Digital Music Libraries

Analysis and Comparison of Feature Sets for Audio Retrieval

eingereicht von

Stefan Leitich

DIPLOMARBEIT

zur Erlangung des akademischen Grades
Magister rerum socialium oeconomicarumque
Magister der Sozial- und Wirtschaftswissenschaften
(Mag. rer. soc. oec.)

**Fakultät für Wirtschaftswissenschaften und Informatik,
Universität Wien
Fakultät für Technische Naturwissenschaften und Informatik,
Technische Universität Wien**

Studienrichtung: Wirtschaftsinformatik

Begutachter: a. o. Univ. Prof. Dr. Andreas Rauber

Wien, im April 2004

Acknowledgements

I would like to thank my parents, who gave me the opportunity to study and great support.

Further on I would like to thank my girlfriend Tizi for emotional and computational support, my sister Lisa and her probands for their patience, Gunther for his pocket calculator, as well as Oliver for extending memory capacities and all open ears for listening to my worries during conducting this work.

Stefan Leitich
Vienna, April 2004

Abstract

Developments in the last years, brought changes in the distribution and user interaction with music. Increased computational power and higher bandwidth on transmission networks made digital music libraries popular for commercial and non-commercial purposes. This created the demand for new applications for accessing music in different ways. Music information retrieval systems aim to satisfy these requirements. To successfully implement music information systems, music content analysis is used to index music pieces in collections. This thesis introduces topics from related scientific domains which have influence on the design of music content analysis approaches. It compares the performance of three different state of the art prototypes in terms of recall and precision values. Evaluation is also done by conducting a user survey.

Kurzfassung

Der technische Fortschritt in den letzten Jahren hat die Verbreitung und die Interaktion von Benutzern mit Musik stark verändert. Durch die Steigerung von Rechenleistung und Bandbreite in Netzwerken sind digital Musikarchive populär geworden. Diese Umstände führten zur Nachfrage nach neuen Anwendungen die Benutzer im Zugriff auf Musikstücke in verschiedenster Weise unterstützen. Die Implementierung von Anwendungen für die Musiksuche wird unter dem Einsatz von Systemen zur Analyse von Musikstücken basierend auf ihrem Inhalt realisiert. Es wird versucht Musikstücke aufgrund ihres Aufbaus und Inhalts zu identifizieren. Diese Diplomarbeit beschäftigt sich mit Themen die Einfluss auf das Design eines solchen Systems haben und vergleicht die Leistungsfähigkeit dreier verschiedener Ansätze zur inhaltsbasierten Indizierung von Musikstücken. Die Leistungsfähigkeit wird anhand von Genauigkeit und Vollständigkeit (Recall, Precision) der Ergebnisse bewertet. Eine weitere Evaluierung der Ergebnisse erfolgt durch eine Benutzerbefragung.

Contents

Chapter 1

Introduction

Common words to introduce the reader in works related to digital music libraries and music content analysis for music information retrieval (MIR) systems are about the problematic of retrieving certain music pieces out of a huge digital music collection. The first digital revolution in the music storage domain was the introduction of the Compact Disc (CD). A medium enhancing the playback quality of music and allowing easy random access to music content on it. The second digital revolution are the possibilities arising through the increase of computational power combined with reduction in size, allowing the construction of portable devices as well as personal computers and computation intensive applications dealing with music content, as well as the existence of new distribution channels through increasing channel bandwidth in the internet and mobile radio services. These circumstances empowered the need for new applications of content processing of music and therefore new approaches for music content analysis to enable the realization of these applications.

The main idea of music content analysis is to find a way to extract characteristic features out of the music input to index music pieces and determine similarity between different music pieces. Similarity can serve to determine a certain music piece by a fragment or to find similar music pieces for example in the terms of musical genre.

Chapter 2 gives an overview of work recently done in the field of music content

analysis.

To understand what factors influence the design and work upon music content analysis systems, *Chapter 3* gives insights in different domains. An overview of general music content processing technologies is given and their impacts on music related activities. Since music information retrieval systems are designed to aid human activities related to music it is important to know about human behavior in these activities. This knowledge makes it possible to define criteria for the design of MIR applications. Looking at rules, which apply for the establishment of musical genre, it gets clear, music genre recognition will be hardly solved perfectly by a straightforward signal processing approach. The emergence of musical genres under musicology and sociology aspects are very difficult processes and have a lot of different influences. The last section in this chapter considers the anatomy of the human ear and psychoacoustic effects the human aural perception owns.

After covering these domains and their effects on the music genre recognition in common the topic gets more technical. In *Chapter 4* its all about signal processing. Two different signal transformation possibilities are shown, the Fourier transformation and the Wavelet transformation. These transformation build the basis for the extraction of specific features out of the music signal. Features to characterize a music piece are manifold and focus mainly on spectral shape, texture, rhythmic or pitch content characteristic of music pieces.

Subsequent to the theoretical discussion of music content analysis related domains as well as signal processing techniques for feature extraction, the experiments conducted with three prototypes are illustrated in *Chapter 5*. The three different prototypes, the music collections the experiments are based upon and the way of distance determination are described and the evaluation results are presented. Evaluation is not done by using the computed feature vectors to algorithmically classify the analyzed music pieces into musical genres. It is done by calculating precision and recall values in two different test settings. In the first case the quality of the query results is measured as how good the system performs in finding different segments of the same music piece. The second test case is about a genre precategorized music

collection and evaluates the recall and precision values for genre recognition. As a total different evaluation the answer sets are presented to users asking them to evaluate the answer set in their perception of musical genre.

The final *Chapter 6* contains insights of the theoretical discussion as well as of the conducted experiments.

Chapter 2

Related Work

The research field spanning the topic - music genre classification - could be referenced as audio signal classification. This consists of extracting relevant features from a sound and using these features to figure out which class this sound is probably most similar to. Thereby research domains like signal processing, spectral analysis, psychoacoustics and machine learning algorithms are encountered. For this thesis it is assumed that the input signal will be music, so there is no need to distinguish between silence, noise, speech or music. A not irrelevant simplification, although there are already promising results in speech recognition or music classification, would be a general classification system, which could route the input signal to the appropriate sound processing application [16].

Music as the input signal can be described by technical or semantic features [24]. As a technical feature would be referred a feature, calculated out of the information representing the musical piece itself. Such an feature would be for example the energy of the signal. Energy is used to discover silence in a signal, as well as dynamic range. It is typically calculated on a short-time window of the signal, squaring the samples and taking the average [19]. Semantic features are annotated meta information about the music piece. This features would be represented by information like title, artist, album or year of production.

Digital stored music can be represented in different ways. Either sample values of the analog signal are coded and stored or a notation containing the kind of instrument, begin/end of a note, pitch and volume is used to encode the musical data. The latter for example would be **M**usic **I**nstrument **D**igital **I**nterface (MIDI) notation.

These different music information representations give different possibilities of processing. The features of the MIDI notation used the research team around Rudi Calibrisi for example, in their work [7] using compression algorithms for music genre classification. The idea behind is, roughly speaking, that two objects are deemed close if it is possible to significantly *compress* one given the information in the other. This concept is used to compute the distance between all pairs of music pieces and building a tree containing this pieces in a way, that is consistent with those distances. Experiments have shown, that a discrimination between three broader musical genres (classical music, rock and jazz) works well, but is not perfect. Other experiments using exclusively classical music show, that the method works really well on smaller sample sets. Sets increasing in size contain pieces of artists, which are scattered all over the tree. This could be because (i) the distances estimated are less accurate; (ii) the number of allocation conflicts in the tree rendering process increases; (iii) the computation time rises to such an extent that the value of correctness turns into a victim of the set time limit. Applications of this research work could be the determination of plagiarism or honest influences between composers or to figure out who is the creator of discovered sheets of music.

Concern of this thesis is, music genre classification based on music as input signal in form of discrete values over time. Scientific publications considering this case are mentioned below.

A good overview of technical features used in music genre recognition and classification is given by George Tzanetakis who concentrates his scientific work on this issue. In his PhD thesis, called, - *Manipulation, Analysis and Retrieval Systems for Audio Signals* [43] -, he does not only list and evaluate technical features used for music genre classification, additionally he presents user interaction variations, which are made possible through applications using such features. He also

developed a framework, **M**usic **A**I **R**esearch **SY**stem for **A**nalysis and **S**ynthesis (MARSYAS)[1] [45] for experimenting, evaluating and integrating common audio content analysis in restricted domains. It consists of a server application, written in C++ and a client done in JAVA™. The server contains all the the signal processing and pattern recognition modules optimized for performance. The client provides a **G**raphical **U**ser **I**nterface (GUI) and communicates the requests for computation to the server and retrieves the results. This breakdown has the advantage of separating computation from interface, so that it is possible to easily implement new frontends for individual demands.

A prototypical music classification system called **Mu**sic **G**enre **R**ecognition by **A**nalysis of **T**exture (MUGRAT) is presented by Karin Kosina in her thesis - *Music Genre Recognition* [25]. Classification is done by using technical features, in detail so called musical surface features and features related to the beat of the song. The feature set is based upon the feature set originally proposed by George Tzanetakis in [48]. Musical surface features are derived from the spectral distribution, from the instrumentation, the timbre and the musical texture. The second category of features used in MUGRAT are beat related features. This contains the calculation of the main beat of a song as well as the second strongest beat, and a number of features concerning the relationship between the first and second beat. A in deep explanation of the here or below mentioned different technical features will be given in the feature extraction section of this thesis.

Due to the fact, that important organizations already worked on standards to provide a framework for describing multimedia data it stands to reason to use these frameworks for their intended purpose. So it was done by the **M**otion **P**icture **E**xpert **G**roup (MPEG) Organization which defined the MPEG-7 standard [23]. An e**X**tensible **M**arkup **L**anguage (XML) like language which already comes up with predefined description schemes for describing multimedia data with meta information as well has the potential to define new description schemes as required for future applications. A classification done with the help of the audio part of

[1] The name is derived from the name of a greek mythology figure.

MPEG-7 [22] is shown by Holger Crysandt and Jens Wellhausen in their paper - *Music classification with MPEG-7* [9]. The predefined features of the standard were used to arrange music pieces of the data set, which consisted of the Top 100 Single Charts of Germany from March 2002, via the **s**elf-**o**rganizing **m**ap (SOM) algorithm. Not all of the features of the vector are able to help a classification, so the most important ones are weighted stronger (mean of *Audio Power* - separates soft, quiet songs from louder ones; mean of *Audio Spectrum Flatness* - separates harmonic from inharmonic songs; *Clearness of Beat* - determine if there is a strong or weak beat). The mean of the *Audio Spectrum Centroid* describes whether the signal is dominated by low or high frequencies and is weighted with zero because music pieces of the same genre are defined not similar, if, for example, one is sung by a female artist and another one by a male artist. Visualization is done by the SOM algorithm, feature vectors with a short (euclidian) distance are placed on the same point or close to each other. It was shown that with the MPEG-7 description of songs a rough classification is possible.

Francois Pachet, head of the music research team at the Sony **C**omputer **S**cience **L**aboratory (CSL) Paris, mentions in his paper - *Content Management for Electronic Music Distribution* [31] - requirements, arising from the actual music market changes, on music content management systems and different programs concerning such needs. The **C**ontent-based **U**nified **I**nterfaces and **D**escriptors for **A**udio/music **D**atabases available **O**nline (Cuidado) music browser is one of his examples and was developed in the Cuidado European project. This browser combines features extracted from the signal with semantic features in form of information about a song and user ratings. The Cuidado Music Browser offers a unique opportunity to compare and assess quality and relevance of this features. It is practical oriented and offers functionality like search through a music database by sound similarity, create musical compilations, build audio summaries from titles and retrieve music according to personal taste.

An already on market application, created and distributed by a company called Mus-

cle Fish[2] is their product SoundFisher. Muscle Fish is a cooperation of a group of engineers who had worked together on electronic musical instrument research and development for Yamaha Music Technologies USA, Inc. The Soundfisher is intended to be used as an sound effects database management system. It calculates technical features to determine the distance between the sound samples in its database. The attributes it extracts are not musically oriented rather than for speech - or speaker recognition, they do not address sound at the level of the musical phrase, melody, rhythm or tempo. Used features are loudness as the root-mean square (RMS) level in decibels, pitch, brightness as the higher frequency content and bandwidth of the signal [52].

An interesting paper entitled - *A comparative Study on Content-Based Music Genre Classification* [27] -, introduces automatic music genre classification via feature extraction through **D**aubechies **W**avelet **C**oefficient **H**istograms (DWCHs). Using two datasets, Dataset A containing 1000 songs over ten genres (this dataset is originally used by Tzanetakis in [48]) and Dataset B containing 756 sound files over 5 genres, music genre classification is evaluated by extracting features like DWCHs, **M**el-**F**requency **C**esptral **C**oefficients (MFCC), **F**ast **F**ourier **T**ransformation (FFT) coefficients, beat related features, pitch related features and combinations of this features. The classification accuracy on any single feature set is significantly better than random guess. The four sets (mentioned above exclusive DWCHs) were combined in every possible way and it could be observed that FFT and MFCC perform best if combined and classification accuracy cannot be considerably improved by adding pitch or beat related features. Main issue of the paper is the comparison of this methods with the DWCH method. The accuracy achieved through the wavelet method exceeds each other feature combination. Classification algorithms used in this paper were Support Vector Machines, K-Nearest Neighbors, Gaussian Mixture Models and Linear Discriminant Analysis.

In 2001 Andreas Rauber and Markus Frühwirth published their work [35][15] about music genre classification. They extracted features from audio signals to enable a

[2] http://www.musclefish.com

SOM to display the relations ie. distances of similarity, between the several music pieces. Their approach is based on the SOMLib Digital Library Project [36] which is capable of automatically organizing digital text documents according to their content. Through adapting the feature extraction process in respect to the audio signal as input values the same principal is used to organize musical content. The audio signal is transformed into the frequency domain using a FFT with about 20 millisecond windows. 17 different frequency bands are used, with each of them represented by 256 coefficients in the frequency domain. Combining these values to a feature vector, 5 seconds of analyzed music are represented by a 4352-dimensional feature vector. Organization in two dimensional space is done in two steps, by first organizing the vectors for the 5 seconds samples of the music pieces on a SOM. The segments are scattered across the map according to their mutual similarity. This makes it possible to find music pieces which have similarity to multiple genres to be located in more than one cluster on the map. As a second step another feature vector is built upon the location of the segments on the first SOM. Counting the segments of a piece of music located in a specific cluster, weighted by the distance to the clusters centroid assembles this feature vector. Training a second SOM using these feature vectors a clustering is obtained, where each piece of music is located in one single location on the output map.

Elias Pampalk continued in his masterthesis, - *Islands of Music; Analysis, Organization and Visualization of Music Archives* [32] -, this work. This work is also published as an article in the New Music Research Journal [37]. He extended the work of Rauber and Frühwirth in respect to psychoacoustical methods to improve the performance. Psychoacoustics is about characteristics of the human perception of sounds, a good overview is given in - *Psychoacoustics, Facts and Models* [53]. Pampalks work gives thought to critical frequency bands we perceive (bark bands[3]) and how the acoustical perception is influenced from each other, this effect is called masking. Further on loudness sensation is considered, the perceived loudness of a tone is depended on its frequency. The human ear is most sensitive to frequencies

[3]in memory of Barkhausen, a scientist who introduced the ”phon”, a value describing loudness level for which the critical band plays an important role.

between 2kHz and 5kHz. Below and above our hearing threshold rises towards to our natural frequency perception limits of 20Hz and 16kHz. He also improved the visualization of the results. The clustered music pieces formerly displayed in table form are now rendered in a two-dimensional map with color coded contour lines. He uses the metaphor contained in the thesis title to visualize the organized music collection. Islands on the map represent groups of similar data items. In the surrounding sea located data items do not belong to specific clusters. Mountain peeks on a Island correspond to the centers of islands. For easy exploration and evaluation of the results, the interface is realized in Hyper Text Markup Language (HTML). The psychoacoustic modifications lead to encouraging results, but are not sufficient for a perfect music genre classification.

In my thesis I try now to further improve the music genre classification results through modifications of the feature extraction process.

Chapter 3

Music Content Analysis

This chapter is meant to give a non technical insight to related aspects of music content analysis. How has this topic evolved through modern digital music storage technologies? Digital processing of music has been a reality for quite a long time now, from studio production to home reproduction and so there is also the logical consequence, that there has to be an impact on music distribution. Although the music industry was not willing to accept this fact at the beginning, problems like illegal music spreading over the internet resulted in heavy losses for them. Right now a rethinking happens and the possibilities of new distribution channels are taken in consideration. This results in the need for supporting technologies. Existing online platforms for direct marketing of songs are for example iTunes[1] and the legal reinstallation of Napster[2].

But now what makes humans capable to differ between different pieces of music or musical genres? A lot of approaches exist for determining the difference between single pieces of music. An important aid for creating such systems is to understand how human beings create such symbols like literal denominations for musical genres and how our auditory perceptive system works. In the sections below answers to the following questions are tried to be given. What kind of information can be

[1] http://www.apple.com/itunes
[2] http://www.napster.com

extracted out of music? What are important criteria for the design of user interfaces through studying human behavior? How are musical genre systems established? What are the important characteristics of the human auditory system?

3.1 Content Processing of Music

The digital revolution on audio did not make that big differences in everyone's music world. Main changes only led to better sound quality and denser distribution carriers and a digital storage of music, while the playback remained in a way the same like using analogue audio carriers. From a sociological view it would be mentioned that the listening process got increasing segmentation, because of a much easier access to single tracks and segments of tracks through the compact disc (CD) technology. This only looks at the consumer sector, in the production environment a lot of changes happened because of new technologies, but this is not of concern here. Through the ongoing progress of increasing computational power, complex mathematical analysis can be done in an reasonable amount of time and so the possibility for new methods of determining and calculating features arises.

3.1.1 Overview content processing technologies

Before mentioning which kind of new applications are possible, because of these changes it is necessary to identify what technologies are available, like described in [1]. The following techniques were developed by ethnomusicologists, specialists of psycho-acoustics, electro-acoustical music composers and sound analysis/synthesis engineers or adapted from other scientific fields to work in a music related area. Some of these techniques are intended to work as real-time application, others are designed to be processed off-line. A future perspective could be that recording or publishing formats already contain meaningful additional non-playback related information for processing applications.

Pitch recognition

Although already successfully done by analogue devices as well as in the digital domain, it is still an active field while trying to achieve higher accuracy. The step from monophonic[3] to polyphonic[4] case has proven to be difficult, multiple pitch recognition in polyphonic contexts delivers results in an single instrument score per detected instrument with errors limited to cases a naive listener would also be challenged. There are still ongoing developments in topics like temporal segmentation associated with pitches.

Beats, rhythm, and dynamics

Time/energy events are an important characteristic in the musical genre recognition field. This concerns attacks, beats, rhythmical structure and phrasing. Beat tracking in real time has received much attention and is today successfully achieved in reasonably complex rhythmical environments. The recognition of full rhythmical or metrical structures is still a challenge but not directly necessary for genre or similarity determination between musical pieces.

Automatic score following

First dealt with through the context of interactive performance, automatic score following is primarily a matching of pitches of music against scores.

Timbre classification

Artificial sound synthesis had the biggest problem in achieving acceptable timbre quality in the beginning. This is the reason why a lot of work has already been done

[3]Melody having a single unaccompanied line [11].
[4]Music comprising two or more relatively autonomous lines [11].

in this field. Studies in psychoacoustics brought scientific results which can be used to develop automatic timbre classifiers.

Temporal segmentation

This is the breakdown of musical content into discrete time segments. Rules were defined, how listeners segment music not only for tonal music, but also in general. Proposals have been done based upon dynamics or on recognized pitches.

Melody extraction and melody matching

Melody information extraction is often called the holy grail of content-based access to music. Querying for melodies (query by humming) is a very popular field at the moment. Promising results are achieved, but a perfect solution does not exist right now. The dominant matching model is contour-based rather than interval-based. This means the matching is done by determining similarities between the query and the dataset in the general form of melody and not by equal tonal intervals. Matching should work in the presence of modifications to the sought melody or being confronted with some non-western music perception models, but these circumstances are really challenging.

Chords, harmony and chord prediction

Assisted analysis of harmony for tonal music can be achieved based on sophisticated pitch detection or of course a score. The challenge lies in determining the right time scale which can easily be done by a human, but it is not easy to figure out the correct basis rhythm of a song by signal processing. Music chord prediction systems specialized upon certain genres of music are working, but a wider set of real music would overextend these systems.

Streams and voices

The ability of the human auditory system to separate different streams or voices in a musical or just acoustic signal and track them over time is very impressive from a technical point of view. Understanding or modeling it has been a key research agenda for the auditory scene analysis community. The approach to achieve a stream-based analysis in the time-frequency domain has proved to be very difficult because of computational complexity and the problem whether and where is the right point in the analysis process to recognize discrete musical objects.

Multifeature classification and retrieval

The content processing topic is the topic this thesis is dedicated to, in the way of indexing whole pieces of music for a retrieval task. But also in a movie production environment, where often specific sounds are searched, or in music productions navigating within a recording multifeature classification and retrieval is used. To see which specific features and bundles are used for such tasks and what they stand for please refer to Chapter 4.2. Although different approaches exist and promising results are achieved this topic is still hot and the demand for applicable solutions is big.

Sonic spaces

A more art related access to the topic are sonic spaces. The idea is to analyze sound documents and calculate multiple features, which are used to give the sound piece coordinates in an three dimensional space (mention that a similar approach is used in [37] as a 2D approach while using the self organizing map algorithm to visualize the output of their Islands of Music application). The user is then enabled to browse through this space using a sound spatializer to perceive the different pieces of music according to their distance.

3.1.2 Impacts on some music activities

If there are new possibilities to analyze music and extract information out of its signal so there have to be new emerging applications. A short overview is given of potentially wide usage and different fields from a consumer's point of view, the broad field of music production excluded again.

Listening and interacting with music

Listening is only at first sight a passive activity. It will be also always the simple perception of oscillating air, but through a technical mediation the listener gets control over what he is listening by choosing and sequencing. The listening experience will get more active for example by extending the acoustic perception with a visual component on a sophisticated level. Lyrics, simplified scores adapted to real time reading and graphical scores which can aid in a pedagogical perspective would be such applications. The main challenge is in the production of new graphical representations that effectively support listening. Interactive listening in its original form happens, if the listener builds the auditory contents he is listening to. At this point the line between perceiving and performing is fading. For such experiments different interfaces exist, ranging from classic mixing consoles to music objects in an abstract version reacting on different obstacles or even interpreting text on a musical alphabet.

Finding, exchanging and selling music

Content-based indexing and retrieval is not only the application this thesis aims at, its also the field receiving most attention (explicit references to work in this area can be found in the Chapter 2). Its appliance ranges from retrieval tasks in large musical repositories like libraries or music catalogs of major labels for music on demand services to the navigation within a single musical recording for navigation from one extract to another similar one. Finding of the right features for effective

classification and distance determination between songs is not the only challenge. A reasonable application also needs a user interface which supports content-based visualization, summarization and auditory browsing. George Tzanetakis focused his PhD-thesis on this whole field, see [43].

Music teaching and critic

The educational aid, audio content processing systems can give, is often underestimated. It has attracted special attention as a practice support for the pupil between his teacher-delivered courses. When trying to develop instrument tuition systems that analyze pupil production and recommend adequate modification or practice, some difficult issues must be addressed. The recorded input signal is of poor quality (no special or high quality equipment should be necessary to make such systems affordable) in maybe noisy environments and simultaneously the features should be processable to compute adequate music pedagogical advice.

Music analysis

Ethnomusicologists are the pioneers of content processing techniques with regard to music transcription. They are faced with the need for a common non-musical notation to analyze all the different musical styles and concepts over the world. So they have an insight to a broad range of different notation models. A system with the ability to transform for instance feature-based signal representations in a score-like representation would be of great interest.

For explicit examples and their references please refer to the original article [1].

3.2 Music information retrieval - Human interaction

The human behavior of seeking music titles someone prefers, creates the requirements on the design of digital library systems. A study dedicated to this topic was

done by Sally Jo Cunningham [10] described in more detail below. In the sound domain an experimental prototype, called ”Sonic Browser” [4] was implemented by Eoin Brazil and Mikael Fernström to test user behavior and elicit the effectiveness of different representation of relations between sounds. Also Tzanetakis deals with this topic in the interaction chapter of his PhD-thesis [43]. It is important to find profound knowledge about human behavior in music retrieval situations, because the risk is high that music information retrieval systems are developed without concerning the requirements of their eventual users.

3.2.1 Human behavior patterns

The following points are the results of a study done by Sally Jo Cunningham [10]. This study was done by using instruments of the empirical social research [40]. The social research knows qualitative and quantitative methods. Quantitative methods are used to ”measure” the accuracy of propositions in a sense of creditableness and objectivity. Here a quantitative method was used, because these methods are suitable for acquiring new insights to a situation. Participating observation and think-aloud protocols were used to attend the test persons in a shopping environment and led to the following results.

Searching

Searching for music - the locating of a desired music media in a collection environment like a shop or a library. The process of searching and browsing figures out to be a casually interleaved activity. Searching for a known-item can easily be followed by relatively unoriented browsing. This behavior is likely encouraged through the disposition of articles, which itself is adjusted through marketing techniques based on studies of human behavior. Music retrieval systems can use multifeature classifications on one hand to give correct answers to explicit defined search queries or on the other also enable the user to browse by presenting objects of low distance at once, based on different features. Features guiding the search are mainly

bibliographic (artist/group name, album name, track title), this indicates the importance of quality bibliographic metadata in a music digital library. The enhancement against the ”real-world” situation should be to give the user the possibility of different search interfaces, like bibliographic or acoustical in terms of query-by-humming, and to give fuzzy answers, because the queries are also formulated with high fuzziness. Different possibilities of sorting for fuzzy answers are also important, because of the partial information of the searcher about his desired items, the manual search through query results can be simplified. Results of a query should not be only explorable in textual format, but also by streamed sound examples. Through the, for bandwidth issues, diminished audio quality and the fragmentary nature of these examples the copyright issues of delivering music samples is solved and it is convenient for customers, because there is no long time delay between the request for a sample and the playback. An important advantage of a digital library also is the information about the availability of searched items. In a shop or library the customer can only infer of not finding a desired item that it is not available, but it gives no information about, if the item is in catalog at all.

Browsing

Browsing in general is differentiated from searching, in that the browsing process is more exploratory and less directed than searching. Browsing is like search by random, walking through a shop or library flipping CD’s and being influenced by what the customer can see. The only clue for customers for browsing seems to be the genre classification of the music catalog in a shop or library. The topic of musical genre definition and classification is discussed later, but the point of classification is of importance for the browsing process as well. Customers prefer a genre classification which suits their needs and own personal view. The loose boundary between different genres makes it difficult but in a physical shop or library environment necessary to assign an item to specific genre, because an CD is an object and can only be displayed on one location at once. A general genre classification scheme would make thinks easier but as noted in [2] this not an straightforward process (more in

Sec. 3.3). The use of a music information retrieval system can or should incorporate aspects like assigning more than one genre to a CD and eventually using more than one genre scheme in its system. Supporting a good browsing function has to be based upon understanding how users define, perceive and describe musical genres.

Collaborative Aspects

Music Shopping is not a task attended by people on their own. It is very often a collaborative task, where the groups of people are friends or families. The collaborative aspect for shopping is big while it is not in a pure information request environment like a library. If a music information retrieval systems should be applied in a commercial way, it will not be easy to support functions which give rich collaborative possibilities. Instruments like instant messaging, chat rooms or a function which gives insight on what other people are listening to right now can try to create a collaborative atmosphere in the digital case. Functions like ”customers who bought (found) this also bought (requested) the following products” [38], user ratings and reviews by users or professionals are only possible in the digital domain and can extend the collaborative experience. This function, especially in a commercial setting, brings the problem that generally unknown items never come up, because people are only linked to sold products. A system, able to determine distance between musical pieces, based upon the content, can very usefully extend this functionality, by presenting ”near-by” pieces of music even if they were not sold yet. A paper dealing with the collaborative idea extended with a small feature based music recommendation system can be found in [6]. From the aspect of the collaborative idea it is to question if a music genre classification without involving the users interests or behavior makes sense and if a system combining these aspects can improve results.

Visual Aspects

The visual recognition of desired items, found by browsing or searching, is most important for users. It is the fastest way to identify known items and speeds up

searches, this is valid for the digital as well for the non digital music library case. In a way also genre specific information is extracted by customers out of the artwork of CD booklets. This is not always possible, but again this completes the palette of deduced information of a CD and the booklet.

3.2.2 User interfaces

The success and usability of a human-interacting application always lies in a great part in its user interface. Through the fact that music information retrieval systems are not yet fully technically mature and are still in a research state the emphasis often lies in finding the right features and their calculation methods for the right application. But the discovering of feature sets expressing certain musical properties of a music piece, also gives the opportunity to implement a new kind of interfaces. Also to mention is, as long as these systems work imperfectly they are complemented by human input, this yields to the realization of a user friendly graphical user interface (GUI). In the following a few interfaces are mentioned, which build upon the concept to map audio files to visual objects with specific properties related to their content. Most of the following kinds of GUI's where mentioned by Tzantekais [43] which also contains the Sonic Browser [4]. The Island of Music project by Pampalk [32] this thesis is based upon is also mentioned.

Browsers

This enlisting of browsers does not claim to be complete, it is intended to show what realizations already were done in a prototypical way. *Standard browsers* provide a simple textual list of audio files of a collection. The user can select one or more files for further processing. Another visual representation are *tree browsers* [4] (see Figure 3.1). The tree can be generated manually or automatically. Lower levels (closer to root) of the tree are more general, getting more and more detailed the higher the tree level. For example first differ between Music and Speech and Speech is subdivided into Male and Female and so on. The *Sonic Browser* or *SoundSpace* is based

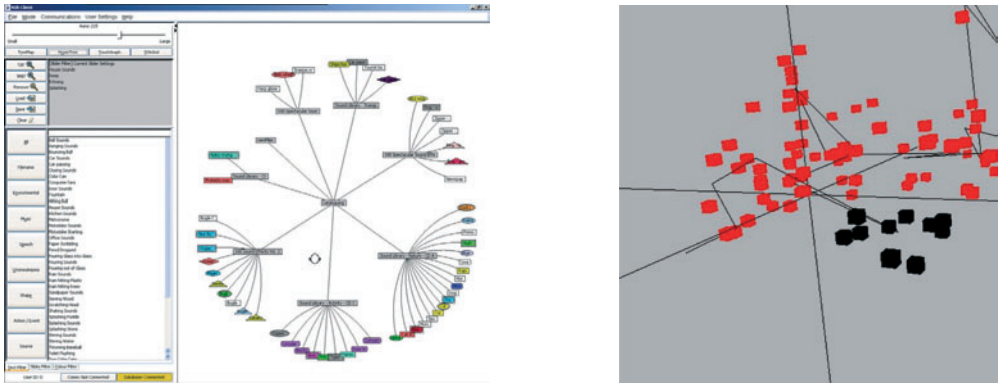

Figure 3.1: On the left a *tree Bowser* visualization is presented, using a tree structure combined with different shapes and colors. The image on the right shows the *Timbrespace Browser* with colored cubes representing different audio objects in three-dimensional space.

on the so called Cocktail party effect. Music objects are arranged in a three dimensional space. The user hears a certain count of the closest objects. Moving closer to an object will increase its volume. So the user can explore a three dimensional space aurally. The *Timbrespace Browser* [47] (see Figure 3.1) locates audio objects in a two-dimensional or three-dimensional space. Objects are not only characterized by their position in space, but they also have certain shapes, textures, colors or text annotations. As underlying principle the so called Principal Components Analysis (PCA) is used. In short this is a dimensionality reduction technique to project large feature vectors into smaller ones like for example three-dimensional coordinates. It has not to be that all three-dimensions are determined through timbre it can also be combined with other features. The large feature vector contains timbral information as the name of the browser let it presume already. A two-dimensional visualization is the *Timbregram* [44] (see Figure 3.2), where the input files are divided in short time frames and depending on timbral features, reduced again by PCA, this short frames are color (or greyscale) encoded. The *islands metaphor* (see Figure 3.2) used by Pampalk in [32] is an approach, which uses the SOM algorithm to locate the pieces of music, represented by the calculated feature vectors, on a

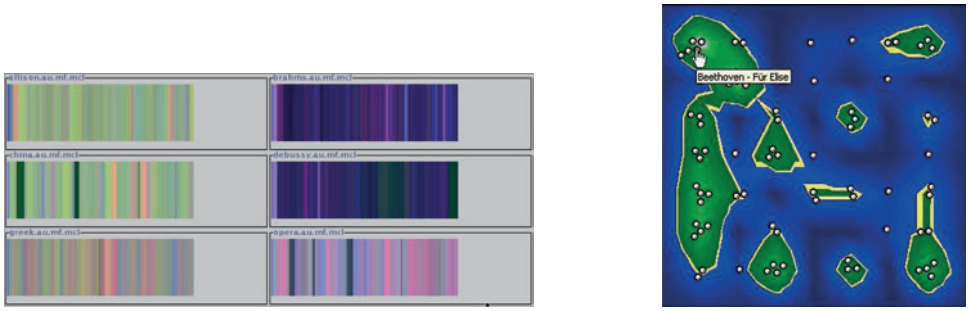

Figure 3.2: The left image shows a *timbregram* containing speech (left) and classical music (right). On the right hand side an example of the *island methaphor* can be seen. Dots on the map represent the location of pieces of music.

two-dimensional map with contour lines. Again the distance between the certain pieces of music expresses their similarity and the contour lines are used to represent density in different areas in the digital music library at all.

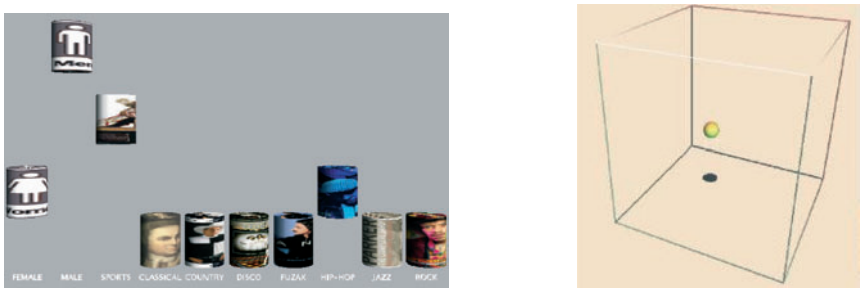

Figure 3.3: Monitoring different genres through a *GenreGram* by the movement of the cylinders along the y-axis is shown on the left. The *TimbreBall* Monitor on the right is used to evaluate features. Sudden jumps can be interpreted as sound texture changes.

Monitors

Monitors visualize audio content information in real time. Commonly known visualizations are simple display of the *waveform* of a audio signal, *spectrums* dis-

playing magnitudes of signal in frequency bands calculated through Fourier Transformation (see Sec. 4.1.1), *spectograms* rendering the same information like *spectrums* in images as gradients, *waterfalls* showing cascading *spectrum* plots in a three-dimensional way, simple metronomes displaying the beats-per-minute (bpm) eventually how strong the beat is and feature plots of user defined features. Tzanetakis [43] also proposes a *GenreGram* [44] (see Figure 3.3) designed for visualizing automatic genre classification results for real-time streams (it also can be used with any kind of audio information, real-time or not). Genres are visualized by moving cylinders. The vertical position indicates the affiliation of an input signal to a genre, the higher - the more accurate a genre determines the current input. It gives valuable feedback to users as well as to classification algorithm designers, viewing different classifications decisions and their strength. This is much more informative than a single decision. The *TimbreBall* [44] Monitor (see Figure 3.3), a small ball in a cube, visualizes in real-time the evolution of extracted feature vectors. The, in the three-dimensional space mapped vectors make the ball move, and sudden jumps of the ball can be interpreted as sound texture changes.

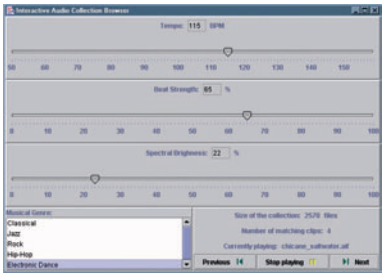

Figure 3.4: This illustrates a *Sound Sliders* query interface. Query parameters are manipulated through shifting the different sliders.

Query interfaces

As this whole thesis concentrates on discrete audio signal as input, query interfaces concerning MIDI related audio content are disregarded. The main idea of the query

interfaces is to fully utilize the already analyzed structure and content of the underlying audio collection the query is matched against. This gives the opportunity to provide instant aural feedback to the user and thus results in an intuitive search process. *Sound Sliders* [51] (see Figure 3.4) are query interfaces which give instant and continuous feedback. As example tempo and beat strength are known for a whole collection. A slider input device lets the user adjust to a certain beat. Music pieces with a related beat are enlisted and, if the query result is larger than one, the top match is played back instantly. The user can skip to the next result by clicking a button. Shifting the slider results in a new query and new results to which the system crossfades automatically from the actual play back. *Sound palettes* are the discrete case of *Sound Sliders*. A palette with a fixed set of visual objects (text, images, shapes) is used to describe discrete attributes of a music piece. The playback of query results is like in the case of *Sound Sliders* instant and continuous. *Loops* are important parts of the last two techniques. The problem of looping a certain piece of music is much often to long. Beat detection can be used to find a adequate playback starting point. It makes no sense to play a song from the beginning in the case it has a long quiet intro and the representing part kicks in later. Crossfading between the pieces makes the listening experiences more pleasant.

3.3 Musical Genre

Before looking on usage and function on categories in music, it may be interesting to take a look at categories in general, like [14]. A very early version of defining the term ”category” was done by Aristotle[5]. According to his idea a category is ultimate and most general predicate, which can be attributed to anything. Categories are used to define entities by relating them to their general essence in the sense of substance, quality, quantity, relation, place, time, position, condition, action or passion. This idea is different from Immanuel Kant's[6] definition. Kant does not see

[5]Aristotle was a philosopher in the ancient Greek; pupil of Plato; tutor of Alexander the Great (384-322 B.C.)

[6]German philosopher of Modern times (1724-1804)

categories as designators of essential modes of reality, but the way we know them. They are universal forms that phenomena must take to become objects of knowledge. He formulated twelve categories under four general titles (quantity, quality, relation, modality) containing thesis, antithesis and synthesis patterns. The modern usage of the term ”category” by cognitive psychologists is different again. Beside this different and complex views, researchers agree in that kind, that categories are classes of objects or events, which humans create to reduce the complexity of the empirical world.

3.3.1 Generic Rules

Musical genre is a set of musical events, whose nature is governed by a definite set of socially accepted rules. Thus it is a set, subsets exist. Musical events can also been situated in the intersection of different subsets. Fabbri [13] uses the term ”musical events” in the sense of: any type of activity performed around any event involving sound, to have an, as broad as possible, definition, so a kind of musical event, not found to be music by objectors, does not exist for his set of rules. The following set of rules for forming genres is intended to show the complexity of the problem and not to resolve the problem. It should help understanding involvements by building a genre system. The rules have no specific hierarchical order. It is to mention that for the description of a single genre some rules are more others less important.

Formal and technical rules

A major role in all musical genres play formal and technical rules. Each genre has its typical form, even if the opposite, in terms of a form being sufficient to define a genre, is not true. Rules exist in written form, theoretical disquisitions or teaching manuals, as well and not less important, rules passed in oral tradition or through models. This is also true for rules referring to performance techniques and instrumental characteristics. For example a transverse flute player in an classical

orchestra and one in a jazz band are certainly on the same level from the point of view about playing at sight and memory, but differ in embouchure, extension, improvisation and interpretation of sheet music. The complexity of a musical event is very high and that is the problem while being confronted with a genre for the first time, because its unclear which specific attributes it omits and which are important. Further attributes are the choice between ”musical sounds” and ”noise”, note systems, conception of musical time, importance of elements like melody, harmony, rhythm and the level of complexity an entire musical system owns. When facing musical genres characterized by the existence of text, the formal and technical rules referring to it should also be taken into consideration. The use of syntax, meter, lexical choices contribute to the identification of a musical genre.

Semiotic rules

All mentioned rules are semiotic, because they are codes creating a relation between the expression of a musical event and its content. In this context here these rules are nearer to the expressions done through musical events. Its about the narrative form, which lies in some genres both in text and in the music, determining different genres. There are rules relating to the communicative functions: referential, emotional, imperative, phatic, metalinguistic, poetic factors of a message. Imperative messages for example mostly relate to dance music, phatic can be used to describe rather ”background music”. Parallel to the musical rules there are also gestural and mimic codes, not only of the performing artist, but also of other interprets, conductors or listeners.

Behavior rules

These rules are on one hand about the psychology of musicians (performers) and their reactions when faced with an audience. On the other hand also about the behavior and psychology of the audience. Each genre is characterized by rules of conversation, smaller or larger rituals. These rules are used to make the circle of

sympathizers of a genre exclusive and quickly expose any intruder, who is not well informed.

Social and ideological rules

Sociological information can get part of genre rules. For example the occupational distribution, typical to listeners of a genre, or age groups and social groups related to a genre can become genre rules. These rules are also used by single individuals taking the chance to deny their affiliation to a certain group by listening to music of a specific genre. Ideological or political connotations of a genre sometimes can exist to.

Economical and juridical rules

Not only now, where the music industry is under pressure, because of illegal music exchange, there are also economical and juridical rules, deciding the survival and success of a genre.

3.3.2 Categorizing Music

The field of rules for building genres is wide spread and sound sometimes very abstract. But as a fact listeners process their perceptions confronting them with cultural cognitive types, to determine what they are listening to [14]. They thus are assigning the perceived music to a genre in their very own genre system. The aim of some music information retrieval systems is to do so in an automatic way.

Used Genre Taxonomies

Existing music information retrieval systems do already use predefined genre systems [2][14]. It is important to say that genre is an intentional concept. Alternatively

genre can be seen as an extensional concept in the meaning of a set of music files sharing certain attributes (features). In idealistic, mathematical worlds these two concepts coincide, in real world they do not. A closer look to different used genre approaches and their classification approaches is given and discussed.

Genre taxonomies are used by record labels, record shops, music charts, music related web sites and online record shops, online radios and of course in music related press and books. Genre description in use is more related to whole albums than on single titles, because industry is selling more albums than singles so far. This will dramatically change with the success of online music retail, which is based rather on titles than on albums. Album oriented Taxonomies are inappropriate for describing titles, except for top level categories (Rock vs. Classical). A comparison of different internet genre taxonomies [2] shows, that there is no conformance between the different genre systems. There are not many shared terms in the different systems, and more important the structure as well as the position of genre classes within the system differ. It was observed that not even terms like ”Rock” or ”Pop” have a common definition between these systems. The semantics of terms for genre denomination are not consistent within a system. They are denoting periods, topics, locations, languages, dance types, artist types and so on. Such a confusion leads to redundancies and thus makes it complicate to be incorporated into automatic classification systems. Interesting is, that it does not have a strong impact on the browsing and searching by users. Like the rules for genre determination, have shown also Pachet and Aucouturier come to the conclusion [2] that musical genres are not only founded on intrinsic properties of music, but also on cultural extrinsic habits.

Music information retrieval projects devote a substantial part of their time in designing a genre system. The time used to *classify songs manually* is enormous too. The manual approach was tried at the Cuidado project [31] and considered genre as an additional metainformation like country, artist type, instrumentation, etc. Combining these it was possible to limit the explosion of genres needed. Additionally the genre system included similarity relations based upon inheritance, string matching and expert knowledge. This approach was canceled because, the bottom genres

were difficult to describe objectively, so no one other than than the designers were able to distinguish between them, and the genre system was very sensitive to music evolution. The development of new genres was difficult to integrate through multiple inheritance issues. Furthermore music evolution induces genre compression (merging of genres) and expansion (splitting into sub-genres). This made the genre system difficult to build and impossible to maintain. The experiment is no proof of impossibility but the Cuidado project changed to a genre system organizing artists rather than titles.

Prescriptive approaches to automatic musical genre clustering attempt to automatically deduce information from the audio signal to perform a genre classification. The works reviewed in [2] all assumed that a genre system is given and the analyzed songs should be assigned appropriately. The genre systems used are all very small, simple and incomplete and look more than a proof of concept than a useful genre labeling system. Problems of this systems are that for different genres, different feature sets are most descriptive. Using the same descriptors for all titles leads to improvable results, but automatic feature selection is a difficult task. The descriptor should vary for different genres. Adding genres, which is necessary because of the permanent development of new genres, or titles to the system will modify the optimal feature set. The training set of sound samples can not be guaranteed to be representative enough for the whole domain to select the right feature set. These automatic systems can, at the time, only present results for a small amount of genres containing no sub-genres. This is a very difficult task, because characterizing differences are harder to find and have to be adjusted to the sub-genre they describe.

Similarity relations used for *automatic genre classification* start with no given genre system. Such an classification approach tries to determine the similarity between different pieces of music by computing intrinsic features, as the descriptive approach does, and tries to determine cultural similarity by data mining principles. For example collaborative filtering is used, based on the idea there are common patterns of taste. Such patterns could be extracted through user behavior, but it is not possible to guarantee that a behavior of taste or shopping is linked to genre

similarity. The problem of collaborative filtering system is, that they return promising results as long as user profiles are small, but with the growth of the profile the results decrease in quality. Other sources for additional information for deducing genre information can be radio programs, because the choice of songs for most radio stations are not arbitrary, or track listing databases, especially tracklistings of samplers/compilations. Drawbacks of such solutions are, that the resulting clusters are not labeled and they only can work for pieces of music occurring in the underlying data sources, unknown songs could not be classified.

3.4 The Human auditory system

The capability to differ between genres, through our aural cognitive system, is astonishing good. R. O. Gjerdigen has shown in an experiment [33] that very short outtakes of a song are enough for humans to assign pieces of music to genres. The probands were students with no remarkable amount of musical education or experience. The average hours of music listening per week they stated was about 24 hours. 10 different genres, acquired in an informal survey before the experiment, were used (only top-level genres, non hierarchical genre system) into which the probands were asked to sort the music titles (Blues, Country, Classical, Dance, Jazz, Latin, Pop, R&B, Rap, Rock). The used sound samples were taken from large internet CD vendors and excerpts were done in the size of 3000ms, and from this one again excerpts of 250ms, 325ms, 400ms and 475ms. The results are interesting. For the 3000ms excerpts the agreement between the genre classification of the CD vendors and the probands was as high as 71%. It also could be observerd over all pieces of music of different lengths, a slightly better performance was done on instrumental pieces of music than on vocal ones. The congruence between the 3000ms results and the 250ms is at about 44%. Still a valuable amount, random guess would lead to a accuracy of 10%. This is astounishing because a time window of such small size, does not let the probands determine features like rhytm, melody and conceptual structure of the song. They have apparentely to perform the classification solely based on

their spectral and timbral perception. Nontheless rhythm related features do contain important genre information.

So lets take a look at the human auditory system. At first a short look on the anatomy of the human ear is done and afterwards a closer look to psychoacoustic principles is given. They play an important role in the prototype used for rhythm related features by this thesis.

3.4.1 Anatomy

The human ear, schematically shown in Figure 3.5, can be divided into three parts [53][8], referred as outer, middle and inner ear. Collecting sound energy and transmitting it through the outer ear canal to the ear drum, is the task of the outer ear.The outer ear protects the middle and inner ear from damage and enables the other parts of our aural system to be located close to the brain with the effect of a reducing the length of the transmitting nerves. The middle ear is the transformer of the signal represented by oscillating air particles into oscillations of the fluid the inner ear is filled by. This is done by the ear drum, working as a pressure receiver over a wide frequency range. It is firmly attached to the long arm of the hammer (malleus). The response of the eardrum is transmitted by the three bones hammer (malleus), anvil (inus) and stirrup (stapes) to the entrance of the inner ear. The entrance of the inner ear is formed out of the footplate of the stirrup (stapes) and the so called oval window. If the oval window is stimulated by the movement of the three middle ear bones the pressure has to be equalized, which the round window is for. In the inner ear the oscillating fluid stimulates hair cells connected to nerves leading this stimulus to the brain.

The human hearing area (as a graph of frequency and sound pressure) is depicted in Figure 3.6. The frequency scale on the abscissa is logarithmic as well as the sound pressure level in dB on a linear scale on the ordinate, because the sound pressure level is logarithmic related to the sound pressure. The unit of sound pressure is the Pascal (Pa). Sound pressure p and sound pressure level L are related by Equa-

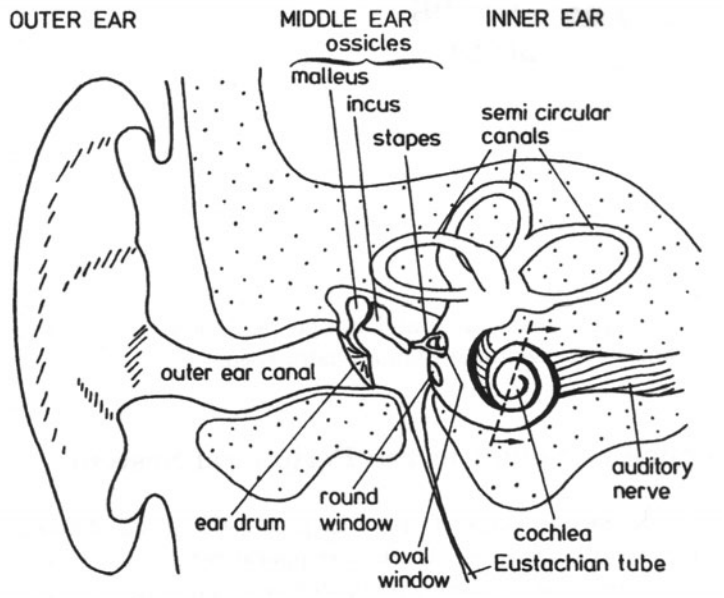

Figure 3.5: Schema of the human ear.

tion 3.1.

$$L = 20log(p/p_0)dB. \tag{3.1}$$

As reference value of the sound pressure level p_0 is standardized to $20\mu Pa$.

The hearing area is the space between threshold in quiet and threshold of pain with a frequency range of about 20Hz to 20kHz. Dashed areas of music and speech depict the allocation of their responding frequency domains. For speech this is 100Hz to 7 kHz and for music about 40Hz to 10kHz.

3.4.2 Psychoacoustic effects

Psychoacoustics, the branch of science, dealing with hearing, the sensations produced by sound is given special attention in this thesis. While work in the audio content processing field mainly does not give special attention, it is done in [37] and in this thesis. This approach does a psychoacoustic related preprocessing to

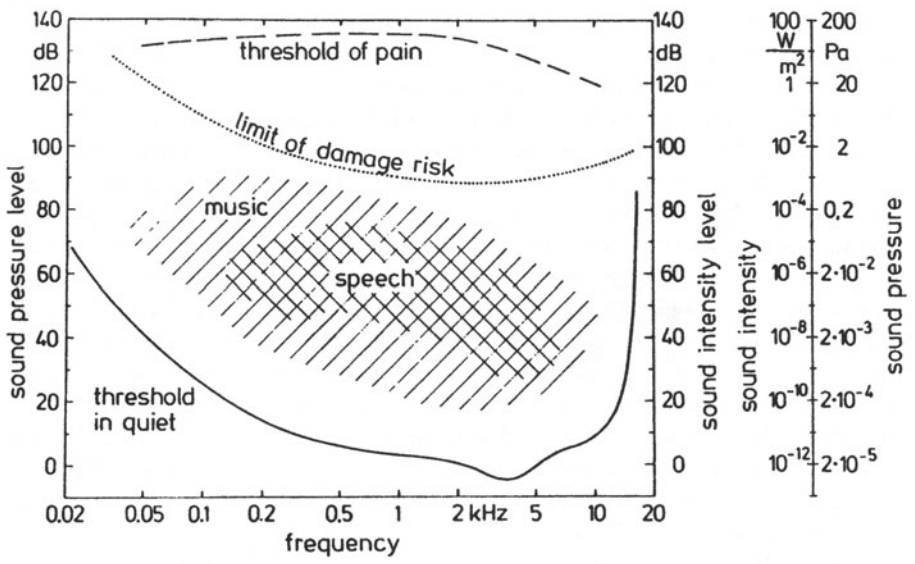

Figure 3.6: Human hearing area.

the underlying input signal (music) before extracting classification relevant information. Effects like masking, loudness sensation, critical bands and roughness are considered. Background information to this effects is given below.

Critical Bands

Critical bands were proposed by Fletcher, as shown in [53]. The assumption is that a pure test tone, masked by noise, is masked through the spectral components of noise lying near the frequency of the test tone. His additional assumption was, if the power of the test tone and the spectral components of the noise lying near by is almost equal, only these spectral components are of interest. Frequencies more far from the test tone do not effect the masking. Under this concept characteristic frequency parts can be defined which are relevant for masking the pure tone. Experiments with different variations of bandwidth determination showed the result (see Table. 3.1), that the critical bandwidths below a frequency of 500Hz are about

100Hz. Above there is a bandwidth of about 20% of the center frequency of the corresponding band. The particular upper cut-off frequency of a band is equal to the lower cut-off frequency of the critical band above. This critical band rate scale describes the property of our hearing system, analyzing a broad frequency spectrum by splitting it into certain frequency bands. Anatomical this effect is associated with the basilar membrane, located at the round window (see Figure 3.5), and its hair cells, which are responsible for analyzing different frequency bands.

Loudness

The eponym of the bark-scale, Barkhausen, introduced the loudness level measure in the twenties of the 20th century. Its level is defined by the sound pressure of a 1kHz tone in a plane wave and frontal incident, that is as loud as the sound. So called equal-loudness curves (Figure 3.7) describe the different loudness levels for different frequencies of pure tones. Loudness can be measured for any sound, but most often it refers to the levels of pure tones. This curves are the result of laboratory measurements. The unit of the loudness level is the ”phon”. The level of phone is given through the 1kHz point of a curve and the corresponding dB level. Threshold in quiet (Figure 3.7) is a loudness curve as well. Its 1kHz point corresponds to 3dB, so the threshold in quiet is referred to as 3phon. The equal-loudness curves for low loudness levels are almost parallel to threshold in quiet. The dip between 2kHz and 5kHz describes the frequency band we are most sensitive to. For low frequencies the curves get more shallow the higher the level. Above 200Hz the curves are nearly the same for different phon values.

For loudness sensation, expressed in ”sone”, the level of 40dB of a 1kHz tone was proposed to give the reference. This expresses the subjective loudness perception of a tone. The pure tone at 1kHz with 2sone is subjectively double as loud as the 1kHz with 1sone. The threshold in quiet, which is the corresponding 3phon curve, equals 0sone. As it was shown through experiments, above 40dB this relation follows a power law (see Equation 3.2).

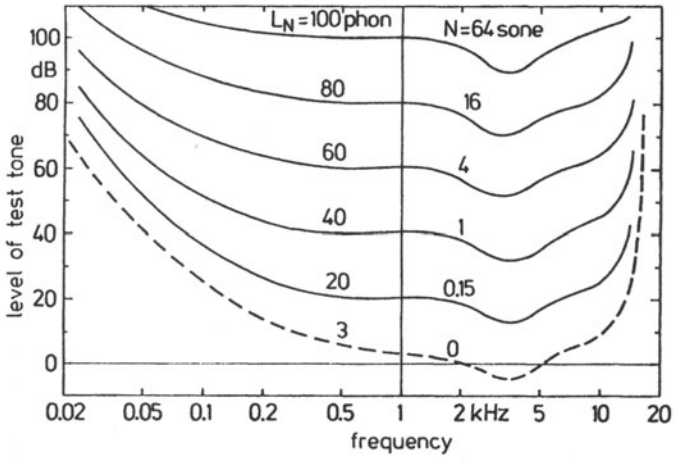

Figure 3.7: Equal-loudness curves of pure tones in a free sound field.

$$L_{sone} = 2^{\frac{1}{10}(L_{phon}-40)} \mid L_{phon} \geq 40 \tag{3.2}$$

Below 40dB, following [3] the loudness sensation is approximated by Equation 3.3

$$L_{sone} = 2^{\frac{1}{40}(L_{phon})^{2.642}} \mid L_{phon} < 40 \tag{3.3}$$

Masking

The effect of masking [53], describes the suppression of the perception of a pure tone by the presence of other tones or by noise. Masking depends on *spectral* as well on *temporal* relation between masker and the masked one. The effect of masking is measured by determining the threshold of masking. This describes the sound pressure level of a test tone, necessary to be audible in the presence of a masker. For frequencies, very different for masker and test tone, this level equals the threshold in quiet. Pure tones masked by white noise show above the threshold in quiet a constant threshold in the frequency range up to 500Hz. For higher test tone frequencies

the the threshold increases for about 10dB per Octave[7]. Masking by narrow band noise, noise with a bandwidth equal to or smaller than the critical-bandwidth, is shown in Figure 3.8. Ascending from low frequencies an steep increase of the masking threshold can be observed. After reaching the maximum the curves show a more shallow decrease against high frequencies. The three curves also show a decreasing maximum for higher center frequencies. Is the difference between the horizontal 60dB line at 250Hz only 2dB, so it is already 3dB at 1kHz and even 5dB at 4kHz. The spectral masking effects concerning two different pure tones are very complex and exceed the frame of this thesis.

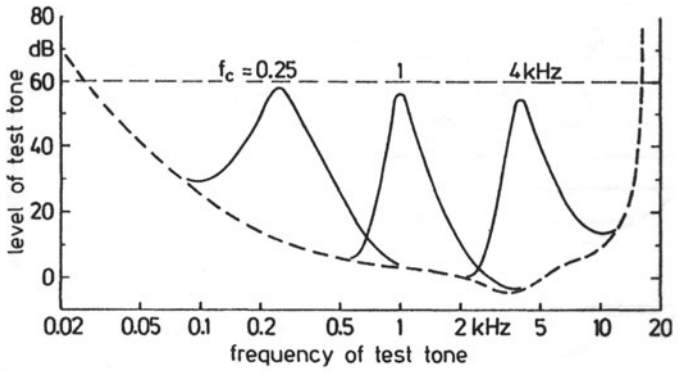

Figure 3.8: Test tone masked by critical-band wide noise at 60dB and center frequencies of 250Hz, 1kHz and 4kHz.

All of the above mentioned spectral effects of masking are analyzed in steady states with long-lasting test and masking sounds. Masking effects do not only relate to spectral properties, but they do also to temporal effects. They can be divided into *premasking*, *simultaneous masking* and *postmasking* effects (see Figure 3.9). The mentioned figure has two time scales on the abscissa. The first zero point describes the onset of the masker. The effect occurring before this point is referred to as *premasking* effect, which are still a lot of uncertainties about. This does not mean the hearing system can hear into the future, but describes the build up time of the

[7] An Octave is a music interval embracing eight diatonic degrees and equals a frequency band with limits having an ratio of 2 to 1.

aural perception. The effect plays a more secondary role, because the time span of interest only lasts 20ms. In *simultaneous masking* there are dependencies on duration as well as on repetition rate of an test tone. These two dependencies show an identical behavior for threshold in quiet and threshold of masking. For an test-tone burst the threshold decreases with 10dB per Decade of the burst duration up to 200ms. Above the thresholds are constant. This behavior can be described by assuming that the hearing system integrates the sound intensity over a time frame of 200ms. The second zero point in the abscissa in Figure 3.9 describes the point, where the masker is switched off. *Postmasking* is strongly dependent on the duration of the masker Δt.

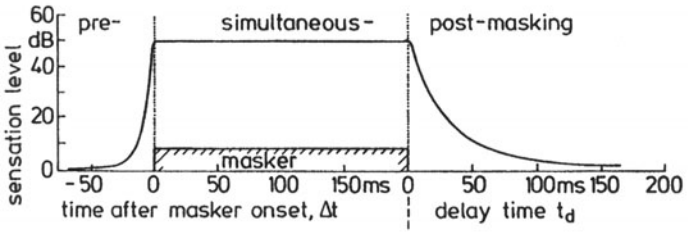

Figure 3.9: Temporal characteristic of masking effects.

In context of this thesis a spreading function is used to incorporate the *simultaneous masking* effects over critical-bands [39][3]. This function describes the influence of the j-th critical-band on the i-th critical-band (see Equation 3.4).

$$S_{i,j} = 15.81 + 7.5(i - j + 0.474) - 17.5\sqrt{1 + (i - j + 0.474)^2} \tag{3.4}$$

Fluctuation strength

The amplitude modulation of a acoustical signal produces, dependent on the modulation frequency, different effects. Lower modulation frequencies (up to 20Hz) produce an effect called fluctuation strength. For higher frequencies (15Hz to 300Hz) the sensation of roughness occurs, see below for more. There is no strong border

between these effects, its more a smooth transition. Fluctuation strength is a sensation which can be analyzed leaving any other effects out of sight. The course of fluctuations strength over modulation frequency can be described as band-pass characteristic with a maximum at 4Hz. This can be seen as an correlation between our hearing system and the nature of speech, because a most common speed of speech at 4 syllables per second can be observed. An increase in sound pressure of the modulated signal also leads to an increase in fluctuation strength.

For the prototype, concerning psycho acoustical effects, used in this thesis a model of fluctuation strength based on temporal variation of the masking pattern is applied [53]. The relationship between the fluctuation strength F and the modulation frequency f_{mod} as well as the masking depth ΔL can be described by the following equation. ΔL describes the difference between the maximum and the minimum of the temporal masking pattern.

$$F \propto \frac{\Delta L}{(f_{mod}/4Hz) + (4Hz/f_{mod})} \tag{3.5}$$

Roughness

Roughness, the sensation caused by amplitude modulated tones with a modulation frequency between 15Hz and 300Hz. This effect reaches its maximum around 70Hz of modulation. From 150Hz to higher modulation frequencies the sensation of three different tones arise, which is not of interest, because of the related high modulation frequency. The most important parameters roughness is dependent of, are the degree of modulation and the frequency of modulation. Above an degree of 25% of modulation roughness gets sensible. The relation between roughness and modulation degree can be aproximated linearly and is independent of modulation frequency. The modulation frequency relation of roughness is bandpass characteristic and the maximum of roughness is only dependend of the carrier frequency. Maximum is reached at a carrier frequency of 1kHz. This can be originated by frequency selectivity of the ear for lower carriers and the limited temporal resolution for higher

ones.

An approximation for roughness [53] would be given through the following equation.

$$R \propto f_{mod} \Delta L \tag{3.6}$$

z	f_l, f_u	f_c	z	Δf_g	z	f_l, f_u	f_c	z	Δf_g
Bark	Hz	Hz	Bark	Hz	Bark	Hz	Hz	Bark	Hz
0	0				12	1720			
		50	0.5	100			1850	12.5	280
1	100				13	2000			
		150	1.5	100			2150	13.5	320
2	200				14	2320			
		250	2.5	100			2500	14.5	380
3	300				15	2700			
		350	3.5	100			2900	15.5	450
4	400				16	3150			
		450	4.5	110			3400	16.5	550
5	510				17	3700			
		570	5.5	120			4000	17.5	700
6	630				18	4400			
		700	6.5	140			4800	18.5	900
7	770				19	5300			
		840	7.5	150			5800	19.5	1100
8	920				20	6400			
		1000	8.5	160			7000	20.5	1300
9	1080				21	7700			
		1170	9.5	190			8500	21.5	1800
10	1270				22	9500			
		1370	10.5	210			10500	22.5	2500
11	1480				23	12000			
		1600	11.5	240			13500	23.5	3500
12	1720				24	15500			
		1850	12.5	280					

Table 3.1: Critical band rate z, lower (f_l) and upper (f_u) frequency limit of critical bandwidths, (Δf_g), centered at (f_c)

Chapter 4

Feature Extraction

The original representation of music as an input signal is useful for playback but it not for processing and classification. In this chapter, I will look at the different possibilities to extract information out of the input signal, which kind of features can be extracted and discussing these features in detail.

As shown in Figure 4.1 the raw audio input signal is transformed via signal analysis techniques and further analyzed in the feature computation process, which results in n attributes $x_1, ..., x_n$ of the feature vector v.

After taking a more detailed look at Fourier- and Wavelet transformation an insight

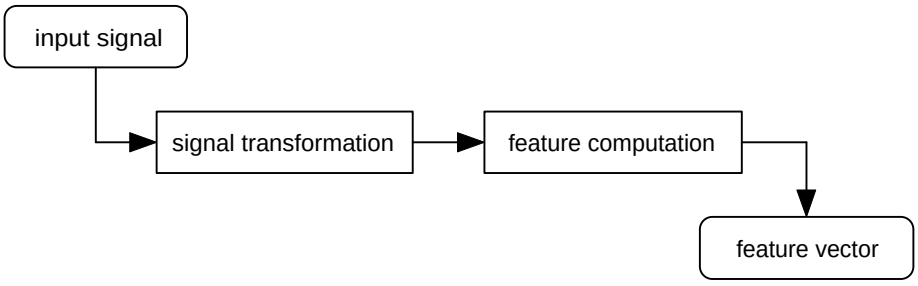

Figure 4.1: Feature extraction process overview.

on different features, used by the music content analysis community, is given.

4.1 Signal Transformation/Analysis

To receive an appropriate computational representation the input signal is transformed from its time-amplitude domain in the time-frequency domain. This is done by using a Fourier based transformation or a wavelet signal decomposition. The motivation behind a transformation is that some properties are much easier computable or are computable at all in the frequency domain.

For example compare the solution of the following problem (calculate the quotient $y = a/b$) in its conventional and through a transformed analytic way. Assume that no calculator is accessible for this task and a high accuracy is required. If this problem has to be solved often with different values its very time-intensive. The alternative would be to do a logarithmic transformation of the analysis $\log(y) = \log(a) - \log(b)$. Using a "classic" log-table the values are quickly transformed and the analysis is a simple subtraction, delogarithmizing via a look at the log-table again gives the final result. The solution of the problem in the transformed may not be really more convenient, but the example should demonstrate the basic idea of transforming a problems solution from one domain in another.

4.1.1 Fourier Transformation

A short overview of the Fourier Transformation (FT) and its variations is given here, for a more detailed explanation see [5][8]. Its idea is based on the fact that any continuous periodic signal can be represented as the sum of properly chosen sinusoidal components. It was first discovered by Jean Baptiste Joseph Fourier (1768 - 1830), a French mathematician and physicist. The Fourier Transformation is not only one dimensional, like the logarithm; transforming a single value y in a single value $\log(y)$, it transforms a function of a variable, defined from $-\infty$ to $+\infty$

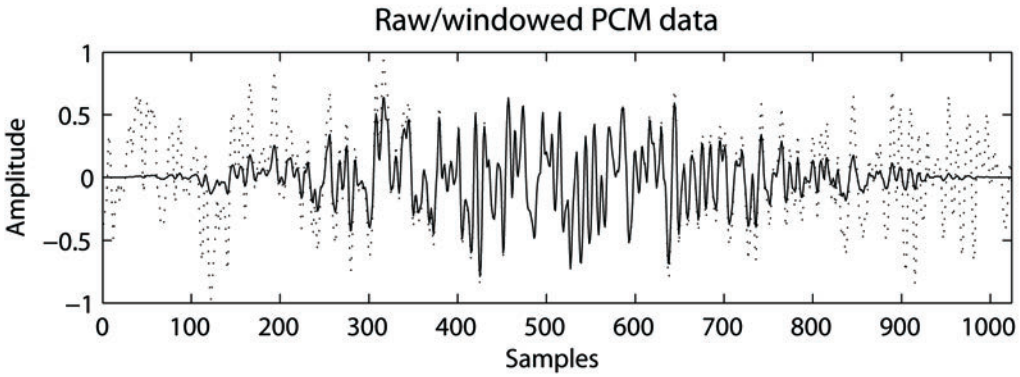

Figure 4.2: Input signal (taken from the drum and bass song *DJ Hype & Zinc - Six Million Ways To Die*) at a sample rate of 44.1kHz, mono and normalized amplitude. The displayed section is 1024 samples (about 23ms) long. The dotted line represents the original signal and the solid line represents the effect of the applied Hanning window function (see Figure 4.3 and Equation 4.8) on the input signal.

into another function of a variable, defined from $-\infty$ to $+\infty$.

This transformation is defined through the Fourier Integral (Equation 4.1), if it exists for all values of f.

$$H(f) = \int_{-\infty}^{+\infty} h(t)e^{-j2\pi ft}dt \tag{4.1}$$

The inverse Fourier transformation converts a frequency domain function $H(f)$ into an time domain function $h(t)$ and is defined by (Equation 4.2).

$$h(t) = \int_{-\infty}^{+\infty} H(f)e^{j2\pi ft}dt \tag{4.2}$$

j denotes the square root of -1 and $e^{jx} = cos(x) + j * sin(x)$.

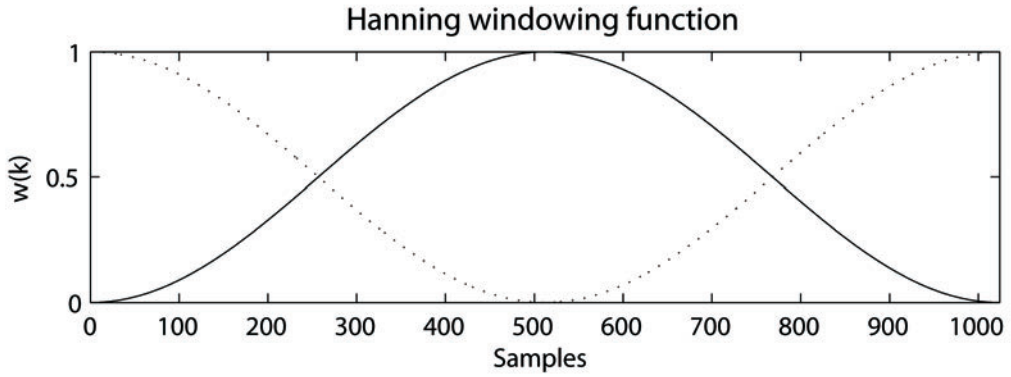

Figure 4.3: The solid line depicts the Hanning windowing function $w(k)$ with a length of 1024 samples. The dotted line shows the second half of the preceding window and the first half of the succeeding window. Summed up the overlapping windows always give 1. The hop size of the windows are always half of the window size (in this case 512 samples).

Discrete Fourier Transformation

Through the discrete representation of the input signal $h(k\Delta t)$ (see Figure 4.2) (k is the index of the consecutively numbered samples, Δt is the time distance between two samples), the need for a discrete transformation evolves. The Discrete Fourier Transformation(DFT) is discrete in time and frequency domain, it is periodic with the sampling frequency of the input signal and it is symmetric about the Nyquist critical frequency f_c. The Nyquist theorem says that the time distance T between two sample values of the input signal can at maximum be $1/2f_c$. At least two samples per period are required to represent a sinusoid signal. A continuous function $h(t)$ is uniquely reconstructable through its sample values, with $T = 1/2f_c$, if the Fourier transformed function is null for all frequencies bigger than f_c. In this case $h(t)$ is given by Equation 4.3.

$$h(t) = T \sum_{n=-\infty}^{+\infty} h(nT) \frac{sin[2\pi f_c(t - nT)]}{\pi(t - nT)} \tag{4.3}$$

The discrete Fourier transformation expressing the relationship between N samples of a time domain function and a transformed frequency domain function with N samples looks like Equation 4.4.

$$G\left(\frac{n}{NT}\right) = \sum_{k=0}^{N-1} g(kT)e^{-j2\pi nk/N} \tag{4.4}$$

where $n = 0, 1, ..., N-1$. The inverse discrete Fourier transformation is then given by Equation 4.5.

$$g(kT) = \frac{1}{N}\sum_{n=0}^{N-1} G\left(\frac{n}{NT}\right) e^{j2\pi nk/N} \tag{4.5}$$

It is important to mention, that the pair of transformation equations (Equation 4.4 and Equation 4.5) demand a periodicity of the time domain function as well as of the frequency domain function, expressed in Equation 4.6 and Equation 4.7. In this case with a period length of N.

$$G\left(\frac{n}{NT}\right) = G\left[\frac{(rN+n)}{NT}\right], \quad r = 0, \pm 1, \pm 2, ... \tag{4.6}$$

$$g(kT) = g[(rN+k)T], \quad r = 0, \pm 1, \pm 2, ... \tag{4.7}$$

Note that the frequency response obtained from the transformation can be complex, although the original signal is completely real. In practice, especially in the case of music analysis it can not be ensured that the analyzed signal is periodic. Much more a certain number of samples is analyzed at once, because the input signal is finite. This equals a multiplication of the signal with a rectangular windowing function. If now the window length does not exactly match an integer multiple of the basic period of the signal, a sudden discontinuity is introduced, which results in high frequency terms in the computed spectrum. Additionally an Direct Current (DC) value is computed, since the average in the time domain of the signal will not be

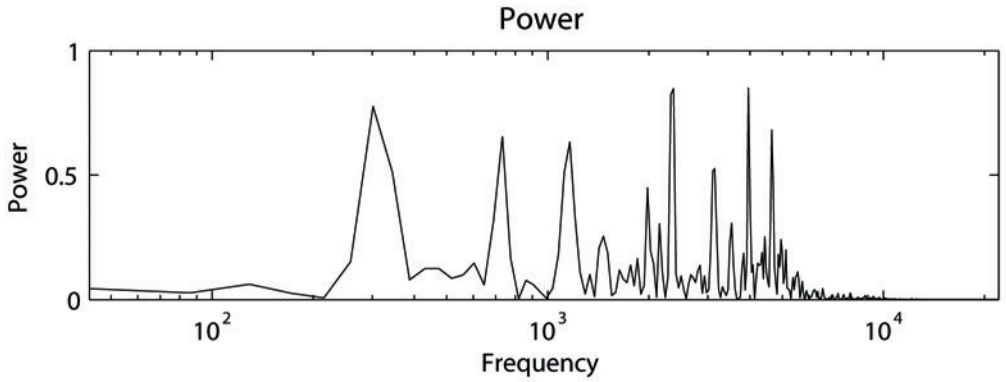

Figure 4.4: Resulting Fast Fourier Transformation of the signal with applied Hanning window of Figure 4.2. It shows the contained frequency spectrum in the respective analysis window of 1024 samples in Figure 4.2.

any longer zero. In order to avoid this problems an appropriate windowing function has to be chosen, like for example the Hanning windowing function (Equation 4.8 and Figure 4.3).

$$w(k+1) = \frac{1}{2} - \frac{1}{2}cos\left(2\pi\frac{k}{n-1}\right), \quad k = 0, ..., n-1 \tag{4.8}$$

Fast Fourier Transformation

The Fast Fourier Transformation (FFT) is a computer algorithm, devised by John W. Tukey and James W. Cooley in 1965, for efficient computation of the DFT. Through their publication of the algorithm in the article ”An Algorithm for the Machine Calculation of Complex Fourier Series” in Mathematics of Computation it came up that a lot of people had similar ideas and solutions for this problem at this time. The advantage of this algorithm is its simplification of the calculating complexity. The number of multiplications to compute the DFT is N^2, while the complexity of the FFT lies by $Nlog_2N$, which results in an enormous reduction of the calculation time and an improvement in accuracy, because of lower rounding

errors. A sample result of an FFT analysis is shown in Figure 4.4.

Short Time Fourier Transformation

The Fourier transformation in the frequency-domain can only give information about which frequencies the signal is composed by but can not tell when this frequencies in time exist. This of course is only relevant in a non-stationary case. In the Fourier case this problem is overcome by using the Short Time Fourier Transformation (STFT). This is like using the normal FT for short time frames, which are multiplied with an windowing function, making the assumption possible, that the signal is stationary in this short time frame.

4.1.2 Wavelet Transformation

The Fourier transformation might be the most popular transformation used in a lot of different scientific fields, but it is not the only possible time-frequency transformation. This section deals with another transformation method very popular in music content analysis, the Wavelet transformation. An in detail explanation of wavelets will not be given in context of this thesis. For further explanations please refer to [20][41][17][34].

The Fourier transformation as well as the Wavelet transformation is a reversible form, and makes it possible to go back and forth between the original signal and the transformed form. Wavelets are functions that satisfy certain mathematical requirements and are used to represent signals or functions. Wavelets process data at different scales or resolutions. This is an important advantage against the STFT, which works with a constant resolution. The signal is decomposed by iteratively passing it through low- and high-pass filters. Higher frequencies are better resolved in time, and lower frequencies are better resolved in frequency. So spectral components of high frequency can be better located in time than spectral components of low frequencies (see Figure 4.5).

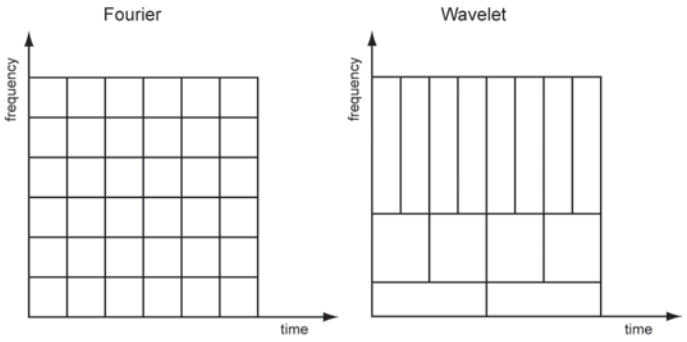

Figure 4.5: Schema of time-frequency resolution in the fourier and the wavelet case.

Continuous Wavelet

The wavelet analysis is done in a similar way to the STFT analysis, in the sense of multiplying the signal with a function, the wavelet. Mathematical the continuous wavelet transform (CWT) is defined by Equation 4.9.

$$\gamma(\tau, s) = \int f(t)\Psi^*_{s,\tau}(t)dt \tag{4.9}$$

The "$*$" denotes the complex conjugation. Equation 4.9 shows the decomposition of the function $f(t)$ into a set of basis functions $\Psi_{s,\tau}(t)$, called the wavelets. s and τ, scale and translation, are the new dimensions after the wavelet transform. The inverse form of the continuous wavelet transform is given through Equation 4.10.

$$f(t) = \int\int \gamma(\tau, s)\Psi_{s,\tau}(t)d\tau ds \tag{4.10}$$

Wavelets are generated from a single basic wavelet $\Psi(t)$, the so called *mother wavelet*, by scale variation and translation (Equation 4.11). The term wavelet itself stands for small wave. This refers to the finite length of the mother wavelet and to the property of being oscillatory. The denomination as "mother" derives from the prototypical nature this function has for the other decomposition functions used in

the wavelet analysis process.

$$\Psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\Psi\left(\frac{t-\tau}{s}\right) \tag{4.11}$$

In equation 4.11, $\Psi_{s,\tau}(t)$ is the mother wavelet, s is the scale factor, τ is the translation factor and the factor $s^{-\frac{1}{2}}$ is for energy normalization across the different scales. The translation factor τ is related to the location of the analysis window, as it is shifted over the input signal. It refers to time information in the transformed domain. The corresponding parameter of frequency as in the STFT is the scale parameter s, which is defined as $1/\text{frequency}$. High scale values correspond to a global view of the signal, while small values of s correspond to a detailed view. Spoken in terms of frequency low frequencies (high scales) correspond to global and high frequencies (low scales) correspond to detailed information.

Note that in these mentioned equations (Equation 4.9,4.10,4.11) the wavelet basis function is not defined. Contrary to the FT where the basis function is always given by a sinusoidal function, no specific basis function is given for wavelets in common. The theory of wavelets defines a framework within it is possible to design wavelets for specific needs.

Discrete Wavelet

The continuous form of the wavelet form does not fit for practical use. The difficulties lie in in the redundancy of the CWT. Also the infinite number of wavelets in the continuous transform is not a desired result. Last, but no least, fast algorithms are needed to calculate wavelet transformation in an acceptable amount of time.

Discreteness of the wavelet is reached by transforming the continuous wavelet representation (Equation 4.11). Discrete wavelets are not continuously scalable and translatable, but only in discrete steps (see Equation 4.12).

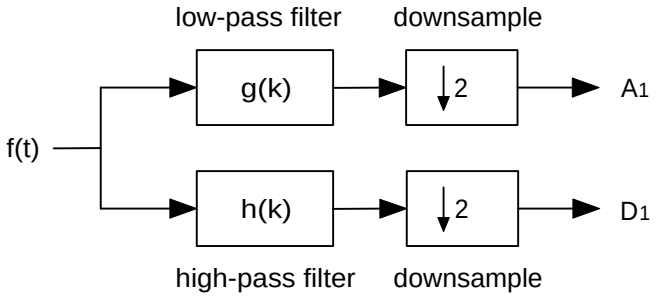

Figure 4.6: Single stage of decomposition in the DWT. In this case ($f(t)$ as input signal) the top level of decomposition. A_n are the approximation coefficients of level n and are used for further decomposition. D_n are the detail coefficients of level n.

$$\Psi_{j,k}(t) = \frac{1}{\sqrt{s_0^j}} \Psi \left(\frac{t - k\tau_0 s_0^j}{s_0^j} \right) \tag{4.12}$$

Usually s_0 and τ_0 is chosen to achieve a dyadic sampling in time- and frequency scale. This especially fits on the characteristics of musical data were a doubling in frequency scale relates to an octave. So usually $s_0 = 2$ and $\tau_0 = 1$. A single wavelet can be said to have a band-pass like spectrum. So a series of dilated wavelets can be seen as a band-pass filter bank.

For a stable reconstruction of a wavelet decomposed signal it was shown that the condition given in Equation 4.13, stating that the wavelet coefficients have to be located between two positive bounds, is necessary and sufficient.

$$A \left\| f \right\|^2 \leq \sum_{j,k} \left| \langle f, \Psi_{j,k} \rangle \right|^2 \leq B \left\| f \right\|^2 \tag{4.13}$$

In Equation 4.13 $\left\| f \right\|^2$ is the energy of $f(t)$, $A > 0$, $B < \infty$ and A,B are independent of $f(t)$.

To definitely remove all redundancies of the wavelet transform they have to be or-

thonormal[1]. Discrete wavelets are orthogonal to their own dilations and translations by special choices of the mother wavelet, in math terms Equation 4.14

$$\int \Psi_{j,k}(t)\Psi^*_{m,n}(t)dt = 1, \quad j = m, k = n \tag{4.14}$$

$$\int \Psi_{j,k}(t)\Psi^*_{m,n}(t)dt = 0, \quad j \neq m, k \neq n$$

The inverse discrete wavelet transformation (Equation 4.15) is given by summing all the orthogonal wavelet basis functions weighted by the corresponding transform coefficients.

$$f(t) = \sum_{j,k} \gamma(j,k)\Psi_{j,k}(t) \tag{4.15}$$

So the redundancy of wavelets is removed, but still there is an infinite number of wavelets. The target is now to reduce the amount of used wavelets but keep the quality of the transformation. This is on one hand done through the spectras of the dilation of wavelets on the other hand by the time intervals of the translation. Every time the wavelet is stretched in the time domain, its band width is halved. To reduce it to a finite number a so called *scaling function* ($\Phi(t)$) is used. The so called *two-scale relation* (see Equation 4.16) expresses the first scaling function in terms of the second one, so this expression states the scaling function at a certain level in dependence of the scaling function at the next smaller scale (more detailed). From this two-scale relation (Equation 4.16) one sees that $\Phi(t)$ is equal to a sum of scaled and shifted versions of itself.

$$\Phi(2^j t) = \sum_k h_{j+1}(k)\Phi(2^{j+1}t - k) \tag{4.16}$$

[1]A rotation (or flip) through the origin will send an orthonormal set to another orthonormal set. In fact, given any orthonormal basis, there is a rotation, or rotation combined with a flip, which will send the orthonormal basis to the standard basis. These are precisely the transformations which preserve the inner product, and are called orthogonal transformations. [12]

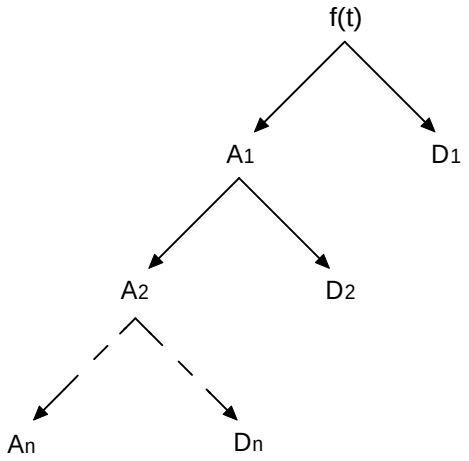

Figure 4.7: Schema of the wavelet decomposition tree, till level n.

The two-scale relation between the scaling function and the wavelet is expressed by Equation 4.17.

$$\Psi(2^j t) = \sum_k g_{j+1}(k)\Phi(2^{j+1}t - k) \tag{4.17}$$

The $h(k)$ weighting factors form a low pass filter and $g(k)$ a high pass filter. $h(k)$ can also be referred to as the *scaling filter* and $g(k)$ as the *wavelet filter*. The indices k and j represent translation and scaling respectively.

The perspective of seeing the DWT as a transformation through a filter bank makes it efficiently computable. Figure 4.7 gives an idea how this transformation process through the filter bank looks like. A single stage in the decomposition process (see Figure 4.6) corresponds to filtering through low- or high-pass filter and subsequently downsampling (by a factor of 2) of the input signal.

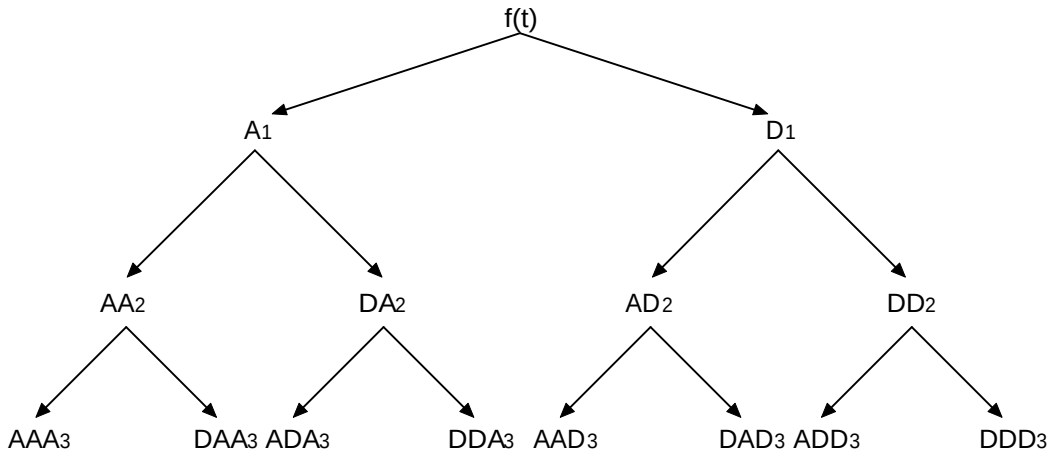

Figure 4.8: Schema of the wavelet packet decomposition tree, till level 3.

Wavelet Packet Transformation

Actually the wavelet transform is a subset of a far more versatile form, the wavelet package transform (WPT). Wavelet packets are particular linear combinations of wavelets. They form bases which retain many of the orthogonality, smoothness and localization properties of their parent wavelets. The coefficients in these combinations are computed by recursively making each computed coefficient the root for a own analysis tree (see Figure 4.8). A music content analysis using WPT for example was done by Grimaldi in [18].

Daubechie Wavelet

The Daubechie wavelet family (dbN) has no explicit expression except the db wavelet of order 1, which equals the Haar wavelet (Figure 4.9), the most simple wavelet. It is discontinuous and resembles a step function. In mathematical terms it is defined by the following conditions (Equation 4.18).

$$\Psi_{db1}(t) = 1 \quad x \in [0, 0.5[\tag{4.18}$$

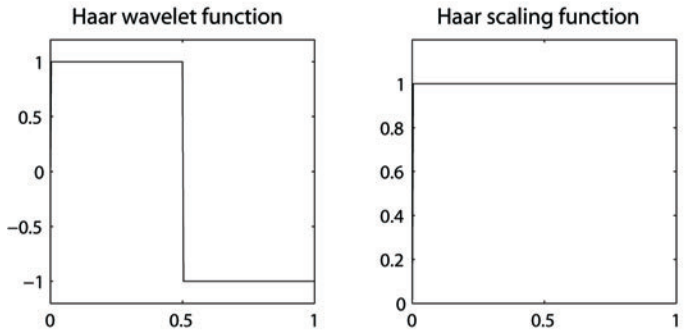

Figure 4.9: Haar wavelet Ψ_{db1} or db1 and scaling function Φ_{db1}.

$$\Psi_{db1}(t) = -1 \quad x \in [0.5, 1[$$
$$\Psi_{db1}(t) = 0 \quad x \notin [0, 1[$$

The related scaling function is formulated as Equation 4.19.

$$\Phi_{db1}(t) = 1 \quad x \in [0, 1] \tag{4.19}$$
$$\Phi_{db1}(t) = 0 \quad x \notin [0, 1]$$

As seen in various papers [27][18][26] the Daubechie wavelet family fits well for music analysis. These articles use Daubechie wavelets 4 (see Figure 4.10) and 8 (see Figure 4.11) with different levels of decomposition.

Figure 4.12 shows the result of a daubechenian DWT of grade 4 (db4) at a decomposition level of 4. The input signal is a 3s segment of the same song used for illustrating the FFT (see Figure 4.2). The tick labels on the x-axis denote the end of the corresponding section. Starting from the right, D_1 denotes the first part of detail coefficients with a length of 66153 samples and a frequency range of 11.025kHz to 22.05kHz, D_2 is 33080 samples long and has a frequency range of 5512.5Hz to 11.025kHz, D_3 is 16543 samples long and its frequency band ranges from 2756.25Hz to 5512.5Hz, D_4 and the approximation coefficients A_4 have a

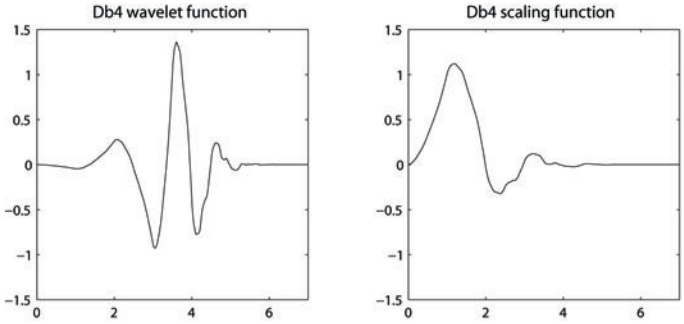

Figure 4.10: Daubechie wavelet Ψ_{db4} of order 4 (db4) and scaling function Φ_{db4}.

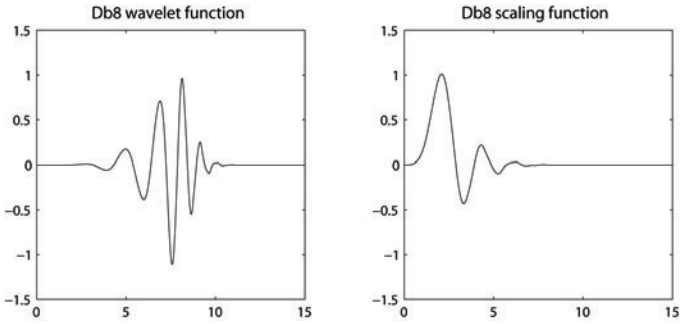

Figure 4.11: Daubechie wavelet Φ_{db8} of order 8 (db8) and scaling function Φ_{db1}.

length of 8275 samples and a frequency range from zero to 2756.25Hz. So the frequency content of the original signal is split up into different frequency bands and those can be therefore analyzed separately.

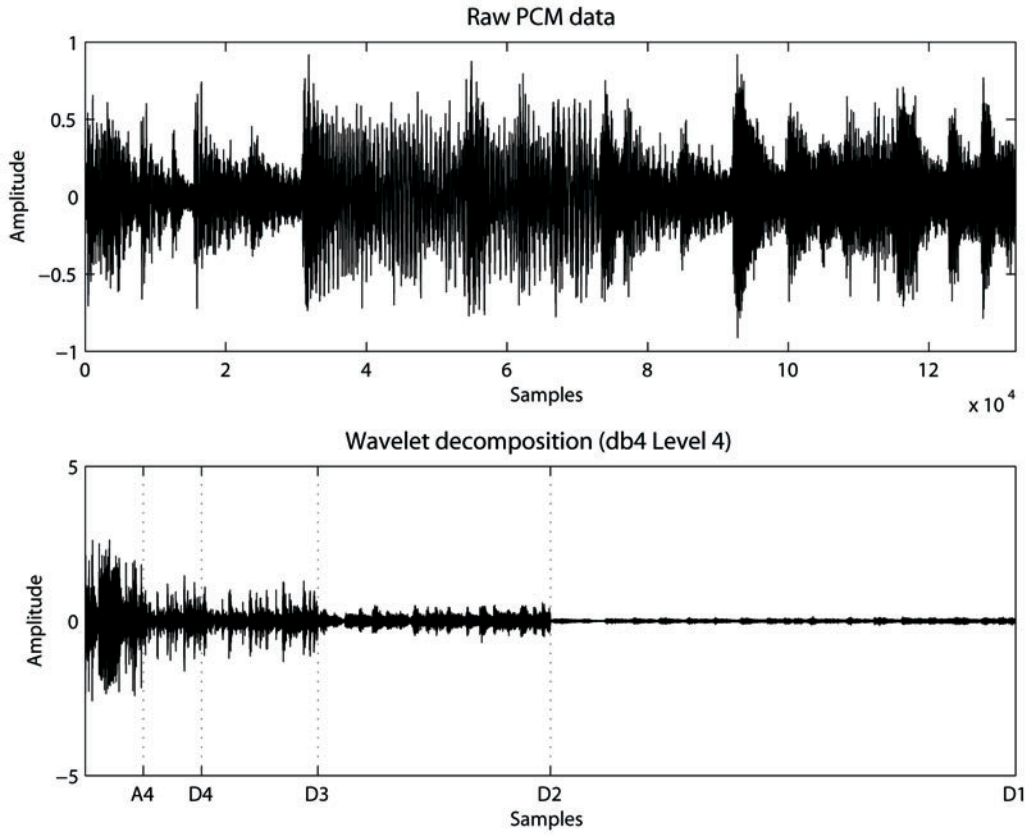

Figure 4.12: Sample DWT analysis of about 3s (132326 samples) section of a song (taken from the drum and bass song *DJ Hype & Zinc - Six Million Ways To Die*) at a sample rate of 44.1kHz, mono and normalized amplitude. Here the db4 wavelet is used with an decomposition level of 4. A_4 are the approximation coefficients at level 4, D_1 to D_4 are the detail coefficients.

4.2 Features

When browsing all the different papers about approaches to music genre recognition, a lot of different approaches are encountered. The typical underlying input signal transformations where presented in the last section (4.1), fast fourier transformation and discrete wavelet transformation. Following the signal transformation (illustrated in Figure 4.1), different kinds of features out of the transformed signal are extracted. Features are the numerical representation of certain characteristics of the analyzed piece of music. The challenges in finding a feature or feature vector for music content analysis applications are the duration of computation, complexity and significance of the resulting feature vector and therefore also appropriate feature vectors vary from their purpose of application. Features are of different nature, and according to Tzanetakis [43] they can be separated into different categories. Spectral shape-, sound ”texture”-, rhythm related-, and pitch related features can be distinguished.

4.2.1 Spectral shape features

For analyzing the spectral shape of an input signal, it has first to be transformed from the time domain into the frequency domain, as already shown in Section 4.1. The resulting frequency information (see Figure 4.4) represents the input, in the extreme case it can be perfectly reproduced through the information, and as might be noticed this information is too detailed. Thus specific features are calculated based upon the given spectral information.

STFT-based features

STFT-based features were and are very common, but get more and more challenge or are complemented by wavelet transform-based features. The STFT is mostly calculated on a window of about 20ms length and a hanning windowing function

(see Figure 4.3). Defined as the center of gravity of the magnitude of the STFT is the *spectral centroid* (Equation 4.20).

$$C_t = \frac{\sum_{n=1}^{N} M_t[n]n}{\sum_{n=1}^{N} M_t[n]} \tag{4.20}$$

$M_t[n]$ corresponds to the magnitude of the FT at frame t and the frequency bin n. Signals with stronger high frequency part, also referred to as "brighter" signal, are characterized through higher *spectral centroid* values. The *spectral centroid* is an important perceptual attribute in the characterization of musical instrument timbre.The *spectral rolloff* (Equation 4.21), a frequency parameter, is calculated as the frequency R_t below which 85% of the magnitude is concentrated.

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 \sum_{n=1}^{N} M_t[n] \tag{4.21}$$

The amount of local spectral change is measured through *spectral flux* (Equation 4.22). It is defined as the squared difference between the normalized magnitudes of successive spectral distributions.

$$F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2 \tag{4.22}$$

$N_t[n]$, $N_{t-1}[n]$ are the normalized values of the Fourier transformation at frame t and the preceding one $t-1$. Thus *spectral flux* results in measuring temporal changes in the frequency domain.

Mel-Frequency Cepstral Coefficients

Mel-Frequency Cepstral[2] Coefficients (MFCC) are a perceptual motivated set of features developed in context of speech recognition. An investigation about their adoption in the music classification domain was done in [28]. These coefficients are computed through a STFT, followed by a filterbank approximating the perceptual resolution properties of the human ear. This filterbank consists of 13 linearly-spaced filters below 1kHz and 27 log-spaced filters above (related to the so called *Mel*-scale). The last step is a Discrete Cosines Transform (DCT) to reduce the dimensionality of the feature vector. Typically 13 coefficients are used for speech recognition purpose, Tzanetakis have found the first 5 to be useful for music content analysis.

MPEG-based features

Most of the input data in music content analysis is represented in the well known MPEG audio compressed format MP3. The idea is to use the compressed data directly for feature calculation. This is time and memory saving, because all other approaches build upon a pulse code modulated (PCM) audio input signal and therefore must be converted at first. The approach was evaluated by Tzanetakis in [46]. Details about the compression format can be found in the International Standardization Organization (ISO)/MPEG standard [21]. To understand the calculation of the features it is important to know, that the signal in the compressed format is splitted in 32 equally spaced subbands and time windows of about 20ms. The computed features are *MPEG centroid*, *MPEG rolloff* and *MPEG flux*, which equal the spectral ones (see Equation 4.20, Equation 4.21 and Equation 4.22) with $N = 32$ frequency bins. Additional the *MPEG root mean square (RMS) power* (Equation 4.23) as a measurement of the loudness is computed.

[2]Cepstral refers to the transform through the *Mel*-scale motivated filter bank and is a pun. Cepstral equals Spectral with the first syllable reversed.

$$RMS_t = \sqrt{\frac{\sum_{n=1}^{32} M_t[n]^2}{32}} \tag{4.23}$$

4.2.2 Texture content features

The ”term” timbre is used to describe all characteristics which differentiate one sound from another of same pitch and loudness. This makes sense in the context of a single spoken syllable or a tone of a single isolated instrument. For more complex audio signals like polyphonic music or speech it can not be applied. But there are important statistical characteristics in the energy distribution in time and frequency, for those the term ”texture” content features will be used. In some cases these textures can be observed visually on a spectrogram, in most cases they can not. The sensation of sound texture features arises on a time section bigger than the usual analysis window of about 20ms usually used for transformation and analysis. In this context the term texture window will be used which is bigger and consists of multiple analysis windows, typically with a length of about 1s.

STFT based texture content features

The texture content features are not directly computed features, but *multidimensional Gaussian distributions parameters (mean (Equation4.24), variance (Equation4.25))*, of the *spectral features* in the analysis window, over the texture window. One feature, which makes it possible to differ speech from music, is computed over the texture window and not over the analysis window. It is called *low energy* feature and is defined as the percentage of analysis windows that have less RMS power than the average RMS power over the whole texture window. Therefore usually speech has high *low energy* values, because of many silent analysis windows.

Wavelet transformation based texture content features

Above mentioned texture content features are computed of STFT information. The wavelet transformation, as mentioned in Section 4.1.2, can also be used to calculate texture content features. The resulting wavelet coefficients of the wavelet transform provide a compact representation of the signal over time and frequency. To further reduce the dimensionality of this information statistical characteristics of the wavelet coefficients are calculated. The *mean* of the absolute value of the coefficients in each wavelet subband, as well as the *ratios of the mean values* between adjacent subbands, is computed and provides information about the frequency distribution of the signal. The amount of change of the frequency distribution over time is represented by the *standard deviation*.

Another technique for computing wavelet transformation based texture content features are wavelet histograms. Thereby the different decomposition coefficients are evaluated through a histogram. Analyzing characteristics of those *wavelet histograms*, like the first three moments of a histogram (*mean* (Equation4.24), *variance* (Equation4.25) and *skewness* (Equation4.27)) give a very compact representation of texture information about a piece of music. This technique is successfully applied in the visual field of content analysis [29].

$$\overline{x} = \frac{1}{N} \sum_{j=1}^{N} x_j \tag{4.24}$$

The mean $\overline{x}$ is calculated as the sum off all values divided by the count of values and estimates the value around which central clustering occurs.

$$Var = \frac{1}{N-1} \sum_{j=1}^{N} (x_j - \overline{x})^2 \tag{4.25}$$

$$\sigma = \sqrt{Var} \tag{4.26}$$

The Variance Var, as well as the standard the deviation σ, characterize the "width" or "variability" of the data around the mean.

$$Skew = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{x_j - \overline{x}}{\sigma}\right]^3 \tag{4.27}$$

The skewness characterizes the degree of asymmetry of a distribution around its mean. A positive value of skewness signifies a distribution with an asymmetric tail extending out towards more positive x and a negative value signifies a distribution whose tail extends out towards more negative x.

4.2.3 Rhythmic content features

As evaluated in [50] rhythmic content features in the sense of *tempo of the main beat* and its *strength* are major descriptors for music content analysis. Extended rhythmic related information, which also is useful, is the *regularity of the rhythm*, the *relations between the main beat and the subbeats* and the *strength of subbeats* in relation to the main beat. The process of automatic beat detection is similar to pitch recognition techniques over a larger time frame. The basis for extracting such kind of rhythmic related features are *beat histograms*.

The schema of an algorithm to compute a *beat histogram* is shown in Figure 4.13. At first the input signal is decomposed in several octave frequency subband by a DWT. An envelope extraction of each of the subband, consisting of a full wave rectification (Equation 4.28), a low pass filtering (Equation 4.29), downsampling (Equation 4.30) and a mean removal (Equation 4.31) is done.

$$y[n] = |x[n]| \tag{4.28}$$

$$y[n] = (1-\alpha)x[n] + \alpha y[n-1] \tag{4.29}$$

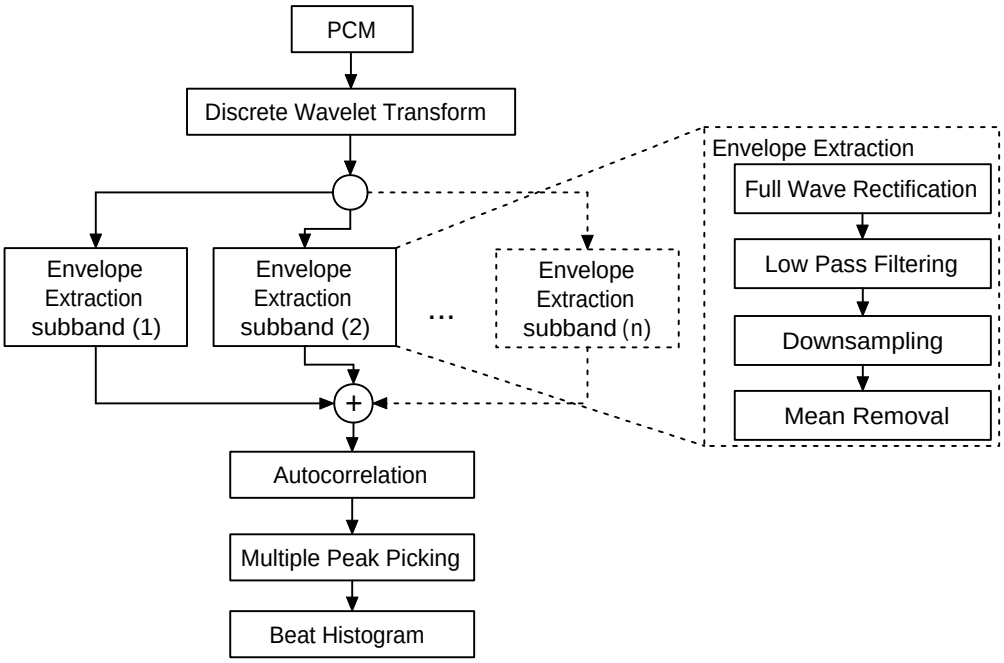

Figure 4.13: Schema of the beat histogram calculation.

$$y[n] = x[kn] \tag{4.30}$$

$$y[n] = x[n] - E[x[n]] \tag{4.31}$$

Afterwards the envelopes of the subbands are summed up and the autocorrelation function (Equation 4.32) of the resulting envelope is computed. The peaks of the autocorrelation function correspond to the various periodicities of the signal. By picking multiple peaks of this function a *beat histogram* is generated, where each bin equals a certain beats per minute (bpm) value. Higher values in a *beat histogram* bin, describe a signal similar to itself and therefore a strong beat.

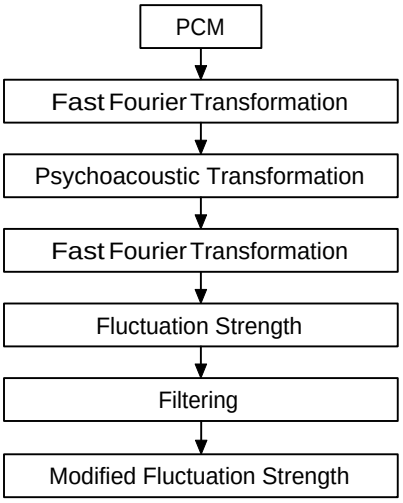

Figure 4.14: Schema of the modified fluctuation strength calculation.

$$y[k] = \frac{1}{N} \sum_{n} x[n]x[n-k] \tag{4.32}$$

Beat histograms provide a good overview of detailed information about rhythmic related features for music content anlysis and applications in the field of genre classification and similarity search.

Another possibility to compute rhythmic related features of a music signal was done in [37]. Here the *modified fluctuation strength* (see Section 3.4.2 - Fluctuation strength) is computed as a representation of psychoacoustic rhythm patterns. It is represented by its value and the corresponding frequency of pitch, expressed in critical bands (see Section 3.4.2 - Critical Bands), as well as the modulation frequency. A schematic view of the calculation process is given in Figure 4.14. The straight forward appearance of Figure 4.14 may not distract from the fact that through using critical bands in frequency representation also a multiband analysis like in Figure 4.13 is done.

Music represented in PCM format is the input signal and is transformed through

FFT of a window of about 20ms in the frequency domain. The frequency representation is undertaken a series of transformations of psychoacoustic aspects. Details to the single steps and motivations of this transforms are found in Section 3.4.2. The first step is a splitting of the signal in critical frequency bands. Afterwards, in the Psychoacoustic Transformation block, the spread critical band rate is computed through applying a masking function to give respect to the influences of adjacent frequency band. The amplitudes are converted in decibel values and through equal loudness curves and loudness sensation conversion sone values are computed. Through the second FFT, which is computed over a window of about 6 seconds of the already computed sone values, the amplitude modulations of loudness in time are extracted and build the basis for the fluctuation strength. The filtering process, consisting of a gradient filter to emphasize sharp edges of the fluctuation strength and a Gaussian filter to enable the distance calculation between *modified fluctuation strength* vectors of songs. At the end of this process the mean values of all analyzed 6 second fragments of a song are taken and build the resulting feature matrix.

4.2.4 Pitch content features

Pitch is defined as the property of a sound and especially a musical tone that is determined by the frequency of the waves producing it i.e. highness or lowness of sound. An computational efficient approach for multipitch detection was proposed by Tolonen in [42]. In this approach not a computational intensive multiband analysis off the audio signal is done, but a high- and a low-pass filter at 1kHz is used. Amplitude envelopes are extracted and summed to compute a Summary Enhanced Auto Correlation Function (SACF). Tzanetakis used this approach to compute *pitch histograms* [49] out of a analysis window of about 20ms again. The three most dominant peaks of the SACF are accumulated into the *pitch histogram*. A histogram bin n corresponds to a musical note of specific pitch (frequency f). As labels the MIDI note numbering is used, following the schema in Equation 4.33.

$$n = 12\ log_2 \frac{f}{440} + 69 \tag{4.33}$$

Folded and *unfolded pitch histograms* are used. While *unfolded pitch histograms* contain all different notes, a *folded pitch histogram* maps the unfolded notes into one single octave (Equation 4.34).

$$c = n\ mod\ 12 \tag{4.34}$$

The folded *pitch histogram*, now contains information about the pitch classes of the harmonic content of the input signal, whereas the unfolded one gives information about the tonal range of the piece of music. To achieve better classification results through *pitch histograms*, the proposed pitch histogram comes under another transformation. For better expressing tonal music relations the *folded pitch histogram* is mapped to a circle of quints (see Equation 4.35). A quint corresponds to 7 semitonal steps.

$$c' = (7c)\ mod\ 12 \tag{4.35}$$

Musical pieces cannot be assigned to a genre on pure pitch information only, but the pitch information gives an idea of certain tendencies to do so. For example jazz or classical music have a higher value of pitch change than rock or pop music. Features which are resulting out of the pitch histograms are for example the bin with the *maximum peak* and its *amplitude* of the *folded pitch histogram*, typically corresponding to the tonic or dominant chord of the music piece. The amplitude gives information about the pitch change rate of the song, lower values represent a lot of pitch changes in the song. In the *unfolded pitch histogram* the *period of the maximum peak* corresponds to the octave range of the musical of the song. The main tonal interval relation can be extracted out of the *interval between the two most prominent peaks* of the *folded pitch histogram* and will be 1 or -1 for pieces with simple harmonic structure.

4.2.5 Other features

A feature which does not perfectly fit in one of the above sections, because it is computed over the plain PCM signal, are the *time domain zero crossings*. This feature is defined by Equation 4.36, where the sgn function is defined through Equation 4.37.

$$Z_t = \frac{1}{2} \sum_k |sgn(x[k]) - sgn(x[k-1])| \tag{4.36}$$

$$sgn[x(k)] = 1, \quad x(k) \geq 0 \tag{4.37}$$
$$sgn[x(k)] = -1, \quad x(k) < 0$$

Time domain zero crossings provide a measure for the noisiness of a signal and makes it possible to differentiate between voiced and unvoiced audio signals. Also signals with a broader frequency spectrum have more *time domain zero crossings* than narrower ones.

The mentioned features sets of different content aspects in the whole Section 4.2 above can of course be combined in an intelligent way to improve classification results achieved through a single feature set of one particular feature domain.

Chapter 5

Experiments

Based on the theoretical discourse in the preceding chapters the following experiments were performed.

Three in music genre recognition popular analyzing approaches were used to conduct evaluation experiments of their performance. Those were the psychoacoustic rhythm patterns of the *SOMeJB* [37] implementation, the genre vector of the *MARSYAS* [45] framework for audio analysis and an implementation of the Daubechies Wavelet Coefficients Histogram approach proposed in [27]. The feature vectors were computed over two test music collections. One test collection is the one used by George Tzanetakis, which evolved into a kind of standard testbed, consisting of approximately 1000 songs divided into 10 genres. The second music collection is a more real world user Mp3 collection of 9360 titles of a variety of genres, including duplicates as different recordings of the same title.The distances between the songs through the extracted features were computed and tried to benchmark the different feature sets. These benchmarks were the recall-precision values of the retrieval perform of the feature sets and a small user study letting human test persons judge the n-nearest neighbors results of a song with respect to genre similarity.

A detailed explanation is given in the next sections.

5.1 Music Collections

An explanation of the composition of the music collections used for the evaluation is given here. Further, the way of distance determination between the computed feature vectors is introduced.

5.1.1 Real World Music Collection

The first music collection used for the evaluation process consists of 9360 different songs, in the MP3 format. These songs are mainly western popular music including classic and jazz songs too. This collection is referred to as ”Real World” collection, because it is a users MP3 collection, which was not especially prepared for scientific research. It contains a wide variety of music titles of this users taste and has artifacts like multiple recordings of the same song from different albums. According to a matching with an open music title database the collection consists of music pieces out of genres like rock, pop, hiphop, alternative, original sound tracks, singer/songwriter, electronica in a widest sense, jazz, blues and a few spoken word tracks. Each of these songs was chopped into 4 different of 30 seconds length each segments (see Figure 5.1). The starting points of these segments in the original audio files are at the beginning of second 30, 45, 70 and 150. So there is a 15 second overlap between the segments starting at second 30 and 45, the one starting at second 70 is nearly appending and 150 is located with a time distance of 90 seconds to the end of the first segment. This gives an overall amount of 37440 segments to analyze. All MP3’s of this collection are in Mono, have a sample word width of 16bit and a sample rate of 44.1kHz.

5.1.2 Tzanetakis’ Music Collection

As second music collection a set of 1203 MP3’s, used by Tzanetakis for his experiments, is analyzed. These sound files have a duration of 30 seconds, are also

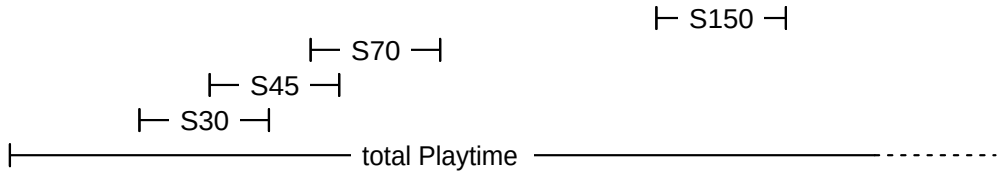

Figure 5.1: Segments illustration, which the Mp3's were chopped into. Segments are named by their starting position in seconds of the total playtime and have an equal length of 30 seconds.

in Mono, do have a sample word width of 16bit, but have a sample rate of 22kHz. The pieces of music are pre-categorized in genres. This genre set consists of the 12 labels ambient, ballad, blues, classical, country, disco, hiphop, jazz, metal, pop, reggae and rock. Basically there are about 100 songs in the average categorized in a genre, for the exact distribution see column n in Table 5.18.

5.1.3 Distance Determination

The determination of distance between the feature vectors of a feature extraction method is computed through determining the Euclidean distance between two features. The Euclidean distance is calculated as shown in Equation 5.1.

$$d_{euclid} = \sqrt{\sum_{i=1}^{N} (p_i - q_i)^2} \tag{5.1}$$

N denotes the dimensionality of the feature vectors p and q to compare.

5.2 Prototypes

The *SOMeJB* [37] prototype was modified in the way of enabling the processing of music as input data in PCM format with a sampling frequency of 44.1kHz. The for-

mer version only supported PCM data sampled with 11kHz, this had the advantage of a smaller feature vector, but since most of the music collections consist natively of Mp3's with a sampling frequency of 44.1kHz, it also had the disadvantage of the very time intensive resampling of the input data.

As an framework for audio analysis, experimenting, evaluating *MARSYAS* [45] can be understood, developed by George Tzanetakis and Perry Cook. It provides functionality for different feature extraction and provides the capability to extract a feature vector for genre classification needs.

DWCH's as underlying technique for genre classification were proposed in [27]. This approach was chosen because more and more research [26][18][30] is done in the wavelet related field for music genre recognition. Based upon the paper [27] an implementation was done to compute the related feature vector.

5.2.1 Psychoacoustic Rhythm Patterns

The *SOMeJB* [37] prototype is based on a psychoacoustically motivated approach, which focuses on the dynamic properties of music. Additionally to the feature extraction it provides genre clustering and a user interface using an *Islands* metaphor. Here only the feature extraction process is of interest. The prototype is implemented in the MATLAB[1] environment.

The schema of the feature extraction process in this prototype is shown in Section 4.2.3, a detailed description can be found in [32]. The modified prototype used for the experiments in this thesis analyzes the complete of pieces of music rather than a selection of 6 second segments. Music content information is represented as a vector of *modified fluctuation strength* of a song, which is a filter modified version of the psychoacoustic parameter, fluctuation strength. The modified fluctuation strength values are computed for 24 critical frequency bands and up to amplitude modulation frequency (AM) of 10Hz. The AM frequency axis is represented by

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

60 values. As Pampalk argues, there is not much activity beyond an AM of 10Hz. Presence of roughness for higher AM values (15Hz-300Hz) is not regarded. The feature vector of *modified fluctuation strength* represents global characteristics of a song and is a time invariant information.

In comparison to the two other prototypes the resulting feature vector is large. It's 1440-dimensions are the vectorization of the computed 24- by 60-dimensional feature matrix. For the second music collection, which contains Mp3's width a sample rate of 22kHz only the first 20 critical frequency bands are taken into account. So the feature vector dimension is reduced to a dimension of 1200, because the underlying feature matrix is 20- by 60-dimensional.

5.2.2 MARSYAS

The MARSYAS [45] framework of Tzanetakis is an general implementation for the extraction of various music content based features. It follows a client-server architecture and is implemented in C++. It is used to extract a feature vector for genre classification needs as proposed in [48].

This application analyses music and extracts features under the aspects of timbral features, rhythmic content features and pitch content features. Features are calculated over *analysis* or *texture windows* respectively to the specific feature (see Sec. 4.2), over each whole piece of music of the collection (see Sec. 5.1). The timbral texture feature vector consists of *means*, *variances* of *spectral centroid*, *spectral rolloff*, *spectral flux* and *zerocrossing rate*. Additional a *low energy feature* and the *first five MFCC coefficients* are computed, excluding the coefficients corresponding to the DC component of the MFCC feature vector. The rhythmic related feature vector, consists of parameters computed through beat histograms. Music genre classification relevant, as found by Tzanetakis [48], are the *first two relative histogram peaks* (divided by the sum of amplitudes), the *ratio of the amplitude of the first to the second histogram peak*, the *periods in bpm of the first two histogram peaks* and the *overall sum of the peak histogram* as indication of beat strength. Pitch

information is represented through information of pitch histograms. These are the *amplitude of the highest peak of the folded pitch histogram*, corresponding to the most dominant pitch class of a piece of music. Further on the *periods of the maximum peaks* of the *unfolded*, as well as the *folded pitch histogram*, the *pitch interval between the two most prominent peaks of the folded pitch histogram* and at least the *overall sum of the histogram*, as measure of the strength of pitch detection.

This results in a compact 30 dimensional feature vector for a piece of music representing information of multiple music content analysis domains (19-dimensional timbral texture feature vector, 6-dimensional rhythmic content feature vector, 5-dimensional pitch content feature vector). A detailed explanation of the different features is given in Section 4.2.

5.2.3 Daubechies Wavelet Coefficients Histogram

The latest development in the music content analysis field has put forth various approaches using wavelets as underlying analysis principle. This implementation is following [27] and was done in the MATLAB environment. It was carried out in this environment, because it provides the appropriate wavelet and statistical tools.

At the beginning a discrete wavelet transformation of the pieces of music is obtained. Contrary to the other two analysis approaches the analyzed window of music has a length of 3 seconds and is located at the 15th second of playtime of the pieces of music. The whole 30 seconds are not processed here. A daubechanian basis wavelet of 8th grade was chosen and a *decomposition of 7 levels* is done. The coefficients in each of the decomposition bands and the last *(level 7) approximation coefficients* are accumulated in histograms. Over these histograms the first three moments are computed. The moments are the *mean*, *variance* and *skewness* of the histograms. The now 24-dimensional feature vector is completed with the computation of the *energy level* in each subband, which is defined as the mean of absolute values in a subband. Unlike to the prototype used in the original paper, the feature vector used here has no additional timbral features for speech recognition added.

The resulting feature vector is 32-dimensional and is, except for the subband energy information, plainly based on statistical information about the wavelet decomposition coefficients.

As the parameters are out of different domains and so have different ranges of values, except for the SOMeJB prototype, the values are normalized with respect to the interval $[0, 1]$.

5.3 Evaluation

The evaluation of the obtained results was arranged as follows. At first recall and precision of retrieval values are computed for both music collections. For the real world music collection the segments starting at second 30 of original playtime of the songs are used as query set and the others (segments starting at second 45, 70 and 150) build the data set, which the queries are conducted on. The second recall and precision evaluation bases upon the music collection of Tzanetakis. Here query and data collection are the same and therefore distances from each to each file are computed. In every case only the best 20 answers, 20 smallest distances, are taken into account. The best match to a query in this case is of course the query itself with a distance of zero and is ignored, because this result has no information.

The parameters of recall and precision, as used in [30], give a good possibility for assessing the obtained retrieval results. Assessment parameters are computed for answer sets with a size of 1, 3, 5, 10 and 20. In case of the real world music collection only the other three time segments of a song in the data set are considered relevant. For the Tzanetakis music collection all answers in the same pre-categorized genre are considered as correct answers. So the evaluation over the real world collection is somehow an assessment of recognizing songs and over the smaller collection it is a genre classification evaluation.

It is important to note that where the number of relevant files are small, the values for the parameters tend to seem very small. Precision and recall are defined by

Equation 5.2 and Equation 5.3 respectively.

$$P_i = \frac{N_{rd}}{N_i} \tag{5.2}$$

$$R_i = \frac{N_{rd}}{N_{rt}} \tag{5.3}$$

N_{rd} denotes the number of relevant titles retrieved, N_{rt} are the number of relevant pieces of music in the whole music collection, N_i stands for the total number of pieces of music retrieved and the index i is the size of the answer set.

The precision values (Equation 5.2) give an idea about the quality of the results of a retrieval task in the sense of setting the number of *retrieved relevant documents* in relation to the *total number of retrieved documents*. For evaluation, the parameter *total number of retrieved documents* is varied. A relative value of about, how much *relevant documents* could be *retrieved* of *all situated relevant* ones in the collection, represents the recall value. In the following tables of results a best case value will be given for comparison to.

As another evaluation method, a small survey was done. People were asked to evaluate the best 10 answers to queries in with respect to musical genre. This was done for queries on the distance results of all three prototypes over results for the real world music collection.

5.3.1 Specific Piece of Music Retrieval

In the Tables 5.1 and 5.2 the results of the recall (R_i) and precision (P_i) evaluation over the real world music collection can be found. The labels DWCH, MARSYAS and RP denote the related prototypes used for the feature extraction.

Explaining the absolute recall values, it is to say that with a number of 28860 MP3's in the data set of the real world collection in the best case all of the three relevant MP3's per query should be in the answer set and therefore the best case are 28860

i	DWCH	MARSYAS	RP	best
20	7438	4135	17147	28860
10	6320	3410	16124	28860
5	5269	2821	14500	28860
3	4489	2410	12924	28860
1	2427	1440	6037	9620

Table 5.1: Absolute recall values for the real world music collection in case of feature extraction by the different three prototypes and in a best case scenario, with answer set sizes (i) 20, 10, 5, 3 and 1.

relevant pieces of music in the answer sets. In case of an answer set with a size of one ($i = 1$) only 9620 relevant pieces of music can be retrieved. Expressing this in relative recall values the best case scenario with $i = 1$ gives a value of 0.3333. As the absolute (see Table 5.1) and relative (see Table 5.2) recall values for the evaluation of the results for the real world music collection show, the rhythm patterns features outperform the other two approaches in every case. The RP achieve a recall rate of about 50-60%, followed by the DWCH with about 20-25% and the MARSYAS with about 10%. This trend is the same for the precision values. Precision values in the best case are smaller than one for answer sets bigger than three ($i > 3$), because there are only three relevant pieces of music in the collection for each query. Table 5.3 shows again clearly the highest values for the RP protoype. The presicion of the results is the best with every examined answer set size i, in almost 45% of all cases the 3 top-ranked retrieved results are the 3 segments from the piece of music used as a query, followed by DWCH and MARSYAS. Note, that the DWCH approach analyses a 3 seconds segment in the pieces of music only, while the other two process the whole length of the input.

To determine the stability of a feature vector over the analyzed segments of a piece of music, a count of the different best ranked segments is done (answer set size $i = 20$). Table 5.4 shows for all three analysis methods that, the close in time the segments are, the better they get recognized. So the segments starting at second 45 of playtime of the original pieces of music are most often the best ranked ones,

R_i	DWCH	MARSYAS	RP	best
20	0.2577	0.1433	0.5941	1
10	0.2190	0.1182	0.5587	1
5	0.1826	0.0977	0.5024	1
3	0.1555	0.0835	0.4478	1
1	0.0841	0.0499	0.2092	0.3333

Table 5.2: Recall values (R_i) for the real world music collection in case of feature extraction by the different three prototypes and in a best case scenario, with answer set sizes (i) 20, 10, 5, 3 and 1.

P_i	DWCH	MARSYAS	RP	best
20	0.0387	0.0215	0.0891	0.15
10	0.0657	0.0354	0.1676	0.3
5	0.1095	0.0586	0.3015	0.6
3	0.1555	0.0835	0.4478	1
1	0.2523	0.1497	0.6275	1

Table 5.3: Precision values (P_i) for the real world music collection in case of feature extraction by the different three prototypes and in a best case scenario, with answer set sizes (i) 20, 10, 5, 3 and 1.

followed by the segments starting at second 70 and then by segments starting at second 150. The relative decrease of the values, the farer the segments are located from the query segment, is smallest for the DWCH approach, this therefore is the most stable one. This makes sense in consideration of the small time frame analyzed by this prototype and the nevertheless competitive precision and recall values. In Table 5.5 the absolute average position of the segments in the answer set ($i = 20$) is listed. The DWCH and MARSYAS ranking of the segments is very similar, whereas the ranking in the RP case is better, but also a bit more spread apart. Through the average position of the three segments (see Table 5.5) the question about the recall values for those segments arise. The following Tables 5.6 to 5.8 depict the corresponding recall values, considering only one segment valid as answer to the query segment. The values in these three tables show, corresponding to the average position (see Table 5.5) and the stability information given in Table 5.4, a similar

segment	DWCH	MARSYAS	RP
45	2827	2032	5378
70	1603	740	1632
150	468	293	921

Table 5.4: Here the absolute number of times a segment is best ranked is shown. A table cell describes the value of how often certain segments were better ranked than the other two relevant ones for the different analysis methods.

segment	DWCH	MARSYAS	RP
45	4.3354	4.5723	2.3708
70	5.0102	5.694	3.7867
150	5.5386	5.7061	4.3277

Table 5.5: This table shows the average position of the relevant segments in an answer set with size $i = 20$, for the different analysis methods.

relative loss of recall for segments located at second 70 and 150 of playtime against the segment located at second 45 for the DWCH and MARSYAS approach. The RP approach comes up with a significant smaller relative loss of recall for the segments located further away in playtime from the query segment, aside from the good recall values themselves.

The Tables 5.9 to 5.11 give information about the distance values in the answer sets. The different column values (except in Table 5.11) can not be compared to each other, because the distance values for RP are not normalized (RP feature vector values are all of the same domain and the dimensionality of the feature spaces differ). While Table 5.9 shows the average distances of answer sets with different size i, Table 5.10 contains the average distances for specific ranks. In Table 5.11 the distances are normalized by considering the distance of rank 20 as 1. As it can be seen in all three tables (Table 5.9- Table 5.11) the relative difference between the values of a prototype is nearly the same for DWCH, MARSYAS and RP. These results make clear that the spacing of distances between ranks or ranges is in no case of analysis method linear. The bigger the rank the smaller the differences of dis-

R_i^{S45}	DWCH	MARSYAS	RP	best
20	0.3842	0.2475	0.7946	1
10	0.3344	0.2093	0.7677	1
5	0.2842	0.1783	0.7262	1
3	0.2476	0.1563	0.6840	1
1	0.1513	0.1057	0.4420	1

Table 5.6: Recall values (R_i) for the real world music collection, considering segments S45 as the only valid answers, in case of feature extraction by the different three prototypes and in a best case scenario, with answer set sizes (i) 20, 10, 5, 3 and 1.

R_i^{S70}	DWCH	MARSYAS	RP	best
20	0.2736	0.1201	0.6034	1
10	0.2288	0.0953	0.5591	1
5	0.1897	0.0755	0.4906	1
3	0.1593	0.0629	0.4204	1
1	0.0830	0.0338	0.1333	1

Table 5.7: Recall values (R_i) for the real world music collection, considering segments S70 as the only valid answers, in case of feature extraction by the different three prototypes and in a best case scenario, with answer set sizes (i) 20, 10, 5, 3 and 1.

tance are, considering range or rank of 20 and a data set size of 28860 MP3's. This behavior is the same for all three prototypes, there are no significant differences.

R_i^{S150}	DWCH	MARSYAS	RP	best
20	0.1369	0.0741	0.4340	1
10	0.1120	0.0597	0.3958	1
5	0.0890	0.0475	0.3324	1
3	0.0726	0.0382	0.2764	1
1	0.0250	0.0144	0.0697	1

Table 5.8: Recall values (R_i) for the real world music collection, considering segments S150 as the only valid answers, in case of feature extraction by the different three prototypes and in a best case scenario, with answer set sizes (i) 20, 10, 5, 3 and 1.

i	DWCH	MARSYAS	RP
20	0.0397	0.2776	0.3274
10	0.0335	0.2363	0.2841
5	0.0283	0.2046	0.2371
3	0.0251	0.1855	0.2008
1	0.0203	0.1553	0.1584

Table 5.9: The average distance value for various answer set sizes (i) and the different analysis methods.

rank	DWCH	MARSYAS	RP
20	0.0493	0.3450	0.3854
10	0.0412	0.2839	0.3474
5	0.0342	0.2395	0.3003
3	0.0293	0.2118	0.2437
1	0.0203	0.1553	0.1584

Table 5.10: The average distance values for specific ranks and the different analysis methods.

rank	DWCH	MARSYAS	RP
20	1	1	1
10	0.8357	0.8229	0.9014
5	0.6937	0.6942	0.7792
3	0.5943	0.6139	0.6323
1	0.4118	0.4501	0.4110

Table 5.11: Normalized average distance values, the distance of rank 20 is considered to be 1, for specific ranks and the different analysis methods.

rank	artist	album	title	seg.
1	Eminem	The Marshall Mathers	B Please II	45
2	Eminem	The Marshall Mathers	B Please II	70
3	Fettes Brot	Amnesie	Sekt Oder Selters	150
4	DJ Tomek		Ich lebe für Hip Hop	150
5	Mobb Deep	Hell On Earth	Give It Up Fast	45
6	Eminem	The Eminem Show	Square Dance	45
7	Fettes Brot	Amnesie	Definition Von Fett	70
8	Wyclef Jean	Masquerade	Party Like I Party	45
9	Limp Bizkit	Chocolate Starfish	My Way	150
10	Farin Urlaub	Endlich Urlaub!	Intro	45
11	Fettes Brot	Amnesie	Mal Sehen	150
12	Fünf Sterne Deluxe		Die Leude	45
13	Wyclef Jean	Masquerade	WThug Like Me	45
14	Wyclef Jean	Masquerade	The Mix Show	45
15	Samy Deluxe	Samy Deluxe	Wickeda MC	45
16	Samy Deluxe	Samy Deluxe	Wickeda MC	70
17	Mary J. Blige	No More Drama	Family Affair	70
18	Mobb Deep	Hell On Earth	Give It Up Fast	70
19	Wyclef Jean	The Ecleftic	Thug Angels	150
20	Rückgrat		Dreckige Rapz	70

Table 5.12: 20 best ranked answers to a query with the segment starting at second 30 of the track ”B Please II” by artist ”Eminem” from the album ”The Marshall Matters” analyzed by the DWCH approach.

Explicit retrieval queries and results

The most interesting question is now, why do the three approaches perform different over genres? This is a difficult question and the answer can not be figured out easily. A closer look to some explicit retrieval queries is done to exemplify the results, but an extensive answer can only be given by analysis of the influences of the different feature vector parameters on the results. Such an investigation is not part of this thesis and is a suggestion for future work.

Table 5.12 to Table 5.14 show the 20 best ranked answers to a query conducted with the segment starting at second 30 of the the track ”B Please II” by artist ”Eminem” from the album ”The Marshall Matters” analyzed by the three different analysis approaches and Table 5.15- Table 5.17 does so for the segment starting at second

rank	artist	album	title	seg.
1	Eminem	The Marshall Mathers	B Please II	70
2	Eminem	The Marshall Mathers	B Please II	45
3	Eminem	The Marshall Mathers	B Please II	150
4	Outkast	ATLiens	ATLiens	70
5	A Tribe Called Quest	Beats Rhymes & Life	Mind Power	70
6	A Tribe Called Quest	Beats Rhymes & Life	Mind Power	45
7	A Tribe Called Quest	Beats Rhymes & Life	Mind Power	150
8	Shaggy		It Wasn't Me	45
9	Shaggy		It Wasn't Me	45
10	A Tribe Called Quest	Beats Rhymes & Life	The Hop	45
11	Outkast	ATLiens	ATLiens	150
12	Cypress Hill (feat. Eminem)		Rap Superstar	70
13	Mobb Deep	Hell On Earth	Nighttime Vultures	45
14	Mobb Deep	Hell On Earth	Nighttime Vultures	70
15	Mobb Deep	Hell On Earth	Nighttime Vultures	150
16	A Tribe Called Quest	Beats Rhymes & Life	The Hop	150
17	Shaggy		It Wasn't Me	70
18	Shaggy		It Wasn't Me	70
19	Shaggy		It Wasn't Me	150
20	A Tribe Called Quest	Beats Rhymes & Life	Phony Rappers	45

Table 5.13: 20 best ranked answers to a query with the segment starting at second 30 of the track ”B Please II” by artist ”Eminem” from the album ”The Marshall Matters” analyzed by the RP approach.

30 of the track ”What's new?” by artist ”Silje Nergaard” from the album ”Port of Call”. The tables contain the raw answers to the queries containing multiple entires of the same song if different segments where found as similar. Also, more than three entries of a song are possible, because, as it was figured out through the experiment, the real world music collection contains sometimes several MP3's of the same title.

The two sample query pieces of music where chosen under the aspect of similar performance following the recall and precision values for the genres hip hop and jazz. Through the size of the real world music collection its impossible to evaluate the results through listening, but by exploring a few results, the setting with this two pieces of music seems representative.

In case of the hip hop song by ”Eminem” the RP (see Table 5.13) results are the

rank	artist	album	title	seg.
1	Eminem	The Marshall Mathers	B Please II	45
2	Mobb Deep	Hell On Earth	Extortion	150
3	Mobb Deep	Hell On Earth	Can't Get Enough Of It	150
4	Eminem	Slim Shady LP	97' Bonnie & Clyde	70
5	Eminem	The Marshall Mathers	Under the Influence	45
6	Wu-Tang Clan	The W	One Blood Under W	45
7	Busta Rhymes	When Disaster Strikes	Turn It Up	45
8	Fettes Brot	Amnesie	Lieblingslied	70
9	Mobb Deep	Hell On Earth	Animal Instinct	70
10	Fettes Brot	Amnesie	Nordisch By Nature	150
11	Wyclef Jean	Masquerade	Oh What a Night	70
12	Eminem	The Marshall Mathers	Drug Ballad	70
13	Morcheeba	Fragments of Freedom	Shallow End	45
14			Get Up	45
15	Busta Rhymes	Extinction Level Event	Iz They Wildin Wit Us & Gettin Rowdy Wit...	45
16	Eminem	The Marshall Mathers	B Please II	150
17	Eminem	The Eminem Show	Drips	70
18	Absolute Beginner	Bambule	Showmaster	70
19	Eminem	Slim Shady LP	97' Bonnie & Clyde	45
20	Outkast	ATLiens	ATLiens	45

Table 5.14: 20 best ranked answers to a query with the segment starting at second 30 of the track "B Please II" by artist "Eminem" from the album "The Marshall Matters" analyzed by the MARSYAS approach.

most homogeneous ones, only number 9 can be found as a little bit misplaced by being finical. It is also important to mention the RP result is the one with the least count of different songs, because of enlisting mostly all segments of a song if one is found relevant. The MARSYAS 5.14 result is also only disturbed by two pieces of music. Number 10 is a more disco affiliated number but with hip hop vocals and number 13 is a mellow pop song with a strong hip hop like beat. DWCH results (Table 5.12) contain the most misplaced answer titles here. Number 9 is a rock song, but consisting of a beat, keys and rap only passage in the 3 second frame of analysis. The same problem arises with number 10, which is actually a punk rock song, but again in the frame of analysis there is a keys only passage. Confusion also brings number 13 and 14, actually hip hop songs too, but with reggae background

music. Over all the results for the hip hop retrieval task are very consistent and reasons for misplacing songs are traceable.

The laid back jazz song ”What’s new?” by artist ”Silje Nergaard” from the album ”Port of Call” causes much bigger confusion. The RP results provided in Table 5.16 come up with the artist ”Tori Amos” three times (Number 1,16,17), who performs with piano and voice like in the query file but has a stronger singer/songwriter association than towards jazz. Also a german songwriter, ”Reinhard May” is enlisted two times, and does definitely not fit in the answer set. A piece of music of ”Queen” is also found similar, which is not a jazz title at all, but listening into it, the misplacement gets understandable. The song is piano and voice only and has a similar mood like the query song. Results of the MARSYAS analysis 5.17 are really bad. Actually, only numbers 13 and 16 do match, all other tracks are classic, mellow pop in the broadest sense, instrumentals or soul music. For the DWCH 5.15 results a similar confusion like for the RP results happen. In the answer set tracks of ”Tori Amos”, ”Rainhard May” and ”Queen” can be found to.

These results show that a numeric only evaluation of the results a music information retrieval system produces is not a good idea and can easily lead to a false estimation of the performance of such a system.

rank	artist	album	title	seg.
1	Tori Amos	Boys For Pele	Way Down	45
2	Tori Amos	Under The Pink	The Wrong Band	45
3	Reinhard Mey	Lampenfieber	Kleiner Kamerad	150
4	Stan Getz	Stan Getz & Oscar Peterson Tri	Detour Ahead	70
5	Marilyn Crispell, Gary Peacock	Amaryllis	Conception Vessel-circl	45
6	Marilyn Crispell, Gary Peacock	Amaryllis	Conception Vessel-circl	150
7	Norah Jones	Come Away With Me	Come Away With Me	70
8	Frank Sinatra	His Greatest Hits	Ol' Man River	70
9	Tony Bennett		I left my heart in San	45
10	Tony Bennett		I left my heart in San	150
11	Stan Getz	Stan Getz & Oscar Peterson	three little words	45
12	Stan Getz	Stan Getz & Oscar Peterson	three little words	150
13	Norah Jones	Come Away With Me	Don't Know Why	150
14	Reinhard Mey	Lampenfieber	Ich bin!	45
15	Queen	Greatest Hits 4	Dear Friend	45
16	Tori Amos		a case of you (joni mitchel)	45
17	Tori Amos		a case of you (joni mitchel)	150
18	Mel Torme		autumn leaves	45
19	Mel Torme		autumn leaves	150
20	Silje Nergaard	port of call	what's new	45

Table 5.15: 20 best ranked answers to a query with the segment starting at second 30 of the track "What's new?" by artist "Silje Nergaard" from the album "Port of Call" analyzed by the DWCH approach.

rank	artist	album	title	seg.
1	Tom Waits		Take it with me	45
2	Elvis Presley	White Christmas	It Is No Secret	150
3	Tom Waits	Alice	I'm Still Here	70
4	Tori Amos		a case of you	70
5	Tori Amos	Under The Pink	Baker Baker	150
6	Tom Waits - gemischt	15-Take It With Me	70	
7	Patricia Barber	night club	invitation	70
8	Silje Nergaard	port of call	every time we say good-bye	70
9	Tom Waits	Alice	I'm Still Here	45
10	Charlie Haden quartet west	haunted heart	deep song	150
11	Charlie Haden quartet west	haunted heart	deep song	45
12	Norah Jones	Come Away With Me	The Nearness of You	45
13	Tori Amos	Boys For Pele	Hey Jupiter	70
14	Charlie Haden quartet	haunted heart	ev'ry time we say good-bye	45
15	Charlie Haden quartet	haunted heart	ev'ry time we say good-bye	150
16	Silje Nergaard	port of call	don't explain	45
17	Silje Nergaard	port of call	don't explain	150
18	marilyn crispell, gary peacock	amaryllis	conception vessel-circl	70
19	Charlie Haden quartet	haunted heart	deep song	70
20	Norah Jones	Come Away With Me	the nearness of you	45

Table 5.16: 20 best ranked answers to a query with the segment starting at second 30 of the track "What's new?" by artist "Silje Nergaard" from the album "Port of Call" analyzed by the RP approach.

rank	artist	album	title	seg.
1	Klaus Nomi		Samson and Delilah (Aria)	70
2	Elvis Presley	White Christmas	Take My Hand, Precious Lord	150
3	Frankie Goes To Hollywood		Power of Love	45
4	Coldplay	A Rush Of Blood In The Head	Amsterdam	150
5	Tori Amos		a case of you	70
6	Tori Amos	Under The Pink	Pretty Good Year	150
7	Faithless	Sunday 8pm	Why Go	45
8	O.S.T	Northern Exposure	Medley	150
9	Pearl Jam	Vitalogy	Better Man	70
10	Chris De Burgh		A Spaceman Came Travelling	150
11	frank sinatra		autumn leaves	45
12	frank sinatra		autumn leaves	150
13	natalie cole	unforgettable with love	smile	70
14	O.S.T	Northern Exposure	Medley	70
15	Randy Crawford	the very best	i stand accused	70
16	O.S.T	Four Weddings And A Funeral	Cant Smile Without You	45
17	Chris Rea		Gone Fishing	45
18	Limp Bizkit	Results May Vary	Let Me Down	150
19	Tom Waits	Heartattack And Vine	On the Nickel	150
20	Georg Friedrich Händel		Concerti Grossi op. 3 N 6 D-Dur	70

Table 5.17: 20 best ranked answers to a query with the segment starting at second 30 of the track ”What’s new?” by artist ”Silje Nergaard” from the album ”Port of Call” analyzed by the MARSYAS approach.

5.3.2 Genre based Retrieval

Tables 5.18 to 5.19 contain the results of the evaluation using the Tzanetakis music collection. Here a fixed answer set size of $i = 10$ is used and the performance in different genres is displayed. Query and data set are the same and all pieces of music of the same genre are considered to be relevant or correct answers to a query. Top ranked in the answers to a query is always the query piece of music itself. Because of this, rank number one is ignored and the absolute ranks 2-11 are interpreted as the answer set of size 10. The number of titles in a genre is enlisted in column n in Table 5.18. The best case scenario calculates as the size of the answer set multiplied by the number of titles in a genre, because the size of every genre is greater than the answer set size ($n > i$). The best case scenario for the recall values are rather small values too, because in a answer set of size 10 only a fraction of all relevant titles situated in the collection can be located. For each query there would be $n - 1$ relevant pieces of music in the collection and this value is always bigger than the answer set size ($i = 10$).

It can be observed that the performance of the prototypes vary from genre to genre. In the average the DWCH approach performs better than the other two, which come up with nearly the same average results for precision and recall evaluation. The best performance is always achieved either by the RP or the DWCH prototype. In the genres labeled ambient, country, disco, hip hop and rock the RP results are the best, with disco and hip hop having the greatest difference to the next best results. The recall and precision values for ballad, blues, classical, jazz, metal, pop and reggae are the best for the DWCH approach and are remarkable better than the other approaches for classical, metal and pop pieces of music. The MARSYAS results are second placed for classical, jazz, metal, pop, reggae and rock. Due to the performance of the feature vectors computed through the DWCH approach it is to mention again that only a 3 second part of the pieces of music is analyzed, while the others are calculated over the whole length.

genre	n	DWCH	MARSYAS	RP	best
ambient	71	238	211	252	710
ballad	45	83	70	79	450
blues	108	314	249	292	1080
classical	110	636	538	481	1100
country	104	297	238	312	1040
disco	102	286	281	375	1020
hiphop	106	304	225	385	1060
jazz	105	345	342	325	1050
metal	104	560	431	336	1040
pop	131	540	401	306	1310
reggae	114	436	387	272	1140
rock	103	177	203	207	1030
all	1203	4216	3576	3622	12030

Table 5.18: Absolute recall values for the Tzanetakis music collection in case of feature extraction by the different three prototypes and in a best case scenario.

R_{10}	DWCH	MARSYAS	RP	best
ambient	0.0486	0.0431	0.0514	0.1449
ballad	0.0429	0.0362	0.0408	0.2324
blues	0.0274	0.0217	0.0255	0.0943
classical	0.0535	0.0453	0.0405	0.0926
country	0.0280	0.0224	0.0294	0.0980
disco	0.0280	0.0275	0.0368	0.1000
hiphop	0.0276	0.0204	0.0349	0.0961
jazz	0.0319	0.0316	0.0300	0.0971
metal	0.0528	0.0406	0.0317	0.0980
pop	0.0320	0.0237	0.0181	0.0775
reggae	0.0341	0.0303	0.0213	0.0893
rock	0.0170	0.0195	0.0199	0.0990
others	-	-	-	-
avg	0.0353	0.0302	0.0317	0.1099

Table 5.19: Recall values for the Tzanetakis music collection in case of feature extraction by the different three prototypes and in a best case scenario.

P_{10}	DWCH	MARSYAS	RP	best
ambient	0.3352	0.2972	0.3549	1
ballad	0.1844	0.1556	0.1756	1
blues	0.2907	0.2306	0.2704	1
classical	0.5782	0.4891	0.4373	1
country	0.2856	0.2288	0.3000	1
disco	0.2804	0.2755	0.3676	1
hiphop	0.2868	0.2123	0.3632	1
jazz	0.3286	0.3257	0.3095	1
metal	0.5385	0.4144	0.3231	1
pop	0.4122	0.3061	0.2336	1
reggae	0.3825	0.3395	0.2386	1
rock	0.1718	0.1971	0.2010	1
avg	0.3396	0.2893	0.2979	1

Table 5.20: Precision values for the Tzanetakis music collection in case of feature extraction by the different three prototypes and in a best case scenario.

	same genre	sounds similar	different
DWCH	17.91	21.21	60.88
MARSYAS	37.17	27.01	35.82
RP	32.11	26.7	41.19

Table 5.21: Percentage of ratings in the categories, same genre, sounds similar and different genre, for best 10 retrieval answers of 12 example songs selected from the real world music collection.

5.3.3 Usability Study of specific Retrieval

As a music information retrieval system based upon music content analysis is intentionally designed for the use by human users, it is obvious and very important to survey those users about their assessment of results such a system is providing. A survey is done in the form of letting users evaluate the answers returned by the three prototypes for query songs. For each feature set 4 songs out of the real world music collection are selected by random, considering a balance of genres. This means that every prototype is used to retrieve results for a classical, a pop, a rock and a hiphop song. The users are presented the 12 ”master” songs and the resulting 10 best answers, and are asked to judge them in the sense of same genre. Not only a similar or unsimilar genre decision is possible, because of the assumption people will have very different opinions on genres. They also have the opportunity to classify a song as similar, but of different genre. That the perception of genre varies from user to user is for example confirmed through different restrictive classification behavior.

> User:*”This is german hiphop! Why is this track placed right in the middle of all english tracks?”*

The answer set is build upon the 20 best answers to the query pieces of music. In case of multiple entries of a title, through enlisting different segments of the piece of music, the best ranked segment is kept and segments ranked behind it are discarded. Subsequent pieces of music to discarded segments are shifted upwards in the answer set. As it came up, while preparing the user survey, multiple entries through of a

title caused by the different segments are not the only problem. The real world music collection contains a lot of titles more than once. Because the segments of pieces of music in the collection are anonymous there is no easy way to compile a top 10 ranking without having double or multiple entries in it. Therefore multiple entries are counted and evaluated only once. Multiple entries concerned the answer sets of all three prototypes of about the same amount.

The group of probands are 11 students with an average age of 26.7 years and is balanced in terms of gender. Over 60% of the surveyed people have some musical education and over 80% would call themselves ”interested in music”. The numerical results (see Table 5.21) of the user survey have to be handled with care. To call a user survey representative a much larger sample of interviewed people would be necessary. Results of the user survey should be seen more as an encouragement to involve users in the performance evaluation process of a music information retrival system. Following the values in Table 5.21 the MARSYAS and RP results are about the same, while the DWCH systems performance is worse. Reasons for these results, being contradictory to the precision values (see Table 5.20) for the evaluation over the Tzanetakis music collection, are the random choice of songs and hence created mismatch of starting position between the prototypes. Another reason is the fact that the users judge a piece of music in at maximum first 10 seconds, while the DWCH approach analyzes begin of second 15 to end of second 17 for feature extraction. Interesting insights in the users formation of opinion upon a piece of music’s genre are observed during the survey. It is also to say that users for almost every track made their decision in about 2 or 3 seconds. Music pieces giving users no clear association to a musical genre are quickly analyzed on a different level than the musical content impression. Through figuring out the artist or the band, known influences for the artist or band, or time period of creation the users find additional information for the genre assignment process.

> User:*”This classic tune sounds very baroque! Is this a cembalo? This title is from a totally different time period and is not similar to the classical query song in my opinion!”*

This is a capability a music information retrieval system based plainly on music content analysis cannot possess. Overall the critiques of the participants for the different answer sets ranged from

> User:*"Is this meant serious? Those songs have nothing in common!"*

to

> User:*"Cool, this would be nice to automatically generate playlists out of my private MP3 collection!"*

Chapter 6

Conclusion

Completing this work the interesting facts the comparison of the three approaches of music content analysis brought up have to be mentioned. Most significant performance differences are the outstanding performance of the RP approach for precision and recall in case of the big music collection and the different segments of a music piece in the collection considered as the correct answers. One has also to mention that the DWCH approach, although only analyzing a small part of the segments does significantly perform better than the MARSYAS approach. A possible reason could be the chosen Euclidean distance measurement, maybe the MARSYAS performance could be improved by weighting the attributes of the feature vector, because they are of different domains. Also, the size of the music collection could matter, it is larger than any other used in literature referred in this thesis. The stability of the DWCH approach, although the precision recall values are not that good, speak for it and it should be possible to improve the results in every aspect by building a feature vector upon the analysis of the whole music piece and not only of a fragment.

Another fact speaking for further research with the DWCH technique are the good results for the genre evaluation over the Tzanetakis music collection. DWCH results can compete with the RP results in this setting with no problem. This speaks for a

good stability of the DWCH approach for genre classification. The average precision for genre recognition is poor for all three prototypes. A precision of about 30% is not enough to base a reasonable application, for example an automatic playlist generation, upon.

The user evaluation of the results is to small to deduce significant numeric conclusions. Of much more interest is the fact, that genre systems people have vary from user to user and it is impossible to build a genre system, which satisfies the requirements of all users. This is also supported by the information about the development of a genre form a musicologists view. Users base their genre assignment decision for a piece of music not only on the music content, but also on various background knowledge they have. These insights also affirmed the importance of involvement of a evaluation through the users these systems are designed for, a numeric only evaluation can be misleading.

The emphasis of an automatic genre classification, which is based on music content analysis only, should concentrate on a successful very coarse top level classification. It is to question if a music content based approach can do a more detailed genre separation. Nonetheless the results are encouraging for future work and it is only a question of time till the approaches are sophisticated enough to effectively fulfill their task.

Bibliography

[1] Philippe Aigrain. New applications of content processing of music. *Journal of New Music Research*, 28(4):271–280, 1999.

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

[3] R. Bladon. Modeling the judgment of vowel quality differences. *Journal of the Acoustical Society of America*, 69:1414–1422, 1981.

[4] Eoin Brazil and Mikael Fernström. Where's that sound? exploring arbitrary user classifications of sounds for audio collection management. In *Proceedings of the 2003 International Conference on Auditory Display*, pages 66–70, July 2003.

[5] Oran E. Brigham. *FFT Schnelle Fourier-Transformation*. Einführung in die Nachrichtentechnik. R.Oldenburg Verlag, München, Wien, 3 edition, 1987.

[6] Hung-Chen Chen and Arbee L. P. Chen. A music recommendation system based on music data grouping and user interests. In *Proceedings of the 10th International Conference on Information and Knowledge Management*, pages 231–238. ACM Press, 2001.

[7] Rudi Cilibrasi, Paul Vitanyi, and Ronald de Wolf. Algorithmic clustering of music. `http://arxiv.org/PS_cache/cs/pdf/0303/0303025.pdf`, March 2003. This is the full paper version of the article published in New Scientist, Technology Research News, Ars Technica, and NRC (Dutch newspaper).

[8] Perry R. Cook. *Music, Cognition and Computerized Sound*. The MIT Press, Cambridge, Massachusetts, London, England, 1999.

[9] Holger Crysandt and Jens Wellhausen. Music classification with MPEG-7. In *Proceedings of SPIE-IS&T Electronic Imaging Vol. 5021, Storage and Retrieval for Media Databases 2003*, Electronic Imaging Proceedings, pages 397–404. The International Society for Optical Engineering, 2003.

[10] Sally Jo Cunningham, Nina Reeves, and Matthew Britland. An ethnographic study of music information seeking: implications for the design of a music digital library. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 5–16. IEEE Computer Society, 2003.

[11] Die Dudenredaktion, editor. *Duden - Fremdwörterbuch*. Der Duden. Dudenverlag, Mannheim, Leipzig, Wien, Zürich, 7 edition, 2001.

[12] Eric W. Weisstein et al. Orthonormal basis. From MathWorld–A Wolfram Web Resource, `http://mathworld.wolfram.com/OrthonormalBasis.html`.

[13] Franco Fabbri. A theory of musical genres: Two applications. *Popular Music Perspectives*, pages 52–81, 1981.

[14] Franco Fabbri. Browsing music spaces: Categories and the musical mind. paper presented at the 3rd Triennial British Musicological Societies' Conference, 1999.

[15] Markus Frühwirth. Automatische Analyse und Organisation von Musikarchiven. Master's thesis, Institute of Software Technology and Interactive Systems, Vienna University of Technology, March 2001.

[16] David Gerhard. *Audio Signal Classification*. PhD thesis, School of Computing Science, Simon Fraser University, February 2000.

[17] Amara Graps. An introduction to wavelets. *IEEE Computational Science and Engineering*, 2(2), 1995.

[18] Marco Grimaldi, Anil Kokaram, and Padraig Cunningham. Classifying music by genre using the wavelet packet transform and round-robin ensemble. Technical Report TCD–CS–2002–64, Trinity College Dublin, November 2002.

[19] Donald E. Hall. *Musikalische Akustik*. Schott Musik International, Mainz, 1997.

[20] Olaf Hansen. *Einführung in die Theorie und Anwendung der Wavelets*. Springer series in information sciences. Logos Verlag, Berlin, 2000.

[21] International Organization for Standardization. Information technology – coding of moving pictures and associated audio for digital storage media at up to about 1,5 mbit/s – part 3: Audio. JTC 1/SC 29 ISO/IEC 11172-3, 1993.

[22] International Organization for Standardization. MPEG-7 multimedia content description interface – part 4: Audio. ISO/IEC JTC1/SC29/WG11 FDIS 15938-4:2001(E), June 2001.

[23] International Organization for Standardization. MPEG-7 overview. ISO/IEC JTC1/SC29/WG11N4980, July 2002.

[24] Wolfgang Klas. *Multimedia II*, lecture notes Metadaten, pages 6.1–6.30. Abt. Multimediale Informationssysteme, Institut für Informatik und Wirtschaftsinformatik, Fakultät für Wirtschaftswissenschaften, Universität Wien, Wien, 2001.

[25] Karin Kosina. Music genre recognition. Master's thesis, Fachhochschul-Studiengang Medientechnick und Design, Hagenberg, June 2002.

[26] Guohui Li and Ashfaq A. Khokhar. Content-based indexing and retrieval of audio data using wavelets. In *IEEE International Conference on Multimedia and Expo (II)*, pages 885–888, 2000.

[27] Tao Li, Mitsunori Ogihara, and Qi Li. A comparative study on content-based music genre classification. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 282–289. Association for Computing Machinery, July 2003.

[28] Beth Logan. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the 1st International Symposium on Music Information Retrieval*, October 2000.

[29] Mrinal K. Mandal, Tyseer Aboulnasr, and Sethuraman Panchanathan. Image indexing using moments and wavelets. *IEEE Transactions on Consumer Electronics*, 42(3):557–565, August 1996.

[30] Vinay P. Namboodiri. Swaram: Segmentation and wavelet based retrieval of music. In *Proceedings of the International Conference on Digital Libraries 2004*, New Dehli, India, February 2004.

[31] Francois Pachet. Content management for electronic music distribution. *Communications of the ACM*, 46(4):71–75, April 2003.

[32] Elias Pampalk. Islands of music - analysis, organization, and visualization of music archives. Master's thesis, Institute of Software Technology and Interactive Systems, Vienna University of Technology, December 2001.

[33] D. Perrot and Robert O. Gjerdigen. Scanning the dial: An exploration of factors in the identification of musical style. In *Proceedings of the 4th Conference of the Society for Music Perception and Recognition*, 1999.

[34] Robi Polikar. The wavelet tutorial. `http://users.rowan.edu/~polikar/WAVELETS/WTtutorial.html`. accessed on March 7th, 2004.

[35] Andreas Rauber and Markus Frühwirth. Automatically analyzing and organizing music archives. In P. Constantopoulos and I.T. Solvberg, editors, *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001)*, Springer Lecture Notes in Computer Science, pages 402–414. Springer Verlag, 2001.

[36] Andreas Rauber and Dieter Merkl. Text mining in the somlib digital library system: The representation of topics and genres. *Applied Intelligence*, 18(3):271–293, May-June 2003. `http://ipsapp008.kluweronline.com/content/getfile/4504/35/4/abstract.htm`.

[37] Andreas Rauber, Elias Pampalk, and Dieter Merkl. The SOM-enhanced JukeBox: Organization and visualization of music collections based on perceptual models. *Journal of New Music Research*, 32(2):193–210, June 2003. `http://www.extenza-eps.com/extenza/loadHTML?objectIDValue=16745&type=abstract`.

[38] J. Ben Schafer, Joseph Konstan, and John Riedl. Recommender systems in e-commerce. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, pages 158–166. ACM Press, 1999.

[39] M. R. Schröder, B. S. Atal, and J. L. Hall. Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of the Acoustical Society of America*, 66:1647–1652, 1979.

[40] Wolfgang Schulz. *Einführung in die Soziologie*. Institut für Soziologie der Sozial- und Wirtschaftswissenschaftlichen Fakultät der Universität Wien, Wien, 4 edition, 1993.

[41] Gilbert Strand and Truong Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, USA, 1996.

[42] Tero Tolonen and Matti Karjalainen. A computationally efficient multipitch analysis model. *IEEE Transactions on Speech and Audio processing*, 8(6):708–716, November 2000.

[43] George Tzanetakis. *Manipulation, Analysis and Retrieval Systems for Audio Signals*. PhD thesis, Princeton University, June 2002.

[44] George Tzanetakis and Perry Cook. 3d graphics tools for isolated sound collections. In *Proceedings of the COST G-6 Conference on Digital Audio Effects*, Verona, Italy, December 2000.

[45] George Tzanetakis and Perry Cook. Marsyas: A framework for audio analysis. *Organized Sound, Cambridge University Press*, 4(3), 2000.

[46] George Tzanetakis and Perry Cook. Sound analysis using mpeg-compressed audio. In *Proceedings of the International Conference on Audio, Speech and Signal Processing*, 2000.

[47] George Tzanetakis and Perry Cook. Marsya3d: A prototype audio browser-editor using a large scale immersive visual and audio display. In *Proceedings of the 2001 International Conference on Auditory Display*, Helsinki, Finland, July 2001.

[48] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), July 2002.

[49] George Tzanetakis, Perry Cook, and Andrey Ermolinskyi. Pitch histograms in audio and symbolic music information retrieval. In *Proceedings of the International Conference on Music information Retrieval*, October 2002.

[50] George Tzanetakis, Perry Cook, and Georg Essl. Human perception and computer extraction of musical beat strength. In *Proceedings of the 5th International Conference on Digital Audio Effects*, Hamburg, Germany, October 2002.

[51] George Tzanetakis, Andreye Ermolinskyi, and Perry Cook. Beyond the query-by-example paradigm: New query interfaces for music information retrieval. In *Proceedings of the International Computer Music Conference (ICMC)*, Gothenburg, Sweden, September 2002.

[52] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaton. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, 3(3):27–36, 1996.

[53] Eberhard Zwicker and Hugo Fastl. *Psychoacoustics – facts and models*. Springer series in information sciences. Springer, Mainz, Berlin, 2 edition, 1999.