical background. Narmour himself states about the principles that they are "innate, hardwired, bottom-up, brute, automatic, subconscious, panstylistic, and resistant to learning" (as cited by [SAPM02]).

4.2.2 Enhancements of the Model

The IR-model inspired many experiments. It was reduced by Schellenberg to its basic principles. He proved experimentally that his simplified version had the same explanatory power as the original [Sch96]. Eerola and North report in [EN00] about an experiment where they assessed an enhanced version of the IR-model. They point out that the traditional information theorist view of complexity does "not address the role of the listener's perceptual system in organising the structural characteristics of music". Therefore they propose an expectancy-based model (EBM) to estimate complexity. Their model for melodic complexity is also inspired by Narmour, but they include some additional features derived from a symbolic representation of the music. In their evaluation they found that *tonality* (modified by metrical position and tone duration), the intervallic principles *registral direction* and *intervallic difference*, and the rhythmic principles *syncopation* and *rhythmic variability* showed significant capability in predicting listeners' complexity judgements. Comparing the prediction accuracy of this model with an information-theoretic and a transition probability model([Sim94b]), they found it to be the best one.

Eerola, Toiviainen, and Krumhansl conducted experiments showing, how the accuracy of the model can be further improved by taking a larger melodic context into account [ETK02]. Since the focus of the original model is limited to two notes at a time only, it neglects the impact of the longer-term melodic evolution (e. g. repetition of motives) on the listeners' predictions of continuation. However, their modifications were aiming more towards a real-time modelling of listeners' continuation predictions than towards a more accurate estimation of the overall complexity.

4.2.3 Conclusion

As a conclusion we can say that the problem of melodic complexity is mostly solved already. With the extended model from Eerola we already have an automatic complexity estimation for melodies that fits our needs. Eerola even made the implementation of his and several other models available in the MIDI toolbox for MATLAB [ET04]. However, there is one problem here. The model, as well as all the other variants presented in this section, requires an accurate symbolic representation of the melody in order to arrive at valid results. Usually, digital music files do not have this symbolic description attached to them. The key problem thus remains in the automatic extraction of the melody from the audio stream in order to apply the existing complexity model in a second step to the transcription. But current methods are still far from reaching such transcriptions from arbitrary music tracks with a sufficient accuracy. The best algorithm for melodic pitch tracking in polyphonic signals in 2004 ISMIR contest achieved only an average accuracy of 75% [GSO$^+$05]. This figure does not even include note segmentation. However, only a few wrong notes in the melody completely mislead the complexity estimation. So unfortunately it is currently not possible to apply the existing models directly in the application context of our research.

4.3 Measures for Rhythmic Complexity

4.3.1 The PS-Measure

Shmulevich and Povel introduced in [SP00] the PS-measure for rhythmic complexity. They state, with reference to [Ess95], that a listener tries to establish an internal clock, when hearing rhythmic music. According to this clock the listener then segments the rhythm pattern and tries to code the segments. So a rhythmic pattern will appear complex, when either the induction of the internal clock is weak or absent, or when the coding of the segments is difficult. So, the algorithm combines the two aspects for the complexity estimation.

The clock induction strength is a weighted combination of the number of clock ticks that coincide with silence and with unaccented events. This shows already, that a rather high abstraction level is assumed here, since the clock grid is assumed to be known and the musical events have to be classified into accented, non-accented, and silent ones.

The coding complexity is estimated by assigning different weights to four different types of segments that can appear when splitting up the sequence of events according to the clock grid. The segment types are *empty segments*, *equally subdivided segments*, *unequally subdivided segments*, and *segments beginning with silence*. An additional weight exists that is added, when two consecutive segments differ from each other. The coding complexity is then simply the sum of all the weights for the whole sequence.

It must be noted that the clock induction strength and the coding complexity when computed this way tend to give higher complexity ratings to longer segments, because there is no normalization. This might be appropriate when isolated rhythm patterns are evaluated, since a longer pattern is supposed to have a higher complexity potential than a short one. However, for complete music tracks some adaptation would be in need. Although rhythm pattern extraction might seem an easier task to solve compared with melody extraction, to date universal methods providing sufficient precision are still lacking (a recent review can be found in [GD05]). The problem is that even the correct localization of rhythmic events, which is already a challenge, still does not suffice for the application of the PS-measure. In addition these events have to be seen in relation to the rhythm grid, which also has to be established first from the audio file.

4.3.2 Danceability

Recently, Jennings et al. published an article on musical genre classification [JIM$^+$04]. What makes this article stand out is the fact that the authors used, exclusively, a feature which has not been considered in music information retrieval research so far. They refer to it as the “Detrended Variance Fluctuation Exponent”, since it originates from a technique called “Detrended Fluctuation Analysis” (DFA). The technique was introduced by Peng et al. [PBH$^+$94] and is intended for the analysis of chaotic time series. It was first applied by Peng in a biomedical context on nucleotide and heartbeat rate time series. Other applications include financial time series (as reported in [Aus00]).

The method intends to identify the *scaling behaviour* of a time series. This is best explained by considering an unbounded time series (i. e. with no upper limit for its values). If we plot a portion of such a series over the time axis, we will need a certain range r_1 of values on the y-axis in order to visualize all the elements falling into the temporal segment t_1. Imagine now that we are zooming in on the plot. This zoom can be expressed as a factor on each axis, for example we can say that the new portion has 0.5 times the length of the previous one, such that $\tau_2 = \tau_1/2$. In this case we can expect that the range of values on the y-axis has changed as well, from r_1 to r_2. By taking the quotient $\frac{\log(r_1/r_2)}{\log(\tau_1/\tau_2)}$ we obtain a scaling exponent α. This exponent can be calculated for different time scales (i. e. for different values of τ_1 and τ_2). If we systematically increase or decrease the time scales and for each step calculate the scaling exponent, we obtain a function $\alpha(\tau)$. From this function we can get an insight into the scaling properties of the time series. For example the scaling can be stable (constant level of α) or it might expose irregularities. Also the general level of α is significant. For white noise it is 0.5, for pink noise it is 1.0, and for brown noise it is 1.5. α levels below 0.5 indicate that the signal is anti-correlated ([Mat01] pp. 41–43).

However original the descriptor of Jennings et al. might be, it is not obvious how it relates to the rhythmic complexity we are interested in. Since their paper focusses on genre classification this aspect was not directly considered. But they state that the strong periodic trends in dance music (as Techno or Brazilian Forró) make it easily distinguishable from high art music by using this feature, because the average α value is clearly lower for the former. “Jazz, Rock and Roll, and Brazilian popular music may occupy an intermediary position between high art music and dance music: complex enough to listen to, but periodic enough to dance to,” they speculate [JIM+04]. Hence, we could think of this feature as a measure of “danceability”, which certainly is an aspect of rhythmic complexity. In the next chapter we will look at further details about first experiments with this descriptor.

4.3.3 Conclusion

With the PS-measure we are facing once more the problem that an existing model for complexity cannot directly be utilized for the intended applications. As with the melodic complexity models already, we again are left with incompatible representations of the music. Namely, the symbolic, abstract rhythm pattern is needed as an input for the model, and the digitized waveform is what we have at hand in the audio file.

The DFA exponent on the other hand is a feature that can directly be applied to the audio file and does not rely on the extraction of other mid- or high-level descriptors as a preprocessing step. Although its validity has not been proved and it certainly does not cover all aspects of rhythmic complexity, it appears to be an interesting and very feasible approach.

4.4 Harmonic Complexity

4.4.1 Preference Rule System

The theory of harmony in music has a very long tradition already. Still, in contrast to melodic complexity, no dominant and well-studied models for harmonic complexity could be identified by the author. Research has been done on the expectations evoked in listeners by harmonic progressions especially on the field of classical music [Sch89]. It turned out, that listeners usually predict a chord that results from a transition considered as common in the given musical context. Yet, to our knowledge no tests have been carried out that correlated the perceived harmonic complexity with the fulfilment or disappointment of these expectations. Temperley supposes that the scores computed with his preference rule system could reveal an estimate for the tension in music ([Tem01] section 11.5, pp. 307–317). The mapping of achieved scores would go from *incomprehensible* (breaking all rules) over *tense* to *calm*, and finally to *boring* (all rules obeyed). Although he does not use the term complexity, this basically reflects what we are looking for. He names four different aspects of this harmonic complexity:

1. The rate at which harmonies change.

2. The amount of harmonic changes on weak beats.

3. The amount of dissonant (ornamental) notes.

4. The distance of consecutive harmonies in a music theoretical sense.

However, his system does not only need a transcription of the chords, it assumes also that the metrical grid is known. So since this high level information is not available in our case, the direct application of Temperley's system is not feasible.

Furthermore, the fourth point has to be addressed carefully when extending the scope from classical music to the different types of modern popular music. Temperley himself considers this tension estimation to be consistent only for particular styles (see [Tem01] figure 11.11 p. 314). He also refers to Lerdahl, who in [Ler96] proposed an even more high-level approach to the estimation of tension taking also larger structural levels into account. But since the application of Temperley's model is already problematic with musical audio signals, Lerdahl's proposal appears even less feasible.

4.4.2 Rewriting Rules

A different approach is taken by Pachet, who proposes the application of rewriting rules [Pac99]. He addresses the effect of harmonic surprise in Jazz music. His argument is that the "rich algebraic structure underlying Jazz chord sequences" has to be considered when talking about expectation and surprise.

His model is based on two ingredients, a set of *typical patterns*, which he relates to the characteristics of the musical style, and a set of *rewriting rules*, according to which a given chord sequence can be transformed. The basic idea is that even a chord sequence that has never been heard before might not be so surprising to a listener who is familiar with these two sets that are sufficient to generate it.

Instead of manually creating the two sets, Pachet applies machine learning techniques in order to obtain them. He uses a string compression algorithm [ZL78] in order to extract the typical patterns. In order to find the rewriting rules he then uses a table method and utilizes the likelihood of occurrence of a hypothesized rule as a basis for finding the best rewriting rules.

Although he is not directly aiming at a complexity estimation, the idea is not too far from our needs. A highly predictable (little surprising) chord sequence could be identified with a low harmonic complexity and vice versa.

4.4.3 Conclusion

The two ideas for harmonic complexity estimation presented above have two drawbacks. First, as with the melodic complexity models already we here again need a transcription of the chords first before we can start to analyze the complexity. We can find many approaches towards chord extraction from musical audio (see e. g. [Kla04], [Bel03]), but so far there are no satisfying solutions. Secondly, either one of the cited methods seems to be appropriate only for a limited set of musical styles, the Preference Rule System more for classical music with traditional rules from harmony theory, the Rewriting Rules more for Jazz music where a diverse set of chord extensions is very common. So none of them is optimal as a general approach to harmonic complexity, because we want to deal with many different musical styles at the same time.

4.5 Other Related Work

In this section we want to take a short look at other approaches to music complexity that take a slightly different angle in addressing the topic.

4.5.1 Information-based Complexity

Pressing in [Pre98] gives a very short outline of three different types of complexity, which he names *hierarchical*, *adaptive*, and *generative* or *Information-based* complexity. Referring to music the first is focussing on the structure of a composition. Pressing mentions Johann Sebastian Bach's *Kunst der Fuge* as an example where the composer exploits hierarchical complexity. "[N]otes function as elements of linear processes operating at different time scales, and are at the same time compatible with a vertically-oriented chordal progression process." The second type of complexity refers to the temporal changes in a musical work, including aspects like the adaption to unpredictable conditions, or the anticipation of changes in themselves or in the environment. Here, Pressing mentions improvisatory performance as an example.

The third type, the *Information-based complexity*, is elaborated in some more detail by Pressing. It is inspired by Kolmogorov complexity, but focuses more on the production of music through a human than on the generation of a string through a computer program. Pressing acknowledges that a pure information theoretic measure falls short in measuring music complexity, because it will always rate random sequences as the most complex ones, which does not go along with human perception. He argues that humans process complexity by developing routines and heuristics that try to overcome the limitations of memory and attention. His approach to complexity is based on the concept that these routines have different difficulty levels, because some are easier to learn than others. So conversely, the complexity of a stimulus is determined by the difficulty assigned to the routines and heuristics that are needed to produce it.

He demonstrates his concept by estimating the complexity of six very short rhythmical patterns. This is achieved by simply applying a processing cost function (cognitive cost) to the symbolic level attribute *syncopation* on quarter-note and eight-note level. His approach shows some similarity with the coding complexity used in the PS-measure for rhythmic complexity (section 4.3.1). Again, the sequence of events is broken up into segments (this time according to two different grids) and a cost is assigned to each segment depending on the type of syncopation that can be found.

This approach is very interesting, because it seems a convincing combination of information theoretic principles and cognitive effects. However, it is purely theoretical and was not evaluated in any experiment with humans. Also there is a crucial point in the identification of the routines and heuristics needed to generate the stimulus, as well as in the assessment of cognitive costs to them. In the case of a complex musical audio signal this is a very hard task to solve.

4.5.2 Complexity of Short Musical Excerpts

In [SWV00] Scheirer directly utilizes the statistical properties of psychoacoustic features of short musical excerpts to model perceived complexity. The excerpts were of only 5 s length, so abstract levels, like melodic complexity or even structural complexity, are not accessible. Consequentially, Scheirer only considers a joint complexity of the excerpts and does not address individual facets.

The following low-level features were identified by him experimentally as the most useful ones for the prediction of human complexity ratings:

1. coherence of spectral assignment to auditory streams

2. variance of number of auditory streams

3. loudness of the loudest moment

4. most-likely tempo

5. variance of time between beats

The computation of these features is extensively described in [Sch00] (chapters 4-6). Scheirer compared the mean complexity ratings of a group of 30 human listeners to the output of his descriptors. He reports that, by using linear regression techniques on these, they were strongly significant in predicting the rated complexities, roughly 20% of the variance in the mean complexity ratings was explained by these five features ($R^2 = 0.186$, $p < 0.001$).

If we compare this to our idea of different complexity facets, features one and two can be assigned to timbral complexity, feature three to acoustic complexity, and features four and five to rhythmic complexity. However, it is questionable whether this approach is still successful when complete tracks are considered rather than only short excerpts. As Scheirer states himself, he is more concerned about complexity as a *musical surface* feature. He defines the musical surface to be "the set of representations and processes that result from immediate, preconscious, perceptual organization of a acoustic musical stimulus and that enable a behavioral response". This musical surface might then help in timbre or genre classification tasks.

4.6 Summary

In this chapter we have seen some evidence for the usefulness of complexity descriptors for music. However, to develop a fully operational set of algorithms for automatic complexity estimation from musical audio signals there are still many open issues. While for the melodic and the rhythmic facet tested models exist in the symbolic domain, they cannot be applied directly due to the lack of reliable extraction algorithms. For harmonic complexity only theoretical suggestions were found, while the timbral, the acoustic, and the structural facet have not been addressed at all so far. In the following two chapters we will therefore proceed with a complete set of operational complexity definitions, a report on first results with practical implementations, and the plans for future steps towards a satisfying solution.

Chapter 5

Own Approaches

In this chapter some of the author's own work on the topic is described. Several of the approaches in section 5.1 have not yet been converted into fully functional implementations, but are presented as a guideline for the direction of future research. Hence, there is some overlap with the next chapter, which is explicitly dedicated to Future Work.

5.1 Operational Definitions

5.1.1 Acoustic Complexity

Under *Acoustic Complexity* we want to capture two different aspects of a musical audio track: the dynamic and the spatial properties. The former are related to the loudness evolution throughout a musical recording, while the latter correspond to the rendering of the auditory scene. From these explanations it is clear already that Acoustic Complexity is not completely intrinsic to the music, but rather to the recording and the performance. Nevertheless, we found it worthwhile to include this complexity facet, because digital music tracks only exist as recorded performances[1] of music. So it is not possible to listen to one without noticing characteristics of the other.

[1] For electronic music the concepts of "recording" and "performing" have to be understood in a slightly wider sense here.

Dynamic Component

The dynamic complexity component relates to the properties of the loudness evolution within a musical track. We refer to it in terms of *abruptness* and *frequency of changes* in dynamic level. Also the *dynamic range* has to be considered here, since a sudden jump by 6dB will be considered more surprising than one by only 1dB. A major design decision for the complexity model is the definition of the time scope. By keeping the frame size small one would find the distinction between dynamically compressed and uncompressed material. The former would have less rapid changes in loudness than the latter and would therefore be considered less complex. With longer windows one could detect fades and dynamic changes between larger segments. The regularity of dynamic changes has to be observed as well. An uncompressed drumloop for example will have many abrupt changes in short-term dynamic level due to the sharp attacks in front of a very low background loudness level. But because of its periodicity these will not be found very complex compared to random (and thus unpredictable) drum hits, if we expect the “good continuation” gestalt principle to hold.

The first step towards dynamic complexity computation is the calculation of an accurate estimate of the perceived loudness. Psychoacoustics are a well studied field and quite reliable measures exist in the literature for loudness estimation of isolated sounds, like sinusoids or noise ([ZF90],[MGB97]). For the complex sounds that form a musical performance however it is a very complicated task. Not only temporal and spectral masking have an influence here, but also subjective components play a role [SN04]. Finally, the playback level of a digital track can not be known, so the loudness can only be approximated.

Spatial Component

For the spatial complexity component we consider only stereo recordings and no advanced multi-channel formats. Currently, two channel stereo tracks form by far the majority of items in digital music file databases. A straight-forward example for spatial complexity thus is the *disparity of the stereo channels.* A quasi mono situation with similar channels reveals less complexity than a recording that has only little correlation between the two channels. But also more advanced aspects are to be considered, such as the movement of the acoustical center of effect within the stereo image, and sound effects changing the spatial impression (e. g. delay and reverberation effects).

Yet, from the computational point of view, these advanced properties are not trivial to measure. While methods exist to calculate the position of a single sound source in space when it is recorded by an array of microphones [SYSPI05], the identification of the center of effect in the stereo panorama is a somewhat different problem. Since usually complex mixtures of sounds are involved in each channel, some kind of source separation technique would need to be applied, but the current state of the art in auditory scene analysis still does not provide the necessary tools for that.

The measurement of reverberation on the other hand has a solid tradition in room acoustics. An overview over several measures of spaciousness in room acoustics can be found in [Gri99]. Usually, these measures take the room impulse response as their input and are thus not suited for a continuous signal. An exception is the InterAural Difference (IAD) introduced by Griesinger, which, as he states, can also be computed as a continuous function for music signals. It is computed according to equation 5.1, where $L(t)$ and $R(t)$ refer to the signal in the left and the right channel, and the equalization eq consists of a low frequency enhancement of 6dB per octave below 300Hz.

$$IAD = 10 \cdot \log_{10} \left(\frac{eq(L(t) - R(t))^2}{L(t)^2 + R(t)^2} \right) \tag{5.1}$$

Still, it must be stated again that the originally intended use of these measures is somewhat different from our needs. The measures are supposed to reveal information about the acoustic properties of a real room when a sound is played back inside. The two channels in equation 5.1 correspond to the signal recorded by two microphones (e. g. inside an artificial head) in the room. We, on the contrary, already have a stereo recording, which was produced with very different techniques, but would simply have to treat it as the recording of the dummy head in a "virtual" room. An extensive discussion of this problem and some experimental results can be found in [Mas02].

5.1.2 Timbral Complexity

There is no clear and precise definition of timbre that could be regarded as a common agreement on the music analysis field. By the American Standards Association ([Ass60] p. 45) the following statement was released: "[Timbre is] that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar." For our purpose we think of timbre as the entity that is the most tightly knitted with sound production (i. e. the source of the sound and the way this source is excited). However, this concept of source should be considered not in a strictly physical sense here. In the case of a group of 20 violinists playing unison for example we would rather refer to the group as the sound source rather than to every individual violin.

We then can derive several specifications of the general ideas of complexity itemized in section 3.1.3. This gives us features like the number of distinguishable instruments or sound textures present in the music, the rate at which the leading instruments change, or the amount of modulation of the sound sources. As reported in [HPD03], source separation and instrument recognition systems for arbitrary polyphonic music signals are not yet available. Nevertheless, since the exact classification of individual instruments is not necessary for our purposes, machine-learning and clustering techniques might be applicable. Aucouturier and Sandler report about applying HMMs for music segmentation [AS01] in a manner that might be useful for timbre complexity estimation as well. If a finite set of timbre states can be assigned to the music, it is also possible to apply techniques from information theory like entropy estimation or compression algorithms. For example the LZ77 algorithm by Ziv and Lempel [ZL77] could be a good choice, since its limited memory buffer resembles in a simplified way human short term memory when processing a musical input [Sny00].

5.1.3 Melodic Complexity

Despite the fact that very well studied models are already available for the assessment of melodic complexity based on a symbolic representation, when it comes to musical audio this can be considered the most difficult of the facets. This is due to melody being a very abstract description which is hard to access from the audio signal. As pointed out in section 4.2.3 there is no extraction algorithm available yet, that can reliably transcribe the melody of any musical audio file.

One strategy could be to focus on melody related features that are more feasible to extract and to try then the development of a new model that can approximate melodic complexity based only on these features. A possible candidate would be for example the harmonic pitch class profile (HPCP) [Gom04]. However, this is not a very straight-forward approach, because the HPCP contains information about all tonal content, not only about the melodic voice. Furthermore the octave information is lost and to arrive at something close to a note segmentation is still a long way to go from that representation. A valid solution thus would be to focus instead on the melody extraction task and to work towards a useable melody transcription in order to be able to apply the already available complexity models. Yet, this is a PhD topic by itself and will not be addressed within the scope of the research presented here.

5.1.4 Rhythmic Complexity

For rhythmic complexity we find a similar picture as for the melody facet. With the PS-measure a model exists already that has been assessed at least preliminary. But again, the model operates on an abstraction level which is not easy to achieve when starting from the audio signal. In this case algorithms exist that are already closer to a reliable extraction than in case of melody extraction (e. g. [DBDS03], [Kla03]), but still the results are not perfect.

Despite being a high-level concept, the rhythm is not as abstract as the melody. Therefore for rhythmic complexity estimation it is more promising to find a circumvention for the explicit and complete transcription than in case of melodic complexity. The danceability descriptor based on the detrended fluctuation analysis is one example (section 4.3.2). Otherones are the rhythmic features used by Scheirer for the complexity estimation of short musical excerpts (section 4.5.2). It seems better to consider the variances in the relative positions of extracted rhythmic events than to rely on their absolute position with respect to a metric grid which also has to be determined first. The extraction is already error-prone and the determination of the grid, based on the extracted events, introduces a second source of error. This second error is much more severe for the complexity estimation, since all events are evaluated with respect to the grid. Several wrongly estimated onsets in contrast will not affect the global result so drastically.

5.1.5 Harmonic Complexity

In section 4.4.3 we discussed already the two drawbacks of the presented methods. As with rhythmic complexity it seems more advantageous to explore alternative options that avoid the need for a precise transcription. A possibility would be the application of the HPCP that we mentioned in section 5.1.3. As explained, this feature is strongly related with the harmonic content of the music and therefore an attractive candidate for our purposes. Several approaches towards chord segmentation and recognition rely on this or similar features (e. g. [Fuj99], [SE03]). It is also interesting from the perceptual point of view, because the concept of pitch classes (i. e. the association of frequencies with the twelve pitch chroma steps in western music) is an important step in the process of harmonyic perception.

A common and important problem (among several others) in chord recognition is the detection of the chord boundaries. However, since we are not interested in a transcription, it might be possible to avoid this step in the harmonic complexity estimation. Instead, the harmonic content represented by the HPCP profile could be analyzed continuously. It could be mapped into the spiral array proposed by Chew that defines a three-dimensional space of harmonic instances [Che03]. The spiral array has the property that the spatial proximity reflects also musical proximity of harmonies to some extent. As the song evolves, the path through this space could be recorded and then analyzed. So the average *speed of movement* within this space could be taken as an indicator for the harmonic complexity.

A second aspect is the *clarity* of the harmonies that appear in the music. In the context of key estimation Gomez defines the *tonal strength* through the strength of the correlation between the mean HPCP profile for the track and the closest key profile [GH04]. In a similar manner she uses the correlation of instantaneous HPCP-vectors with prototypical chord profiles for a *chord strength* measure. Both measures are related with our concept of harmonic complexity and could be combined with the spiral array approach.

5.1.6 Structural Complexity

Musical structure forms one of the highest levels of abstraction in content analysis. It is unique compared to the other facets in the sense that all of them are potentially relevant for its computation. We want to refer to structure on a rather macroscopic level (i. e. in terms of intro, verse, and chorus rather than motive or theme).

Along with our remarks in section 3.1.3 we can identify attributes of structural complexity such as the *number of distinguishable parts*, or the *level of periodicity* of their appearance. It would also be desirable to estimate the *dissimilarity* of consecutive parts. Very contrasting parts following each other would be surprising and thus probably enhance the perceived complexity.

Once more, we have to face the fact, that before we can perform any structural complexity processing, first the structure itself has to be extracted. Various approaches to this problem have been taken and are still explored (see e. g. [CV03], [SBM$^+$02]). The general purpose solution has yet to be found.

5.2 Implementations

Following the review of computing different complexity facets of musical audio, in this section two implementations by the author are described in detail. Both have been programmed in MATLAB and were tested on a large music collection. The evaluation results are reported in section 5.3.

5.2.1 The Dynamic Component of Acoustic Complexity

As mentioned in section 5.1.1, the computation of instantaneous loudness is a key issue for estimating the dynamic component of acoustic complexity. For the reason of computational efficiency the implementation described here uses a simplified loudness model that was described by Earl Vickers in [Vic01]. Some modifications and additions have been made in order to make the algorithm fit for the desired task.

As a first step the algorithm applies a very simplified “B” weighting function to the audio signal in order to account for the human ear’s transfer function. For efficiency reasons this weighting is done by a first-order Butterworth high-pass filter with a cut-off frequency of 200 Hz. So actually only the low end of the weighting function is approximated. In the case of music signals this is tolerable, since they have much more energy in the bass and mid-range compared to the high frequencies.

The pre-emphasized signal $x_{pre}(n)$ is then fed into a root-mean-square level detector with an integrated smoothing function. This level detector consists of a running average V_{ms} (eq. 5.2) that is downsampled according to the chosen frame length N.

$$\begin{aligned} V_{ms}(n) &= c \cdot V_{ms}(n-1) + (1-c) \cdot x_{pre}^2(n) \quad \text{, with} \\ c &= e^{-\frac{1}{\tau F_s}} \end{aligned} \tag{5.2}$$

F_s corresponds to the sampling frequency and τ is the time constant for the smoothing (35 ms in this implementation). The downsampling is done according to equation 5.3. We chose a framesize of 200 ms corresponding to $N = 8820$ for a sampling rate of 44.1 kHz. This time span roughly resembles the energy integration function of the human hearing system.

$$V_{rms}(i) = \sqrt{V_{ms}(N \cdot i + N - 1)} \tag{5.3}$$

The instantaneous level is then converted to dB by calculating

$$V_{dB} = 20 \cdot \log_{10}\left(V_{rms}(i)\right). \tag{5.4}$$

In order to avoid silence at the end or the beginning of the track to have an effect on the complexity estimation, successive frames with a level below -90 dB are deleted when they appear at either end. Afterwards, the global loudness level L according to Vickers is calculated. L is a weighted average of all M instantaneous level estimates, where the louder ones are assigned a higher weight:

$$\begin{aligned} L &= \sum_{i=0}^{M-1} w(i) \cdot V_{dB}(i) \quad \text{, with} \\ w(i) &= \frac{u(i)}{\sum_{j=0}^{M-1} u(j)} \quad \text{and} \\ u(j) &= 0.9^{-V_{dB}(j)} \end{aligned} \tag{5.5}$$

The emphasis on the loud frames is grounded in psychoacoustic findings. So for example Zwicker and Fastl [ZF90] suggest that the loudness of a dynamically changing sound can be characterized by the loudness level which only 5% of the

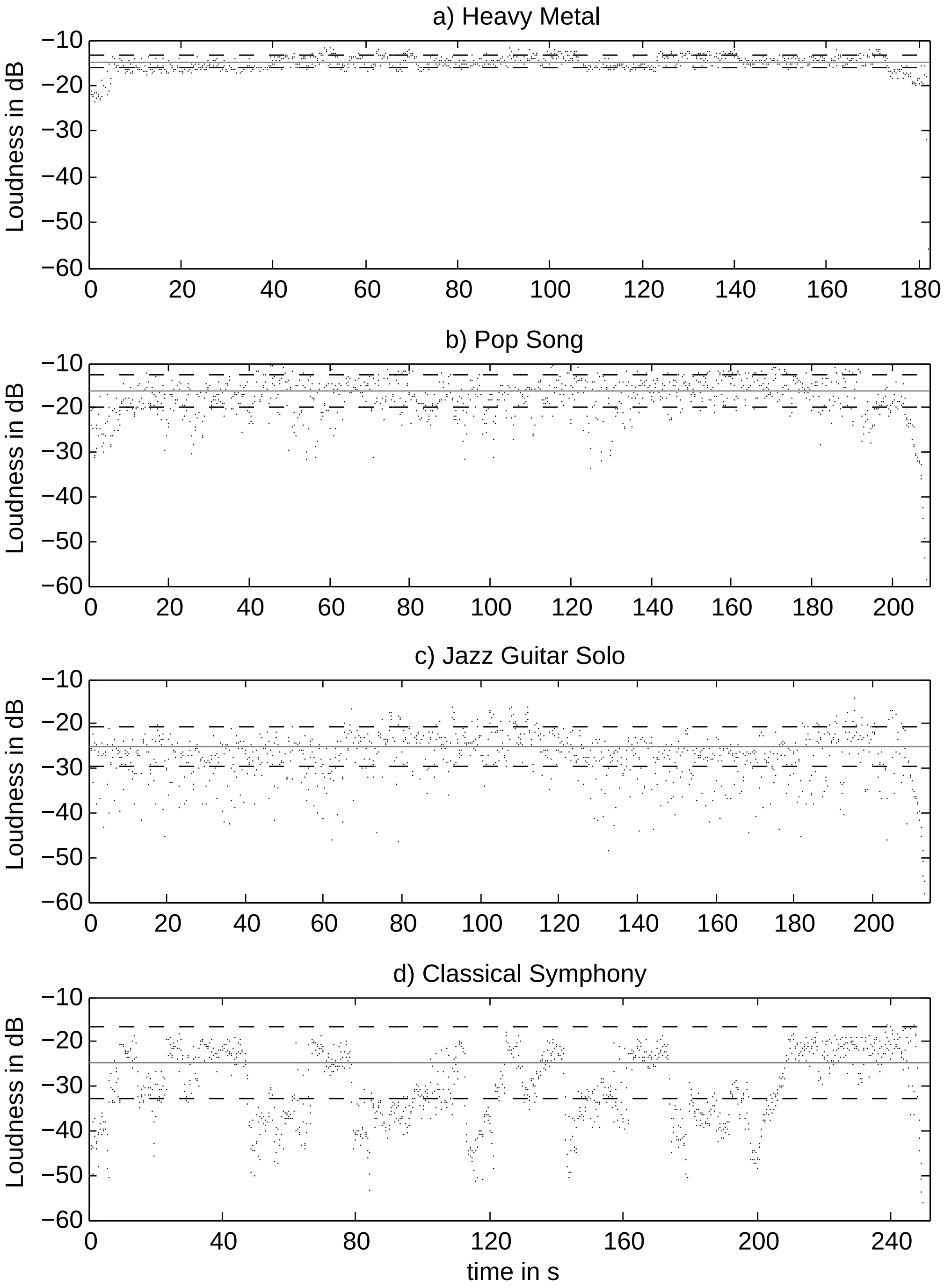


Figure 5.1: Instantaneous loudness (dots), global loudness (grey solid line), and average distance margin (dashed lines) for four example tracks.

frames exceed. In this implementation a variation of Vickers *dynamic spread* is used as the final dynamic complexity measure. It is simplified in the sense that periodic loudness variation and the suddenness of changes are not considered. Instead it comprises basically the mean distance from the global loudness level (eq. 5.6) so that high values correspond to higher complexity and vice versa.

$$C_{dyn} = \frac{1}{M} \sum_{i=0}^{M-1} |V_{dB}(i) - L| \tag{5.6}$$

Figure 5.1 shows the computational results of the algorithm for four prototypical music tracks. In figure 5.1 a we see a highly saturated heavy metal track with a basically flat dynamic level. The global loudness was estimated as -14.9 dB and the average deviation from this level is only 1.37 dB. Figure 5.1 d shows another extreme, a recording of classical music that changes in loudness between -50 dB and -20 dB. The algorithm estimates a clearly lower global loudness of -24.8 dB and an average deviation of 8.05 dB. In between there is a pop song (figure 5.1 b) moderately varying in loudness with a global level of -16.4 dB and an average deviation of 3.71 dB. The jazz guitar solo from figure 5.1 c is recorded at very low volume. The estimated global loudness is only -25 dB. Despite some singular drops down to -50 dB due to short pauses between played notes, the average deviation amounts only to 4.5 dB and is thus considerably smaller than that of the classical recording.

5.2.2 Danceability

The implementation described here is following the description by Jennings et al. in [JIM+04]. When exact specifications were missing, the author tried to integrate reasonable solutions. The experimental findings obtained with this implementation on a large music database where reported at the 118th AES Convention [SH05].

As a first step the audio signal is segmented into non-overlapping blocks of 10 ms length. For each block the standard deviation $s(n)$ of the amplitude is computed. The values $s(n)$ resemble a bounded, non-stationary time series, which can be associated with the averaged physical intensity of the audio signal in each block (see figure 5.2). In order to obtain the unbounded time series $y(m)$,

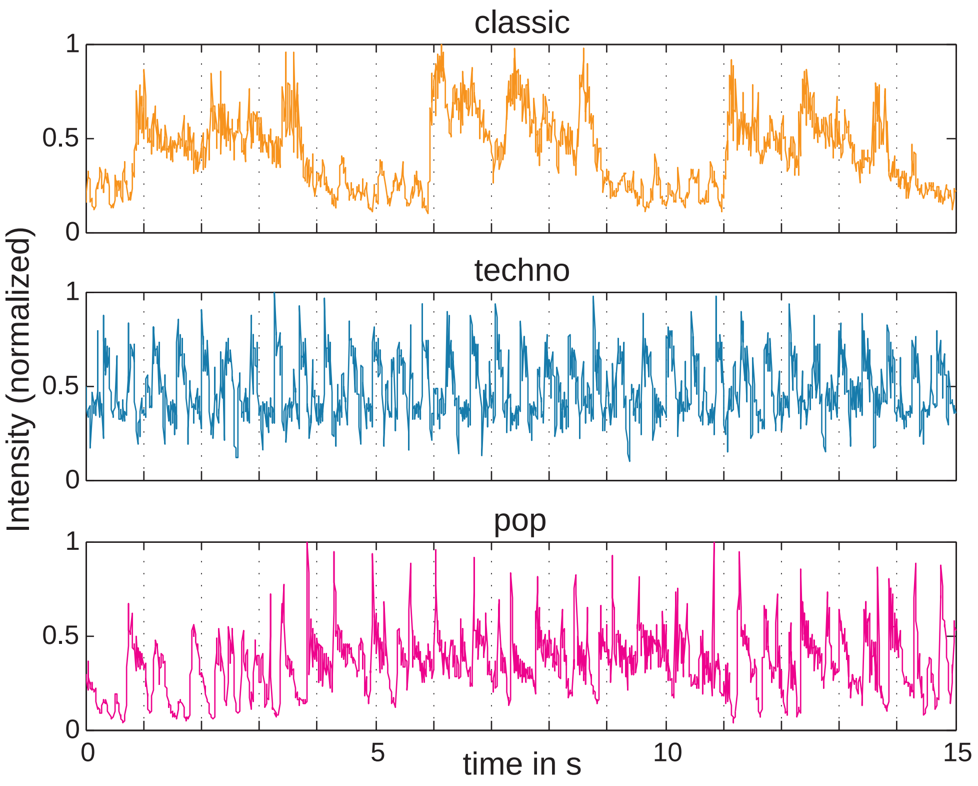


Figure 5.2: Excerpts from the time series $s(n)$ for three example pieces from different musical genres.

$s(n)$ is integrated:

$$y(m) = \sum_{n=1}^{m} s(n) \tag{5.7}$$

This integration step is crucial in the process of DFA computation, because for bounded time series the DFA exponent (our final feature) would always be 0 when time scales of greater size are considered. This effect is explained in more detail in [Pen05].

The series $y(m)$ can be thought of as a random walk in one dimension. $y(m)$ is now again segmented into blocks of τ elements length. This time, we advance only by one sample from one block to the next in the manner of a sliding window. There are two reasons for this extreme overlap. First, we obtain more blocks from the signal, which is of interest, since we will obtain better statistics from a larger number of blocks. Secondly, we avoid possible synchronization with the rhythmical structure of the audio signal, which would lead to arbitrary results depending on the offset we happen to have. However, performing the computation in this manner the number of operations is increased enormously.

From each block we now remove the linear trend $\hat{y}_k$ and compute $D(k, \tau)$, the mean of the squared residual:

$$D(k,\tau) = \frac{1}{\tau}\sum_{m=0}^{\tau-1}\left(y(k+m) - \hat{y}_k(m)\right)^2 \tag{5.8}$$

We then obtain the detrended fluctuation $F(\tau)$ of the time series by computing the square root of the mean of $D(k, \tau)$ for all K blocks:

$$F(\tau) = \sqrt{\frac{1}{K}\sum_{k=1}^{K} D(k,\tau)} \tag{5.9}$$

As indicated, the fluctuation F is a function of τ (i. e. of the time scale in focus). The goal of DFA is to reveal correlation properties on different time scales. We therefore repeat the process above for different values of τ that are within the range of our interest. Jennings et al. [JIM+04] use a range from 310 ms ($\tau = 31$) to 10 s not specifying the step size in their paper. Relating these time scales to the musical signal they are reaching from the beat level through the bar level up to a level of simple rhythm patterns.

The DFA exponent α is defined as the slope on a double log graph of F over τ (eq. 5.10) as shown in figure 5.3 for the three example tracks. It therefore makes sense to increase τ by a constant multiplication factor rather than a fixed step size. Apart from giving equally spaced supporting points on the logarithmic axis it also reduces the computational operations without affecting the accuracy gravely. We chose a factor of 1.1 giving us 36 different values for τ covering time scales from 310 ms to 8.8 s.

For small values of τ an adjustment is needed in the denominator when computing α (see [BGH+95]) giving us the following formula for the DFA exponent:

$$\alpha(i) = \frac{\log_{10}\left(F(\tau_{i+1})/F(\tau_i)\right)}{\log_{10}\left((\tau_{i+1}+3)/(\tau_i+3)\right)} \tag{5.10}$$

As τ grows, the influence of the correction becomes negligible. In case that the time series has stable fractal scaling properties within the examined range, the double log graph of F over τ is a straight line making $\alpha(i)$ a constant function. We find a constant value of 0.5 for a completely random series (white noise), a value of 1 for a series with $1/f$-type noise, and 1.5 for a Brown noise series (integrated white noise) [Pen05].

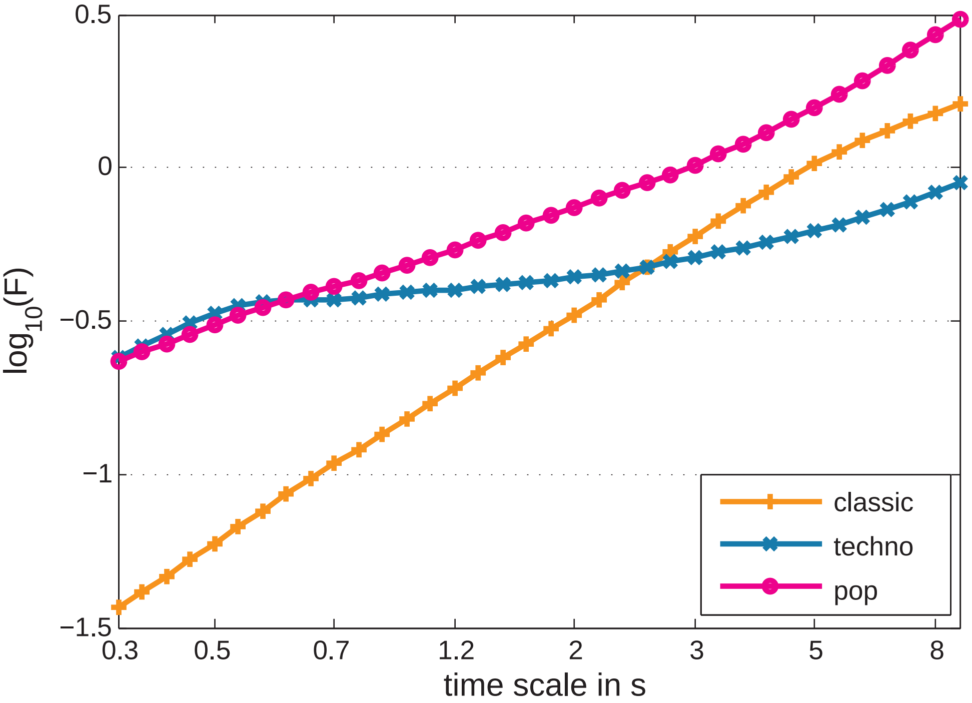


Figure 5.3: Double logarithmic plots of mean residual over time scales.

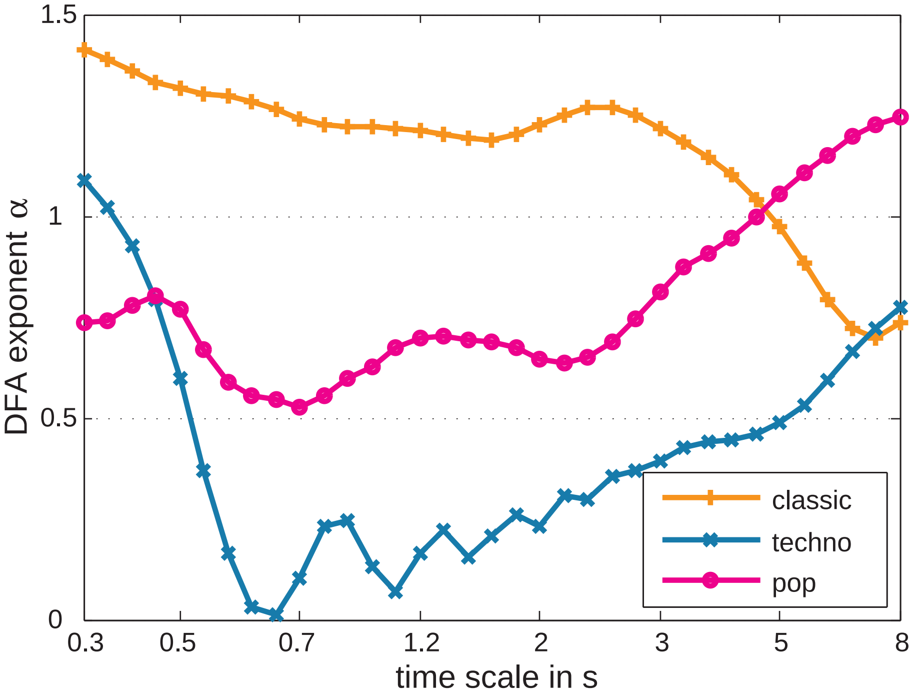


Figure 5.4: DFA exponent functions for the three example tracks from figure 5.2.

For music signals normally we do not have stable scaling properties (see figure 5.4). Unlike heart rate time series, for example, there is much more variance in $\alpha(i)$ for music. Nevertheless, we can find that music with sudden jumps in intensity is generally yielding a lower level of $\alpha(i)$ than music with a smoother varying series of intensity values. Thus, music with pronounced percussion events and emphasized note onsets shows lower α values than music with a more floating, steady nature. There is also a relationship between strong periodic trends and the α function. Figure 5.4 shows the evolution of α over the different time scales for three musical pieces. As can be seen, the DFA exponent varies significantly within each single piece.

The most stable scaling behavior is found for the classical piece at short time scales, in contrast, the pop piece shows an intermediate, and the techno piece shows a high instability. This is due to the presence of a strong and regular beat pattern in the two latter cases (see figure 5.2). In the techno piece the periodic beat dominates the intensity fluctuation completely since intensity variations on larger time scales are negligible in comparison. This strong periodic trend deteriorates the scaling properties of the series and causes α to drop significantly. Towards larger time scales however, the influence of the periodic intensity variation fades off and α raises back towards its normal level. In the pop music piece there is also a regular beat, but it is less dominant than in the techno piece. As can be seen in figure 5.2, there are also some noticeable changes in intensity on a larger time scale. Still, α is clearly decreased by the periodic trend. Towards larger time scales, we can observe the same effect as in the techno piece. For the classical piece no dominant, regular beat pattern can be identified in the time series. Thus, the scaling properties are not affected in the corresponding range. But in contrast to the other two examples the series reveals a larger scale pattern in some parts, which can also be seen in figure 5.2. This causes α to drop in the upper range.

In order to arrive at an indicator for the danceability and thus a certain aspect of the rhythmic complexity the α values have to be further reduced. While different ways are thinkable to do this, in the first implementation simply the average α level was computed for each track. A high value refers to a high complexity (not danceable), a low value refers to a low complexity (highly danceable).

5.3 Evaluation

The evaluation of complexity descriptors is not easy. The ideal condition for evaluation is of course a solid ground-truth annotation against which the performance of the extraction algorithms can be measured. Unfortunately these ground-truth annotations are rather troublesome to come by. Especially for complexity we have the problem that, other than for genre or tempo for example, available annotations usually do not cover this issue. The manual annotation of one's own testing material needs a lot of time and resources and might not be feasible in some cases.

The two descriptor implementations described above were evaluated by subjective means on the basis of a large music collection. This was done by randomly picking tracks at different complexity levels and judging the danceability and the dynamic complexity in direct comparison by listening. A formal user study hwas not been carried out yet.

For the danceability descriptor a more objective evaluation was done. General statistical methods and machine learning methods were applied in order to explore relations between the semantic labels, or certain artists and the DFA exponent. The rationale behind this is to prove a systematic variation of the DFA exponent subject to certain semantic attributes assigned to the music. More details about the methodology and the findings are provided in the following sections.

5.3.1 The Dataset

The two descriptor implementations described above were computed on a large music collection. A data set of 7750 tracks from MTG-DB [CKF$^+$04], a digital music collection from the MTG lab, was used in the experiment. Each track refers to a full piece of music. The dataset also contained annotated semantic labels for each item, which were obtained from music experts and afficionados, and had been manually assigned to the tracks. In our experiments we used the artist names and also "tone" labels consisting in abstract attributes that are associated with the music, such as "Rousing", "Sentimental", or "Theatrical". The list of "tone" labels is composed of a total of 172 different entries. In the statistical analysis only a subset of 136 labels were considered, because the

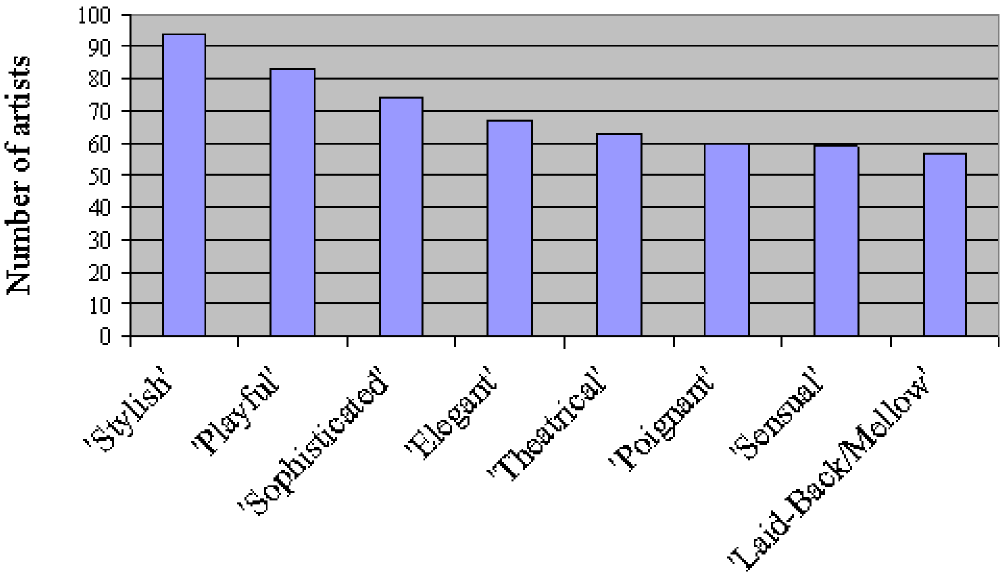


Figure 5.5: Top-eight labels with the highest number of assigned artists.

remaining ones appeared less than 100 times each. It must be noted that these labels are originally assigned to the artists and not to the individual tracks. Therefore a certain degree of fuzziness has to be accepted with these descriptions when working on the track level. The data set contained very different, mostly popular styles of music from a total of 289 different artists. A maximum of 34 labels were assigned to a single artist, while the average was 11 labels per artists. Figure 5.5 shows a bar plot of the eight labels that were assigned to the highest number of artists. The average number of artists sharing a label was 18.

5.3.2 Results

By manual random checks it was found that the complexity estimations at the extreme ends were the most consistent ones. Comparing the tracks from these regions with each other and with the intermediate ones the underlying concept of the descriptors immediately became apparent. This effect can be easily seen in figure 5.6, where the 60 highly danceable techno music tracks can be almost perfectly separated from the 60 non-danceable film score tracks only by considering their average α value. For the dynamic complexity the low end contained tracks that were either highly saturated or simply low quality recordings, while at the other end we found mostly high art music recordings and professionally mastered tracks. The fine grain ranking within a local region however did not appear

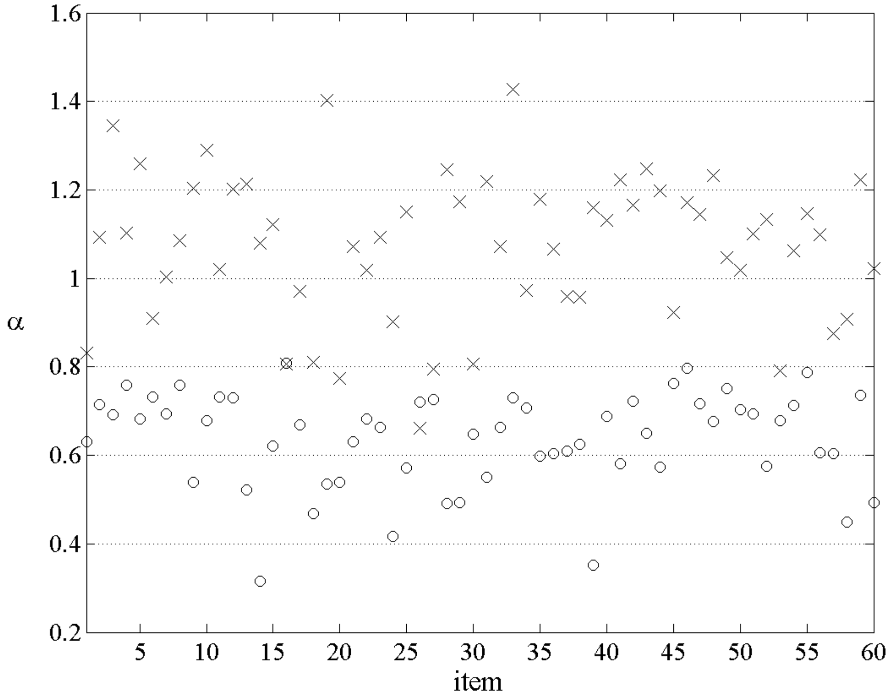


Figure 5.6: α-levels for 60 techno (o) and 60 film score tracks (x), unordered.

comprehensible in many cases. This was especially noticeable in the very dense area of intermediate values. So rather a coarse classification into 3–5 complexity levels than a continuous ordering of the entire collection was achieved. This is a comfortable number of discrete classes for a semantic descriptor considering human memory capabilities.

The results of the statistical tests for the danceability descriptor sustain the findings from manual random evaluation. Strong coherence of high statistical significance was found for several of the "tone" labels that are semantically close to the concept "danceable" or "not danceable" respectively. For example the labels "Party/Celebratory" and "Energetic" in the context of music have a clear relation with danceability, whereas "Plaintive" and "Reflective" appear more appropriate descriptions for music that is not well suited for dancing.

The results reveal a consistency on a high abstraction level even exceeding the aspect of danceability. Figure 5.7 shows how the distribution of some labels on the deciles starting from the lowest to the highest α values in the collection. A strong skew is apparent here with certain labels being highly over-represented either in the highest or the lowest deciles.

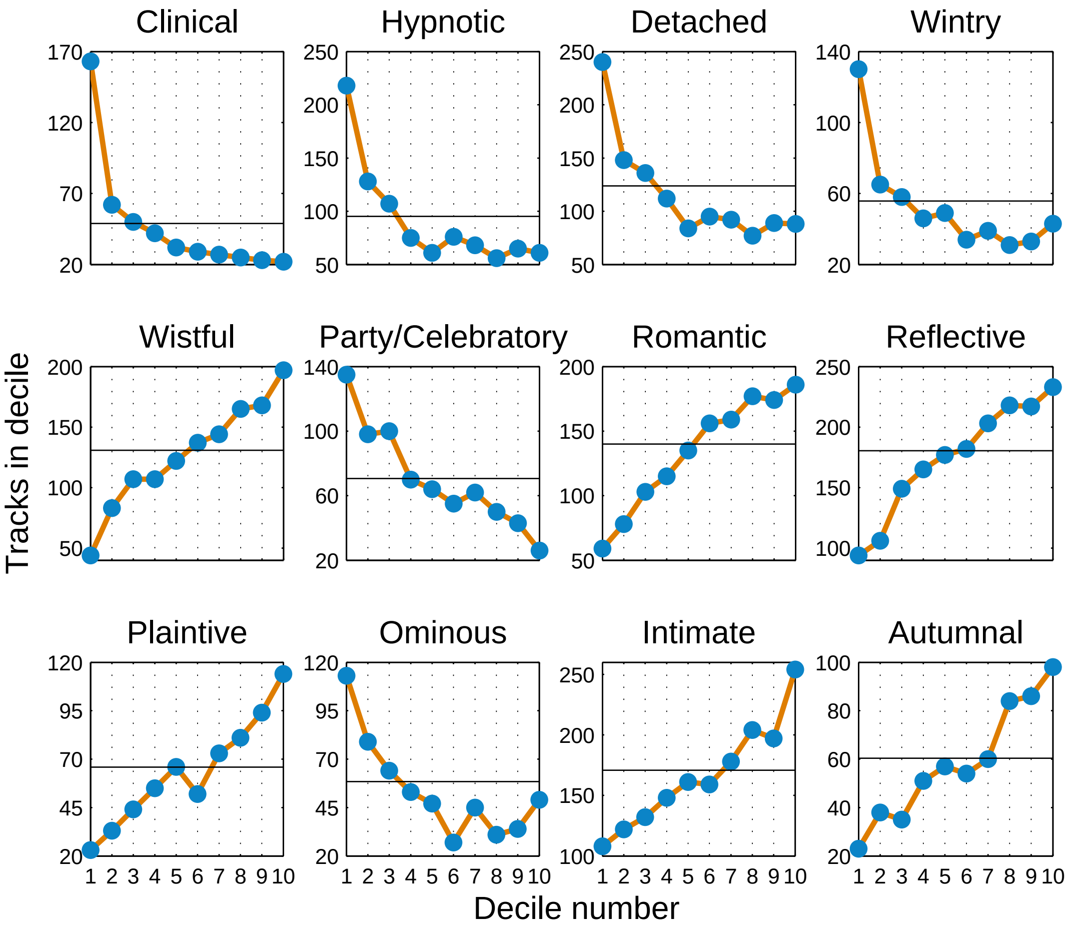


Figure 5.7: Distributions on deciles for the twelve labels with most significant deviation from equal distribution (solid horizontal lines).

The distribution of α values on the whole collection was normal with a mean of 0.863 and a standard deviation of 0.022. The tracks assigned to each label were tested for significant deviations from this distribution with the generalized t-test (eq. 5.11). Only those with normal distributions were considered. Of these, 24 showed a significantly higher and 35 a significantly lower mean value.

$$-2.58 < \frac{0.863 - \bar{\alpha}_{label}}{\sqrt{\frac{0.022^2}{7750} + \frac{\sigma^2_{label}}{n_{label}}}} < 2.58 \qquad (5.11)$$

The value $\bar{\alpha}_{label}$ is the mean α value for the considered label, σ^2_{label} is the corresponding variance, and n_{label} is the number of tracks having this label assigned. Table 5.1 shows the ten labels that yielded the highest significance in their deviation from the global distribution in either direction.

When looking at the two lists of labels a certain affinity can be noted in many cases on either side. The group of labels for higher α wakes associations of *Soft-*

Label	$\bar{\alpha}$	n	**Label**	$\bar{\alpha}$	n
Party/Celebratory	0.796	706	Romantic	0.911	1399
Clinical	0.761	489	Wistful	0.909	1308
Hypnotic	0.804	951	Plaintive	0.922	659
Energetic	0.820	898	Reflective	0.901	1805
Visceral	0.808	422	Calm/Peaceful	0.908	1102
Trippy	0.824	998	Autumnal	0.916	604
Outrageous	0.781	102	Intimate	0.897	1709
Exuberant	0.839	1383	Stately	0.908	730
Irreverent	0.830	657	Gentle	0.892	1327
Sparkling	0.790	116	Elegant	0.886	2506

Table 5.1: The ten most significantly deviating labels in each direction.

ness, *Warmness*, *Tranquility*, and *Melancholy*. For the others we might form two subgroups, one around terms like *Exuberance* and *Vehemence*, the other around *Tedium* and *Coldness*. Comparing labels from both lists with each other, we can identify several almost antonymous pairs, for example: “Clinical” – “Intimate”, “Outrageous” – “Refined/Mannered”, “Boisterous” – “Calm/Peacefull”, “Carefree” – “Melancholic”.

In the machine learning experiments two artist classification tasks were tried. It must be stated again here, that the “tone” labels mentioned above originally also belong to the artists and thus only indirectly to the individual tracks. In a two class decision experiment we used 238 Frank Sinatra tracks and 238 tracks from nine other artists who either had the labels “Brittle” or “Outrageous” assigned to them. For the artist “Sinatra” a total of 18 labels were listed in our data set, among them “Romantic”, “Calm/Peaceful”, and “Sentimental”. From the results of the statistical analysis we would expect the Sinatra songs to be distributed around a greater value of α than the other ones. The classes should therefore be separable up to a certain degree. It must be noted, that among the nine selected artists we find also assigned labels like “Whistful” and “Reflective”, which are linked to higher α values as well (see table 5.1). The classification with a decision table yielded a success rate of 73%, which is clearly above the 50% chance level.

In a second experiment we used three different classes: 108 tracks composed by Henry Mancini, 65 tracks composed by Bernard Herrmann, and 52 tracks of dance music with a strong beat. We purposely did not use the labels for selecting the artists in this case. Mancini and Herrmann were both film music composers, but while Herrmann mainly uses ”classical” orchestration, Mancini often arranges his music in a jazz–like style. We had to select dance music from different artists, because there was no single one with a sufficient number of tracks in the collection. In terms of the DFA exponent we would expect to find the highest values associated with Herrmann's music, because it is the least danceable in general terms. Intermediate values can be expected for Mancini's tracks, since there are many which at least possess a pronounced beat. The lowest values should be found for the dance music, which has strong and regular beat patterns. With a success rate of 67% the classification reached again clearly above chance level (48%). Furthermore the confusion matrix shows exactly the expected situation:

	Predicted class		
True class	Herrmann	Mancini	Dance
Herrmann	41	23	1
Mancini	18	84	6
Dance	1	25	26

While the classes “Herrmann” and “Dance” are almost perfectly separable, “Mancini” takes an intermediate position showing a considerable overlap with both neighboring classes. The decision table identified the optimal thresholds between the classes to be $\alpha_{HM} = 1.0$ and $\alpha_{MD} = 0.7$.

5.3.3 Concluding Remarks

Summarizing we can say that there a is strong evidence for the danceability descriptor being related with semantic properties of the music tracks it was computed on. From our experiments and observations the hypothesis put forward by Jennings et al. seems to be valid, and the DFA exponent can be considered a good indicator for the danceability of music. However, this should be seen rather in the broad sense, classifying music into a small number of categories from “extremely easy” over “moderately easy” to “very difficult” to dance to. Currently, a fine grain ordering by the DFA exponent inside such a category is not beneficial. Due to subjectivity effects such an ordering might not prove useful anyway. By averaging the DFA exponent function we used an extremely compact and simple representation in the experiments. It should be possible to improve the results by a more sophisticated reduction of the function α. It can further be concluded from our results, that the DFA exponent shows to be meaningful also in revealing even higher level attributes of music. It thus might form a valuable addition for different music classification tasks, like artist identification or musical genre recognition.

Chapter 6

Future Work

We have seen in the preceding chapters that for a complete solution to the problem of automatic complexity estimation many parts are still incomplete or missing. In this chapter we will describe a more detailed road map of planned and possible future research. In several cases the author already has done initial tests, which can be used as a basis for further exploitation. We will first look at the tasks that are to be covered by the PhD thesis and then list a few issues that are out of scope for the thesis, but still worthwhile to be explored scientifically.

6.1 PhD Thesis

6.1.1 Filling the Gaps

With the algorithms described in section 5.2 we already have partly covered two complexity facets: the rhythmic and the acoustic one. Nevertheless, the two algorithms are not yet in their final version and might be further enhanced during the research work for the PhD thesis. In a similar manner as in section 5.1 we will now go through the different complexity facets while putting emphasis on the concrete implementation steps.

Acoustic Complexity

The current implementation of the dynamic component already captures a certain part of what was listed as relevant in chapter 5. Lacking so far are the aspect of abruptness in loudness changes and the consideration of periodicities.

A possible way to address this would be to perform a linear prediction on the loudness envelope and to estimate the average prediction error as a measure of complexity. Also, the derivative of the loudness envelope could be considered for further analysis. Furthermore it has to be tested in which way the different values are finally combined, because the goal is only a single number reflecting the dynamic component of acoustic complexity.

Considering the spatial component we already made reference to Griesinger [Gri99] and the InterAural Difference measure (eq. 5.1 on page 38). A second option is a measure based on the magnitude squared coherence (MSC), which is proposed by Wittkop ([Wit01] chapter 4) as a method for diffusiveness estimation of acoustical situations with binaural hearing aids. The MSC for a two–channel signal (channels x and y) was originally proposed by Allen et al. [ABB77] as

$$MSC(f,n) = \left(\frac{|\Phi_{xy}(f,n)|}{\sqrt{\Phi_{xx}(f,n)\Phi_{yy}(f,n)}} \right)^2, \qquad (6.1)$$

where Φ denotes the smoothed Fourier Transform of the correlation function[1], and f and n are the frequency and the time-frame indexes of the short-term spectra derived from the signal. The smoothing is done by a first order low–pass filter along the time dimension. Wittkop proposes to collapse the frequency resolution to critical bands and introduces weighting and transformation functions in order to arrive at a compact and stable long-term coherence estimation.

Rhythmic Complexity

With the danceability descriptor we have covered already one part of the rhythmic complexity facet. As mentioned, the danceability estimation can still be optimized, especially with respect to the intermediate range. One approach is the optimization of the parameters in the algorithm, mainly the step size and the range of the time scales τ. Another option is a more detailed analysis of the α function. Instead of simply averaging all the values there might be more intelligent ways to reduce the representation and arrive at a danceability estimate. Also different pre-processing can be considered as an option. The algorithm could be applied in different frequency bands. Instead of the intensity envelope, a different time series, which is related to rhythmical events might be used (e. g.

[1] The correlation is computed in frequency domain, while the final step, the inverse transformation back to time domain, is left out (see [PFTV92] pp. 545–546).

the synthesized output of an onset detector).

Apart from the danceability estimation, which operates originally on a very low abstraction level, it would probably be beneficial to explore approaches that take a simple but perceptually relevant form of a rhythmic transcription as an input. While the PS-Measure is demanding already a too abstract representation in form of the notation of isolated rhythm patterns, other methods could be applied on an intermediate level. Assuming a reasonably reliable onset detection algorithm as a preprocessing unit it would be interesting to explore the variation of the inter-onset intervals. Also an entropy coding method can be applied to the extracted onset series, presumably best after a temporal quantization. Recurring onset patterns would then decrease entropy and indicate a lower complexity.

Timbral Complexity

Entropy coding could also be the solution for the timbral complexity estimation. At least after some initial experiments with unsupervised clustering and Hidden Markov Models, the methods from information theory appeared the most promising in comparison. The key problem here is to find the right audio features and to transform them into a string of discrete symbols from a finite alphabet. Since the amount of features that are "somehow related" to timbre is enormous, the difficulty lies in selecting the combination that best resembles the human perception of timbre. At the same time the selection should cover all perceptually relevant parts and remain very compact in order to keep the resulting alphabet size reasonable. For the quantization process (which is unavoidable, because the timbre features will be continuous) a hysteresis function has proved to be useful in initial experiments. By having different thresholds for the upward and downward transition between two partitions the effect of an oscillation for values close to the borderline is avoided.

A systematical problem with timbral complexity is further the one of dominant singing voice or speech. In terms of signal analysis a singer who produces different vowels is also producing different timbres. In fact, vowels (or in general phonemes) can only be differentiated by their different timbres. On the other hand, a human listener who clearly perceives the different timbres will not realize it, because he recognizes that they encode speech. We can illustrate this effect, if we think about estimating the timbral complexity of a monotonous voice reading a text and for example of a filtered sawtooth waveform where the filters cutoff frequency is slowly oscillating. A naïve human will tend to attribute the higher complexity to the latter in direct comparison, since the frequent timbre changes in the former are not explicitly recognized as such. For a computer this is difficult to achieve, because instead of speech the former is a sequence of very different, fast changing timbres, whereas the latter is very predictable and only moderately changing. By using a low temporal resolution it can be tried to avoid this effect to a certain degree. However, a true solution would need the identification and isolation of language through the computer even within the polyphonic mix of a musical recording. This is still quite far from becoming feasible. More realistic is at least the automatic distinction between passages with singing voice and those without (e. g. [NW04]). Based on this discrimination it could be tried to disregard the timbre changes due to the voice by simply skipping these segments or assigning them a fixed timbre symbol.

Harmonic Complexity

For harmonic complexity it is also safer not to rely on a very high level description on the input side. Although the automatic extraction of chords is making progress (see for example [HS05] or [DC04]) results are still not accurate enough to allow for an automated complexity estimation based on high level musical rules. What seems more appropriate instead is again the application of a complexity measure on an intermediate abstraction level.

As described in section 5.1 the harmonic pitchclass profile (HPCP) seems a good candidate as an audio feature in this context. We can identify two properties of this feature that could contribute to a harmonic complexity estimation. The first one is the clarity of the chords, with a pure triad being of lowest complexity, while extra notes like the seventh or the ninth increase complexity. As mentioned, this could be captured by using correlation or distance measures for the HPCP-vectors and prototypical triad vectors in the line of tonal strength [GH04].

The second property focusses on the harmonic changes for consecutive vectors. If the changes are expectable, because the perceived distance between consecutive chords is small this would lead to a low complexity level, whereas harmonic sequences containing perceptually large jumps would be considered more complex. So the difficulty here is the mapping of the HPCP-vectors onto a

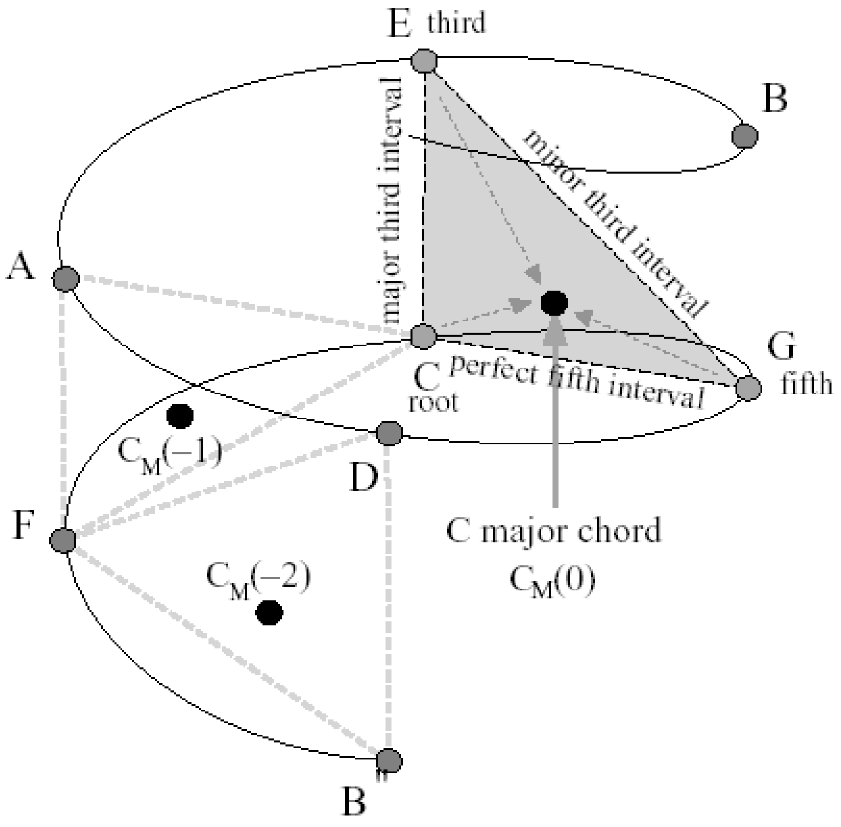


Figure 6.1: Construction of the chord spiral from the pitch spiral [Che00].

representation that allows for an easy calculation of perceived distances between chords. The spiral array model developed by Elaine Chew [Che00] seems to be an appropriate candidate for this task. Although the distances do not perfectly match the perceptual ones in all cases, the model is already coming very close to our needs. Especially interesting here is the underlying hierarchical structure. As shown in figure 6.1 by starting with a spiral of pitch classes (ordered as in the circle of fifth) a second spiral representing the major triads can be constructed within the first one. This is done by taking the center of gravity of the triangle spanned by the tonic, the fifth and the third as a reference point for the chord. The same can be done for minor triads. Chew examined different height gains of the spiral and different weights for tonic, fifth, and third in her dissertation in order to obtain satisfying distances.

The important point for our application is that the chord spiral is derived from the pitch class spiral. Although we cannot expect to find ideal triads in the HPCP-vector, we still have a pitch class representation of the musical content in this feature. So by calculating the center of gravity for an HPCP-vector in the pitch class spiral we arrive at a certain chord representation. Furthermore we can hope, that the distances between these chord representations do relate to some degree with the perceived distances. This way, by calculating the average speed when moving through the spiral, we can have an estimate of the harmonic complexity without the need for an explicit chord recognition or even segmentation.

Structural and Melodic Complexity

These last two facets of complexity will be treated with low priority only, since both have to rely on the extraction of a very abstract descriptor. Particularly for melodic complexity this "preprocessing" step is the most important missing link in the chain at the current state of the art. Considering the level of performance of today's melody transcription algorithms an application of the existing complexity models appears realistic only in several years from now.

Regarding the extraction of music structure we can hope to arrive earlier at acceptable results. While the number of different parts or the change rate are then very straightforward complexity estimations, the periodicity and the dissimilarity need to be addressed more carefully. For the periodicity again an entropy estimation could be used, but we have to expect very short sequences here. Hence, even the number of repetitions might already be a useful indicator. The dissimilarity forms a much harder problem. Since the problem of music similarity estimation has not been solved completely (despite a lot of research), also the opposite, the absence of similarity is difficult to quantify. Furthermore, for our application an absolute measure would be needed, because finally we want to compare dissimilarities between segments of one track with the dissimilarities between the ones from other tracks. As a workaround it could be considered to simply use the novelty scores from the segmentation process that appeared at the segment transitions. However, the notion of these is somewhat different from the real intention, because then only the local region around the transition and only the features used by the segmentation algorithm would have an influence.

6.1.2 Evaluation Strategies

One important aspect of any research involving the development of models or algorithms is the evaluation of the achievements. This evaluation can be seen under two slightly different aspects. One is the direct validation of the models' prediction against a reliable ground-truth. The other is the proof that the application of the model or the algorithm yields a noticeable benefit in the desired context. As mentioned in section 5.3 already it is very costly to obtain a solid ground-truth for all proposed complexity facets. Several listening tests with a group of subjects and a sufficient number of tracks would be needed. Considering that we are aiming at a complexity estimation for full tracks and not for short excerpts gives rise to even more difficulties, since this prolongs the annotation process.

For these reasons, this strategy will only be followed in singular cases during the PhD research work. Rather the author plans to utilize indirect measurements by using relevant existing annotations together with machine learning as well as statistical analysis methods. As demonstrated with the danceability descriptor this approach can give direct insights into the fitness of a particular descriptor for its application to music retrieval tasks. Apart from being more application oriented another advantage is that the potential of the large music collection with its diversity of semantic annotations owned by the MTG can readily be exploited.

However, additional listening tests on a smaller scale might be considered. Instead of asking the listeners to annotate their complexity impressions there is also the possibility to let them directly evaluate the performance of the algorithms. For example clusters of tracks could be built according to the extracted complexity values and the participants could then rate the consistency of the clusters, or the comprehensibility of the underlying concept. While this is a nice way to validate the implemented model it has the disadvantage that further optimization of the model requires again a listening test in order to know whether a measurable improvement was achieved.

6.1.3 Efficiency Aspects

The intended application of the descriptors developed in the course of this research are about large music collections. This is a point worth mentioning, because for big collections with a large number of tracks the advantages of the depicted interaction enhancements do “kick in” noticeably. But on the other hand this means that the computational cost for calculating the descriptors has to be reasonable. Traditionally file based (off-line) computation opposed to stream based computation always has the advantage that real-time is not the ultimate limit. However, scalability has to be considered when we are dealing with collections that contain a large number of tracks. Already a collection with 1000 titles would mean more than two full days of “real-time” computing if we assume an average track duration of 3 minutes. Hence, computational efficiency is a very relevant issue when the final usability is considered. Nevertheless, it will only play a minor role during the research within the PhD thesis scope. The main focus lies in providing functioning models rather than speed-optimized implementations. But efficiency will certainly be taken into account when alternative or simplified ways for the computation exist.

One example is the danceability computation. As described in section 5.2 the computation of this descriptor is relatively costly, because it involves a detrending operation in a sliding window. In order to reduce the number of operations it is therefore worthwhile to investigate whether a bigger hop size than 1 sample still leads to acceptable results. Apart from this option it is also of interest to compare the output of the described approach with the one based on a method published by Willson and Francis [WF03]. They claim that the detrended fluctuation analysis (DFA) that forms the basis of the danceability descriptor is “essentially equivalent” to a special spectral analysis of the signal based on a conventional Fourier Transformation. Since with the Fast Fourier Transform (FFT) a highly efficient implementation is at hand, this would mean a much faster computation of the descriptor.

Another aspect in this context are synergy effects. Because we address the different facets of complexity individually, it is possible that in some models the same intermediate step (e. g. a FFT or the extraction of a low-level descriptor) is part of the computation. It is therefore useful to identify these commonalities in order to avoid repeated processing. This does not end with the set of complexity facets. In a final application where also other semantic descriptors are involved an intelligent data management can improve the efficiency considerably. However, this already reaches beyond the scope of the PhD thesis.

6.2 Beyond the Horizon

Once a reliable complexity estimation is available there are several ways to proceed. One interesting topic is the learning of user preferences from listening habits or from the compound of a personal music collection. As pointed out in section 3.2 already, this could lead to the establishment of user "complexity profiles" capturing the preferred levels of complexity for the different facets. With Berlyne's theory [Ber71] of optimal arousal potential and the reported experimental findings from section 4.1 in mind, such a profile could contribute to the prediction of musical preference for new music. In combination with other descriptors a completely automatic music recommendation engine could be developed, which not even needed to be triggered actively by the user. The selection of the appropriate descriptors (which might very well vary from one user to the other) and their weighting are the key points here. This is one task future research could address.

A second aspect is the refinement of the complexity estimations. As stated in chapters 2 and 3, the models developed in the context of this research are meant to reflect a "common sense" judgement of complexity facets on a rather rough scale, nuances remain unrevealed. However, they also might be useful to consider, for example if a collection is otherwise very homogeneous. Future research could try to optimize and enhance the models towards this goal.

It is likely though that this optimization is only useful when the (limited) universality constraint is given up or at least relaxed. In order to arrive at a more detailed level of complexity estimations it would probably be necessary to take an individual user's background into account instead of assuming a generalized one. We would give up the objectivity and comparability of the descriptor. The complexity descriptors of the same track might no longer be the same when computed with complexity models of different users. But on the other hand we could think of virtual music agents, that gather musical knowledge together with the user and adapt their complexity judgements according to his knowledge. They could give truly personalized and individual recommendations. But this approach has little in common with what is the defined goal of the presented research. The complexity models would need to be purposely designed for the adaptation to individual musical experience and knowledge. Only a variation of parameters would not do the job here. So this is definitely a task that future research has to take care of.

Finally, it should be mentioned that complexity models could even be a useful integration within certain algorithms that aim at a high level semantic description of a different kind. One example is the music segmentation, which could make use of changes in an instantaneous complexity description. But also for melody extraction the complexity evaluation of different alternatives might make sense when there are several candidates for the continuation. In a Viterbi-like manner the algorithm could try to estimate the resulting complexity for different "paths" and give preference to one with a certain level of complexity (e. g. avoiding extreme values). Similar ideas are thinkable for tempo tracking or chord transcription. In fact, Pressing and Lawrence [PL93] are using a very similar approach in their auto-notation program "Transcribe". The program tries to visualize a MIDI or audio input in the optimal way with respect to cognitive criteria such as the cost functions for producing syncopated rhythmic patterns.

Conclusion

In this document the concept of a multi-faceted music complexity descriptor set has been presented. We saw that there are useful applications for this type of descriptors in the field of digital music collection interaction. Some psychological and musicological studies have been reviewed that give evidence to the significance of music complexity with respect to music preference. This alone already makes music complexity an interesting means to navigate through a collection or to select tracks for a user. Also other practical aspects speak in favor of the proposed complexity descriptors as their compactness in representation (storage and visualization) and their intuitiveness (querying).

By reviewing former research on this field it was shown that no satisfying solution for the automatic estimation of music complexity exists to date. The proposed models for melodic, harmonic, and rhythmic complexity that can be found in the literature rely on an accurate symbolic representation of the music, which is not possible to reach automatically from the musical audio signal at present. While some previous work can be adapted and several related findings can be combined and enhanced, there are still many gaps to fill in order to arrive at a complete and fully functional solution that satisfies our needs. Therefore, a set of operational definitions for the different complexity facets has been presented in order to clarify in which direction and by which means the author is working on the topic. Preliminary implementations of the rhythmic and the acoustic facet of complexity have been described and experimental results with them were reported. Especially for the tested rhythmic complexity measure they appear to be very encouraging. The implemented danceability descriptor reveals clearly a link to high-level musical attributes, which was demonstrated by statistical analysis and machine learning experiments.

In the chapter on future work the next steps that are to be taken were depicted. The focus will lie on filling in the missing parts, especially regarding

Acoustic, Rhythmic, Timbre, and Harmonic complexity. For each of these facets a practical and realistic outline of an algorithmic approach has been given, while experiences from preliminary experiments were taken into account. Since the evaluation with an annotated ground-truth is problematic for this descriptor set due to the high cost of manual annotation, a more application driven approach has been given the preference. As in the mentioned experiments with the implementation of the danceability measure, the focus will lie more on statistical analysis and machine learning methods than on formal listening tests with a group of subjects. However, in individual cases the latter will also be considered when feasible. Also the computational efficiency of the proposed methods will be considered in the future development, since for large collections this point is highly relevant. Yet, as we will be able to do only the first steps in this domain, performance in terms of processing speed will take a subordinate position to performance in terms of functionality.

List of Figures

Bibliography

[ABB77] J. B. Allen, D. A. Berkley, and J. Blauert. Multimicrophone signal-processing technique to remove room reverberation from speech signals. *Journal of the Acoustic Society of America*, 62(4):912–915, 1977.

[AS01] J. J. Acouturier and M. Sandler. Segmentation of music signals using hidden markov models. In *Proceedings of the 110th AES Convention*, Amsterdam, NL, 2001.

[Ass60] American Standards Association. American standard acoustical terminology. Definition 12.9, 1960.

[Aus00] M. Ausloos. Statistical physics in foreign exchange currency and stock markets. *Physica A*, 285:48–65, 2000.

[Bel03] J. P. Bello. *Towards the automated analysis of simple polyphonic music: A knowledge-based approach.* PhD thesis, Queen Mary University of London, 2003.

[Ber60] D. E. Berlyne. *Conflict, Arousal, and Curiosity.* McGraw-Hill, New York, Toronto, London, 1960.

[Ber71] D. E. Berlyne. *Aesthetics and psychobiology.* Appleton-Century-Crofts, New York, 1971.

[Ber74] D. E. Berlyne. The new experimental aesthetics. In *Studies in the new experimental aesthetics: steps towards an objective psychology of aesthetic appreciation.* Halsted Press, New York, 1974.

[BGH$^+$95] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C.-K. Peng, M. Simons, and H. E. Stanley. Long-range correlation properties of coding and noncoding DNA sequences: Genbank analysis. *Physical Review E*, 51(5):5084–5091, 1995.

[Bre90] A. S. Bregman. *Auditory Scene Analysis.* The MIT Press, Cambridge, London, 1990.

[Che00] E. Chew. *Towards a Mathematical Model of Tonality.* PhD thesis, Massachusetts Institute of Technology, Cambridge, USA, February 2000.

[Che03] E. Chew. Thinking Out of the Grid and Inside the Spiral - Geometric Interpretations of and Comparisons with the Spiral Array Model. Technical Report 03-002, University of Southern California, 2003.

[CKF$^+$04] P. Cano, M. Koppenberger, S. Ferradans, A. Martinez, F. Gouyon, V. Sandvold, V. Tarasov, and N. Wack. Mtg-db: A repository for music audio processing. In *Proceedings of 4th International Conference on Web Delivering of Music*, Barcelona, Spain, 2004.

[CV03] W. Chai and B. Vercoe. Structural analysis of musical signals for indexing and thumbnailing. In *Proceedings of the 3d ACM/IEEE-CS Joint Conference on Digital Libraries*, Houston, Texas, 2003.

[DBDS03] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler. Complex domain onset detection for musical signals. In *Proceedings of the 6th International Conference on Digital Audio Effects*, London, UK, 2003.

[DC04] C. Derboven and M. Cremer. A system for harmonic analysis of polyphonic music. In *Proceedings of the AES 25th International Conference on Metadata for Audio*, London, UK, 2004.

[Edm95] B. Edmonds. What is complexity? - the philosophy of complexity per se with application to some examples in evolution. In Heylighen and Aerts, editors, *The Evolution of Complexity*. Kluwer, Dordrecht, 1995.

[EN00] T. Eerola and A. C. North. Cognitive complexity and the structure of musical patterns. In *Proceedings of the 6th International Conference on Music Perception and Cognition*, Newcastle, UK, 2000.

[Ess95] P. Essens. Structuring temporal sequences: Comparison of models and factors of complexity. *Perception and Psychophysics*, 57(4):519–532, 1995.

[ET04] T. Eerola and P. Toiviainen. Midi toolbox: Matlab tools for music research. http://www.jyu.fi/musica/miditoolbox/, 2004.

[ETK02] T. Eerola, P. Toiviainen, and C. L. Krumhansl. Real-time prediction of melodies: Continuous predictability judgements and dynamic models. In *Proceedings of the 7th International Conference on Music Perception and Cognition*, Sydney, Australia, 2002.

[Fin89] L. Finnäs. How can musical preference be modified? a research review. *Bulletin of the Council for Research in Music Education*, 102:1–58, 1989.

[Fuj99] T. Fujishima. Realtime chord recognition of musical sound: A system using common lisp music. In *Proceedings of the International Computer Music Conference*, Beijing, China, 1999.

[GD05] F. Gouyon and S. Dixon. A review of automatic rhythm transcription systems. *Computer Music Journal*, 29(1):34–54, 2005.

[GH04] E. Gomez and P. Herrera. Automatic extraction of tonal metadata from polyphonic audio recordings. In *Proceedings of the AES 25th International Conference on Metadata for Audio*, London, UK, 2004.

[Gom04] E. Gomez. Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing. Special Cluster on Music Computing*, 2004.

[Gri99] D. Griesinger. Objective measures of spaciousness and envelopment. In *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction*, Rovaniemi, Finland, 1999.

[GSO$^+$05] E. Gomez, S. Streich, B. Ong, R. P. Paiva, S. Tappert, J.-M. Batke, G. Poliner, D. Ellis, and J. P. Bello. A quantitative comparison of different approaches for melody extraction from polyphonic audio recordings. Submitted to IEEE Transactions on Speech and Audio Processing, 2005.

[GV99] A. Gammerman and V. Vovk. Kolmogorov complexity: Sources, theory and applications. *Computer Journal*, 42(4):252–255, 1999.

[Hey75] R. G. Heyduk. Rated preference for musical compositions as it relates to complexity and exposure frequency. *Perception and Psychophysics*, 17(1):84–91, 1975.

[HPD03] P. Herrera, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1), 2003.

[HS05] C. Harte and M. Sandler. Automatic chord identification using a quantized chromagram. In *Proceedings of the AES 118th Convention*, Barcelona, Spain, 2005.

[JIM$^+$04] H. D. Jennings, P. Ch. Ivanov, A. M. Martins, P. C. da Silva, and G. M. Viswanathan. Variance fluctuations in nonstationary time series: a comparative study of music genres. *Physica A: Statistical and Theoretical Physics*, 336(3-4):585–594, May 2004.

[KL05] D. Kusek and G. Leonhard. *The Future of Music*. Berklee Press, 2005.

[Kla03] A. Klapuri. Musical meter estimation and music transcription. Presented at Cambridge Music Colloqium, 2003.

[Kla04] A. Klapuri. *Signal Processing Methods for the Automatic Transcription of Music*. PhD thesis, Tampere University of Technology, 2004.

[Ler96] F. Lerdahl. Calculating tonal tension. *Music Perception*, 13(3):319–363, 1996.

[Mas02] R. Mason. *Elicitation and measurement of auditory spatial attributes in reproduced sound*. PhD thesis, University of Surrey, February 2002.

[Mat01] L. Matassini. *Signal analysis and modelling of non-linear non-stationary phenomena*. PhD thesis, Bergische Universität Gesamthochschule Wuppertal, Wuppertal, Germany, July 2001.

[MGB97] B. C. Moore, B. R. Glasberg, and T. Baer. A model for the prediction of thresholds, loudness, and partial loudness. *Journal of the Audio Engineering Society*, 45(4):224–239, April 1997.

[Nar90] E. Narmour. *The analysis and cognition of basic melodic structures*. The University of Chicago Press, Chicago and London, 1990.

[NH97a] A. C. North and D. J. Hargreaves. Experimental aesthetics and everyday music listening. In *The social psychology of music*, pages 84–103. Oxford University Press, Oxford, 1997.

[NH97b] A. C. North and D. J. Hargreaves. Liking, arousal potential, and emotions expressed by music. *Scandinavian Journal of Psychology*, 38:45–53, 1997.

[NH97c] A. C. North and D. J. Hargreaves. Music and consumer behaviour. In *The social psychology of music*, pages 268–289. Oxford University Press, Oxford, 1997.

[NW04] T. L. Nwe and Y. Wang. Automatic detection of vocal segments in popular songs. In *Proceedings of the 2004 International Symposium on Music Information Retrieval*, Barcelona, Spain, 2004.

[OH04] B. Ong and P. Herrera. Computing structural descriptions of music through the identification of representative excerpts from audio files. In *Proceedings of the AES 25th International Conference on Metadata for Audio*, London, UK, 2004.

[OO05] M. G. Orr and S. Ohlsson. Relationships between complexity and liking as a function of expertise. *Music Perception*, 22(4):583–611, 2005.

[Pac99] F. Pachet. Surprising harmonies. *International Journal of Computing Anticipatory Systems*, 4, February 1999.

[Pam01] E. Pampalk. Islands of music: Analysis, organization, and visualization of music archives. Master's thesis, Vienna University of Technology, Vienna, Austria, 2001.

[Par04] R. M. Parry. Musical complexity and top 40 chart performance. Technical report, College of Computing, Georgia Institute of Technology, 2004.

[PBH$^+$94] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. Mosaic organization of DNA nucleotides. *Physical Review E*, 49:1685–1689, 1994.

[Pen05] C.-K. Peng. Fractal mechanics in neural control: Human heartbeat and gait dynamics in health and disease. Online Tutorial, 2005. http://www.physionet.org/tutorials/fmnc/.

[PFTV92] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge, 1992.

[PL93] J. Pressing and P. Lawrence. Transcribe: a comprehensive autotranscription program. In *Proceedings of the 1993 International Computer Music Conference*, pages 343–345, Tokyo, Japan, 1993.

[PRC00] F. Pachet, P. Roy, and D. Cazaly. A combinatorial approach to content-based music selection. *IEEE MultiMedia*, 7(1):44–51, 2000.

[Pre98] J. Pressing. Cognitive complexity and the structure of musical patterns. In *Proceedings of the 4th Conference of the Australasian Cognitive Science Society*, Newcastle, Australia, 1998.

[SAPM02] E. G. Schellenberg, M. Adachi, K. T. Purdy, and M. C. McKinnon. Expectancy in melody: Tests of children and adults. *Journal of Experimental Psychology*, 131(4):511–537, 2002.

[SBM$^+$02] D. van Steelant, B. de Baets, H. de Meyer, M. Leman, J.-P. Martens, L. Clarisse, and M. Lesaffre. Discovering structure and repetition in musical audio. In *Proceedings of Eurofuse Workshop*, Varenna, Italy, 2002.

[Sch89] M. Schmuckler. Expectation in music: Investigation of melodic and harmonic processes. *Music Perception*, 7:109–150, 1989.

[Sch96] E. G. Schellenberg. Expectancy in melody: Tests of the implication-realization model. *Cognition*, 58(1):75–125, 1996.

[Sch00] E. D. Scheirer. *Music-Listening Systems.* PhD thesis, Massachusetts Institute of Technology, Cambridge, USA, 2000.

[Sch01] E. D. Scheirer. Structured audio, kolmogorov complexity, and generalized audio coding. *IEEE Transactions on Speech and Audio Processing*, 9(8):914–931, 2001.

[SE03] A. Sheh and D. Ellis. Chord segmentation and recognition using em-trained hidden markov models. In *Proceedings of the International Symposium on Music Information Retrieval*, Baltimore, USA, 2003.

[SH05] S. Streich and P. Herrera. Detrended fluctuation analysis of music signals: Danceability estimation and further semantic characterization. In *Proceedings of the AES 118th Convention*, Barcelona, Spain, 2005.

[Sha48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27, 1948.

[Sim94a] D. K. Simonton. Computer content analysis of melodic structure: Classical composers and their compositions. *Psychology of Music*, 22:31–43, 1994.

[Sim94b] D. K. Simonton. Melodic structure and note transition probabilities: A content analysis of 15,618 classical themes. *Psychology of Music*, 22:31–43, 1994.

[SM75] L. Steck and P. Machotka. Preference for musical complexit: Effects of context. *Journal of Experimental Psychology: Human Perception and Performance*, 104(2):170–174, 1975.

[SN04] E. Skovenborg and S. H. Nielsen. Evaluation of different loudness models with music and speech material. In *Proceedings of the AES 117th Convention*, San Francisco, USA, 2004.

[Sny00] B. Snyder. *Music and Memory.* The MIT Press, Cambridge, London, 2000.

[SP00] I. Shmulevich and D.-J. Povel. Measures of temporal pattern complexity. *Journal of New Music Research*, 29(1):61–69, 2000.

[Sta01] R. K. Standish. On complexity and emergence. *Complexity International*, 9, 2001. http://journal-ci.csse.monash.edu.au/ci/vol09/standi09/.

[SWV00] E. D. Scheirer, R. B. Watson, and B. L. Vercoe. On the perceived complexity of short musical segments. In *Proceedings of the 2000 International Conference on Music Perception and Cognition*, Keele, UK, 2000.

[SYSPI05] H. F. Silverman, Y. Yu, J. M. Sachar, and W. R. Patterson III. Performance of real-time source-location estimators for a large-aperture microphone array. *IEEE Transactions of Speech and Audio Processing*, 2005. Accepted for Publication.

[TBB00] B. Tillmann, J. J. Bharucha, and E. Bigand. Implicit learning of music: A self-organizing approach. *Psychological Review*, 107:885–913, 2000.

[Tem01] D. Temperley. *The cognition of basic musical structures.* the MIT Press, Cambridge, London, 2001.

[Vic01] E. Vickers. Automatic long-term loudness and dynamics matching. In *Proceedings of the AES 111th Convention*, New York, 2001.

[VSV$^+$03] M. Vilermo, S. Streich, M. Väänänen, K. Linzmeier, B. Grill, and Y. Wang. Perceptual optimization of the frequency selective switch in scalable audio coding. In *Proceedings of the AES 114th Convention*, Amsterdam, Netherlands, 2003.

[Wal73] E. L. Walker. Psychological complexity and preference: A hedgehog theory of behaviour. In D. E. Berlyne and K. B. Madsen, editors, *Pleasure, reward, preference.* Academic Press, New York, 1973.

[WF03] K. Willson and D. P. Francis. A direct analytical demonstration of the essential equivalence of detrended fluctuation analysis and spectral analysis of RR interval variability. *Physiological Measurment*, 24:N1–N7, 2003.

[Wit01] T. Wittkop. *Two-channel noise reduction algorithms motivated by models of binaural interaction.* PhD thesis, Universität Oldenburg, Oldenburg, Germany, 2001.

[WS02] Y. Wang and S. Streich. A drumbeat-pattern based error concealment method for music streaming applications. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Orlando, USA, 2002.

[Wun74] W. Wundt. *Grundzüge der physiologischen Psychologie.* Engelmann, Leipzig, 1874.

[ZF90] E. Zwicker and H. Fastl. *Psychoacoustics – Facts and Models.* Springer, Berlin, Germany, 1990.

[ZL77] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977.

[ZL78] J. Ziv and A. Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.