Music Information Retrieval

Conceptual Framework, Annotation and User Behaviour

Micheline Lesaffre

Thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Art Science
Thesis supervisor: Prof. Dr. M. Leman
Faculty of Arts and Philosophy, Department of Art, Music and Theatre Sciences
Ghent University 2005-2006

Volume 1

Chair of the dissertation committee

Prof. Dr. Hendrik Pinxten,
Department of Comparative Science of Culture.

Doctoral dissertation committee

Prof. Dr. Francis Maes
Department of Musicology, Ghent University.

Prof. Dr. Ir. Jean-Pierre Martens
ELIS, Department of Electronics and Information Systems, Ghent University.

Dr. Ir. Leon Van Noorden
TU/e, Eindhoven University of Technology and IPEM, Ghent University.

Dr. Frans Wiering
ICS, Institute for Information and Computing Sciences, Utrecht University.

Candidate's Declaration

I certify that the thesis entitled: "Music Information Retrieval: Conceptual Framework, Annotation and User Behaviour", submitted for the degree of doctor of art science, is the result of my own research, except where otherwise acknowledged, and that this thesis in whole or in part has not been submitted for an award, including a higher degree, to any other university or institution.

This thesis was written in the context of the Musical Audio Mining (MAMI) project, which was funded by the Flemish Institute for the Promotion of Scientific and Technical Research in Industry.

Acknowledgments

I was fortunate to have been able to carry out this study at the Institute for Psychoacoustics and Electronic Music (IPEM), the research centre of the department of Musicology at Ghent University, between 2001 and 2005.

When I arrived at IPEM in 1999, I did not know what to expect from my research at that time. Writing my doctoral thesis in a language which is not my own has been quite an experience. It goes without saying, however, that the process of doing research and struggling to finalize a dissertation remains a collaborative work. I therefore owe many people a debt of gratitude.

First and foremost, I wish to express my deepest appreciation to my supervisor, Prof. Dr. Marc Leman, head of the wonderful research team at IPEM. He deserves special credits for continuously providing me with valuable ideas, guidance and experienced criticism during my research period. Thanks for opening the doors to a life-changing experience.

Further, I would also like to thank the members of the exam committee, Prof. Dr. Francis Maes, Prof. Dr. Ir. Jean-Pierre Martens, Dr. Ir. Leon Van Noorden, and Dr. Frans Wiering for reviewing the present manuscript. I would like to express my gratitude to the members of my training committee, Dr. Dirk Moelants, and Prof. Dr. Ir. Jean-Pierre Martens who left their valued stamp on my work.

My sincere thanks also go to all my colleagues in the IPEM team for their support and co-operation. There are two persons I would like to mention in particular: Liesbeth de Voogdt, M.D., whose help in conducting the experiments and whose enthusiastic willingness to discuss statistics with me have enriched my work, and Frank Desmet, M.D., who was of great support in different aspects of database technology.

Many thanks should also go to my dearest friend, Dr. Peter van Poucke, for bringing light into my life at moments when I absolutely needed it. I would like to extend these warm feelings to my parents and all my special friends for all the good times we have together.

Finally, there is no way to express the love and gratitude that I feel for my partner and soul mate Erwin Desmet, who has encouraged me throughout my life to hang on to this cosmic joke we live in.

Content

List of original articles IX
Abbreviations X
Abstract XI

INTRODUCTION 3

1 PROBLEM SPECIFICATION 11

1.1 THE MIR SYSTEM AND ITS COMPONENTS 11
1.1.1 Databases 12
1.1.2 Actions 14
1.1.3 Content 14
1.1.4 Users 15
1.1.5 Queries 16

1.2 PROBLEMS OF MIR RESEARCH 17
1.2.1 Digital music databases and MIR systems 17
1.2.2 Conceptual framework 20
1.2.3 Music annotation 21
1.2.4 User context 21
1.2.5 Summary of MIR problems 22

1.3 RESEARCH CHALLENGES AND PROCESSES 23
1.3.1 The process of categorization 24
1.3.2 The process of annotation 24
1.3.3 The process of contextualization 25

1.4 RESEARCH QUESTIONS 25

1.5 APPROACH AND METHODOLOGY 25

2 TAXONOMY AS CONCEPTUAL FRAMEWORK 29

2.1 BACKGROUND 31

2.2 A MUSIC INFORMATION RETRIEVAL FRAMEWORK 32
2.2.1 Definition of terms 33
2.2.2 Framework components 33

2.3 THE MAMI TAXONOMY 35

2.4 A MULTI-LEVELLED FRAMEWORK 36
2.4.1 Description levels 37
2.4.1.1 Low-level descriptors 38
2.4.1.2 Mid-level descriptors and concepts 38
2.4.1.3 High-level descriptors and concepts 39
2.4.2 Concept categories and features 39
2.4.2.1 Spatial features 40
2.4.2.2 Temporal features 40
2.4.2.3 Spatial-temporal features 41

2.5 CONSTITUENT MUSIC CATEGORIES 43
2.5.1 Descriptor types 43
2.5.2 Taxonomy of music categories 44

2.6 SYSTEM REQUIREMENTS 49
2.6.1 Flexibility 49
2.6.2 Consistency 50
2.7 CONCLUSION 51
3 ANNOTATION OF MUSIC 53
3.1 BACKGROUND 55
3.2 THE SCOPE OF MUSIC ANNOTATION 57
3.3 AUTOMATIC VERSUS MANUAL ANNOTATION 58
3.4 ANNOTATION PROBLEM SPECIFICATION 59
3.5 MANUAL ANNOTATION FRAMEWORK 60
3.5.1 Context dependencies 60
3.5.2 Computer modelling 62
3.5.3 Representation levels 63
3.6 CONCLUSION 64
3.7 CASE STUDY 1: ANNOTATION OF VOCAL QUERIES 65
3.7.1 Annotation strategy 65
3.7.2 User-oriented annotation 65
3.7.3 Model-oriented annotation 67
3.7.4 Methodology 68
3.7.5 Results 69
3.7.6 Conclusion 70
3.8 CASE STUDY 2: ANNOTATION OF DRUMS 71
3.8.1 Annotation strategy 71
3.8.2 Annotated features 72
3.8.3 Annotation methodology 73
3.8.4 Results 74
3.8.5 Conclusion 75
4 SPONTANEOUS USER BEHAVIOUR 77
4.1 WAYS OF QUESTIONING 79
4.1.1 Textual-based approaches 80
4.1.2 Audio-based approaches 80
4.2 FIELDWORK 82
4.3 SURVEY OF MUSIC SEARCH BEHAVIOUR 84
4.3.1 Pilot study 84
4.3.2 Set up, design and methodology 84
4.3.3 Questions and findings 84
4.3.4 User groups 87
4.3.5 Comparison with other findings 89
4.3.6 Conclusion 89
4.4 EXPERIMENT ON SPONTANEOUS VOCAL QUERY BEHAVIOUR 90
4.4.1 Previous experimental studies 91
4.4.2 Music stimuli 94
4.4.3 Set up, method and procedure 96
4.4.4 Experiment part one 97

4.4.5 Experiment part two ... 98
4.4.6 Experiment part three ... 99
4.4.7 Output ... 99
4.4.8 Analysis strategy ... 100
4.4.9 General aspects of the queries ... 102
4.4.10 Segment specific aspects of the queries ... 102
4.4.11 Syllabic queries ... 104
4.4.12 Analysis by subjects ... 106
4.4.13 Stimuli related aspects ... 110
4.4.14 Effects of memory use ... 111
4.4.15 Discussion ... 112
4.4.16 Conclusion ... 114

5 USER CONTEXT ... 117

5.1 BACKGROUND IN USER CONTEXT ... 119
5.2 GLOBAL SET UP OF THE STUDY ... 120
5.3 SURVEY METHOD ... 121
5.3.1 Survey set up ... 121
5.3.2 Survey design ... 122
5.3.3 Survey procedure ... 122
5.4 QUESTIONS AND FINDINGS ... 123
5.4.1 Global background ... 124
5.4.2 Music background ... 127
5.4.3 Genre ... 128
5.4.4 Taste ... 130
5.4.5 Favourites ... 132
5.4.6 Qualities of favourite titles ... 135
5.4.7 Relations ... 140
5.5 CONCLUSION ... 143

6 DESCRIPTION OF HIGH-LEVEL FEATURES ... 145

6.1 BACKGROUND IN ANNOTATION OF HIGH-LEVEL FEATURES ... 147
6.1.1 Emotion-based music information retrieval ... 147
6.1.2 Commercial applications ... 148
6.1.3 User studies ... 148
6.1.4 Attribution of affect and structural content ... 149
6.2 DESCRIPTION OF MUSIC QUALITIES ... 149
6.2.1 Attribution of affect and emotion ... 150
6.2.2 Description of structural qualities ... 150
6.2.3 Description of involved activity ... 151
6.2.4 Impact of familiarity ... 151
6.3 METHOD ... 151
6.3.1 Subjects ... 151
6.3.2 Stimuli ... 152
6.3.3 Design ... 153
6.3.4 Procedure ... 155
6.4 STATISTICS ... 156
6.4.1 Data exploration ... 156

6.4.2 Analysis ... 156
6.5 RESULTS ... 157
6.5.1 Comparison of datasets ... 157
6.5.2 Influence of subject related factors ... 158
6.5.3 Effect of familiarity with the music excerpts ... 159
6.5.4 Factor analysis of expressive features ... 160
6.5.5 Factor analysis of structural features ... 161
6.5.6 Unanimity among subjects about structural features ... 162
6.5.7 Relationships between perceived qualities ... 163
6.6 CONSISTENCY WITHIN ANNOTATIONS ... 166
6.6.1 Set up ... 166
6.6.2 Analysis ... 167
6.6.3 Results ... 167
6.7 DISCUSSION ... 170
6.8 CONCLUSION ... 172
7 THREE APPLICATIONS ... 175
7.1 APPLICATION 1: USER INTERFACE TAXONOMY ... 177
7.1.1 Music categories ... 178
7.1.2 Query methods ... 179
7.1.3 Use modes ... 180
7.1.4 Feedback ... 181
7.1.5 Detailed overview ... 182
7.1.6 Conclusion ... 193
7.2 APPLICATION 2: DATA MANAGEMENT ... 193
7.2.1 Approach ... 194
7.2.2 Top-down approach: conceptual design ... 195
7.2.3 Preliminary set up: tables and relationships ... 196
7.2.4 Bottom-up approach: database normalization ... 198
7.2.5 Query builder ... 198
7.2.6 Conclusion ... 199
7.3 APPLICATION 3: SEMANTIC MUSIC RECOMMENDER SYSTEM ... 200
7.3.1 Background ... 200
7.3.2 Approach ... 202
7.3.3 Design and procedure ... 202
7.3.4 Fuzzy logic functions ... 207
7.3.5 Conclusion ... 210
DISCUSSION AND CONCLUSION ... 211
List of figures ... 227
List of tables ... 228
Content of the electronic appendix ... 230
References ... 231

List of original articles

This thesis is based on the following articles, referred to in the text by Roman numerals. (PDF files of these articles are included in the electronic appendix).

I. Lesaffre, M., Moelants, D., Leman, M., De Baets, B., De Meyer, H. & Martens, J.-P. (2003). User behaviour in the spontaneous reproduction of musical pieces by Vocal Query. In R. Kopiez, A.C. Lehmann, I. Wolther and C. Wolf (Eds.) *Proceedings of the 5th triennial European Society for the Cognitive Sciences of Music Conference (ESCOM)*, 208-211, Hanover.

II. Lesaffre, M., Leman, M., Tanghe, K., De Baets, B., De Meyer, H. & Martens, J.-P. (2003). User-dependent taxonomy of musical features as a conceptual framework for musical audio-mining technology. *Proceedings of Stockholm Music Acoustics Conference (SMAC)*, 635-638, Stockholm.

III. Lesaffre, M., Tanghe, K., Martens, G., Moelants, D., Leman, M., De Baets, B., De Meyer, H. & Martens, J.-P. (2003). The MAMI Query-by-Voice experiment: collecting and annotating vocal queries for music information retrieval. *Proceedings of the International Conference on Music Information Retrieval (ISMIR03)*, 65-71, Baltimore,

IV Lesaffre, M., Leman, M., De Baets, B. & Martens, J.P. (2004). Methodological considerations concerning manual annotation of musical audio in function of algorithm development. *Proceedings of the International Conference on Music Information Retrieval (ISMIR04)*, 64-71, Barcelona.

V. Lesaffre, M., Moelants, D., & Leman, M. (2005). Spontaneous user behaviour in vocal queries for audio-mining. In: Walter B. Hewlett & Eleanor Selfridge-Field (Eds.), Music Query: Methods, Models, and User Studies, *Computing in musicology 13*, 129-146. Published by CCARH and the MIT Press.

VI. Lesaffre, M., De Voogdt, L., & Leman, M. (2005). Relations between listener's background, their description of expressive qualities and structural features of music. (In preparation).

Abbreviations

CBMA	Content-Based Music Analysis
CBMIR	Content-Based Music Information Retrieval
ICASSP	International Conference on Acoustics, Speech and Signal Processing
IPEM	Institute for Psychoacoustics and Electronic Music
IR	Information Retrieval
ISMIR	International Conference on Music Information Retrieval
MAMI	Musical Audio Mining
MDL	Music Digital Library
MIDI	Musical Instrument Digital Interface
MIR	Music Information Retrieval
MIREX	Music Information Retrieval Evaluation eXchange
MIRS	Music Information Retrieval System
MHI	Musical Habits and Interests
USIT	USer Interface Taxonomy
PMQ	Perceived Musical Qualities
RDBMS	Relational Database Management System
SeMuRes	Semantic Music Recommender System
SMR	Symbolic Music Representation
QbV	Query-by-Voice
WYHIWYG	What You Hum Is What You Get

Abstract

Understanding music is a process both based on and influenced by the knowledge and experience of the listener. Although content-based music retrieval has been given increasing attention in recent years, much of the research still focuses on bottom-up retrieval techniques. In order to make a music information retrieval system appealing and useful to the user, more effort should be spent on constructing systems that both operate directly on the encoding of the physical energy of music and are flexible with respect to users' experiences.

This thesis is based on a user-centred approach, taking into account the mutual relationship between music as an acoustic phenomenon and as an expressive phenomenon. The issues it addresses are: the lack of a conceptual framework, the shortage of annotated musical audio databases, the lack of understanding of the behaviour of system users and shortage of user-dependent knowledge with respect to high-level features of music.

In the theoretical part of this thesis, a conceptual framework for content-based music information retrieval is defined. The proposed conceptual framework - the first of its kind - is conceived as a coordinating structure between the automatic description of low-level music content, and the description of high-level content by the system users. A general framework for the manual annotation of musical audio is outlined as well. A new methodology for the manual annotation of musical audio is introduced and tested in case studies. The results from these studies show that manually annotated music files can be of great help in the development of accurate analysis tools for music information retrieval.

Empirical investigation is the foundation on which the aforementioned theoretical framework is built. Two elaborate studies involving different experimental issues are presented. In the first study, elements of signification related to spontaneous user behaviour are clarified. In the second study, a global profile of music information retrieval system users is given and their description of high-level content is discussed. This study has uncovered relationships between the users' demographical background and their perception of expressive and structural features of music. Such a multi-level approach is exceptional as it included a large sample of the population of real users of interactive music systems. Tests have shown that the findings of this study are representative of the targeted population.

Finally, the multi-purpose material provided by the theoretical background and the results from empirical investigations are put into practice in three music information retrieval applications: a prototype of a user interface based on a taxonomy, an annotated database of experimental findings and a prototype semantic user recommender system.

Results are presented and discussed for all methods used. They show that, if reliably generated, the use of knowledge on users can significantly improve the quality of music content analysis. This thesis demonstrates that an informed knowledge of human approaches to music information retrieval provides valuable insights, which may be of particular assistance in the development of user-friendly, content-based access to digital music collections.

Introduction

Research area

In this thesis the research is presented that I have been doing during the past four years in the area of Music Information Retrieval (MIR)[1]. Researchers have spent over half a century focusing on text indexing and retrieval. As a consequence, the discipline of Information Retrieval (IR) has become well-known through web search engines like Altavista and Google. So much has been accomplished in the realm of text retrieval that other types of information such as images, film and music have called for attention. Whether we like it or not, music nowadays is part and parcel of our daily lives. Our involvement with music is reflected in a network of commodities and services, but the methods that allow interaction with music do not always meet user needs. Users of search and retrieval systems have recognized that conventional text-based searching does not provide the best method for accessing items of interest. Therefore, search and retrieval of specific musical content (e.g. emotion, melody) has become an important aspect of system development. Future music information retrieval systems will fundamentally alter the way people experience and interact with music. The challenge of making a modest contribution to the development of music information retrieval has been appealing to me.

Background

The earliest forms of music information retrieval were printed reference works or manual music retrieval systems arranged by title, composer or genre. Kassler (1966) came up with the idea of applying automatic information retrieval techniques to music. Yet, intensive research toward automated music retrieval has only been conducted during the past two decades. It came to the fore as a response to the growing interest in digital music collections, online services and interactive music search and retrieval systems. Music information retrieval research based on classical information retrieval mainly relies on textual query input. Yet, specifications like title, composer or style, however, require that users know the music that they desire. Experience with the Internet during the last decade, however, has shown indeed that most user queries do not involve looking for known items. Rather, the most common query is to discover items of which the user is ignorant (Huron and Aarden, 2000). For these kinds of queries, classic reference information has proved to have a limited value. Information technology thus becomes increasingly interesting when music information processing goes beyond text-based into content-based methods. Accordingly, much current research into music information retrieval deals with the characteristics of the music itself rather than with information about the music. Musical features are used to search databases of musical content by means of musically expressed queries.

[1] For more information on the MIR research community, see the proceedings of the International Symposium on MIR (http://www.ismir.net).

Among the very alluring alternatives of the last years is *vocal querying* in which users search the required musical piece by singing parts (the so-called *query-by-humming* or *WYHIWYG*[2] paradigm) and query-by-emotion. The average users take advantages of such query method because prior bibliographic knowledge of a piece of music is no longer needed. What is more, these are more natural ways of searching. Nevertheless, their practical application still faces many theoretical and technical barriers. Reasons for the existence of these barriers are often heaped together and explained by the fuzzy idea of the *complexity of music*.

Currently, music information retrieval is known as the interdisciplinary research area dedicated to retrieving information *about* and *from* music. The process of music information retrieval deals with both organizing and extracting music information. Music information retrieval research then entails actions, methods and procedures for searching huge amounts of digitally stored musical data. My study is situated in the subdomain of music information retrieval that regards systems that permit a user to interact with the musical audio stream, also referred to as *musical audio mining* (MAMI) (Leman, 2002). Data mining on musical audio aims at extracting from the musical audio stream the kind of high-level content users can deal with in interactive music systems. Moreover, content-based MIR systems allow searching for some musical characteristics or all of the music itself, whether melody, rhythm or harmony alone or all aspects at once. It is allocated to search and retrieve music by means of content-based text and musical audio queries.

Aims

My research is above all *user-centred*. It is a top-down approach that is aimed at improving the efficiency of what P. Lamere describes as *Search Inside the Music*[3]. I am aware that much work has already been done in the field of music information retrieval, but some aspects still are virgin territory. A wide variety of bottom-up retrieval techniques has been developed from the motivation that they are needed for handling huge amounts of digital music. Few techniques, however, have shown to be effective. Very often, the difficulties encountered stem from a mismatch between the system and the user. More specifically, there is no theoretical background that clearly associates low-level music description (quantitatively) with the way in which humans understand music (qualitatively). Several authors within the music information retrieval community (e.g. Futrelle, 2002; Uitdenbogerd, 2002) have been commenting on the need for user-centred approaches. My aim was not to continue the research on the low-level aspects of music information retrieval that can be characterized as processes that deal with *content extraction*. Instead, I have focused on user behaviour and aspects of cognition and emotion that are high-level semantics. I

[2] What You Hum is What You Get.
[3] During a talk at Sun Labs on April 28 2005, MP3 retrieved from http://research.sun.com/sunlabsday/talks.php?id=53

characterize the process that deals with these features as *content addition*. The advantage of my approach is that it bears close similarity to music perception, which is an area that is often underestimated in music information retrieval system development.

In fact, the issue that becomes pressing is that there is no agreement on a methodological ground that could result in the ideal, *fully integrated* music information retrieval system. A fully integrated music information retrieval system gives access to any kind of music, is suited for any indexing method, is sensitive to any kind of user, covers all kinds of music information a user might desire and anyone can contribute to it. It acts as a connective tissue filling in the spaces between the researches from the different communities involved.

This dissertation is not about a user-centred research *versus* system-centred research approach. On the contrary, it is intended to provide methodological and empirical ground in view of linking between both. This thesis is the result of the challenge of working out a framework that will have a beneficial effect on music information retrieval system development. To achieve this goal I attempt to provide deeper insight into three of the many problems that stand in the way for the accomplishment of a fully integrated approach to music information retrieval. The weaknesses I touch upon are: (1) the lack of a conceptual framework, (2) the shortage of annotated musical audio databases and (3) deficiency of information on music information retrieval users' behaviour.

The first problem regards the requirement of a methodology that supports linkage between bottom-up and top-down approaches to music information retrieval. I am not aware of the existence of a framework that could maintain a fully integrated approach to music information retrieval. However, incorporation cannot be successful if there is no agreement on the elements that constitute the whole. What I set out to do is to supply a structure for dealing with carefully selected concepts of music. The kind of framework that I explore, concerns both humans as well as machines dealing with music content. Secondly, I tackle the problem that not enough research has been done on databases with real music. This means that I found myself having to generate a reference collection and develop a methodology for the annotation of music. The third issue relates to the fact that I could find no empirical user data available for research purposes. On top of that, the literature scarcely reports on responses from real users to carefully crafted questionnaires assessing their context (e.g. personal background, spontaneous behaviour, habits, musical skills, perceptual limitations). The only way to deal with these concerns was by doing heaps of manual work there where automatic tools (still) fail. Therefore, I had to gather reliable data based on real music and involving real humans.

Motivation

My motivation for conducting elaborate empirical investigations is double. To begin with, I believe that with fully operational and, ideally, fully integrated, music information retrieval systems an attractive future is in development for everyone interested in music, from

amateur to musicologist, audio researcher, composer and performer. My second motivation is rather personal. It has to do with the lifelong fascination for music that I have, together with a bizarre chain of events. Not in the least being part of the research team led by Prof. Dr. M. Leman made me take up the challenge. The work described in this thesis expands on the IPEM Toolbox and MAMI[4], two directly related research projects at Ghent University in which I was involved. The idea underlying both projects was that the whole music community would benefit from a methodology that supports the development of new tools for a fully integrated approach to interactive music systems.

The IPEM Toolbox (1999-2002) provides foundations and tools for new music analysis that is perception-based. It started from the observation that a sonological analysis of musical sound, based on human perception, would provide better understanding of the music components that contribute to musical information processing. Computer modelling of the human auditory system was used to implement feature extraction tools that apply to different musical parameters such as rhythm, pitch, roughness and energy. The auditory model developed at ELIS[5] (Martens and Van Immerseel, 1990; De Mulder et al., 2004) provided the basis for the development of higher level feature extraction tools. A representational framework was presented that includes acoustical, perceptual and cognitive description levels. The manual that comes with the toolbox documents both the concepts of the implemented models and the usage of the functions[6]. During the course of 2004, the IPEM Toolbox software was released as an open source package.

The MAMI project (2001-2005) has just come to an end. It was a data-mining project that focused on musical audio recognition. An initial taxonomy, in this thesis referred to as the "MAMI taxonomy", has been defined as a referential framework for research perspectives. New ways of searching an audio archive have been investigated and tools have been developed that deal with musical content and associated processing. Software, test collections and annotation material are available online[7].

Thesis outline

This thesis consists of two bound paper volumes and an electronic appendix.

Volume one of this thesis consists of seven chapters that reflect three research procedures: theoretical, empirical and practical.

In the first research procedure (chapters one to three), the subject matter is approached from a theoretical viewpoint. The components of a music information retrieval system, the

[4] The IPEM Toolbox has been developed at IPEM, research centre of the department of Musicology, with support from the Research Councel of the University (BOF) and the Fund for Scientific Research of Flanders (FWO). The MAMI project was also conducted at IPEM and was supported by the Flemish Institute for the promotion of Scientific and Technical Research in Industry.
[5] http://www.elis.UGent.be
[6] http://www.ipem.UGent.be/Toolbox
[7] http://www.ipem.UGent.be/MAMI/Public

problems of music information retrieval research and the research challenges are discussed in chapter one. Next, in chapter two, a taxonomy as conceptual framework is presented. Then, in chapter three, music annotation is defined and a model for manual annotation of musical audio is worked out. The annotation strategy is tested in two case studies concerning annotation of vocal melodies and annotation of drums.

The second research procedure (chapters four to six), is concerned with the empirical investigation that has been carried out. Chapter four is a study of user behaviour in the context of query-by-voice applications. First, fieldwork in the realm of music distribution and a survey among university students dedicated to users' music search behaviour is reported. Then a study into spontaneous user behaviour in vocal queries is discussed. An extensive study that focuses on the annotation of qualities of music is presented in the next two chapters. Chapter five focuses on the first part of the investigation, which is a survey of music information retrieval users' musical habits and interests. An experiment with participants in the survey concerning the description of high-level musical features is discussed in chapter six.

The third research procedure (chapter seven) is practical. It is dedicated to the development of three applications based on the theoretical and empirical background.

Finally, conclusions and suggestions are formulated.

Volume two, consists of evidence tables, figures and illustrations that refer to the text in volume one. Reference to volume two is made in volume one by means of the symbol ❷, followed by the page number between brackets [e.g. (❷ p. 33)].

The **electronic appendix**, referred to by means of the symbol ❸, is included in the back cover of volume two. A list of its contents is incorporated at the back of volume one. The electronic appendix comprises, among other things, the original articles on which this thesis is based (pdf), lists with music stimuli used in the experiments and demos of the applications discussed in chapter seven.

1 Problem Specification

Interest in music information retrieval has been steadily growing as is evidenced by the numbers of MIR-related papers at international conferences (e.g. ICASSP) and the fifth year existence of ISMIR[8], a conference solely focused on music information retrieval. Many of the techniques used in music information retrieval have their roots in more traditional areas such as speech recognition, psychoacoustics and audio compression.

The main motivation for research on musical information retrieval is that traditional information retrieval is based on bibliographic descriptions that do not provide access to the music itself, but to metadata, such as the title of the song, the name of the composer or the year of its creation. In contrast with this, music content would address the musical idea represented in the score, the gestures of the performer playing an instrument or the auditive result. In order to efficiently interact with large collections of music, it is necessary to develop tools that can deal with music content. Automated access to music content is an idea that intrigues many musicologists, computer scientists, librarians and music lovers. However, the different research communities involved have their own research goals in mind and often it seems that there are as many approaches to music information retrieval as there are researchers.

Music content can be specified in many ways and the many different characteristics of music, and of how it is perceived, offer a wide variety of approaches to building a music information retrieval system. Nevertheless, the general requirements that every such system faces are the following:

- it needs to be able to deal with large musical audio databases;
- its accuracy of music recognition must be sufficient;
- its computational complexity should be acceptable;
- its feedback should be subjectively meaningful.

In this chapter I define the *research questions* as used in this dissertation. To begin with, I give some background of my understanding of the music information retrieval *system components*. This is followed by a brief overview of the music information retrieval *problems* that I address. Next I explain the *goals*, the *philosophy* and the *methodology of my approach.*

1.1 The MIR system and its components

The ultimate task of a music information retrieval system (MIRS) that achieves automatic content-based retrieval of audio is the accurate transfer of musical information from a database to a user. The database can be stored both locally and distributed and can be retrieved both locally and via a network. In any case, interaction with databases requires user devices that can handle the input query and provide the requested information. Music information retrieval systems deal with techniques and mechanisms for representing music

[8] A cumulative list of past ISMIR papers is available at http://www.ismir.net/allpapers.html

queries, searching collections of music for possible matches, and retrieving those results that best match user queries. In what follows a global sketch of the MIRS architecture is given, describing the constituent parts and related problems.

A music information retrieval system basically consists of (1) a *database* and can deal with (2) *actions* that retrieve (3) *musical content* requested by (4) a *user* who provides a (5) *query* to the system. Figure 1 shows a generalized representation of a MIR system architecture that starts from annotated music. It consists of target structures with the annotated audio database and abstract representations (left), and query structures with query input and abstract representations (right). A user (middle) provides a query. He/she has particular preferences and expects an accurate query response (bottom). The action task (e.g. feature extraction, similarity matching) consists of retrieving the required information (e.g. a list of titles or a music excerpt) using query input information. In what follows, a description of the constituent parts of a music information retrieval system is given.

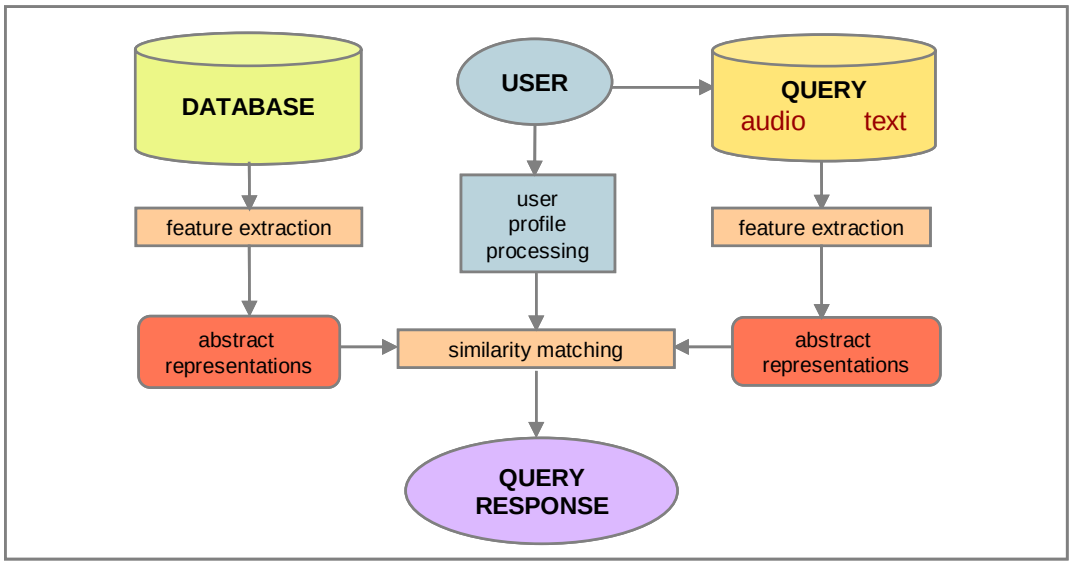


Figure 1 : Generalized representation of a MIR system architecture.

1.1.1 Databases

A digital music database[9] stores *musical* data and is essentially a type of *multimedia database* which is designed to both contain and search music. This database may store digital copies of music score sheets (e.g. The Lester S. Levy Collection of Sheet Music[10])

[9] In the MIR literature a “digital music database” is often referred to as “digital music library” (DML) and this is usually restricted to scores and bibliographic information. I prefer to use the first description because, in its traditional sense, a library is a collection of books and periodicals, whereas the focus here is on databases also including real music. Besides, in computer science a “library” also means a collection of subprograms used to develop software.

[10] The Lester S. Levy Collection of Sheet Music at the Milton S. Eisenhower Library of The John Hopkins University contains over 29.000 pieces of popular American music spanning the period 1780 to 1960. http://levysheetmusic.mse.jhu.edu

as well as digital audio (e.g. Variations2 Music Library at the University of Indiana[11]) and video (e.g. live performances).

According to the definition provided by the Association of Research Libraries[12], a digital library is an entity that is (1) not stand-alone, (2) technology-driven, (3) linked to other libraries in a transparent fashion, (4) universally accessible, (5) not limited to the digitization of existing print documents and (6) contains content that can only exist in a digital environment. However, it is not evident to simply transpose this definition to "digital libraries with music content". Digital music libraries have a variety of purposes and functions (e.g. preservation, improving accessibility to the material, integrating various formats of music into one collection, providing music education tools) and due to copyright issues it is not at all easy to make them "universally accessible". Besides, most of the so-called digital music libraries are prototypes still under development (e.g. MELDEX, SEMEX).

Digital music databases may contain many kinds of music information and each type has its own storage format. Although the term *music* data is vague, the results of searching are quite dependent on how this music data is stored in the database (e.g. as score notation or as audio stream). In the first case it is spatially stored, in the second case temporally.

To complicate matters further, databases containing musical audio make use of different audio file formats that are *symbolic, sampled* or *compressed*. All of these representations can be approached in terms of the amount of musical structure they retain, and their ability for faithful reproduction.

Symbolic formats, like MIDI, represent music information such as note durations, pitch and intensities. A large amount of musical structure is captured, but the resulting representation is limited. It is for example unable to capture the nuance of a live performance.

Sampled formats*,* such as Pulse Code Manipulation[13] (PCM), represent music by periodically sampling the data, quantizing the sample, and encoding the quantized value digitally. PCM is unable to explicitly represent any musical structure.

Compressed formats, such as MP3 encoding, use psychoacoustic models of human hearing to discard irrelevant or imperceptible data from a PCM bit stream. MP3 produces a perceptually comparable, but distinct bit stream significantly smaller than the "raw" PCM data. Because of the lossy transformation inherent in these encoding schemes, they retain even less of the original structure than the input PCM. Nevertheless, compressed formats are often used because they decrease the amount of disc space required to store the music.

[11] Variations2 provides online access to selected recordings and scores from the Indiana University Cook Music Library for use by IU Bloomington students, faculty and staff. http://variations2.indiana.edu
[12] http://www.arl.org
[13] PCM: raw uncompressed digital audio encoding.

In a nutshell, digital music databases impose new content-based methods that can deal with real musical audio.

1.1.2 Actions

The task of a music information retrieval action (e.g. a vocal query) is to retrieve the desired music or music information. Actions are transformation processes or techniques that support search of and retrieval from a digital database. Transformation processes define similarity measures, and find occurrences and variations of a musical fragment within a collection of music documents. Actions include computational methods for classification, clustering and modelling (e.g. feature extraction for monophonic and polyphonic music, similarity and pattern matching). In most of the MIR systems, the action starts with description of the features used for matching. This process of feature extraction is bottom-up, which means that it takes low-level representations and identifies higher level features. Given the fact that features at one level may build upon features at any other level, the system performance is very much influenced by the adequate choice of feature sets.

1.1.3 Content

In the domain of content-based music information retrieval, the analysis process typically starts from sound as physical energy. The core task of content-based MIR systems is to allow users to search musical pieces using music content as a search key. This content will be based on the user's description of musical experiences. Content-based music analysis thus relates to the transformation of sound energy into semantic variables or abstract notions associated with the piece. Of course, the inevitable use of adjectives and metaphor generates ambiguity of meaning.

In discussions of signification and meaning in music, the non-referential character of music is often stressed (e.g. Cumming, 2000; Shepherd and Wicke, 1997), but, content and meaning in music is much debated and there is no clear definition of what "musical meaning" really is. Music could even be considered as a phenomenon without meaning, or at least with formal musical structures that are void of denotations. Hence, musical meaning arises rather from the associations which the music (i.e. travelling sound waves) evokes in the listener. Therefore, music content by its nature is hard to define and a host to paradox.

The definition of content in general and music content in particular, depends on the approach. Most often, music content has been viewed from two angles: the viewpoint of *physics* (i.e. engineering and computer science) and the viewpoint of human *cognition* (i.e. psychology of music). As a physical phenomenon, music content simply refers to the measurable elements in the sound energy of music, i.e. its formal structure and everything that it contains. From the viewpoint of engineering and computer science, music content consists of a collection of facts or data that are stored and used by a computer program. In this sense, content does not necessarily entail meaning. In fact, information that is heavily

loaded with meaning and information that is pure nonsense, can be exactly equivalent. Contrarily, from the perspective of the psychology of music, meaning is highly relevant. As a cognitive phenomenon, music content relates to the meaning derived from the ideas expressed in the music, the story told or the subject dealt with.

Although this well-known dichotomy has been discussed a lot in the literature on information retrieval in general and music information retrieval in particular (e.g. Martin, 1998), the concept of music content is often used in a vague and confusing way. Whatever is retrieved from an electronic format is always restricted in one way or the other. For example, what musicologists mean by “music” may be the score, and what a listener perceives as the same “music” may be something entirely different, such as an emotion. However, it may be assumed that the musicologist and the average user share some higher-order understanding of features of music.

The nature of these features and how they should be defined and measured is, however, largely unknown. Even when it is possible to represent every sound event and all levels of detail, different approaches share different opinions about what exactly are important features of music. Ideally, music information retrieval systems should be able to interpret content, depending on multiple levels of approach. The problem of content-based music analysis is that linkage between measurable elements and meaning is a difficult task because it highly depends on subjective interpretation.

The intriguing thing is that music as a phenomenon offers a means to study issues that are related to subjective experience, cognitive structures and physiological processes as well. Music is traditionally analysed using a complex theoretical framework, starting from a symbolic representation (e.g. a score). Thus far, most of the research in the area of content-based music analysis has been bottom-up. It has principally been done from a viewpoint of computer science and electronic engineering. In other words, little attention has been paid to perception and cognitive characteristics of music. Since I was involved in the IPEM Toolbox and MAMI research projects I have become more and more aware of the importance of music analysis based on human perception. From this research experience I have learned that music information retrieval system development will benefit a lot from knowledge of how people perceive the music they hear and what meaning(s) they attach to it.

1.1.4 Users

Potential users of music information retrieval systems have different needs and they draw upon various ways of expressing themselves. Consequently, a music information retrieval system has to deal with different groups. User groups such as novices and musicians, or adolescents and middle aged people each understand music in a different way and as a consequence, they will act differently. Upon whose perception then should a system be

based and what about our knowledge of that perception? According to what level of ability should a system be implemented?

Although indexing and retrieval techniques are being investigated all the time, it has not been thoroughly researched how a given group of people prefers to locate music, what strategies they would employ and what expectations they have. It is likely that a music information retrieval system will have to offer multiple user interfaces to be of use to different user groups (e.g. inexperienced, professionals). Offering the opportunity to make combination searches is another way to meet user variation. For example, users could retrieve a musical piece by combining a melody input (e.g. a pre-recorded excerpt or a melody sung in a microphone) and a keyword-based description (e.g. style, mood).

1.1.5 Queries

In general a query is a form of questioning. In music information retrieval a query is a statement of information needs. The concept of a query can be approached from a system-oriented or from a user-oriented viewpoint. In database management systems, for example, system-oriented queries are the primary mechanism for retrieving information from a database. Queries then consist of questions presented to the database in a predefined format (e.g. SQL). The definition of music information retrieval system architecture as presented in this thesis is user-oriented. A query is regarded as the methods by which a user of a search engine or database system enters information.

Depending on their knowledge of the piece they are looking for, there are two ways in which people may search for music information. The first way is the actions they undertake when they know what they are looking for. For example, they know the answer (e.g. title, performer) but want to find more information about the subject (e.g. composer, music). In that case the use of textual input is the common query method. Text-based queries allow specification of titles, performers, composers, etc. The second way involves people not knowing exactly what they are looking for. In that case they often encounter problems identifying a piece of music of which they can only remember a fragment. Alternatively, they might prefer to hum a melody or to use a recorded fragment.

The old way solution is to visit a music shop and ask the person behind the counter. An approach that is becoming more common is to pose the question in a newsgroup on the Internet, including contextual information such as lyrics or where the piece of music was heard. It may even include an audio file containing a recorded or sung portion of the piece. In future applications, human computer interaction will be improved by singing, playing or notating music. Dealing with these multiple inputs poses many challenging problems for both their combination and the low-level transcription needed to transform an acoustic signal into a symbolic representation.

Summarizing, it may be said that databases, actions, content, users and queries form the main constituent components of a music information retrieval system.

1.2 Problems of MIR research

Music information retrieval research still faces many problems. This section considers some of the problems encountered in music information retrieval research that are addressed in this thesis. The focus is on MIR systems and the lack of databases with real music, the need for a conceptual framework, requirements for annotation of music and the shortage of user-oriented studies. During the course of this dissertation work, the scarcity of literature on conceptual structures, annotation methodologies and user behaviour in music information retrieval context was striking. An overview of the literature on these problems is given at the beginning of each of the relevant chapters, dealing with these issues.

1.2.1 Digital music databases and MIR systems

In the context of music information retrieval, computers are used to write, edit, record and perform music. Thanks to CD ripping and other applications, it is easy to create a digital database of one's music collection the size of which would have been unthinkable only a decade ago. In the same time period, computers have become enormously faster, and both disk storage for all that data as well as networks to move it around have become cheaper. More complex algorithms can be run on bigger data sets. In particular, statistical learning approaches to music are ideally suited to this technology. Detailed models can now be trained on massive data sets, using extracted audio signal features. As a consequence, there exist many databases with musical information on the web (e.g. All Music Guide, RISM) and content-based search for sound has been a topic of interest for over a decade. Nonetheless, but few research results have led to applications that are available to the public and most systems are prototypes.

Although the term “digital music library” is being widely used to describe various online venues that provide music, within the music information retrieval community there is no agreement on its definition. As mentioned before[14], “music database” seems to be a more appropriate description in view of a fully integrated approach to music information retrieval. In music information retrieval research, music collections typically consist of sets of short fragments, mostly incipits of (monophonic) melodies which are represented as strings of pitches, pitch intervals, contours, and durations in an electronic score format such as MIDI or Humdrum. Usually the musical audio signal is converted into a symbolic description like MIDI because it is a standardized format that it is flexible to use. However, MIDI files do not fully capture the richness of music content. In order to build efficient user-friendly systems, it is necessary to get a detailed view of musical information contained in real sound examples, both of existing pieces and of imitations of these pieces used in queries for music content. So, to achieve the goals of a fully integrated MIR system, focus on

[14] See 1.1.1: Databases.

databases with real music and on users that interact with real music is needed. Therefore, in the context of the MAMI project the MAMI test collection[15] was created.

Parallel with the approaches to music databases, systems that are designed to offer users the possibility to search by content vary widely. The development of such systems is concerned with techniques for representing music queries, searching collections of music for possible matches and retrieving results that best match the queries. Some systems are being developed for searching musical audio, others for searching notated music. Frequently, they have the disadvantage of being developed without a clear link with the nature and content of the music the systems are supposed to deal with. Systems that focus on musical audio have a rather narrow scope and the range of musical data onto which the tools can be applied is limited as well[16].

There are very few commercial applications of content-based music information retrieval systems (CBMIR) that deal with musical audio[17] (e.g. Shazam) and only a small number of projects are available online for testing. Typke et al. (2005) give an overview of CBMIR systems, both for audio and for symbolic music notation. Typke maintains a list that provides an overview[18] of CBMIR systems including information on query formulation, stored data and indexing techniques. In what follows some examples of CBMIR systems that allow audio input are discussed.

Meldex/Greenstone[19], the New Zealand Digital Library's Web-based retrieval system (Mc Nab et al., 1996; Bainbridge et al., 2004), was among the first WYHIWYG systems to offer an interface online. Meldex allows for searching in a large collection of folk song scores of which a part is accessible by humming-based retrieval. It allows query-by-humming in the form of a URL given by the user. The query is returned to the user in music notation together with a ranked list of musically notated songs similar to the query. In addition text-based search is possible against file names, song titles, track names and lyrics. The system uses string matching techniques based on similarity of melodic contour. A limitation of the meldex system is that it is based on scores rather than on musical audio. As a consequence, the transcription of the queries into the internal representation of the database introduces imperfections in user input.

Another example is Themefinder[20] (Huron, 1999). It provides search access to a database containing tens of thousands of themes. The Themefinder database contains monophonic incipits for about 20.000 themes from classical and folksong repertoires represented in the Humdrum **kern notation (i.e. notes, rests and bar lines as in Western music notation).

15 See 4.4.2: The MAMI database.
16 See proceedings ISMIR 2001-2005.
17 At ISMIR 2003 Avery Wang amazed attendees with a demonstration of the Shazam audio search engine. Shazam offers the automatic identification of recordings via mobile phones using audio fingerprinting. http://www.shazam.com
18 http://mirsystems.info
19 MELDEX stands for melody index. http://www.nzdl.org.musiclib
20 http://www.themefinder.org

Text searches, as well as musical searches are possible. Queries are based on pitch information and can be made in many forms, e.g. as intervals, scale degrees, pitch classes or contour patterns. In addition, the search may be further refined by entering the name of the composer and a time signature. Results are delivered in the MuseData format, which is a file format from which MIDI files can be generated. Although Themefinder offers multiple retrieval techniques to the user it is limited in its application to thematically represented Western music.

Musipedia[21], previously known as Tuneserver (Prechelt and Typke, 2001) is a system that relies on melody recognition. It uses the Parsons code for melodic contour (indications of rise or fall in pitch). The user can directly use this code or may search using recorded queries (.wav files) that contain whistled or sung fragments from a melody. The Parsons code is then calculated, as it is needed for a search. However, in order to use the applet to record a tune some computer preparations are required that might put the inexperienced user off. The advantage of Musipedia is that the music collection that contains melodies and musical themes (pitch contours) can be edited and expanded by the user.

The following list summarizes the characteristics in the majority of the digital music databases that are currently used in music information retrieval system development:

- digital music collections consist of short fragments, mostly only the beginning of a piece;
- the files are for the most part monophonic, i.e. one single voice in the form of a melody fragment;
- annotations are manually added;
- frequently used annotations are contour, interval and duration;
- representation is limited to a string of pitches or pitch intervals;
- search algorithms are based on string editing (i.e. comparison of a symbolic sequence query with a set of symbolic sequences of music);
- a symbolic format such as MIDI and Humdrum[22] is used;
- databases are often limited to one particular genre, such as folk music.

In order to overcome the problems of digital music databases, the participants of the informal session on MIR Tested / Evaluation Issues at ISMIR 2001[23] decided to explore the possibility of creating standardized MIR test collections, retrieval tasks, and evaluation metrics[24]. The purpose was to formulate perspectives on what needs to be done to create “meaningful” MIR test collections, retrieval tasks, and evaluation metrics. Thus far, several projects are attempting to create test beds to suit the needs of music information retrieval research but there remain a lot of questions to be solved concerning file formats, metadata and copyright issues (Reiss and Sandler, 2004).

21 http://musicpedia.org
22 Humdrum is a set of general-purpose software tools intended to assist music researchers.
23 http://ismir2001.indiana.edu
24 http://www.ohsu.edu/jcdl

As far as music information retrieval systems are concerned, there still are many music system development approaches used on different data collections and they all have various shortcomings.

Major inadequacies are that MIR systems:

- are often difficult to use;
- only apply to a specific music genre or type and thus leave out large amounts of music;
- do not apply to real music.

To summarize, musical audio databases and MIR systems still need a lot of improvement in order to provide the user with an easy to use tool that retrieves any music he/she desires.

1.2.2 Conceptual framework

In the information society, there is no longer a lack of music information, on the contrary, the information-based technology might be in danger of drowning in an ocean of unstructured data. How will music information retrieval research avoid going under in data management problems?

In the digital domain, system developers are confronted with content, document and knowledge management and taxonomies are an increasingly popular method of organizing information in this digital environment. There are now many software solutions to assist in automating the taxonomy building process, such as Autonomy, Inxight, Taxomita or Quiver. The "Taxonomy Warehouse"[25] by Synapse, for example, aims to provide a comprehensive directory of taxonomies, thesauri and classification schemes that can be browsed by category. However, browsing this list shows that most existing taxonomies have been developed in the domains of general business, banking and commerce. Until now taxonomies for music description have not been conceived of as operational instruments related to audio. In order to integrate taxonomies in a music information retrieval system based on audio, an intermediate step should be taken that focuses on the establishment of a conceptual framework that supports the implementation of operational taxonomies. In short, music information retrieval researchers must have a framework that defines content classification requirements.

Up to now, the music industry, Internet music retailers, copyright companies and many others have designed music taxonomies for genre classification (Pachet, 2000). Most of these taxonomies are created to meet the particular aims of their creators, or guide their consumers, without having any foundation within a network of musically relevant concepts. As a result of this particularity, these taxonomies show many constraints and inconsistencies. Moreover, it is not evident how these general taxonomies can be related to sound characteristics.

[25] http://www.taxonomywarehouse.com

What is needed instead is an instrumental taxonomy that relates sound energy to musical meaning. Such a tool would not only describe musical features at higher levels of abstraction but also enables users to specify their queries using content descriptions in addition to or apart from musical input. Such a taxonomy is considered a controlled language tool meant to define contents, structure and boundaries as a framework for musical audio mining. It establishes relations between macroscopic characteristics that make up the overall structure of a musical piece or fragment (high-level) and microscopic features that influence the identity of a sound (low-level).

1.2.3 Music annotation

Annotation of music is another problematic area of music information retrieval research. The concept of annotation is well-known in the speech community, where a large set of annotated databases has been constructed that proved to be useful for the development of algorithms for speech recognition. The Linguistic Data Consortium (LCD)[26], for example, provides a list of tools and formats that create and manage linguistic annotations. Linguistic-based tools may be useful for simple annotation tasks, but in general, they are not satisfactory. The main problem is that music content description may require many new formats of description, which often go beyond the commonly used metadata-based indexing. A simple example is the annotation of the beat, which the annotator may want to tap along (and thus annotate) while listening to the played music.

In the music domain, there are but a few initiatives that address this problem of annotation, such as SIMAC[27], MAMI and SEMA[28]. Apart from the few annotation tools, music annotation research today is mainly restricted to the domain of music notation (e.g. MPEG-4, MPEG-7 and MPEG-21). The media industry and researchers involved in content-based music analysis are actively discussing the needs for music representation.

Most annotation and analysis tools (e.g. Timeliner) have paid attention to linguistic and symbolic oriented annotation, but the picture of music annotation is rather dispersed. There is no clear methodology, nor is the problem domain very well described. As a consequence, a well-founded approach to the annotation of music is lacking.

1.2.4 User context

The success of music information retrieval technology primarily depends on both assessing and meeting the needs of its users. Although music has many functions, the social and psychological ones are the most important. Consequently, it can be expected that the most useful retrieval systems will be those that facilitate searching according to these social and psychological functions. Typically, such indexes will focus on stylistics, mood, and similarity

[26] http://www.ldc.upenn.edu/annotation/
[27] http://www.semanticaudio.org/
[28] http://www.ipem.ugent.be

information. User studies will provide opportunities to better understand the salient issues and so better design appropriate access tools.

There is a scarcity of literature reporting on the analysis of the real-world needs and uses of music information retrieval and digital music databases (Downie, 2003; Byrd and Crawford, 2002; Futrelle and Downie, 2002). Besides, experimental research often examines exactly those aspects of music (such as key sense or pitch perception) that the subjects have been taught for.

Although acknowledged in the past it seems that less research is focusing on user implications in general and user behaviour in particular. In his conference report on ISMIR 2004, Droetboom (2004)[29] remarks that few user studies (6%) have been conducted and much of the research being reported was therefore disconnected from user needs. To overcome the risk of developing systems that are not suited to the users, a priori research on user behaviour with regard to music information retrieval is required. Therefore it is essential to define who the potential MIR users are, their needs and the strategies they will most likely use.

1.2.5 Summary of MIR problems

Although a substantial number of research projects have addressed music information retrieval, the user-centred approach is still in its infancy. A summary of findings supports the research objectives of this dissertation.

To my knowledge:

- a conceptual framework for MIR purposes is still inexistent;
- systems for automated MIR are usually developed without a clear link with the nature and content of music;
- MIR research scarcely tackles the problem of databases containing real music;
- there is no metadata (created by humans) accessible for testing and training;
- there is no annotation methodology available to the MIR community;
- methods for evaluation are at a very primitive stage of development;
- few studies have been undertaken focusing on user behaviour;
- research barely involves real users: much experimental research relies on results from experts;
- user specific interests and the search context are usually neglected;
- the comprehension of music at higher levels is far from clear.

Many problems to be faced are due to the *nature of music* itself. Quite some of these have to do with human perception and cognition of music. To improve retrieval tools and thus the overall satisfaction of a user, it is necessary to develop methods that are able to support the user in the search process. For example, by providing additional user adapted information

[29] http://www.dlib.org/dlib/december04/droettboom/12droettboom.html

about the search results as well as the data collection itself and especially by adapting the retrieval tool to the user's needs and interests.

1.3 Research challenges and processes

The goal of this dissertation is to present a model which can serve as a conceptual framework to stimulate successful interdisciplinary activity between cognitive musicology and computer science (i.e. algorithm development), in order to establish fully operational and user-friendly MIR systems. More specifically, the objective is to propose both empirically grounded insight into high-level features of music, based on user behaviour, and to provide a conceptual framework. This model should ideally contribute to building music information retrieval systems that are attractive to the user by supporting various ways of user interaction. To begin with, the challenges of achieving these aims are summed up. From there on the research processes involved are described.

Challenges facing this initiative are:

- constructing a conceptual framework for user-oriented content-based search and retrieval technology for digital music databases;
- defining a method for music analysis that represents the manual annotation of musical features, needed as a reference for developing automatic systems for music information retrieval;
- providing a theoretical framework for the standardized evaluation of results generated by music information retrieval research tools;
- providing knowledge on high-level music content for the development of applications that rely on automatically extracted content from music;
- providing a foundation for new music analysis that goes beyond analysis limited to music represented by a score or in an electronic format.

The requirements for dealing with the challenges summed up above delineate the three research processes that are reflected in the eight chapters of this dissertation: *categorization, annotation* and *contextualization*.

The first process, **categorization**, is theoretical. It deals with the understanding of the musical surface in terms of categories of events. It involves the elaboration of taxonomies in view of a conceptual framework for user-dependent music information retrieval.

The second process, **annotation**, is methodological. It refers to the way in which our understanding of a domain (i.e. human interaction) is structured in terms of another domain (i.e. algorithm development). It involves the manual annotation of musical audio and the creation of annotated musical audio databases.

The third process, **contextualization**, is both theoretical and methodological. It clusters concepts into specified application-oriented relationships. It involves conceptual modelling and definition of user context.

1.3.1 The process of categorization

The model presented in this thesis defines major constituent features of music and relations between them from a perceptual and cognitive viewpoint. The central theoretical question is: “what is the part of both user participation and tools in the scientific results and conclusions?” In other words: “what role can or will the user play in music information retrieval system development?”

The process of analyzing the theoretical question of categorization starts from a three level structure of the music phenomenon:

- **acoustical level:** real sound, observed independently of hearing;
- **perceptual level**: registration of sound and perceptual organization in the consciousness;
- **semantic level**: experience and the meaning of the organized aural sensations.

This three tier approach was chosen in order to determine the role of the music information retrieval technology in the different levels of the realization of a music information retrieval system.

The *acoustical* level of the phenomenon has to do with the technical means for the processing of the audio stream. Within music information retrieval research, a wide range of automated feature extraction tools are being developed and tested (e.g. for pitch and rhythm).

The *perceptual level* of the phenomenon is dealt with by different disciplines including music theory, psychology and sociology. At this level methodologies are applied that analyse the acoustical material through the perceptual understanding of music. Thus, the tools used at this level are mainly based on perceptual categories (e.g. interval structure, duration and time signature, legato/staccato articulation). The most intangible level for automated analysis is the semantic one.

Semantic analysis mainly relies on the unique features of user intuition. The major problem at this level is the quantification of the surveyed parameters.

1.3.2 The process of annotation

Up to now there is a general lack of research material in the form of digital music databases and annotations of music. Access to large music databases requires interoperable data representations and methods for handling new description formats for querying and delivering musical data. Musical audio databases are still in a stage of simple annotation methodology. As an alternative, a general framework is defined for manual annotation of musical audio. This framework has been approached from three viewpoints related to *context dependencies, computer modelling* and *representation*. Manual annotation methodologies toward *user-oriented* and *system-oriented* annotation have been developed and tested. The annotation methodology and its outcome are expected to have a positive

influence on the improvement of information search and retrieval processes and to greatly improve system usability.

1.3.3 The process of contextualization

Today, users of interactive music search and retrieval systems mainly communicate by using controlled vocabulary and keyword searching in English. The use of this common language, for example, has facilitated global communication but suffers from the inability of accounting for subtle nuances in queries provided by a user whose native language is not English. However, there are more important reasons why a user would want other ways to access music databases. A search process can indeed be frustrating when it retrieves records that are too few, too many or have nothing to do with the subject matter in question. The possibility for *natural behaviour* on the part of the user and overall *user-friendliness* of the system are believed to be major contributors towards efficient system use. The most intangible level for music information retrieval is the semantic one. This is the level where the connotations arise, essentially because of social and cultural determinants. At this point knowledge of *users' context, intuition* and *appraisal* of music becomes crucial.

1.4 Research questions

In attempting to come to grips with the research challenges and processes discussed above following research questions (RQ) come to the foreground.

RQ #1: What is a suitable taxonomic scheme?
RQ #2: What is a suitable reference database?
RQ #3: What are appropriate annotation methodologies?
RQ #4: What are suitable ways for users to interact with MIR systems?
RQ #5: Who are the potential users of MIR systems?
RQ #6: What is the effect of users' background on quality description?
RQ #7: What are fitting music qualities?
RQ #8: Are users capable of making reliable judgments?
RQ #9: What holds an appropriate user interface?

1.5 Approach and methodology

This thesis argues for a *user–centred* approach to music information retrieval system development. Attempting to understand how people "can" and "want" to interact with music information retrieval systems, implies the development of a conceptual model for observing the multiple aspects relevant to person-music interactions. A model aims at representing a phenomenon or a set of phenomena as accurately as possible. It requires critical evaluation (theory) and arguments in support of theoretical assumptions (practice).

My research approach is *phenomenological*. During the research process I examine the syntax of the music and the semantic meaning of the experience after which I make a

critical evaluation of the process as a whole. This approach allows an interpretation of the meaning of the experience and the effects of the music. This thesis starts from a user-centred approach to music information retrieval, assuming that music has a content based in musically external factors (such as human emotions or affects) and inherent symbolic qualities within certain musical parameters. Thinking of music as solely being a spatial and temporal configuration of sound elements might lead to the apparent logic but misleading conclusion that the scientific study of music is similar to the scientific study of sound. I adhere to the viewpoint that music does not exist just as a configuration of sound. Music is a type of sound which humans respond to in a certain way. The characterization of certain sounds as musical and giving meaning to them cannot be separated from the response that people have whilst making music or listening to music.

It might seem that carrying out experiments, observations and analyses of human response to music would provide the right answers, but the issue is not that simple. The difficulty with music and the question of how people emotionally respond to it is that the only description of music that average listeners can make is a purely subjective one. That is, music is characterized as music because of the emotional response to it. Classical musicologists might be tempted to work around this difficulty by giving a formal description of music according to the rules of music. These rules would include all the content of music theory, with descriptions of concepts such as scales, timbre, tempo, melody and harmony. All of these concepts are certainly relevant to any scientific attempt to understand what music is, but it is obvious that they alone cannot provide a complete description of what music is or what it means. Today computer science has opened new ways for dealing with music. However, the ideal music information retrieval system will not go beyond the problem of defining meaning in music. Nevertheless, it will have the advantage of being able to fulfil many desires people have toward music.

Every research is based on a set of assumptions regarding the topic under study. The assumptions discussed in this thesis differ from those typically made in music information retrieval. I concentrate on two factors of primary interest that have scarcely been studied: the abilities of *real users* and their perception of musical features when listening to *real music*. The problem is tackled at two levels: one level is *theoretical*, the other *practical*. For the first level the speculative or deductive method is used. For the second level the empirical or inductive method is used. As a consequence the research is subdivided in simultaneous developing research processes, which are categorization, annotation and contextualization. At each stage of the research process new assumptions have been tested.

In order to answer the theoretical questions, requirements for a music information retrieval taxonomy as a conceptual framework are defined, starting from three levels of the music phenomenon: acoustical, perceptual and semantic. These are considered to be the basic conditions for constituting the model. During the course of the investigation specific

taxonomies are developed and a framework for manual annotation of music is defined. This structure has been approached from three viewpoints related to context dependencies, computer modelling and representation.

The empirical research for this study is multi-faceted. The results from a range of research methods have been triangulated to create a more solid and comprehensive representation than could be expected from the utilization of just one method. Among these methods are: unstructured interviews, web-based surveys, construction of test datasets, annotation of music and an application of database management. The practical components of this study focus upon groups of people who are assumed to be potential MIR system users. That is to say, people who are both interested in music and in using network applications.

To summarize, in the table below an overview is given of the three research processes categorization, annotation and contextualization that constitute this dissertation. The global idea is synthesized according to applied framework, theoretical and practical approaches.

RESEARCH PROCESSES	FRAMEWORK	THEORY	PRACTICE
CATEGORIZATION	Taxonomy as conceptual framework	Acoustical Perceptual Semantically	
ANNOTATION	User-oriented & system-oriented framework	Context Modelling Representation	Case studies: - annotation by experts - annotation by users
CONTEXTUALIZATION	Application-oriented framework	Socio-cultural background Musical background Appraisal	Interviews, surveys Annotation of qualities Applications

Table 1: General model of the research processes.

2 Taxonomy as Conceptual Framework

While humans tend to be good at recognition tasks, the automatic recognition of music is a well-known problem which currently can only be solved within a limited context. The major difficulty is that for achieving efficient recognition, interpretation of music content is needed. Automatic interpretation will be possible only of there is comprehensive knowledge available on constituent music features. Therefore, a model is required that carries the representation of this information. In this section (based on original article II) a framework is presented that aims at conceiving the necessary conceptual structures for music description. The preliminary design for this framework was the *MAMI taxonomy*, created for the Musical Audio Mining (MAMI) project in which I was involved[30]. This framework, based on a taxonomy of music terms, was the driving force behind the gradual exploration of the research questions in this thesis. It has been used to determine gaps in the music information retrieval research domain and it has also been the starting point for setting up the experiments discussed in chapters four, five and six. During the course of this investigation the MAMI taxonomy has been refined on basis of experimental studies and developed into a conceptual framework for content-based music information retrieval. The ultimate goal of the conceptual framework is: to provide a *user-dependent structure* for description of music that would considerably facilitate the development of music information retrieval systems. This chapter specifies the aims and requirements for a music information retrieval framework and step by step defines the criteria that set the conceptual framework.

2.1 Background

In the music information retrieval domain, there is no commonly agreed definition of taxonomy for music description. Most existing automatic classification systems relate to genre taxonomies that have been designed from a commercial viewpoint (Aucouturier and Pachet, 2003). Since these taxonomies have been developed to please the consumer they are incoherent and show many limitations.

Pachet and Cazaly (2000) compare the Internet genre taxonomies used by Allmusic[31] (531 genres), Amazon[32] (719 genres) and Mp3[33] (430 genres). They found that there is no consensus in the naming of genres used in these classifications: only 70 words are common to the three taxonomies. More importantly, there is no shared structure and even largely used terms like “Rock” or “Pop” do not have common definitions, i.e. they do not denote the same set of songs.

Casey (2002) developed a system for generalized sound classification and similarity that has been incorporated into the MPEG-7 standard[34]. The method that Casey developed involves training statistical models to learn to recognize the classes of sound defined in a

[30] See the Introduction of this thesis and http://www.ipem.be/MAMI
[31] http://www.Allmusic.com
[32] http://www.Amazon.com
[33] http://www.mp3.com
[34] In 2001 MPEG-7 became ISO/IEC 15398 standard for multimedia content description tools for applications ranging from content management, organization, navigation and automated processing. http://www.chiarliglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

taxonomy. A controlled-term structure of the musical instruments hierarchy is presented. It aims at providing semantic relationships between instrument categories. MPEG-7 allows flexibility in defining taxonomies using controlled terms. However, MPEG-7 is insufficient as a standard for audio-based musical content. Kim and Sikora (2004), for example, compared MPEG-7 Audio Spectrum Projection (ASP) features for sound (e.g. speaker, bird, horn, violin, telephone) recognition with Mel-frequency cepstral coefficients[35] (MFCCs). For classification they use three different audio taxonomy methods. Their results show that MFCC features yield better performance than MPEG-7 ASP and that the latter is also more time and memory consuming.

To summarize, in the music information retrieval domain there is a lack of taxonomy structures and taxonomy applications are scarce as well. Most taxonomies have not been conceived as structures related to musical audio, and when so they are insufficient for content-based music analysis. As a consequence it is hard to implement a taxonomy-driven music information retrieval system.

2.2 A music information retrieval framework

It is almost impossible to begin any music search activity without dealing with categories such as “type”, “genre” and “style” or affective terms like ”mellow” or “hard”. Communication about music implies reference to a more or less detailed structure. Moreover, understanding music assumes being able to think in terms of categories or events. What I propose as an alternative to the lack of taxonomy structures is a conceptual framework which aims at:

- supporting the structuring, classification, modelling, and representation of the concepts and relationships pertaining to subject matter of interest to the music information retrieval community;
- enabling the music information retrieval community to employ the same terms in the same way;
- comprising sets of terms that the music information retrieval community could agree to use to refer to concepts and relationships;
- specifying the meaning of the terms.

The conceptual framework thus comprises different systematic approaches to description. In particular, it contains issues related to categorization, classification, controlled vocabularies, ontologies, taxonomies and thesauri. All these terms refer to ways of organizing information into categories. In the literature they are often confused and used in many different ways by different communities.

[35] Mel-frequency cepstral coefficients are coefficients that represent sound. They are most commonly used as the standard analysis techniques for automatic speech recognition.

2.2.1 Definition of terms

In what follows a synthesizing description of terms used for structuring information is given in alphabetic order.

Categorization is the process of associating a document with one or more subject categories. Categorization is generally not rigorous. It is rather based on similarity than on systematic arrangement.

Classification is mainly used in libraries, where specialists enter metadata (e.g. composer, date, title) for a document, apply subject categories to it, and place it into a class for later retrieval. Classification schemes generally focus on providing a single cataloguing.

A **controlled vocabulary** is a set of terms that people agree to use and the meaning of which is unequivocally understood. The vocabulary has detailed definitions for each term.

Ontology is a philosophical term for the study of the categories of things within a certain domain. It provides a logical framework for academic research on knowledge representation.

Taxonomy is the organization of a particular set of information for a particular purpose. The term comes from biology, where it is used to define the single location for a species within a complex hierarchical scheme.

A **thesaurus** is a non hierarchical set of related terms describing a set of documents. Thesauri include synonyms and more complex relationships, such as broader or narrower terms, related terms and other forms of words.

To summarize, the differentiation between terms used for defining the structuring of information is not fundamental and up to interpretation. The major difference is the way that meaning is specified. A broad definition has been given in order to clarify the context in which the terms are used in this thesis.

2.2.2 Framework components

The conceptual framework for music information retrieval that is envisaged aims at defining concepts and relationships between concepts that will be used for content-based description of music. Figure 2 shows the general scheme of the structural components for the development of the model for content-based music information retrieval. The MAMI taxonomy structure is the starting point to the conceptual framework. It comprises concepts, classes and relationships between concepts and classes. The elaborated conceptual framework is the intermediate structure between multiple taxonomies and the model that envisages taxonomy-driven music information retrieval system. This model is built on both fundamentally rigid structures (e.g. concepts) and unique relationships (e.g. individual associations).

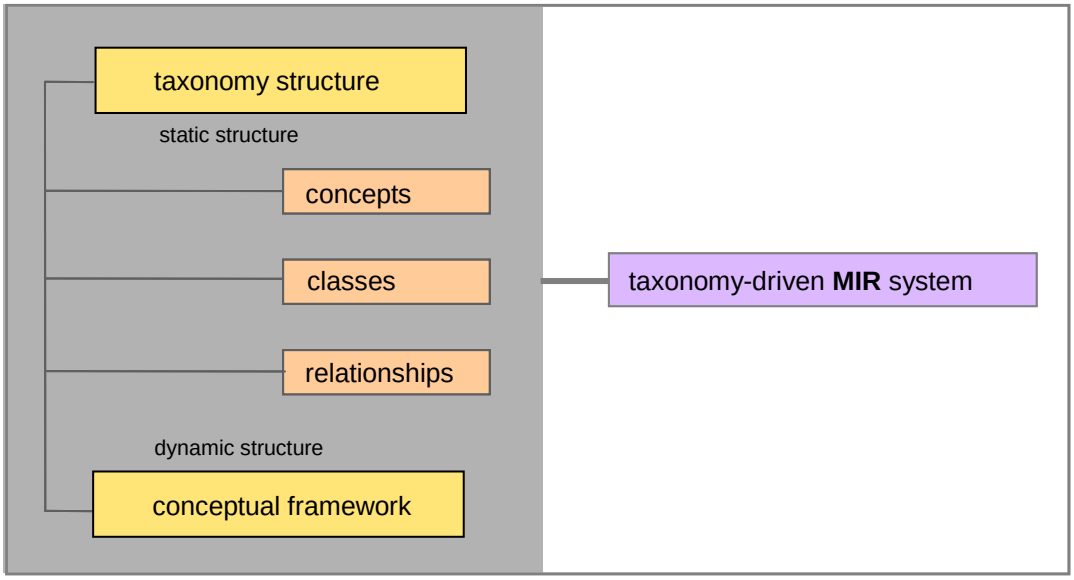


Figure 2: Components of a Music Information Retrieval (MIR) framework.

In the global domain of information retrieval, taxonomies are supposed to help users to quickly and easily find the information they desire.

A distinction has been made between taxonomy as a structure (i.e. a conceptual vocabulary) and taxonomy as an application that drives a system. Historically, taxonomy structures have been used by biologists to classify plants or animals according to a set of natural relationships[36]. In content management and information architecture, taxonomy applications are developed for the automatic organization of information. Taxonomy structures may be user-oriented or system-oriented, whereas taxonomy applications are only "visible" to computer programs. Currently, there are a number of applications that can help create taxonomies, although the amount of automation can vary from simply adding an item up to automating the entire classification process.

Categorizing is a thing that humans do and organizing structures for music have been used for ages. However, while the act of grouping (e.g. genres, instruments) is meant to order things, applied to technology, categorizing often creates more confusion than order.

Until now, in music information retrieval, the emphasis on content classification has always been more on feature extraction (bottom-up) rather than on user needs (top-down). In this dissertation, the taxonomy specification is approached in view of accommodating user-friendly music information retrieval applications. In other words, it should strive to contribute to accommodate the needs of multiple user groups with different search behaviour.

[36] Taxonomy has only recently emerged from the domains of biology, book indexing and library science into information search and retrieval. Taxonomies that describe classification structures are now widely used in various research areas such as natural language processing, information retrieval, database design, artificial intelligence, the semantic web, web services, software system design, and cognitive sciences. Even though taxonomy has become a commonly used term, there are a lot of misconceptions and taxonomical requirements are frequently underestimated.

In the next section, the goal of the MAMI taxonomy, which was the earliest stage in the process of developing the conceptual framework, is explained.

2.3 The MAMI taxonomy

The musical audio mining project (MAMI) concentrated on technology that allows users to search and retrieve music by means of content-based text and audio queries as well. Though the idea of such system is promising, there is a need for a better understanding of both the concepts involved and of the role of user preferences. The development of taxonomies for different aspects of music information retrieval systems, aims at bridging the gap between system development and user interactive interfacing. There exists no elaborated conceptual definition that could function within this area.

The purpose of designing a musical audio mining taxonomy is twofold. Firstly, the MAMI-taxonomy is meant to present a structure, according to which theoretical discussions can be focused, developments evaluated, research conducted, and data meaningfully compared. Secondly, it is a methodology in progress that should support the implementation of an entirely automated system for content-based music information retrieval.

The MAMI taxonomy meets an important demand from many researchers working in the field of music information retrieval. Indeed, the taxonomy names and organizes music concepts into categories that share similar characteristics. The object in view is a dynamic structure of music-related concepts and relationships between these concepts.

The intended model, based on the MAMI taxonomy, will differ from existing approaches because it:

- supports structure, content and applications;
- is customized to reflect goals of content-based MIR research;
- relies on a combination of human endeavour and algorithms;
- applies to multiple content representations (e.g. textual, musical audio).

As this study envisages a fully integrated approach to music information retrieval, the taxonomic structure that is presented aims at describing music that:

- is representative for the cultural diversity of the actual music production;
- allows users to communicate with a system in a natural way.

Although there are many software tools[37] that can assist in developing bottom-up taxonomy they need to be managed by human experts who understand the edges of the taxonomy envelope. The absence of a widely accepted conceptual framework for music information retrieval is one of the most important weaknesses in the research domain. It does not only make it difficult to state hypotheses and predictions clearly, it also makes it hard to select

[37] http://www.searchtools.com/info/classifiers-tools.html

appropriate concepts when designing experiments. The problem concerns the multiplicity of categories used to describe music and the lack of methodical investigation that clarifies the concepts.

The basic idea of the MAMI taxonomy is that the issues that music information retrieval deals with can be arranged in a hierarchy from less to more complex. Hence, music is characterized by multiple levels of content descriptors, from low-level acoustical features up to high-level structural descriptions and qualitative specifications. A scan of the literature shows that many researchers agree that music allows description at different representation levels (e.g. Zannos, 1999; Gødoy, 2001; Leman, 2002; Tanzi, 2003; Vinet, 2003). What taxonomies will do is providing descriptor labels for these levels. While most research regards the lower and highest levels of description, more elaboration is required that focuses on integrating between the levels. Such a linking taxonomy requires findings from empirical investigation with listeners/users.

To sum up, a taxonomy is needed for multiple reasons:

- it helps to delineate the conceptual relationships that exist within and between various levels in the multitude of music descriptions;
- it provides a common language for system developers who tend to come from different disciplines;
- it provides a common language to users who interact with these systems;
- it reduces complexity.

2.4 A multi-levelled framework

A content-based music information retrieval taxonomy will be multi-dimensional, characterized by a multi-levelled conceptual framework that contains sets of more or less domain-dependent concepts. The levels under consideration are based on a model presented in the IPEM Toolbox (Leman et al., 2001). The representational structure developed for the IPEM Toolbox focuses on perception-based music analysis. This framework represents a bottom-up approach to music analysis taking human perception as the basis for feature extraction and higher level conceptualization. The model distinguishes between three description levels, namely low (involving acoustical and sensorial properties of content), mid (associated with perceptual objects), and high (associated with cognition). Constituent parts of the conceptual framework that I present in view of music information retrieval are *description levels, concept categories* and *features*.

A **description level** defines the degree in which content is related to the musical audio signal. The three level approaches (i.e. low, mid, high) from the IPEM Toolbox have been adapted and refined with more detailed characterization of the respective levels.

A **concept category** is a class of items with common attributes. Useful categories, for instance, could include the organization of genres (e.g. pop, classical), styles (e.g. baroque,

romanticism), structural entities (e.g. part of a symphony, melody phrase) and sound sources (e.g. piano, voice).

A **feature** is defined as a property extracted from the musical signal. The most commonly used features are pitch, melodic profile[38], note duration ratio sequence[39], and interval sequence[40]. Recently, research that involves real music has also focused on descriptors related to timbre (GOASEMA project[41]) and rhythm (MAMI project). The selection of relevant features is a decisive factor for the performance of a system that handles real music. In what follows a more detailed definition of the distinct description levels, concept categories and features is given.

2.4.1 Description levels

A representation of the three description levels of the MAMI taxonomy is presented in figure 3. Definition of the descriptors and content according to these levels is given below.

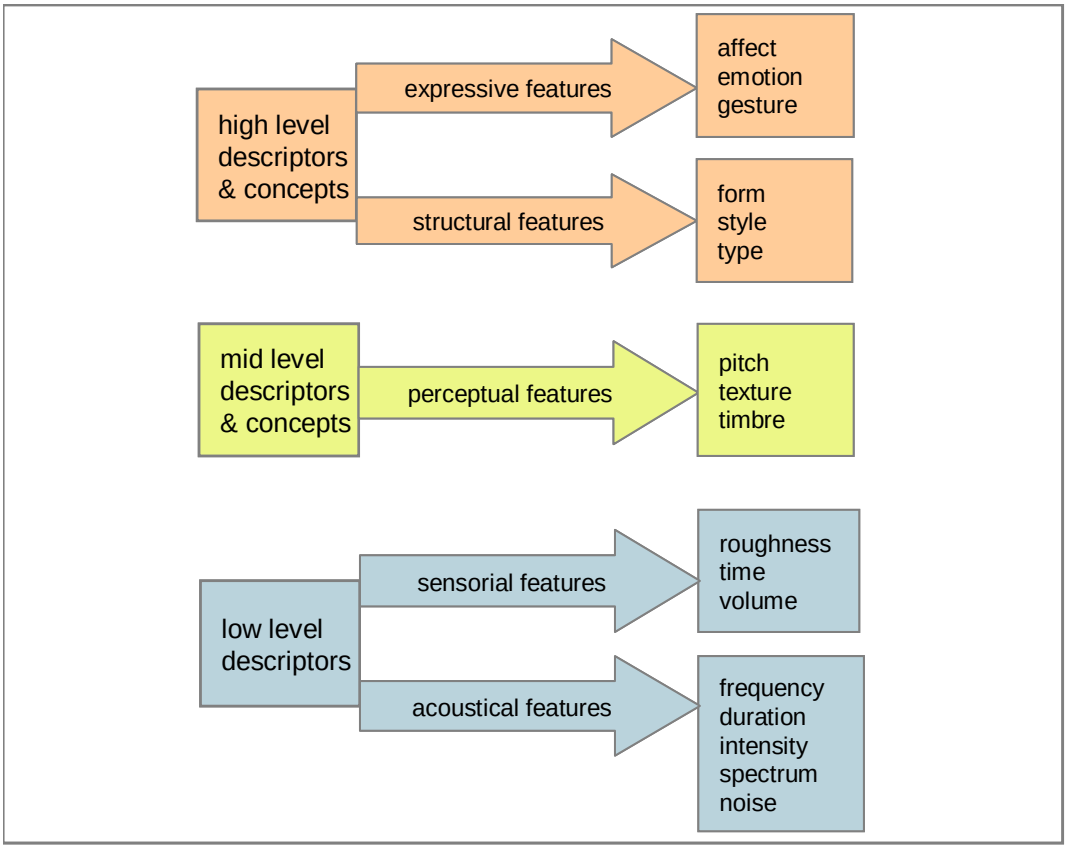


Figure 3: Description levels of he conceptual framework.

For defining entities at low, mid and high description levels a distinction is made between descriptors and concepts. In the context of music information retrieval and handling of

[38] In most of the cases the simple Parsons code is used indicating higher, lower or similar notes (i.e. UDR).
[39] That is the ratio of sounding durations between two consecutive notes (e.g. 1/1 3/1 1/6).
[40] That are numbers denoting interval size (e.g. 0 +2 -3).
[41] http://www.ipem.ugent.be/2004GOASEMA

digital databases, a descriptor is defined as a vocabulary term assigned to a musical feature. Descriptors are unique identifiers that specify quantified music content. A concept, on the other hand, is an abstract notion that exists in the human mind and that attempts to capture the essence of content. Concepts are independent of the terms used to label them but may vary from one culture to another.

2.4.1.1 Low-level descriptors

Low-level descriptors typically define content that is close to the acoustical or sensorial properties of the audio signal. These descriptors are typically related to temporal features of the signal and local (non-contextual) properties. Low-level descriptors are obtained from frame-based analysis of the acoustical wave. Distinction is made between features related to *acoustical* and features associated with *sensorial* properties.

Features related to acoustical properties:

- are *physical descriptions* of the audio stream in terms of frequency, duration, intensity and spectrum;
- have *unambiguous* values.

Features associated with sensorial properties:

- involve *low-level processing* which is situated in the periphery of the human auditory system;
- relate to *perceptual characterizations* such as roughness, onset time and loudness;
- relate to *echoic memory* processes;
- allow *symbolic description* of sound in terms of low-dimensional parameters;
- have *equivocal* values.

2.4.1.2 Mid-level descriptors and concepts

Mid-level descriptors involve time-space transformations and context dependencies within specified time-scales. Time-space transformations allow for the specification of the musical signal content in spatial terms (e.g. timbre, pitch and chords) and temporal terms (e.g. beat, meter and rhythmic pattern). Content formation at this level has to take into account a time frame of maximum 3 seconds (Leman et al., 2004) which is needed to form representations of the musical present or the “now”. Mid-level concepts relate to *perceptual* properties.

Features associated with perceptual properties:

- involve listeners *experience*;
- involve *cultural determination*;
- processing is assumed to be situated in the periphery of the auditory system;
- relate to *perceptual characterizations* such as texture, timbre and pitch;
- have ambiguous values;

- address *keywords* that identify content in terms of objects that have a defined beginning and ending (segments);
- depend on *context*.

2.4.1.3 High-level descriptors and concepts

High-level descriptors typically involve learning and categorization beyond the representation of the "now". These descriptors need to deal with structure and interpretation. They may be related to *cognitive* as well as to *emotional* and *affective* issues concerning content. High-level concepts are related to expressiveness. To the present day, however, knowledge of how high-level descriptors may be associated with high-level concepts (e.g. perceived affect qualities) is still insufficient.

The features at the high description level are:

- determined by the *cultural* context;
- related to *long-term memory* processes that keep track (among other things) of statistical properties in the signals through learning, attention and context-dependencies.

The high-level concepts describe semantic features. Indeed, they convey a defined meaning that is not necessarily directly associated with the signal properties, but possibly with properties of subjective feelings and interpretations. To clarify this point, a further distinction is made between *primary* and *secondary* semantic features.

Primary semantic features are related to high-level context, and are characterized by:

- *expressive concepts* that describe relations between perceived features in terms of emotion and affect (e.g. romantic, solemn, passionate);
- *metaphoric concepts*: i.e. from the domain of kinaesthetic (e.g. body movement, such as "pressing", "flickering" and "floating" (Laban, 1963)) and from the domain of synaesthetics (e.g. visual, olfactory and tactile concepts).

Secondary semantic features are related to high-level form and structure, and are characterized by:

- *grouping* of perceived features on the basis of segmentation and similarity;
- hierarchical relations between perceived features.

2.4.2 Concept categories and features

Across the description levels of the conceptual framework it is possible to define concept categories or classes and features to which descriptors apply. It is for example possible to consider *classes of spatial-temporal structure*, or *genres and styles*[42]. A model of

[42] In music management systems the term genre is often used interchangeably with style (e.g. AMG Allmusic, http://www.allmusic.com). In this thesis, genre and style are not treated as synonyms. Musical genres are categories which might share a certain style or music language and which have certain elements in common. Cope (1991) defines musical style as "the identifiable characteristics of a composer's music which are recognizable from one work to another".

descriptors that define the spatial-temporal structure class has been worked out. It is the major category to define structuring and grouping of sound. The descriptors that identify spatial-temporal structures describe content in terms of space and time features. These descriptors are important in view of the segmentation of a musical piece into fragments and in view of the perceptual grouping of features. Segmentation refers to the dividing of musical audio into time zones. Structuring then means organization in the spatial and temporal domain. Music is essentially related to the dimensions of space and time in that all music features are structured in spatial and temporal hierarchies.

However, one should keep in mind that space and time are interrelated. Space is defined as the *vertical or synchronic domain* where juxtaposition[43] of objects (e.g. pitches) is organized, and time is the *horizontal or diachronic domain* where sequences of objects (e.g. rhythmic figures) are organized. Besides, all characteristics of sound vary in space and time, and hierarchies of groupings over space and time can be formed. Therefore, a subdivision is made between three concept classes of descriptors which apply to spatial features, temporal features and spatial-temporal features.

2.4.2.1 Spatial features

The subclass of concepts and descriptors which apply to spatial features consists of descriptors that specify features in the frequency domain. They span over the three description levels (low, mid, high) and are defined to have a restricted extension in time (<3 seconds).

Possible concepts and descriptors for spatial features are:

- **low-level descriptors** that characterize for example the presence of pitch, noise or silence, spectral density, centroid and texture related properties such as roughness;
- **mid-level descriptors** that allow for the characterization of musical objects in the vertical domain. Those objects, typically, have a defined beginning and ending and are characterized by a set of attributes. An object can be a partial, a pitch (i.e. a fused collection of partials), a chord (i.e. a juxtaposition of pitches), or a tone centre (i.e. a juxtaposition of chords over the three seconds time frame). Of course, different types of pitches, chords and tone centres can be distinguished;
- **high-level concepts** that involve semantic descriptions such as “harmonic”, “dissonant”, “pleasant”, “sensual” to any of the low and mid-level descriptions.

2.4.2.2 Temporal features

The subclass of concepts and descriptors which apply to temporal features consists of concepts that specify features in the temporal domain. They span over the three description

[43] Juxtaposition refers to the fact that objects are placed together, so that the perception of differences between them is emphasized.

levels as well (low, mid, high), but are defined to operate within the three seconds time window.

Possible concepts and descriptors for temporal features are:

- **low-level descriptors** that characterize for example onset, offset, attack characteristics and transient/stable behaviour;
- **mid-level descriptors** that allow for the characterization of beat, tempo, meter and short rhythmical entities.
- **high-level concepts** that involve semantic descriptors such as “punctuated rhythm”, “fast, slow tempo”, “allegro”, “vivace”, “largo”, “energetic rhythm” and “swing”.

2.4.2.3 Spatial-temporal features

The subclass of concepts and descriptors, which apply to spatial-temporal features, consists of descriptors that specify features in the spatial temporal space beyond the three seconds time window, that is, beyond the phenomenological “now”. Considered concepts and descriptors are related to monophonic or polyphonic features.

Concepts and descriptors related to monophonic features are:

- **low-level** descriptors that characterize for example single voicing such as one channel or one track, a pure tone or a frequency content having harmonic – enharmonic structures;
- **mid-level descriptors** that allow for the characterization of motifs or figures defined by pitch configurations such as interval and contour information (“gap-filling”, “centripetal”, “centrifugal”);
- **high-level concepts** which involve semantic descriptions that allow for the characterization of figures as being “pregnant”, “sad” (e.g. lamento), as evoking objects, events or processes in the real world (e.g. bird, train, person), as possessing style properties (e.g. belcanto, recitative).

Concepts and descriptors related to polyphonic features are:

- **low-level** descriptors that characterize for example multiple voicing such as multi-channel sounds or multi-track;
- **mid-level descriptors** that allow for the characterization of rhythmic properties such as homorhythmic and polyrhythmic;
- **high-level concepts** which associate semantic descriptors that allow for example style characterizations related to genre specifications (e.g. jazz, bossanova or beat) and melody/accompany relationships (e.g. basso continuo, ostinato).

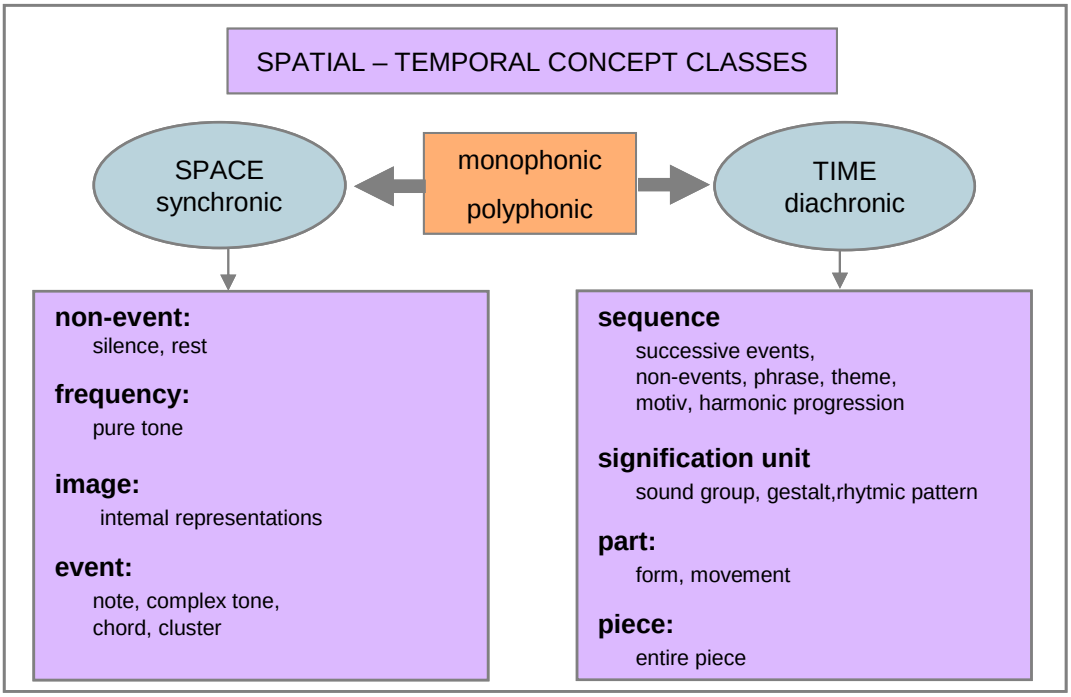


Figure 4 : Concept classes specifying spatial-temporal structure features.

Figure 4 is a schematic representation of subclasses of structural concepts and descriptors. Their relationship to low, mid and high description levels is described above. Here a distinction is made between *synchronic concept classes* and *diachronic concept classes*, ranked according to the increasing complexity of information in the synchronic and diachronic domain. Synchronic concept classes refer to *spatial sound grouping* concepts and descriptors; diachronic concept classes apply to *temporal sound grouping* concepts and descriptors.

Synchronic concepts or spatial sound grouping concepts require a description of the organization of features in the frequency domain.

Following sub-classes are distinguished:

- non-event: silence or a rest;
- frequency: pure tones;
- images: reflect features of the sound as internal representations of auditory information;
- events: a note, complex tone, chord, cluster, coloured noise and white noise.

Diachronic concepts or temporal sound grouping concepts require a description of features in the temporal domain that is the trajectory (e.g. dynamics, articulation) of individual sounds through time/frequency space.

Following sub-classes are distinguished:

- sequence: a succession of events and non events, melody, phrase, theme, motif, harmonic progression etc.;
- signification unit: a sound group, Gestalt or rhythmic pattern;

- part of a composition: a form, a movement of a symphony etc.;
- piece of music: entire composition.

To summarize, in order to achieve the goals of a fully operational music information retrieval system, music content needs to be categorized and labelled. What used to be a normal task to librarians, has only recently been introduced in the domain of information management and retrieval. Although there are numerous ways to approach the design of a taxonomy some points are for certain: the music information retrieval community needs well-founded information structures, must use standardized terminology and should be aware of the importance of user participation and feedback.

2.5 Constituent music categories

The structure of the multi-levelled framework discussed in previous section distinguishes between two types of descriptors of musically relevant auditory phenomena: *local* and *global* descriptors. This section discusses constituent music categories that are defined by global descriptors. These categories are considered the most important music components in view of linking between low-level features of music (i.e. local descriptors) and the query input (i.e. auditory or other) provided by users of music information retrieval systems. The model that includes six categories is built on the empirical research that will be presented in the next chapters of this thesis. In what follows the local and global descriptor types and the constituent music categories are described and situated within the multi-levelled conceptual framework.

2.5.1 Descriptor types

This part clarifies the differences between local or non-contextual, and global or contextual descriptors. The distinction involving local and global descriptors is based on the internal representational framework of the IPEM Toolbox that reckons with the size of the time frame that content formation has to take into account. Local descriptors are derived from music content within time scales shorter than three seconds, whereas global descriptors are derived from musical context dependencies within time scales of about and beyond three seconds (Leman et al., 2001).

Local or non-contextual descriptors are situated at the lower levels of the conceptual framework. At the acoustical and sensorial levels the description of music content is approached from the viewpoint of physics, automatic (auditory-based) feature extraction and low-level processing. Acoustical phenomena, such as frequency and intensity, are regarded as objective terms. Sensorial features involve characterizations such as onset time and loudness and are somewhat equivocal. Local descriptors characterize quantifiable properties of sound rather than the internal experience humans have when listening to these sounds.

In this framework, there is no clear boundary between local and global descriptors. The threshold is defined by the periphery of places in space or time where quantifiable phenomena flow over into subjective phenomena.

Global or contextual descriptors are situated at the mid and high-levels of the framework that comprise perceptual, structural and expressive features of music. These features refer to the internal subjective experience of music. From a subjective viewpoint a distinction can be made between not value laden phenomena (e.g. pitch as the subjective sense of highness or lowness) and value laden phenomena (e.g. cultural values, metaphors, adjectives).

From the perceptual perspective, auditory phenomena such as pitch and timbre refer to subjective experiences but are not value laden. Timbre, for example, is the sense of tone-colour by which one sound may be distinguished from another sound of the same loudness and pitch. It is subjective but not highly linked to pleasure or delight or fear.
From a psychological perspective, value laden statements refer to the opinions, beliefs and feelings of an individual. Such phenomena refer exclusively to the induced subjective experience of stimuli and may be expressed by adjectives such as “beautiful” and “tender”. These are the kind of terms that a not musically experienced user of music information retrieval systems most likely would rely on.

2.5.2 Taxonomy of music categories

Very often, a definition of music lists the elements that make up music under that definition. There are many definitions of music and they thus account for multiple lists with different elements. In this thesis, it is not aimed at going into the discussion about the definition of music, or at providing a detailed understanding of every element of music. Attention is drawn on music categories that are essential in view of the music information retrieval framework that is envisaged.

Although there is a lack of consensus, in Western music[44] theory, the major constituent elements of music which are often distinguished are melody, harmony, rhythm and timbre. The cognitive psychologist Gardner (1993), for example, denotes that the main elements of music are pitch, rhythm, and timbre. Pitch conceived as a unit in time and space, however, is constituent for melody and harmony. Pitches are organized both horizontally (relations over time or “melody”) and vertically (relations over space or “harmony”) according to a prescribed pattern. All these music categories are structural components of a piece of music. They are situated at the perceptual (mid) and structural (high) levels of the conceptual framework.

The empirical research that has been conducted in my study involves global descriptors of music content which means that perceptual, structural and expressive features are included

[44] Western music is considered as music of European and European influenced cultures

as well. As a consequence, the model that is build on this research considers melody, harmony, rhythm, sound source (e.g. instruments, voice), dynamics and expressivity as major candidates for establishing a treatable compromise between the different levels of the conceptual framework. Apart from the four basic categories (i.e. melody, harmony, rhythm and timbre) often used in the traditional approach to Western music theory, I consider dynamics and expressivity as essential conditions for music information retrieval system development. These two categories are indispensable because they relate to concept classes and features *across* the five distinct description levels (i.e. acoustical, sensorial, perceptual, structural and expressive) of the elaborated conceptual framework.

In what follows I propose a taxonomy in which constituent music categories include six elementary classes: melody, harmony, rhythm, timbre (i.e. sound source) dynamics and expression. Figure 5 situates the six constituent music categories for music information retrieval within the multi-levelled conceptual framework structure as it was presented in figure 3. As shown in this figure, the concept classes relating to audio-structural properties are distinguished according to acoustical, sensorial, perceptual and structural descriptors. In music information retrieval, however, many researchers hang on to the bottom-up approach and study feature selection in view of the automatic extraction of the low and mid-level audio-structural features. In Pickens (2001), for example, an overview is given of feature selection techniques for music information retrieval.

<table>
<tr><th colspan="2">STRUCTURE</th><th colspan="2">CONCEPT LEVEL</th><th colspan="5">MUSICAL CONTENT CATEGORIES AND FEATURES</th></tr>
<tr><td rowspan="6">CONTEXTUAL</td><td rowspan="6">GLOBAL DESCRIPTORS</td><td rowspan="2">HIGH II</td><td rowspan="2">EXPRESSIVE</td><td colspan="5">expression</td></tr>
<tr><td colspan="5">affect
experience</td></tr>
<tr><td rowspan="2">HIGH I</td><td rowspan="2">STRUCTURAL</td><td>melody</td><td>harmony</td><td>rhythm</td><td>source</td><td>dynamics</td></tr>
<tr><td>key
profile</td><td>tonality
cadence</td><td>patterns
tempo</td><td>instrument
voice</td><td>trajectory
articulation</td></tr>
<tr><td rowspan="2">MID</td><td rowspan="2">PERCEPTUAL</td><td colspan="5"></td></tr>
<tr><td>successive
intervallic
pattern</td><td>simultane
intervallic
pattern</td><td>beat
i o i</td><td>spectral
envelope</td><td>dynamic
range
sound level</td></tr>
<tr><td rowspan="4">NON-CONTEXTUAL</td><td rowspan="4">LOCAL DESCRIPTORS</td><td rowspan="2">LOW II</td><td rowspan="2">SENSORIAL</td><td colspan="2">pitch</td><td>time</td><td>timbre</td><td>loudness</td></tr>
<tr><td colspan="2" rowspan="2">periodicity pich
pitch deviations
fundamental frequency</td><td rowspan="2">note-
duration
onset
offset</td><td rowspan="2">roughness
spectral flux
spectral-
centroid</td><td rowspan="2">peak
neural-
energy</td></tr>
<tr><td rowspan="2">LOW I</td><td rowspan="2">ACOUSTICAL</td></tr>
<tr><td colspan="2">frequency</td><td>duration</td><td>spectrum</td><td>intensity</td></tr>
</table>

Figure 5: Conceptual framework for music information retrieval.

Melody, harmony and rhythm attach to the concept class describing spatial-temporal structural properties. Melody, harmony and rhythm define music content both in terms of spatial and temporal structural properties[45].

Timbre is part of the concept class that describes sound sources (e.g. instrumentation) or sonic material. In view of instrument and sound recognition taxonomy examples have been presented (e.g. Martin and Kim, 1998; Casey, 2002).

The dynamics category relates to loudness features and to the many aspects of movement and musical gesture. It is an important subclass to the concept that describes the many aspects of movement in music and musical gesture. Dynamics levels and accents are natural indicators for emotion and affect as well (Kamenetsky et al., 1997). In their survey on music information retrieval systems Typke et al. (2005) refer to the importance of dynamic behaviour of sound for extracting perceptual relevant features of music. Characteristics of dynamics might be helpful to calculate time differences for loudness, pitch and tone (i.e. brightness).

Expression, on the other hand is a high-level inter-subjective category (the object is music, the interpretation is emotion) which cannot be directly extracted from the music, but which could be modelled on the relationships between extracted structural features and perceived qualities.

Figure 6 represents the six constituent music categories according to expressive qualities and spatial-temporal structure.

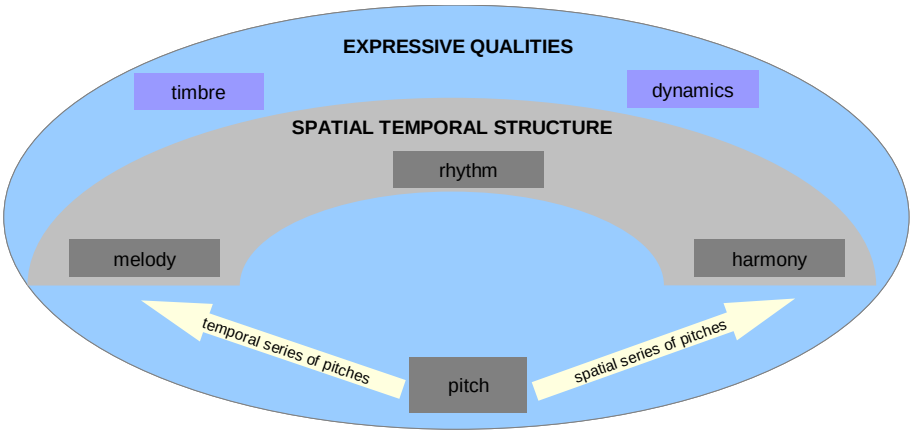


Figure 6: Constituent music categories of the expressive and structural layers.

In what follows the structural and qualitative properties of the six constituent music categories that define the global descriptors is given. The aim is not to give a clear definition of each class but to give a global characterization. In chapter seven, an

[45] See 2.4.2 Spatial-temporal features.

application[46] is presented in which this taxonomy of music categories is used as the basic interface to the user of a music information retrieval system.

Melody may be defined as a temporal series of individual pitches occurring successively in such a way that the order of pitches constitutes a recognizable entity. Melody characteristics include pitch range (distance between lowest and highest pitch), shape (melodic contour), and movement. Specific pitch patterns constitute a scale. Most music cultures have scales that vary remarkably in complexity. Music that has a strong melodic component takes on a linear character. The distance between any two different pitches is known as an interval. Melodic pitches rise and fall (melodic curve) and are perceived as high and low. Upward-moving melodies build tension and create the effect of goal-oriented forward motion. Downward-moving melodic lines dissipate tension. The melody range may be defined as the distance between the lowest and highest pitches of a melody. If there are a large number of notes between the lowest and highest pitches, the melody is said to have a wide range. If only a few notes separate the lowest and highest pitches, the melody is said to have a narrow range. Melodies with wide ranges have an animated character and contribute to musical motion while melodies with a narrow range are less animated.

Harmony is an important organizing characteristic of most Western music. Characterizing features are chords (pitches in a vertical pattern), scales and keys (a hierarchy of pitches), harmonic progression, modulation and consonance or dissonance. Harmony basically addresses a succession of chords. A chord is by definition three or more notes that consist of intervals of a third and that are played simultaneously. A triad is the most common chord form. It is built on the first, third, and fifth notes of the scale. Triad chords may be built on all seven notes of the scale (with the eighth note a repeat of the first). The organization around tones is known as tonality. Tonality provides the basis for the establishment of keys. Pieces built on the tones of the major scale are referred to as major and those built on a minor scale are said to be in a minor key. Harmony contributes to the effect of motion in music. Since harmony is derived from scale pitches, harmony automatically assumes gravitational attraction. Harmonic progressions are designed with this gravitational effect in mind. Consonance appeals to a state of stability or rest. Consonant are all perfect intervals, as well as all major and minor intervals that do not contain adjacent pitch classes. Dissonance creates a state of instability or motion. All sevenths, diminished and augmented intervals are considered dissonant.

By its simple definition, **rhythm** is the arrangement of durational patterns or varying tone lengths. Rhythm is mainly characterized by beat, or pulse (the most basic unit of musical time), tempo (the rate of speed at which music moves) and meter (the organization of beats into groups). The regular pulsations of the music are called the beat. The beat is what you clap your hands to or the rate at which people tap their feet to the music. Stronger beats are referred to as accented. Measures of music divide a piece into time segments. Time

[46] See 7.2: Application 1: User interface taxonomy.

patterns in music are referred to in terms of meter. Two beats to a measure is duple meter, while three beat measures indicate triple meter. Four beats to a measure is known as quadruple time. Six beats to a measure is representative of time that can be divided by three. When the melody falls on notes that occur between beats it is said to be syncopated time. Along with rhythm comes the concept of tempo. Tempo identifies the rate of speed of the beat of music and is measured by the number of beats per minute.

Timbre or tone colour is defined by the quality that allows differentiating between sounds of the same level and loudness generated by musical instruments or human voices. Timbre refers to the characteristic sound created by a particular musical instrument (including the voice) or combinations of instruments. The perception of timbre quality is determined by the frequency content or harmonic profile produced by the sound source. It is the spectral-temporal pattern of a generated sound indicating the way the energy in the system is distributed between different frequency components and the way that distribution is changing over time. Specification of timbre qualities is related to musical source and performance. For instance, a trumpet and saxophone have different timbres, even when they are playing exactly the same notes.

The term for gradations of amplitude (loudness and softness) in music is **dynamics**. They are a part of the articulation of accent in music. Distinction is made between accents of loudness (i.e. dynamic accent) and accents of length (i.e. agogic accent). Many dynamic indicators are used such as fortissimo, forte, mezzo-forte, mezzo, mezzo-piano, piano, pianissimo respectively meaning very loud, loud, medium-loud, medium, medium-soft, soft and very soft. Other dynamic features include crescendos, decrescendos and diminuendos. Changes in dynamics can be achieved gradually or immediate. Dynamic levels and accents are natural indicators for emotional mood. Marches, fanfares, and triumphal music tend to be loud while love songs and lullabies tend to be soft. Variations in musical dynamics present the expression given to music by diverse modifications in volume.

Music is experienced as structures that evoke meaning. Expressive characteristics of music depend on the spatial-temporal relationships between the constituent music categories described before. **Expressivity** in music causes the perception of emotive / affective qualities. Music is able to stimulate a wide range of emotional (e.g. excitement, boredom, joy, and sadness) and physical responses (e.g. toe tapping, hands clapping and shouting). Expressivity in music can bring to mind special memories and all kinds of associations that may be visual or tactile as well. The effect of music on emotions depends on context, in this thesis defined as personality dimensions and social cultural situations. As a consequence of context dependencies it is not self-evident that a given piece of music has the same emotional outcome in people. This implies that there is no explicit method that can explain all the cognitive processes involved when people listen to music.

To summarize, a brief description of constituent music categories makes it clear that each category uses various acoustical features. From a perceptual viewpoint, a listener

integrates those features into a whole. Research on context effects of melody recognition, for example, leads us to believe that melodies are represented psychologically as dynamic forms rather than as a group of distinct musical features. A slight modification in only one feature is enough to radically transform the way in which a melody is perceived (Bigand and Pineau, 1996). This supports the idea that more knowledge is needed on the kind of music that interests the average users of a music information retrieval system and on how they perceive this music. The objective is that the proposed conceptual framework meets the requirements of taxonomy-driven content-based music information retrieval systems. The multi-levelled framework as described before motivates a number of requirements for such a structure. In the next section these prerequisites are identified.

2.6 System requirements

With regard to the implementation of an operational taxonomy it is very unlikely that one single protocol could achieve a compromise between the diverse requirements of the many music features involved in content-based music information retrieval. It is therefore necessary to keep in mind the worst-case scenarios in order to minimize the range of protocols needed. Standardizing the way in which applications define their requirements is a necessary action towards reduction. Concept definition and classification is a first step towards standardization. In the context of content-based music information retrieval *flexibility* and *consistency* are considered as the central requirements.

2.6.1 Flexibility

Although the taxonomy structure is rigid, the taxonomy system should be flexible enough for a range of abilities. It should adapt easily to different conditions as they occur. Consider that taxonomy system implementation starts from a simple form of music taxonomy, such as a controlled vocabulary. It also comprises a list of terms and mutual relationships in terms of definitions. However, existing definitions (i.e. ANSI[47]; Willemze, 1975) appear to have a lot of shortcomings: they are general, verbal and do not provide a straightforward connection with acoustics. More specifically, the problems encountered have to do with the nature of descriptive language, the user's background, connecting functions, the interdisciplinary approach, subjectivity and the lack of instrumental descriptions.

A The nature of descriptive language

On the meaning of many terms there is no general agreement. The ambiguous use of terms in the literature does not facilitate the development of taxonomies. It is important to realize that certain terms refer to acoustic measures (e.g. frequency, intensity), whereas others refer to perceptual and cognitive ones (e.g. pitch, loudness). Cognitive measures are of a much higher level, and mostly incorporate related effects from multiple acoustic phenomena.

[47] American National Standards Institute.

B The user's background

Advanced music terminology is only applicable to a restricted class of users such as musicologists and experienced musicians. In fact, persons with no musical background using a musical search engine may simply choose to sing or hum a fragment of the music they wish to retrieve. Or, they probably will use natural language (e.g. Downie, 2002) to describe characteristics of the desired musical piece. Musicologists on the other hand might want to determine a particular melodic phrase or rhythmic structure and could use professional jargon, relying on theoretical definitions, to make descriptions

C Connecting functions

As taxonomies are often represented by hierarchical trees, additional meaning is specified via the signification of the hierarchical link. In a traditional taxonomy this meaning is "generalization" or "specialization" or "is a kind of". Nowadays taxonomy often refers to other kinds of hierarchies with other meanings for the links such as "containing", "part of", "instance of", "broader than", "related", "synonymous" or "associated". However, there may be many different meanings to a link and taxonomy may refer to relationship schemes other than hierarchical ones, such as network structures.

D Interdisciplinary

The concepts of a taxonomy structure must have the same meaning throughout the different levels of abstraction. In music information retrieval the concept "dissonance", for example, is sometimes used in the sense of "sensory dissonance", or as "cognitive dissonance", depending on the disciplinary background of the research team.

E Subjectivity

With the increase of the level of abstraction the significance of a term becomes more and more ambiguous. At the highest level a word can mean several things and one thing can be referred to by a number of words. "Romantic" as in romantic music, for example, can refer to a style period, a composition style or a particular mood.

F Lack of instrumental descriptions

Most definitions rely on phenomenological distinctions and lack any instrumental account which allows a straightforward connection with the audio stream.

2.6.2 Consistency

Most existing taxonomies for the Internet are used as concept structures and are not carried by a reasoning engine. They typically have been designed to directly meet the user's needs. No consistency is maintained, and users are supposed to quickly understand the system from a few examples. There is no agreement on definitions, models, methods or results. The music information retrieval and music digital library research communities have acknowledged the need to establish more rigorous approaches to evaluation and acquire standardized collections of music, standardized retrieval tasks and standardized retrieval metrics. Downie (2003) points out the need for a comprehensive evaluation paradigm for

music information retrieval systems. Research teams have difficulties to scientifically compare their approaches because there has existed no standard music collections against which techniques could be tested, no standardized sets of performance tasks and no standardized evaluation metrics.

A maximum of consistency must be achieved in order to:

- improve precision in content-based retrieval of audio information;
- build meaningful relationships internally in a database;
- establish meaningful comparison of results;
- build user-friendly interfaces.

2.7 Conclusion

In this chapter the process of developing a conceptual framework for content-based music information retrieval has been presented. First, it has been explained why such framework is needed. After the definition of the MAMI-taxonomy, on which the model is based, the description levels, concepts and descriptors of the conceptual framework have been discussed in detail. A model has been elaborated of descriptors that define the spatial-temporal structure class, the major category to delineate structuring and grouping of sound. Next, descriptor types have been distinguished and a taxonomy of constituent music categories has been presented. On the basis of the conceptual framework “flexibility” and “consistency” have been identified as central requirements for a content-based music information retrieval system.

The definition of a taxonomy and conceptual framework constitutes but the first step in a user-centred methodology. In order to turn this framework into a working model it is necessary to take account of listeners and users. A second step therefore, involves listeners and users in defining the complex relationships between low-level descriptors and high-level descriptors. This step uses music annotation experiments as a bottom-up method to define these relationships. In the next chapter the problem of music annotation is discussed and a model for manual annotation is defined.

3 Annotation of music

In research on musical audio-mining, annotated music databases are needed which allow the development of computational tools. These tools extract from the musical audio stream the kind of high-level content that users can deal with in music information retrieval contexts. The notion of annotation is ill-defined, both in the syntactic and semantic sense. As a consequence, annotation has been approached from a variety of perspectives (but mainly linguistic-symbolic oriented ones), and a general methodology is lacking. This chapter, based on original articles I and IV, presents a taxonomy of music features, a framework and a methodology in view of manual annotation of musical audio. The model presented regards annotation in function of a computational approach to music search and retrieval that is based on algorithms that learn from annotated data. First the background in music annotation is sketched. Then the concept of annotating music is defined. Next, different annotation types are considered. After specification of the difficulties faced, a general framework is presented that aims at providing acceptably accurate solutions to music annotation. Finally, two case studies in manual annotation of vocal queries and of drums are summarized. They were set up to test both the annotation methodology and new algorithms developed for the MAMI project. The taxonomies used for the case studies are presented.

3.1 Background

Since the background of audio annotation is mainly situated in the domain of speech research, it is straightforward to investigate to what extend linguistic-based tools may be useful for music annotation. Many tools have been developed for creating linguistic annotations. Annotation graphs, for example, provide a comprehensive framework for constructing linguistic annotations. Annotation Graph Toolkit[48] (AGTK), is a formal framework for representing linguistic annotation of time series data. Handschuh et al. (2003) describe a framework for semantic annotation for the semantic web. Bigbee et al. (2001) review capabilities of multi-modal annotation by examining how linguistic and gesture analysis tools integrate video data. They point to the increasing importance of multi-modal corpora and interfaces in future tools. A critical problem is the lack of standards that are able to mediate among the variety of existing tools and purposes (Bird and Liberman, 2001).

The music domain is more demanding because it has to deal with information associated with both frequency (e.g. pitch, timbre) and time (e.g. onset, tempo). In extending the concept of annotation to music analysis in general, it appears that the literature on music annotation is mainly concerned with linguistic and symbolic description. Few studies have investigated methods for the time synchronous annotation of an audio stream. The media industry and researchers involved in content-based music annotation are actively discussing the needs for music representation. In 2004 the Moving Picture Experts Group

[48] http://agtk.sourcefourge.net

(MPEG)[49] started a new activity aimed at the systematic support of symbolic forms of music representations by integrating symbolic music representation (SMR) into MPEG multimedia applications and formats. The decoding and rendering should allow the user to add annotations to SMR events. Here annotations are considered as audiovisual objects, music representation elements or simple URL-links. The idea is that the annotation format will be normative, as well as the way annotations are issued by the end-user.

Apart from the few annotation tools, today music annotation research is moreover mainly restricted to the domain of music notation, like the development of tools for adding specific interpretation symbols to a score such as bowing, fingering, breathes and simple text. Efforts in this context relate to the development of standards for music description, such as MPEG-4, MPEG-7 and MPEG-21. The media industry and researchers involved in content-based music analysis are actively discussing the needs for music representation.

Nevertheless, there are few systems available that allow the use of visual symbols or icons. Acousmograph (GRM)[50], for example, is similar to a sonogram, offering the user the opportunity to select part of a graph and listen to the chosen image. Another example is Timeliner[51], a visualization and annotation tool. It enables users to create their own annotated visualizations of music retrieved from a digital library. Timeliner has been integrated into Variations2 (Notess 2004), the Indiana University digital music library[52]. The tool enables students and instructors to create visual representations (bubble diagrams) of the formal structure of musical works contained within the digital library. Timeliner annotations describe musical events or processes taking place in a bubble or associated with a time point. In Variations2, Timeliner is used as a pedagogical tool. Accessing Variations2 is only possible with an Indiana University network ID.

Recently a C++ Library for Audio and Music (CLAM)[53] has been enriched with “Annotator”[54], a manual edition tool that allows visualizing and editing of low-level and high-level description of virtually any kind. The idea is to offer a tool for fine-tuning, editing and testing description extraction algorithms.

In general, music databases that have been annotated using existing annotation tools, mainly focus on metadata description, and less on the description of musical content as such. Donald Byrd maintains a list[55] as a continuing work-in-progress that surveys candidate MIR collections. Many of these databases, however, already start from symbolic representation of music (scores). The Repertoire International des Sources Musicales (RISM)[56], for example, documents musical sources of manuscripts or printed music, works

49 http://www.chiarligione.org/mpeg/
50 http://www.ina.fr/grm/outils_dev/acousmographe/
51 Timeliner was created by the Charles Stark Draper Laboratory in 1982 and was used to emulate the actions of the Space Shuttle Crew in timelines and on-board crew procedures. In 1992 it was chosen as the User Interface Language (UIL) for the Space Station program.
52 http://variations2.indiana.edu/research/
53 The CLAM library was published in context of the AGNULA IST-2001-24879 project.
54 http://www.iua.upf.es/mtg/clam
55 http://php.indiana.edu/~donbyrd/MusicTestCollections.html
56 http://www.rism-ch.ch/

on music theory and libretti stored in libraries, archives, monasteries, schools and private collections and provides images of musical incipits. The RISM Music Manuscript Database is linked to three other databases providing additional information to specific content: Composer, Library Sigla and Bibliographic Citations. The Meldex Digital Music Library[57] (McNab et al. 1996) handles melodic or textual queries and offers access to songs in two ways. The results of a query are visualized or presented as an automatically compiled list of metadata, such as titles.

At the Centre for Computer Assisted Research in the Humanities (CCARH)[58], MuseData has been designed to represent both notational and sound information (MIDI). The Real World Computing (RWC) Music Database (Goto et al. 2003) has been built in view of meeting the need of commonly available databases for research purposes. It consists of four databases containing popular music, classical music, jazz music, and royalty free music. Another two component databases were added with musical genre and musical instrument sounds. RWC contains music in both MIDI and audio form and provides lyrics of songs as text files. Standard MIDI files are generated as substitutes for scores, when no scores are available.

Automatic music annotation research involving genre, instrumentation, rhythmic style and emotion has been reviewed by Turnbull (2005). It has been observed that it is hard to compare results because researchers do not use standard data sets[59] and there is no standard labelling scheme for music.

To summarize, from scanning the literature on music annotation we learn that:

- tools and formats used for linguistic annotation are too limited in scope;
- music content description requires new description types;
- annotation research is mainly restricted to music notation;
- annotated music databases mainly focus on metadata description;
- there is a lack of methods for the annotation of a musical audio stream.

3.2 The scope of music annotation

Annotation in the broad sense is the activity of annotating something. An annotation may be any type of data that represents another type of data and is often referred to as metadata (data about data). *Music annotation* is an open term that integrates *any information* (textual, visual or auditory) that can be added to music. It refers to the act of describing content using appropriate space-time markers and labelling. In *music description* on the other hand, distinctive music characteristics are described in a way that is close to the original music.

[57] http://www.nzdl.org
[58] http://www.ccarh.org
[59] Though standard music data sets exist, such as the Real World Computing (RWC) data set, they are usually small or expensive. http://staff.aist.go.jp/m.goto/RWC-MDB/

The activity of annotating generates additional information that may be useful in contexts of information retrieval and data-mining, either as indices for retrieval or as training data for the development of computational tools. In that respect, annotation is a broad field that covers semantic content as well as labelling and segmentation. In the domain of music description, annotation pertains to metadata and music features that users might find particularly relevant in the context of music information retrieval. Metadata describes an asset and provides a set of attributes that are used to further classify or consume content.

MAMI research mainly focuses on the manual annotation of musical audio in function of the development, through computational learning, of tools for the automatic generation of similar annotations from the audio. Up to now, there has been a general lack of training data and the methodology for manual annotation of musical audio in function of algorithm development has been largely underestimated and underdeveloped.

Annotation of music is quite unlike speech annotation because it is less determined in terms of its content. Unlike speech sounds, music is not defined by a limited set of lexical entities. Instead, its syntax typically depends on multiple constraints that allow a great and almost unlimited variety of forms and structures. Moreover, its semantics are non-denotative and depend on subjective appreciation. Consequently, the process of manual annotation is rather complex because it comprises multiple annotation levels and different possible types of content description. The challenge of seeking common ground in the diverse expressions of music annotation has not been addressed thus far. Manual annotation of musical audio indeed raises questions that point to the nature of musical content processing, the context of music information retrieval, and the relationship between natural and cultural constraints involved in musical engagement.

3.3 Automatic versus manual annotation

Automatic music annotation uses low-level audio content to describe high-level musical concepts. In music information retrieval research, there are two components that constitute an automatic music annotation system, namely feature extraction and learning or modelling. Feature extraction uses techniques such as digital signal processing (DSP). It implies the design of a useful representation extracted from the low-level musical audio signal. In the approach of annotation in view of learning and modelling, a number of models have been applied to music annotation, such as k-nearest neighbour (KNN) classifiers, support vector machines (SVM) and neural networks (NN).

There exist a number of automatic music annotation prototype systems that attempt to classify samples of low-level audio into categories based on high-level concepts including genre, instrumentation, rhythmic style and emotion (e.g. MARSYAS[60]). However, there are still a lot of shortcomings which necessitates more investigation on manual annotation.

[60] MARSYAS is a software framework developed by G. Tzanetakis, that can be downloaded at http://sourceforge.net/projects/marsyas

Most commercial systems (e.g. Apple iTunes, Amazon, AMG Allmusic and Moodlogic) use human experts or collaborative filtering to annotate music. AMG Allmusic[61] for example, has created a large database of songs and musical meta-data. The meta-data has been compiled by a large group of experts. A user can search the database by artist, song, instrument, moods, themes, time period and country. In his review of automatic music annotation Turnbull (2005) has observed that "most systems pose the annotation problem as a supervised learning problem" [62]. One drawback of this technique is that it requires a classifier be trained for each individual musical concept (e.g. genre, instrumentation, emotion). As said before, there is a lack of databases with annotated music developed to perform such computational learning. Researchers tend to construct small data sets that cannot be shared due to copyright laws.

Manual annotation of musical audio is considered as an intermediate step toward better automated annotation tools. In this study manual annotation is approached in function of the development of tools for the automatic generation of similar annotations from the audio. Annotate the data to test a system is often done somewhat carelessly and that does not bode well for reporting and comparing results or for training automatic information retrieval systems. Besides, in many cases, where the data is related to the perception of music or human interaction with the music, a good collaboration between musicologists, cognitive psychologists and music information retrieval algorithm developers is required. The work on query-by-voice, for example, is such a study where perception and production plays a role. Music similarity is another area in which there is a need to combine perceptual studies with the algorithmic approaches.

3.4 Annotation problem specification

The major task of content-based music information retrieval is to investigate issues that are needed to make a connection between the musical audio stream and content descriptions that are easy to use and understand for non-experts. The main problem is that audio streams are physical representations, while user-friendly descriptions pertain to high-level human information processing capabilities. These involve a complex set of goal-directed cognitive, affective and motor actions. Humans typically process music information in terms of purposes, goal-directed actions, values and meanings. They handle a subjective (first person) ontology that is very different from the objective (third person) ontology of physical signals[63]. A major goal of manual annotation of music is therefore to provide data that allows computational systems to learn the task of annotation and thus to build bridges between first person and third person descriptions of music. Modelling based on imitation

[61] http://www.allmusic.com/

[62] Supervised learning is a machine learning technique for creating a function from training data that consist of pairs of input objects (typically vectors) and desired outputs.

[63] See for a more detailed account M. Leman: "Being involved with music. A theory of embodied music cognition and mediation technology" (book in press).

learning is considered a candidate to cope with the gap between the measurable quantities of an audio signal and the intentionality of subjective qualities.

3.5 Manual annotation framework

The definition of a general framework for manual annotation of music is regarded from three viewpoints related to (1) context dependencies, (2) computer modelling and (3) representation.

3.5.1 Context dependencies

There are at least three observations to keep in mind when dealing with music annotation, namely the intentional nature of human communication, the requirements of music information retrieval contexts, and the development of mediation technology.

To begin with, since music annotation aims at making the link between the musical audio stream and levels of content description that allow humans to access the information stream, it is necessary to take into account the highly focused level of human communication. This level is called the *cultural level* because the implied ontology is based on human learning, subjective experiences and symbolization. As this level is characterized by goal-oriented behaviour and intentional attitudes (thoughts, beliefs, desires,…) its descriptions are very different from objective or nature-driven descriptions that pertain to physical signals. As a consequence, two methodologies are involved that are *naturalistic* approaches and *culturalistic* approaches.

Naturalistic approaches are studied in the natural sciences. They aim at developing tools that extract nature-driven descriptions from audio. These tools are objective in the sense that they start from physical “energies” and rely upon “universal” principles of human information processing. The resulting descriptions have an inter-subjective basis and do not involve the subjective goal-directed action ontology on which human communication patterns typically rely. Examples are the extraction of pitch from a sung melody, or the extraction of genre or timbre classes from polyphonic audio.

Culturalistic approaches are studied in the human sciences. In contrast to naturalistic approaches, they tend to describe music in terms of its signification, its meaning, value, and role as cultural phenomenon. Thus far, cultural determined content description has been based on textual and visual descriptors that are strongly linguistically-symbolically oriented. Reference is often made to subjective experience and historical and social-cultural interactions.

Some culturalist musicologists who look at music from the cultural perspective tend to claim that links between subjective and objective descriptions of music are impossible (Leman, in press). Yet, there is no strong evidence for such statement. The main argument draws on the idea that signification is an arbitrary activity that depends on the cultural environment,

history and personal taste. Association and signification can indeed be random, but we believe that there are at least certain aspects of descriptions, including those at the high semantic (first person) levels, that are not completely arbitrary. Given a proper analysis of the goals, they can be very functional in music information retrieval contexts. When aiming at describing music semantically, this hypothesis is rather fundamental. Scepticism can only be refuted when the proof has been given of a working system.

A second observation closely connected to the first point, is that music descriptions serve a goal that is largely determined by the MIR context. Out of a myriad of possible natural objective and subjective descriptions, we should select those that serve the particular goals of a particular context. Hence, research in audio-mining is not purely a matter of bottom-up signal processing and making the link between third person and first person descriptions. At a certain moment, a thorough analysis has to be made of things such as the retrieval context, the economic and ethical value, and the purpose. Decisions may have to be taken about the context-based bias of the whole enterprise, including the work on manual annotation. Reference can be made to the DEKKMMA-project[64], where researchers are confronted with a large audio database of Central African music, and where little experience is available of how and what people would tend to search in such database. Analysis will have to clarify that it is likely that users who are unfamiliar with the Central African idiom behave very differently from users who know the music.

A third observation is concerned with the technology used for music mediation. Mediation here refers to the ways in which streams of musical information are transmitted and to the tools that may be used to specify and retrieve content. The most recent developments seem to go in the direction of networked wireless mobile devices that give access to a large amount of databases (Baumann, 2004). Such technologies may imply certain constraints on the possible types of musical content specification.

To sum up, given the aims of a music information retrieval system, annotation should take into account at least three different types of context, namely *culture, user,* and *mediation*. Starting from that background, Figure 7 shows the general framework for annotation of musical audio. This framework incorporates the naturalistic and culturalistic approaches that focus on human information processing and social-cultural context (adapted from Leman, in press).

[64] http://music.africamuseum.be

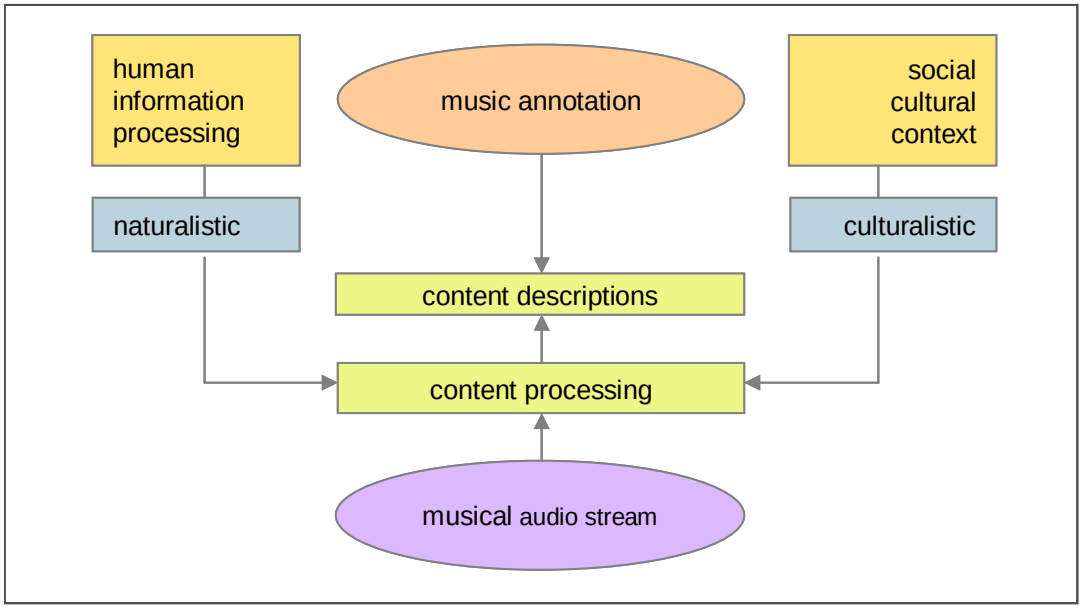


Figure 7 : General framework for annotation of musical audio.

3.5.2 Computer modelling

Apart from the general framework in which annotation has to be carried out, the computational modelling approach needs careful analysis. In the past, the problem of manual annotation of musical content has often been considered from the viewpoint of *Cartesian* modelling. The strategy consists in the specification of a set of pre-defined feature extractors with limited scope to clear meaning in a restricted context (most often) of stimulus-response experimentation. It is then hoped that through combination of a selected set of weighted pre-defined features, high-level semantic knowledge can be predicted. However, it turns out that this strategy has a number of limitations (Leman, 2004), such as a complicated semantic interpretation when features are summarized or combined (linearly as well as non-linearly). If no significant meaning can be given to these features it may be better to give up the idea of working with many local descriptors and look for optional methods. Pachet & Zils (2003) explore an alternative method for extracting automatically high-level music descriptors. Their approach is based on genetic programming. All the same, straightforward imitation learning based on an appropriate level of manual annotation may be of help.

Taking into account the multiple ways in which users can engage with music, annotation should extend the possibilities of descriptions that are linguistic-symbolic with non-symbolic and non-linguistic forms of description. This draws on the understanding that the interaction between subjective experience and objective description is a dynamic process constrained by both natural and cultural determinants. Somehow, levels of annotation in between what is considered to be natural and cultural processing should be chosen. An ecological approach indeed regards any response to music as the result of a complex interaction with the subject in its social, cultural and physical environment. Levels of annotation can be

addressed that lie on the borderline of objective/subjective descriptions and that form the connection points with first person and third person descriptions of music.

This calls for an investigation into new forms of annotation based on the mimetic and gesture capabilities of human communication. Performing a manual annotation in the form of motor action subsumes the interconnectedness between culture-based and nature-based computational music research approaches.

3.5.3 Representation levels

Annotation comprises diverse types of description, depending on the purpose and the level of annotation. Various types of music annotation are related to syntactic, semantic, structural and articulation elements. Figure 8 shows the distinctive representation levels and associated annotation methods.

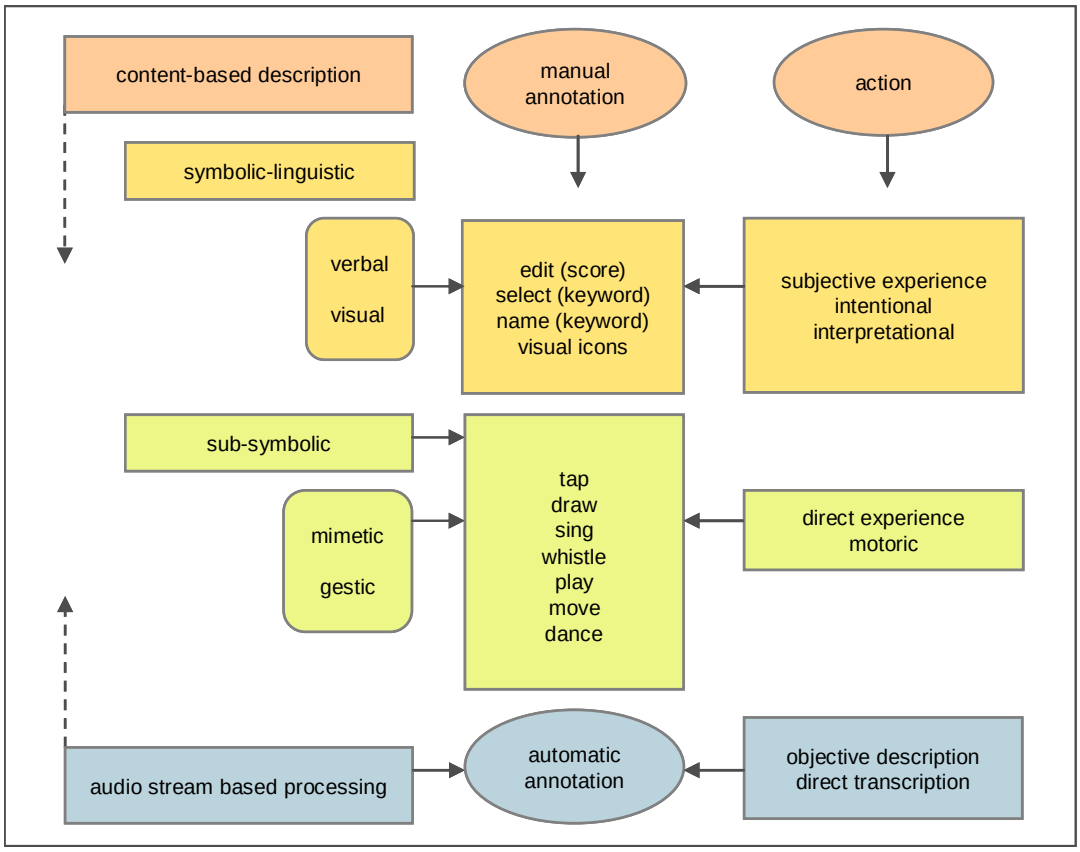


Figure 8: Representation levels and associated annotation methods.

Symbolic-linguistic-based annotation mainly focuses on the description of structural and semantic units. The user's interaction with symbolic representation relates to verbal and visual descriptors. This could include a score, or conventional music notation machine code (e.g. MIDI, SASL). The main problem of this approach is that symbols are deprived of any semantic connotation and that they require human interpretation based on the inter-

subjective semantics. Annotation thus relies on subjective experience of the creator or user who performs an interpretational and intentional action.

Sub-symbolic-based annotation is mainly based on multi-dimensional spaces. Some studies focus on representation forms such as sonification and visualization. Toiviainen (2003), for example, explores the additional value of visual data mining for large music collections. Through investigation of multiple music dimensions he has found that some musical features are more natural oriented and other more cultural. Pampalk (2003) presents a visualization method that uses self-organizing maps for grouping similar pieces pertaining to different music genres. Manual annotation involved the placing of the pieces on a map according to personal style.

Mimetic-based annotation is related to imitative aspects of motor actions. Imitation behaviour in general is a topic of growing interest within the cognitive sciences (Leman and Camurri, in press). The process of imitative learning is founded on perceiving real world behaviour and learning from it through action. Applied to music research, imitation is a means of analyzing the perception of similarity aspects within music through motor responses. The now popular query by voice paradigm for example supports the idea of retrieving music by vocal imitation of representative melodies (e.g. Birmingham, 2002; Lesaffre, 2003; Pauws, 2002).

Gesture-based annotation accounts for representation as the result of multi-modal gesture-based interaction and emotional expressiveness (Leman and Camurri, in press). It differs from mimetic annotation because it does not need to be learned or rehearsed. Gesture annotation involves motor action as a physical manifestation of the sonorous, such as body movement or dancing. Modelling gesture annotation takes into account time dependencies. Consequently, it mainly applies to the representation of rhythmic features. At the gesture level two annotation forms are distinguished, namely sound-producing action (e.g. tapping, hitting and stroking) and sound-accompanying action (e.g. body movement and dancing).

3.6 Conclusion

In this chapter is has been argued why manual annotation of musical audio is needed. A manual annotation framework has been characterized from three viewpoints: the first related to context dependencies, the second to computer modelling and the third to representation. New annotation methodology based on this model has been tested in two case studies.

A case study on the manual annotation of vocal queries and another one on the annotation of drums have been set up to test both the manual annotation methodology presented in the general framework and new algorithms developed for the MAMI project. In both studies, the manual annotations were performed by experts.

In what follows the annotation strategy is described and the results are summarized.

3.7 Case study 1: annotation of vocal queries

An experiment was set up by which subject were requested to produce vocal queries by singing into a microphone. The goal, set up and analysis of the experiment output are discussed in chapter four. In this section attention is drawn to the methodology of manual annotation applied to the vocal queries gathered by the experiment.

3.7.1 Annotation strategy

The annotation strategy was focused on two objectives, one *user-oriented*, and one *modelling-oriented*. The user-oriented approach aimed at providing content about the spontaneous behaviour of users taking part in the vocal query experiment. The model-oriented approach aimed at providing detailed descriptions of queries in order to build a referential framework for testing automatic transcription models.

User-oriented annotations were carried out for a set of 1148 queries taken from the user responses to the songs used in the experiment[65]. The user-oriented annotation focused on the analysis of long-term versus short-term memory effects, the different vocal methods used and the differences between subjects and their relation to musical education, age and gender. The model-oriented annotations were carried out for 32 queries, excerpts of the popular songs “Blowin’ in the Wind” (B. Dylan), “Walk on the Wild Side” (L. Reed), “Sunday Bloody Sunday” (U2) and “My Way” (C. François). These annotations provide detailed descriptions of low and mid-level acoustical features, as described in the next section.

3.7.2 User-oriented annotation

User-oriented annotation (detailed in ❸) was performed for the following features*: timing, query method, performance style, similarity with the target* and *syllabic structure*. Table 2 shows the taxonomy structure for manual annotation of user-oriented features.

A Timing and query method

As timing characteristics, the total length of the recording and the start, end and length of the actual query were collected. Vocal queries may contain a mix of different query methods such as humming (H), singing syllables (S), singing lyrics (T) and whistling (W). In addition to those methods percussion (P) (e.g. tapping along with the drum) and comments (C) (spoken comments made by the subjects) are found. Studying these six methods in more detail required segmentation into homogeneous parts. Segmentation in temporal units according to the methods used resulted in a set of 2114 segments. Segment annotations at the temporal level indicate the starting and ending time as well as the total duration of the segment. The number of segments in a query (according to the methods used) is counted.

[65] See 4.4.13: Stimuli related aspects, table 15.

USER-ORIENTED ANNOTATION	
Query timing	Query start time (in seconds) Query end time (in seconds) Query total time (in seconds)
Query method	Hum (H) Sing syllables (S) Sing text (T) Whistle (W) Percussive (P) Comment (C)
Query segmentation	Query method
Query segment timing	Segment start time (in seconds) Segment end time (in seconds) Segment total time (in seconds)
Performance	Melodic (M) Rhythmic (R) Intermediate (I)
Similarity degree	Not (0) Little (1) Moderate (2) Rather (3) Much (4) Very much (5)
Syllabic structure	Onset Nucleus Coda

Table 2: Taxonomy for user-oriented annotation of vocal queries.

B Performance style

Furthermore, distinction is made between three different performance styles focusing on melodic, rhythmic or intermediate interpretations. A performance style is considered to be melodic when a clear succession of different pitches or melodic intervals is observed. A segment is annotated as rhythmic when no clear pitch intervals are noticed (as in a spoken text or a percussive sequence). An intermediate category is used to classify segments where a sense of pitch is present, but without a clear melody (e.g. using a reciting tone).

C Target similarity

Then, to each of the segments, a relative similarity rating is given on a six-point scale, ranging from not recognizable (0) to sounding similar (5) to the target song (presented textually in part 1 and aurally in part 2 from the experiment). The estimation of the degree of similarity between query segment and target is focused on melodic and rhythmic properties. Aspects of timbre or use of lyrics are neglected. This estimation, obviously, is subjective and therefore the similarity measures are only used to compare large sets of data. That is to compare the efficiency of the different methods, the performance of different groups of users and the effects of differences in memory recall.

D Syllabic structure

As 766 segments out of 2114 have a syllabic content, a major part of the annotation work is related to the analysis of syllabic queries. Syllables are considered as non-semantic vocal events, containing a vowel, which can be preceded and/or followed by a consonant or a complex of consonants. Syllables are analysed according to their structural components: the onset (initial consonant or complex of consonants), the nucleus (vowel) and the coda (final consonant). The 766 syllabic segments in the database contain a total number of 14748 syllabic units.

3.7.3 Model-oriented annotation

The model-oriented annotation (examples in ❸) is performed on a set of homogeneous query segments containing singing syllables or whistling, and a set of heterogeneous queries containing a mixture of methods. The features investigated are: event, onset, frequency, pitch stability, query method including sung words or syllables. Table 3 shows the taxonomy structure for manual annotation of model-oriented features.

MODEL-ORIENTED ANNOTATION	
Onset	Point in time (in seconds)
Onset sureness	Not pronounced (0) Clear (1)
Event	Duration (in seconds)
Frequency	Numerical (up to 1 Herz)
Pitch stability	Up (U) Down (D) Stable (S) Fluctuating (F)
Query method	Sing (S) Text (T) Whistle (W)
Syllables	Text
Lyrics	Text

Tabel 3: Taxonomy for model-oriented annotation of vocal queries.

A Event and onset

At the local temporal level, events are determined. An event or object is characterized by its duration. It starts at the moment in time defined by the beginning of an onset and ends at the moment in time where the onset of the next event or non-event begins. From a conceptual point of view, an onset is considered a moment in time defined by the beginning of an event. This definition accounts for successive events pertaining to monophonic files. The point of onset is based on both auditory (listening with headphones) and visual (looking at the waveform) perception. Moreover, a sureness quotation is given (0, 1) which distinguishes between clear onsets (1) and less pronounced onsets (0). An onset is considered less pronounced when successive notes with the same or different pitch are performed smoothly with no explicit separation (e.g. legato performance).

B Frequency and pitch stability

For pitch annotation of vocal queries, frequency was assigned in Hertz with a resolution of one Hertz. Taking into account that the query files in the data set are real audio and not MIDI encoded a resolution greater than a semitone makes sense. It facilitates adaptation to users who sing too high or too low with frequency deviations that do not coincide with a semitone. Additionally, for each object, an extra segmentation was annotated at the moment in time that the frequency of a note reaches stability. Four semantic labelled categories are used: up (U), down (D), stable (S) and fluctuating (F). A frequency is considered stable when the distance between the lowest and highest frequency within an object is smaller than 5 Hz. As sung melodies, especially when produced by untrained voices, move within a limited frequency range, this threshold rarely exceeds a quartertone.

C Query method

Within each object the query method (S, T or W) are annotated. For sung queries words or syllables are added.

3.7.4 Methodology

A manual annotation methodology has been worked out using existing software tools. The annotation task was performed by music experts working at IPEM.

A Method of user-oriented annotation

The user-oriented features are annotated using Excel, SPSS and CoolEdit. The latter was used for editing (i.e. segmentation) of the queries. The annotations were saved in an Excel file. Using SPSS statistical analysis was performed. The results[66] of that investigation have been applied to determine the set of queries for model-oriented annotation.

B Method of model-oriented annotation

For model-oriented annotation the open source PRAAT program for speech analysis (Boersma and Weenink, 1996) is used in combination with PureData, a graphical computer music system, and Matlab, a tool for doing numerical computations with matrices and vectors. An advantage of using this program is that the annotator has a visual image of the signal on the screen (see figure 9: top tier) and has the ability to listen repeatedly to any fragment of the signal. Another advantage is that labelling and segmentation of the sound files is stored in a TextGrid object that is separated from the sound. Once the labelling is done is can easily be converted to a MIDI format, the format that is used by most transcription systems. Boundaries are marked by the places in time where an event, a non-event or a pause for breath begins. The TextGrid object is written into a formatted ASCII-text file. Time points are labelled on multiple tiers that are time synchronized for different annotation levels as shown in figure 9.

[66] See 4.4: Experiment on spontaneous vocal query behaviour.

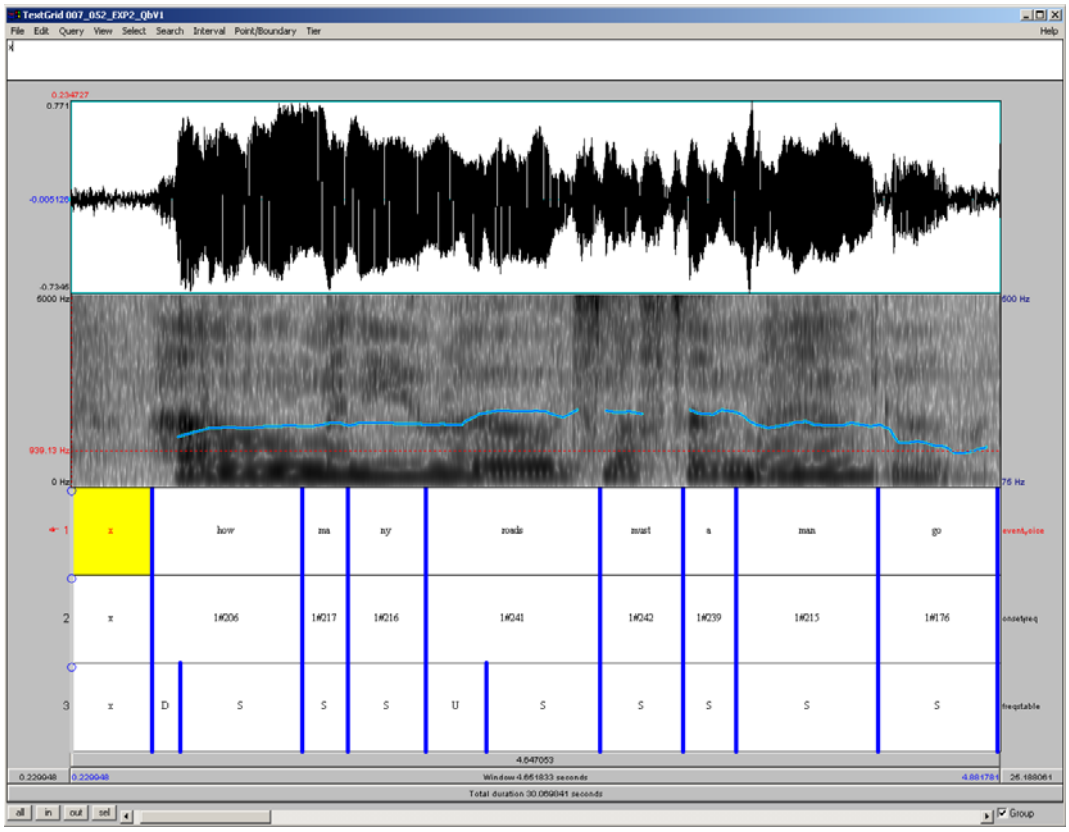


Figure 9: Vocal query annotation using PRAAT.

The tiers from top to bottom show (1) the sound waveform, (2) the pitch contour of a sound as a function of time, (3) used lyrics or query method, (4) sureness quotation and frequency and (5) pitch stability.

Tier objects representing non-events such as breathing or a clapping door are labelled with an ‘x’. The query method is labelled with H (humming), P (percussion), W (whistling) or C (comment). For sung queries the lyrics or nonsense syllables have been annotated. Frequency estimation was done aurally, directly comparing the sung tones with pure tones generated by a frequency generator implemented in Pure Data (Puckette, no date). Then a final check is done, reproducing the annotated frequencies simultaneously with the original sounds, using a Matlab script. To each event, an extra segmentation descriptor has been added defining pitch dynamics in terms of stability. This feature refers to the steady part of a note where all the harmonics become stable and clearly marked in the spectrum. Stability annotation is conforming to four semantic labelled categories: up, down, stable and fluctuating. A pitch is considered stable when the distance between the lowest and highest frequency within an event is equal to or smaller than 5 Hz.

3.7.5 Results

This part summarizes the results for both user-oriented and model-oriented annotation.

A Results for user-oriented annotation

User-oriented annotations were used for statistical analysis. The metadata provided insight in the structure of a query-by-voice search. Some basic characteristics of vocal queries and several user categories could be distinguished. A detailed description of the findings is included in chapter four.

Summarizing, it has been shown that:

- vocal queries have a mean length of 14 seconds;
- six query methods are distinguished;
- most common query methods are singing lyrics or syllables;
- least common query methods are humming and percussion;
- the use of query methods is user dependent;
- the choice of syllables is user dependent;
- there are effects of age, gender and musical experience;
- the duration of a query is dependent on memory;
- similarity with the target is dependent on memory.

B Results for model-oriented annotation

The model-oriented annotation methodology and the annotated queries have been used as a reference framework for testing existing pitch to MIDI models and for the improvement of an acoustic module of a vocal query system. This module is based on a melody transcription algorithm developed for the MAMI project. Results of this investigation are discussed by Clarisse et al. (2002) and De Mulder et al. (2003). Clarisse et al. establish that most transcription systems are incapable of accurately transcribing singing sequences of naïve singers. They present a new auditory model based transcriber of vocal melodic queries. This model is extended by De Mulder et al. (2004). The annotated metadata used have proven to be valuable as reference for evaluation and training or parameter estimation. Experiments have shown that their system can transcribe vocal queries with an accuracy ranging from 76 % (whistling) to 85 % (humming), and that it clearly outperforms other state-of-the art systems.

3.7.6 Conclusion

Annotation of music collections is often seen as a necessary step for the development of music information retrieval systems. Candidate music information retrieval test collections, their characteristics and availability have for example been described by Byrd (2003). Although a lot of music information retrieval research focuses on vocal query paradigms, none of these collections contain human produced queries stored as audio. Researchers tend to work with databases containing small sets of queries with tunes obtained from small-scale experimental studies. As a consequence query sets with sung tunes are far from elaborate.

The case study on annotation of vocal queries gave rise to an annotated query database with melodies that are representations of all possible vocal query methods. Taxonomy structures containing features for user-oriented en model-oriented annotation of vocal queries have been designed. An investigation concerning the way a user spontaneously behaves when performing vocal queries has shown that the use of manually annotated queries yields a better conceptual understanding of some of the recurring issues encountered in an undertaking of music content retrieval. Manual annotation of model-oriented features was beneficial for the development of a melody transcription tool for the MAMI project. Given these results, the annotated vocal query database is made accessible[67] to researchers. It brings basic material within the reach of music information retrieval researchers.

3.8 Case study 2: annotation of drums

The second case study reports on a manual annotation methodology that was set up for gathering annotations for drums in 52 real-world music fragments containing different drum event types. Research on drum detection is relatively new and there is a real need for reference material to train and test drum detection algorithms. The aim was to provide ground truth to train and test algorithms that can automatically detect drum events in polyphonic music. Background and details on the set up of this study are described in (Tanghe et al., 2005). In the context of the MAMI project, algorithms have been designed to localize and label drum events, based on a model consisting of three main parts: onset detection, feature extraction and feature vector classification. However, such system cannot be fully worked out without ground truth data that are needed for several reasons. First the system contains a machine learning algorithm that requires to be trained in a supervised way on data that represents the true task of drum detection. Second, in order to obtain a parameter combination that gives a good overall performance of the system, the annotations are used for parameter optimization. Finally, the annotations are useful for comparing various drum detection systems against each other in a systematic way.

3.8.1 Annotation strategy

In this study annotation was performed by ten experienced drummers or percussionists aged between 23 and 57 years. Half of the group was autodidact and the other half was musical educated at a conservatory. The task could not be performed by musicologists because it requires a more than average familiarity with drum sounds. Motivated percussionists were needed who could perform the task rigorously.

For example, unlike monophonic melody lines, drum events can overlap. This means that for a specific position in time multiple drum events can occur. A person that is not

[67] The database including queries and annotation files can be accessed from the public section of the MAMI project website: http://www.ipem.ugent.be/MAMI/public/data/QbVExperiment

experienced with drum sounds could easily miss one or more simultaneously occurring events. Therefore, the annotators were selected from the pool of participants in the online inquiry described in chapter five employing following criteria: they play a percussion instrument and they estimate themselves as having a high-level of musicality. A few were also recruited through other connections simply because they were known as good drummers.

The annotators have been asked in advance to answer a short list of questions about the music styles/genres they are familiar with. Other questions were related to their acquaintance with sequencer software/hardware, their opinion about the importance of various drum sound types and the appropriateness of different methods for entering a drum sequence into a computer.

All these answers have been taken into account while setting up the annotation task and selecting the music. A dataset of 52 music fragments, digitally extracted from various commercial music CD's[68], was generated. The selection criterion was finding a good compromise between style diversity and annotator preferences. Having as much different styles and genres as possible is important to evaluate algorithms for robustness and flexibility, but making sure that the annotators are familiar with the music is also significant because it provides more reliable annotations.

Since most of the annotators took part in the online survey (see chapter five) that aimed at recruiting a large group of subjects willing to participate in diverse annotation experiments, the music they had then specified as being their "favourite" has also been taken into account. A list of the music fragments involved is included in ❸.

Additionally to the set of real music fragments, three reference fragments were included: a very simple, self made reference file containing clear drums of various types and no music, and two recordings of MIDI files with drums and music. Since these audio fragments were generated from a symbolic representation, it is possible to compare the manual annotations with the true events. This would provide an idea about the annotation quality.

3.8.2 Annotated features

The goal of the annotation task was to provide reliable ground truth data that represents the positions in the sound files where drum events occur, together with labels specifying the types of drum events. The annotation task included editing of 18 types of drum events (see table 4) and tapping the beat. An elaborated list of drum types is justified because information provided by the annotators beforehand and some preliminary tests showed that a rather diverse list of drum types was preferred.

[68] A disadvantage of this choice is that the music itself along with the annotation data cannot be distributed due to copyright restrictions.

Full name	Label	Note
Bass drum	BD	36
Snare drum	SD	40
Open hi-hat	OH	46
Closed hi-hat	CH	42
Ride cymbal	RC	59
Crash cymbal	CC	57
Low tom	LT	45
Mid tom	MT	47
High tom	HT	50
Claps	CP	39
Rim shot	RS	37
Splash cymbal	SC	55
Shaker	SH	70
Tambourine	TB	54
Wood block	WB	77
Low conga	LC	64
High conga	HC	63
Cow bell	CB	56
Other drum	-D	75

Table 4: Taxonomy of drum types, labels and MIDI notes.

3.8.3 Annotation methodology

The annotation methodology was set up by K. Tanghe[69] who developed the software for the drum detection for the MAMI project. A detailed description of the set up is given in Tanghe et al. (2005). Cakewalk Sonar 2.2, a standard music production software package with multi-track audio and MIDI sequencing capabilities, was used. Figure 10 shows a screenshot of the annotation interface.

Each annotator was presented with a Sonar multi-track project consisting of one audio track for the music fragments, one MIDI drum track for the drum annotations, and one MIDI drum track for the beat annotations. The music fragments were placed after each other on the audio track. The top half of the screen showed the three tracks and standard controls, and the bottom half showed a drum grid view of either the drum track or the beat track.

The annotations were stored in a MIDI file where a MIDI “note on” message per annotated drum event encodes both the position and drum type. The MIDI tracks were connected to a virtual sound module, which means that the annotators could actually listen to their annotated events by playing back the annotation track through the sound module. Visual editing and MIDI keyboard recording as well was possible. As an alternative to entering all the drum events one by one into the drum grid view a MIDI keyboard, with drum labels stacked to the keys, was attached to the annotation computer. A small document with

[69] From 2001 to 2005 K. Tanghe was a member of the MAMI–research team.

guidelines ❸ was verbally presented to the annotators and was left at their disposal in the annotation room as a reference.

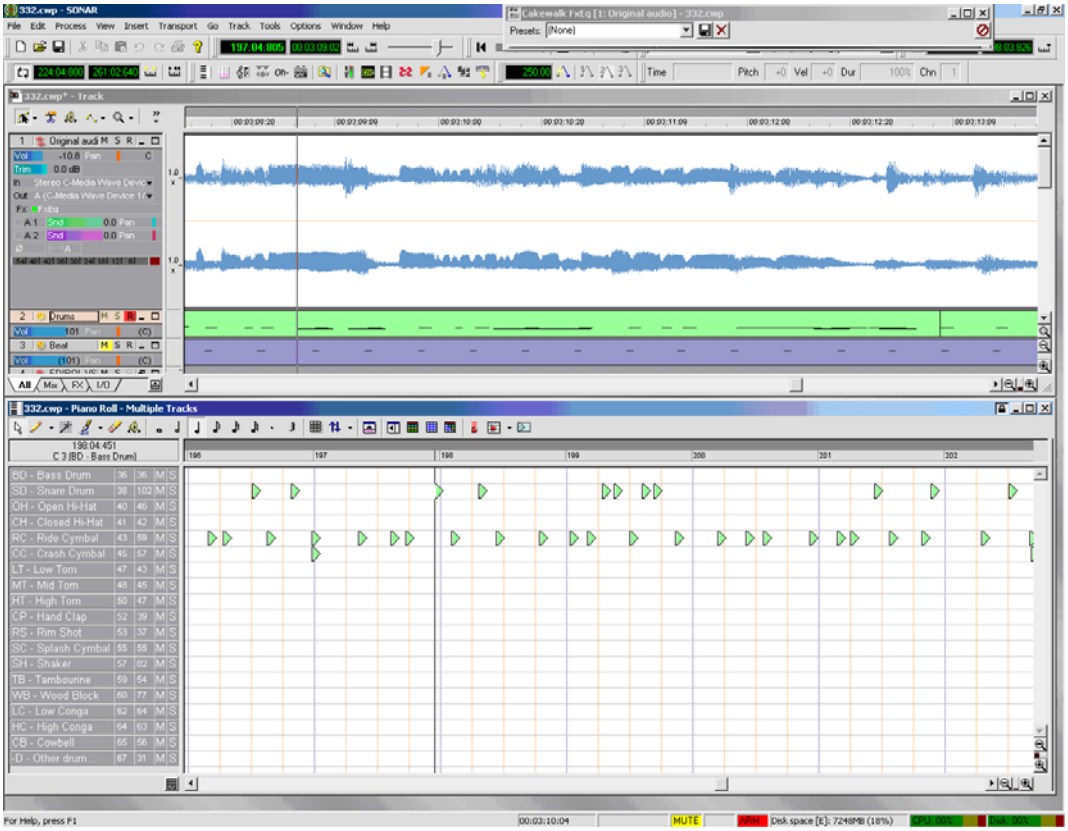


Figure 10: Multi-track sequencer and beat annotation set up.

3.8.4 Results

The annotators reported that some drum types and articulations are difficult to annotate as a single percussive element:

- brushes with their typical “dragged” effect;
- “flammed” drums (typically the snare drum) where two hits of the same drum type are deliberately played almost at the same time, creating the sensation of a ghost note occurring slightly before the main note;
- hi-hat sounds are sometimes mixed up with shaker events.

Annotation was further complicated by sounds being post-processed with audio effects and the dynamic processing of bass frequencies (which make it hard to perceive the bass drum and the bass lead as separate entities).

Manual evaluation of the annotations revealed following mistakes:

- double events;
- single missing events;
- spurious events (caused by copy/paste actions).

The reference material[70] generated by this manual annotation task has been used for development and testing algorithms developed for the automatic detection of drum sounds in polyphonic audio.

3.8.5 Conclusion

The case study for manual annotation of drum events provided insight in characteristics of drum events. Besides it generated ground truth data for drum algorithms developed during the MAMI project. Apart from this, the reference material has proved to be useful in a broader way. It was used as a test collection in the Music Information Retrieval Evaluation eXchange (MIREX)[71] contest track of the ISMIR 2005 conference. The goal of this annual contest is to compare state-of-the-art algorithms and systems relevant for music information retrieval. In this context, ground truth data is needed in order to perform objective evaluations. The drum detection algorithm developed by K. Tanghe on basis of this test collection was ranked second in the contest.

From the viewpoint of the annotators involved in both case studies a suggestion could be made towards tool developers. Being able to see the annotations and the waveform at the same time, together with the possibility to actually hear the annotated events proved to be essential for the accuracy of the annotation process. From this experience and from the belief that manual annotation is a major process in the development of automatic annotation methods, there is a need for easy-to-use and flexible tools dedicated to manual music annotation[72].

The next three chapters will focus on the user. Experimental investigation regarding search behaviour, vocal querying behaviour, background of music information retrieval system users and annotation of music qualities will be presented.

[70] The drum annotation files ca be accessed from http://www.ipem.ugent.be/MAMI/Public/Data/DrumAnnotation
[71] http://www.music-it.org/mirexwiki
[72] Recently (20050721) CLAM Annotator has been released; however in view of finishing this dissertation in fall 2005 the tool could not be tested.

4 Spontaneous User Behaviour

One of the major issues in music information search and retrieval is the problem associated with initiating a query. Users find it hard to generate a good query because their initial information may be vague (i.e. *I do not know what I am looking for but I will know when I find it*). A lack of high-quality interfaces for query formulation and a limited knowledge of the query methods which users would tend to use, has been a barrier to the development of effective retrieval technologies. Therefore, investigation is needed that concentrates on more natural query methods that could capture users' desires and tailor the retrieving accordingly.

This chapter (based on original articles I and V), focuses on users' search behaviour and on vocal query methods, that is the strategies and modalities to pose a query. To begin with, a brief overview is given of different ways of questioning an annotated musical audio database. Then, tentative research leading up to a large experiment is reported. The study of natural user behaviour comprises four parts (1) a field interview, (2) the testing of a simple questionnaire, (3) a web based survey and (4) an experiment on spontaneous vocal query behaviour. This study has been carried out between winter 2001 and spring 2002.

The objective of the whole process was to get an idea of (1) the methods that customers and music lovers prefer to use when they search music, (2) whether they feel a need for new search methods and if so (3) what methods they suggest, (4) what kind of information is used to find a piece of music and (5) what kind of information does the user want as an answer to the search process and (6) how do people behave when they spontaneously reproduce music using any vocal query method.

4.1 Ways of questioning

In this section a literature-based overview of the most important existing and candidate query methods for multimodal music information search is given. A global distinction is made between text-based information retrieval approaches and audio based approaches.

In traditional information retrieval, a query is a statement of information needs. It typically consists of keywords combined with Boolean operators (i.e. AND, OR, NOT) and other modifiers. Many libraries, for example, allow search and retrieval based on the title of the piece, the name of the composer, the year of production and different other meta-data search categories. However, searching for music by identifying its content (e.g. a melody) is considered a more natural way than searching by its meta-data (e.g. title), because it is assumed that content is usually more memorable and a more robust feature of the musical work (Chai, 2001). In user interfaces, voice commands have shown to be a quick and fairly natural way to interact (Sorsa and Halonen, 2002). Studies have been carried out which look at the kinds of music information needs people have (Lee, 2004). However, we do not currently know in which way people who are willing to use new retrieval technology, would want to interact with such a system.

Selfridge-Field (2000) has discussed the key differences between querying a music collection and querying a text collection. She remarks that most useful music queries will be fuzzy and that suitable ways for searching are likely to be as heterogeneous as the repertoires themselves. Possible query modes depend on the database type and on the information retrieval system used and multimodal querying of musical audio databases poses a big challenge to the design of a music information retrieval system.

In short, looking at the amount of research that focuses on new music information retrieval systems it can be expected that novel ways of querying are likely to be introduced in the near future.

4.1.1 Textual-based approaches

A Query-by-Text

In general, the retrieval of text documents is based on controlled vocabularies and keywords. These are hierarchical lists which have been designated to a database to represent the major concepts and relationships between concepts contained within that database. These lists constitute the database ontology, that is, the set of what exists in the database. The lists can change from database to database. The hierarchical nature of the lists benefits search strategies by allowing broad concepts to be narrowed down in a manner that stays consistent within that framework. This type of searching is called *subject searching*. It displays only those records that match exactly to those terms entered within the manner that the database has been set up. Keyword searching allows entering a search term that describes the term as it is used in information source records. This may be expanded with the help of a thesaurus, which allows the use of synonyms and variations of search terms. The disadvantage of this method is that users have to know what they are looking for.

B Query-by-Note

These systems are being developed wherein searchers construct queries consisting of pitch and/or rhythm information (e.g. Pickens, 2000; Doraisamy and Rüger, 2002). Input methods for query-by-note systems include symbolic interfaces as well as both physical (MIDI) and virtual (Java based) keyboards. The disadvantage of this method is that it can only be used by people who have a background in music theory and that it can only account for Western music.

4.1.2 Audio-based approaches

A Query-by-Humming

The so-called query-by-humming and query-by-singing systems allow users to interact with their respective search engines via queries that are sung or hummed into a microphone (e.g. Birmingham et al., 2001; Haus and Pollastri, 2001). Until recent, researchers talked about *query-by-humming applications* and *hummed queries* or *sung queries* without making

a distinction between different vocal query methods that have been used. Moreover, the constraints these methods impose to the user have not been seriously investigated or taken into account. As a consequence, the act of so-called *humming* has been conceived of rather confusingly. Sometimes the terms *query-by-melody* (e.g. Shalev-Shwartz, 2002) or aural querying (e.g. Birmingham et al., 2001; Pardo et al., 2003) have been used as well.

A disadvantage of query-by-humming is that until now it has only been applicable when the audio data is stored in symbolic form such as MIDI. Although progress has been made over the last few years (e.g. De Mulder et al., 2005), transcribing audio signals into symbolic form is still problematic. Furthermore, this way of querying is of course limited to music genres with pronounced (singable) melodies.

B Query-by-Voice

As an alternative to the mixed used of terms, the generalized concept query-by-voice was introduced by Lesaffre et al. (2003) to describe any query produced by the voice and the vocal organs (i.e. vocal chords, tongue, lips, teeth). The idea of a query-by-voice system is that it takes any vocally produced query as input and compares it to a music database. The system then returns a ranked list of music closest to the input query.

C Query-by-Beat boxing

Beat boxing is a type of vocal percussion, where musicians use their lips, cheeks and throat to create different beats. It originated as an urban art form. The partakers of early eighties hip-hop culture could seldom afford beat machines, samplers or sound synthesizers. Generally, the musician imitates the sound of a real drum set or other percussion instrument, and there are no limits to the sounds that can be produced with the mouth. Currently, the use of beat boxing as a query mechanism (query-by-beat boxing) for both retrieval and browsing are being explored. Tools are being developed to retrieve music with a microphone by beat boxing (Kapur et al., 2004).

D Query-by-Tapping

The query-by-tapping method presented by Jang et al. (2001), requires a user input in the format of tapping on a microphone. The extracted duration of notes is then used to retrieve the intended song in the database. Since there is no singing or humming involved, no pitch information is used in the retrieval process. Although it is claimed that beat information is an effective means of retrieving the required song from a large music database, it was found that tapping clips longer than 15 seconds reduces system performance because most people cannot keep tempo precisely enough. For better performance, it is suggested that query-by-tapping should be combined with query-by-humming.

E Query-by-Example

These systems take pre-recorded music in the form of CD tracks or MP3 files as their query input (e.g. Haitsma and Kalker, 2002; Harb and Chen, 2003). A musical fragment is used

as a query and the result is a list of other musical pieces ranked by their content similarity (e.g. Tzanetakis, 2002).

To summarize the above approaches, table 5 shows the main query methods currently under research in the music information retrieval community. A distinction is made between verbal description, symbolic description, physical action and an excerpt as reference. Along with the query method categories some examples of query techniques are given.

TEXT-BASED APPROACHES		
Verbal description	Keyword searching	
	Alphabetic list selection	
	Hierarchical list selection	
Symbolic description	Notation of pitch or rhythm information	
AUDIO-BASED APPROACHES		
Physical action	Move along with the rhythm	
	Tap the beat	
	Vocally reproduce a melody	
	Vocally reproduce a rhythmic pattern	
	Instrumentally reproduce a melody	
	Instrumentallyreproduce a rhythmic pattern	
Reference excerpt	Prerecorded example	

Table 5: Overview of query methods under research.

4.2 Fieldwork

In addition to a literature-based survey of possible query methods a study has been set up in order to obtain a better understanding of user behaviour. In this study people were simply asked about their query methods for music search.

To start with, investigation in the field of music distribution was done in the form of individual face-to-face interviews asking open questions. This means that during a personal conversation in situ respondents gave verbal answers to the questions of the interviewer. Afterwards, the questionnaire was filled out by the interviewer. The target population consisted of sellers and customers of the music department of FNAC[73] shops in Ghent and Leuven[74]. There were four sellers and eleven customers involved. The outcome of the interviews was taken as a starting point for the design of a systematic survey[75].

It is important to remark that in FNAC shops there exits a well thought-out sales strategy that has been worked out in order to attract clients and build strong customer relations. Trained sellers provide information in such way that the majority of the customers rely on this service. So, people's main search method when visiting a FNAC music store is asking

[73] Fédération Nationale d'Achats des Cadres.
[74] Ghent and Leuven are medium-sized cities in Belgium. In 2001 Ghent had 224.685 inhabitants and Leuven had 88.581 inhabitants.
[75] An attempt was done to get statistical information on consumer behaviour from studies performed by FNAC. Whatever, I could not dispose of this inside information.

the person at the information desk. Therefore, the interview mainly considered the search methods used by the person at the information desk to help the clients and the kind of questions that are asked.

At the time the interviews took place (November 2001), the sellers at FNAC disposed of paper catalogues for searching information (e.g. Bielefelder Katalog Klassik for classical music) and computers. Notwithstanding the fact that by then, most catalogues were already available online, FNAC employees did not have access to the Internet at that time. The software they used had limited search possibilities, namely the CD barcode, reference number, CD title[76], performer or composer. In fact, they had to rely on their sales training combined with their personal experience and background.

From interviewing sellers we learned that:

- the majority of the customers turn to the sellers for advice;
- they seldom ask for a reference catalogue;
- CD sales are down because people copy and download music;
- customers tend to be difficult: when they decide to buy a CD or a DVD, they want to be sure they made the right choice;
- the right choice is music of lasting value (e.g. classical, expressing an enduring memory);
- there are three types of customers: those who want to buy music they have heard in a particular place or situation, those who look for music that is similar to a particular genre, style or mood and those who want to buy it as a present;
- people who know exactly what they want are a minority. These people use the shop's classification system (e.g. genre, artist).

How do people make clear what they are looking for?

- two methods are commonly used, specifically verbalization and singing;
- verbal descriptions are either direct (e.g. "I want the new album by x") or fuzzy (e.g. "I am looking for a particular CD and I know it has a red cover");
- comparison is also frequently used (e.g. "It sounds like …").

From interviewing subjects in the field of distribution we learn that:

- the majority of the customers come with the intention of buying a CD or DVD, and less than a quarter of them just want to look around;
- most people count on help and suggestions from the seller;
- the kind of information they use (in hierarchical order) is the name of the composer /artist, title, CD label, genre, instrument and period;
- people know the music they buy (in hierarchical order) from friends, going out, radio, television and club life.

[76] Only CD title, not titles of songs or tracks. A full title was needed; searching for keywords was not possible.

4.3 Survey of music search behaviour

Taking into account the information obtained from the above interviews, a questionnaire ❸ was designed and a pilot study was set up in order to avoid response bias in the survey that was to follow.

4.3.1 Pilot study

The participants in the pilot study were nine students in the musicology department at Ghent University who attend the course “Psychology of music”. The study was structured into two parts. First, a lecture/demonstration introduced the participants to the project goals and survey methodology. Then, they manually filled in a completion form including questions about their personal profile (e.g. gender, age), musical background (e.g. level of music education, activity), search methods (e.g. shop, Internet, library), information type (e.g. title, composer), web based method (e.g. text, sing) and Internet activities. The purpose was to get as many comments, suggestions and criticisms as possible from people who are acquainted with the field. The questionnaire and the outcome were discussed in the classroom. The feedback from the musicology students was intended to further refine the overall conception of the survey.

4.3.2 Set up, design and methodology

Using the knowledge of the pilot study, then, a self administering survey of users’ music search behaviour was set up and presented in the form of a web-based questionnaire. A simple CGI script that stores HTML form input to survey data files, which can easily be imported into statistical programs, was used. Participants were 49 students in Bachelor of Art Science at Ghent University. The survey was conducted in January 2002 in a university PC room during the course “Informatics applied to the arts”.

Each subject sat in front of a PC where the questionnaire was available online. Methodology and goals of the survey were verbally introduced by the experimenter. Answering the 18 questions took 15 minutes on the average. The survey consisted of two parts. In the first part personal questions were asked (e.g. age, gender). The second part focused on gathering information concerning musical background (e.g. education, methods for listening to music, listening time, musical education and genre preference), Internet activities (e.g. preferred searching methods). The types of questions asked were open, closed and quantitative questions.

4.3.3 Questions and findings

For analysis, the returned data were imported into the Statistical Package for the Social Sciences (SPSS version 10.0). Frequency, mean and standard deviations were calculated. In this section the results are summarized, tables and graphs are included in (❷ pp.7-25).

For ordinal data, that is for questions that involve ratings on a score from 0 (not) to 5 (very much), percentages have been calculated on the sum of ratings for each variable.

A Personal background

There were 49 respondents, all students in art science, aged between 18 and 35 years with an average of 20,6 years. Around three quarter of the subjects were female (77,6%) and approximately a quarter were male (22,4%).

B Music behaviour

Q # 1: How many hours per day do you listen to music?

Participants answered that they listen between 1 and 9 hours per day to music with 3,7 hours on the average.

Q # 2: Do you listen to music actively (attentively) or passively (music as a background)?

The way participants listen to music is more or less equally distributed: 49% say they listen actively and 51% say they listen passively.

Q # 3: Do you play an instrument?

A bit more than half of the participants do not play an instrument (55,1%).

Q # 4: What is the highest music education you have accomplished?

Almost half of the participants are not musical educated (44,9%). From the trained participants 42,9% went to a music school[77] and 12,2% learned music by self-study. No participants went to the conservatory.

Q # 5: How many CD's do you buy per year?

Participants buy between 0 and 250 (only one person) CD's per year, with 16,2 CD's on the average.

Q # 6: Which radio stations did you listen to last week for at least one hour?

Participants in the survey prefer listening to public[78] radio stations. The broadcasting they most frequently listen to is Studio Brussels (39,5%) followed by Klara (17,1%) and Radio 1 (15,8%). Some subjects (3,9%) say they do not listen to the radio and several (6,6%) prefer other radio stations than those in the list.

[77] The categories music school and conservatory should be regarded from the point of view of music education as it is thought in Belgium. Specialized music schools are secondary (music school), high school (conservatory) and academic (university). All these educational institutions mainly study classical music.

[78] At the time the survey was set up, there were two kinds of radio stations in the Flemish part of Belgium: public, commercial and local. There are 6 public stations (Radio 1, Radio 2, Klara, Donna, Studio Brussels and Radio Vlaanderen Internationaal) which were part of the Flemish Radio and Television, and approximately 330 local radio stations. The list included 5 public and 2 local radio stations.

Q # 7: Which genres did you listen to during the last week for at least one hour?

Youth culture music such as pop (22,9%) and rock (22,0%) are the most favoured genres. Nevertheless, jazz (19,5%) and classical (13,6%) are also quite popular genres among participants.

Q # 8: How do you get in contact with the music that you buy?

The radio (26,3%) is the first medium by which participants discover the music that they buy. Second means is recommendation by friends (23,0%), followed by finding it by chance (16,7%), read about it in the media (16,0%), heard while going out (13,0%) and the Internet (4,8%).

Q # 9: How do you search for music that you do not know?

In hierarchical order, the preferred methods for searching music is trying to deduce the title from the text (33,9%), finding additional information using the Internet (23,2%), asking a seller in a store (19,3%), search a music catalogue (8,9%) and sifting through play lists (8,6%).

Q #10: What kind of information do you use to find a piece of music?

The information types participants usually rely on to find a piece of music are performer (25,8%), title (22,3%), genre (21,8%), composer (15,9%), record label (6,1%), year/period (4,6%) and instrument (3,5%).

Q #11: Say you would like to buy a CD with the 5th Symphony of Beethoven, but you do not know the title or the composer. What kind of music description do you find suitable for making clear what you are looking for?

The kinds of music description features that participants are likely to use for describing the music they desire are above all genre (38,7%) and rhythm (22,9%). Instrument (13,2%), pitch (12,6%) and emotion (12,6%) are almost equally valued.

C Internet activities

Q #12: How many hours per week on the average are you online?

Participants spent between 1 and 20 hours per week online, with 4,8 hours on the average.

Q #13: Give a rating to following Internet activities according to the degree of your practice.

Most practiced Internet activities among participants are purposive search for information (37,4%)[79] and emailing (36,1%). Several participants make use of the Internet for recreation (14,7%) and chatting (9,8%). Very few draw on the Internet for buying things.

[79] Given that we deal with university students and that the survey was not anonymous, here, some political correctness might be involved.

Q #14: Assume that following search methods are available.
Give a rating according to the degree of what you think of their suitability for finding music using the Internet.

According to their assumed suitability Internet search methods are ranked as follows: textual information (e.g., title, composer, performer) (51,2%), a pre-recorded example (21,5%), singing or humming into a microphone (12,4%), staff notation (11,4%) and rhythm annotation (3,5%).

Q #15: What type of information would you like to retrieve, using an intelligent search system?

The top three of the information type that they would like to retrieve is the title (28,8%), composer (28,6%) and a sound file (22,0%). Less interest goes to retrieving a score (9,7%) or recording information (10,9%).

Q #16: Do you download music from the Internet?

Almost half of the subjects (49%) say that they download music from the Internet. However, knowing that this is an illegal practice might have influenced the objectivity of the answers.

Q #17: Do you buy less CD's because of the Internet?

20,4% of respondents said they buy less CD's in a shop because of the Internet.

Q #18: What is the main reason for not buying CD's online?

The number one reason why participants do not buy music online is because they find it unsafe (28,6%). Almost a quarter of the subjects (24,5%) gave other reasons than those in the list, such as "it is much more fun to wander around in a music store", "I miss the contact with the seller" and "the charm of ferreting around in drawers gets lost". More discouraging factors are that it is too technical (22,4%) and that they cannot see and touch the product in advance (20,4%). Only two participants reported that they find it too expensive.

4.3.4 User groups

It was checked whether patterns of search behaviour differ by user groups. Some remarkable differences are found related to gender and expertise level. For defining the expertise level a new binary variable was created based on the way participants listen to music (actively, passively), whether they play an instrument or not and on their music education. The reported results are based on the sum of rating scores 4 and 5.

Detailed bar graphs for each score (from 0 to 5) and stacked bars per user group are included in (❷ pp.26-49).

A Differences within gender

General search methods (Q#9)

Male participants (27,3%) make more use of the Internet than female participants (15,8%) to search for music information. Women (15,8%) on the other hand explain their needs to sellers in a store whereas men say that they never do this.

Information type (Q#10)

Among female participants around three quarter (76,3%) use the performer as information type. That is more than twice as much as men do (36,4%). More than half of the male participants use genre (54,5%) and composer (45,5%).

Music description (Q#11)

While about a quarter of the participants finds rhythm a suitable feature for describing music, there is little difference found by gender. More men (18,2%) than women (5,3%) seem to find emotion a suitable characteristic. The score for genre is remarkably higher for women (68,4%) than for men (18,2%).

Internet activities (Q#13)

The pattern of Internet practice across gender groups reflects that more female participants (73,7%) use the Internet for purposive information search than men (45,5%). Women more often email (68,4%) than men (45,5%).

Internet search method (Q#14)

Male participants are not at all interested (0%) in tapping the rhythm or uploading a score as Internet search methods. Besides, over 70% of men and women as well give a high score for the textual search method. They show interest in uploading a music fragment as example (men 18,2%; women 28,9%) and singing (men 18,2%; women 5,3%).

Internet feedback (Q#15)

Here there is some noticeable difference between men and women. The most wanted feedback information is the title (men 81,8%; women 94,7%) and composer/performer (men 90,9%; women 92,1%). The next information desired is an example in the form of an audio fragment (men 63,6%; women 57,9%).

B Differences within expertise level

General search method (Q#9)

Expert participants do not use catalogues or play lists (0%) when looking for music, they rather explain things to a seller in a shop (19%) or use the Internet (19%). Some novice participants, on the other hand, use catalogues (10,7%) and play lists (7,1%).

Information type (Q#10)

Experts (38,1%) make more use of “composer” as information type than novices (21,4%). Novices then are more interested in using “performer” (75%) than experts do (57,1%).

Music description (Q#11)

Novices (67,9%) clearly find genre more suitable than experts (42,9%) for describing music.

They also give higher rating scores for instrument (novices 14,3%; experts 4,8%) and emotion (novices 10,7%; experts 4,8%). Experts prefer rhythm and pitch.

Internet activities (Q#13)
Novices indulge more in recreational activities than experts do (novices 17,9%; experts 4,8%). Experts are busier with emailing (novices 53,6%; experts 76,2%) and chatting (novices 3,6%; experts 19,0%).

Internet search method (Q#14)
Textual search outranks by far the other search methods (novices 89,3%; experts 81,0%). Although the difference is rather small, experts show more interest in using an example, a score and singing, but they are not likely to tap the rhythm.

Internet feedback (Q#15)
Novices and experts both clearly prefer title and composer/performer as feedback information. Experts show more interest in retrieving an example (novices 50,6%; experts 71,4%) or a score (novices 3,6%; experts 28,6%).

4.3.5 Comparison with other findings

Even though this is a small study with narrow scope, the findings are generally in agreement with those reported by Lee (2004). Lee's study on music information needs, uses and search behaviours comprises some questions that are similar than the ones in my survey. The questionnaire in Lee's survey is larger and has been conducted two years later than the one reported above. In both studies, the participants are recruited from a university population, but besides students, Lee also included staff. The findings of Lee are based on 427 user responses. From both studies it is clear that the top ranked music genres are pop and rock, in Lee (2004) this is followed by classical and alternative, and in my study this is followed by jazz and classical (alternative was not an option). The studies also agree on preferred search methods. One of the remarkable findings is that users state explicitly that traditional meta-data such as title, composer or performer continue to play an important role in the music information-seeking process. Finally, from both surveys it can be concluded that participants are likely to rely on recommendations from other people in their search process.

4.3.6 Conclusion

It is clear that, as a way of getting insight into the user's query methods for music search and retrieval, interviews and surveys are but limited methods. Moreover, the range of locations in which the interviews took place was narrow and the amount of subjects interviewed was small. Nevertheless, interviews and surveys are indispensable at the initial stage of the research process. They yield the most general features of the landscape they are designed to reveal. Therefore, conclusions from this small study cannot be generalized but are indicative for further research.

It seems that people visiting music stores mainly trust on the knowledge of the person behind the information desk. At this stage of the investigation it has been shown that the impact of personal (physical) contact should not be underestimated.

The study also revealed that due to the narrow scope of the sample of respondents in the survey the participants were homogeneous in their vocation (they were all university students in the department of Art Science at Ghent University). Of course, this skews the sample considerable but it does present us with an interesting opportunity. Participants were all familiar with the Internet and had a broad interest in arts, which is in accordance with characteristics that could be expected from potential users. Besides, the respondents were not at all homogeneous in their musical background. So, as far as music is concerned, the sample is more representative of a broader public.

Summary of the main findings:

- people often search for similar pieces of music (sounds like), because they prefer a certain style or because they are in a particular mood or putting together a program for a particular occasion;
- the preferred methods when they know the music they want are asking the seller in a store or browsing the Internet.
- the most frequently used information types are: title and genre;
- when they do not know what they search they find genre and rhythm the most suitable features;
- making use of the Internet, textual search or giving an example would be the favourite methods;
- in hierarchical order, what they would like to retrieve from a database are: title, composer and a musical fragment.

At the time that this explorative investigation was conducted, most of the subjects involved in the interview and survey never heard before of “vocal querying”. They found the idea of having such a system at their disposal very appealing. Some of them were even so enthusiastic that they volunteered to take part in the query-by-voice experiment that is described in the next section.

4.4 Experiment on spontaneous vocal query behaviour

This section contains the description and results of an experiment (see original articles I, III and V) in which spontaneous behaviour while performing vocal queries has been investigated. The study was conducted in 2002 at IPEM, Ghent University and aimed at analyzing the musical responses from subjects in a query-by-voice (QbV) experiment that allows *maximal freedom* for the participant. In the first chapter of this thesis it was noted that until recently the role of the user in music information retrieval systems has often been neglected. However, it is has become clear that the user's musical memory and capabilities of verbally describing and imitating music are determining factors for implementing efficient

systems. Little is known about (1) the effect of memory on music recall, (2) performance skills in reproducing an audio query and (3) what typically sung patterns are.

The goal of this study is to determine characteristics of naturally expressed vocal and (textual) queries. It has been set up to investigate (1) the effect of long-term and short-term memory on spontaneous vocal querying behaviour, (2) the query methods people prefer and (3) differences between subject groups defined by music education, age and gender. Important information in favour of the development of systems for content-based access to digital collections of music has been derived.

Performing a vocal query from memory is an experience that is a direct result of memorization. A MIR system must deal with significant errors due to human inaccuracies in both recall and performance. The user may not correctly recall a theme, or may not have the vocal skills to produce a good imitation of a tune. In many experiments (e.g. Lee 2004) people's recognition of well-known melodies has been investigated. However, we must not only know how people recognize melodies but also what is the most comfortable way for a user to reproduce a tune. The query-by-voice experiment provided the structure for the material that can be expected as input to systems for music search and retrieval that allow vocal querying.

4.4.1 Previous experimental studies

In this section, a scan of the literature on melodic query experiments is given. It reveals interesting background on datasets, error models, influence of memory and user implications.

A Datasets

A query-by-voice driven music information retrieval system will typically consist of a database with target files. These are the files which users search through and retrieve using vocal queries. For research purposes, small databases with target files are generated. Query files are to be matched with these target files. Researchers usually work with databases containing small sets of queries obtained from small-scale experimental studies. These sets are used as test-beds for the development of reliable query-by-voice systems. Most of them, however, are rather limited in scope and do not take into account spontaneous user behaviours. Research groups have developed their own particular sets typically consisting of user provided sung melodies following strict rules (e.g. Mc Nab, 1996; Kosugi, 2001). For experimentation purposes, some teams also generate synthetic queries (e.g. Meek and Birmingham, 2002) or test the retrieval performance of their system with fragments extracted from MIDI files in an experimental database (e.g. Doraisamy and Rüger, 2002). Query sets with sung tunes, however, are far from elaborate and there is a need for thorough studies related to the user's singing preferences and habits (Downie, 2003). I am not aware of experiments described in the literature that directly investigate the *spontaneous* behaviours and query method preferences of users.

The literature on vocal queries reflects a variety of topics that are more or less related to user behaviour. However, many researchers do not seem to consider it an important aspect of their research. In some studies, the query methods used are not even clearly reported. The generalized use of the word “humming” leads to confusion about how to interpret experimental studies. Lu et al. (2001), for example, describe their query test set as hummed files, but they mention that their participants were using lyrics and syllables. In what follows, on overview of previous studies is given, following three aspects of experimental design namely (1) the development of error models, (2) the effect of memory and (3) constraints imposed on the user.

B Development of error models

Error models deal with query imprecision as a consequence of human errors such as singing out of tune, wrong tempo, wrong notes and insertions or deletions of notes. McNab et al. (1996) were the first to investigate how correctly people sing well-known tunes, focusing on melodic aspects such as melodic contour and intervals. A lot of groups have been using the findings of McNab for performance evaluation of new models (e.g. Carré et al., 2001). Lindsay (1996) described a psychological experiment he set up to study errors in musical performance and the natural musical response of people that sing from memory. On the whole, however, the attention for error models is rather limited.

C Effect of memory

Both McNab et al. (1996) and Lindsay (1996) addressed short-term memory for music. The subjects in the McNab experiment, for example, practiced the singing of each song in order to refresh their memories and to decide what key to sing in. In Lindsay’s study, subjects were requested to repeat specifically designed five note phrases that were presented over headphones. Although the output should be representative for what can be expected as input to a query-by-voice system, there are some very clear constraints. Only a small number of participants were involved, the singing was rather unnatural and the experiments relied on direct memory responses. However, one may assume that a great percentage of users of a query-by-voice system will rely on long-term memory for melodies, rather than on short-term memory. In a small-scale experiment Meek and Birmingham (2002) ask five subjects to sing passages from four well-known songs, twice from memory and twice after hearing a piano rendition of the passage. The queries served to test the ability of their model but the set was too small to investigate the impact of memory as well. In studying long-term memory of absolute pitch, Levitin (1994) asked subjects to reproduce the pitch of the first few notes of their favourite rock ‘n’ roll song from memory. His findings show that subjects may be rather good at this task, indicating that their long-term memory preserved a stable and accurate representation of musical pitch. Levitin concluded that the subjects heard the correct pitches in their heads but couldn’t always reproduce them. So, if someone was a little off-key, it was probably due to his or her singing ability. In short, it is remarkable

that within the music information retrieval community very few studies investigate the impact of long-term memory.

D Constraints imposed on the user

The most common constraint that is imposed on users concerns the query method. In all but one of the experiments and systems for which the input method is specified in the literature, subjects are asked to sing syllables. The exception is the software system Tuneserver developed in 1997 by Prechelt and Typke (2001) that required whistled input, which was then transformed into Parsons's Code for further processing. This was the situation at the time the query-by-voice experiment has been conducted. From 2002 on the Tuneserver search engine was known as Melodyhound and from then on users could contribute new melodies. However, melodies were not editable and only the composer name, title and Parson Code could be stored. In 2004 Melodyhound was adapted once more, forming the basis of Musipedia[80].

The reason for why the use of specific syllables is often required is that it facilitates the segmentation of the sung melodies. Methods like singing lyrics (Lu et al. 2001) or the use of less clearly articulated syllables (Kosugi et al., 2000) are explicitly rejected because they are too difficult to process. Sometimes the input possibility has been strictly limited to one single type of syllable, e.g. [ba] (Brøndsted et al., 2001), [ta] (Kosugi et al., 2000) or [fa] (Pauws, 2002). Other systems suggest a range of syllables such as [la] or [da] (McNab, 1996), and [na], [ta] or [pa] (Haus and Pollastri, 2001).

Apart from imposed syllables other restrictive factors towards the users have been introduced. In Lindsay's (1996) experiment, for example, subjects were asked to watch a loudness meter on the computer screen. They had to sing as loud as they could, without letting the meter go into red. Kosugi et al. (2000) asked their subjects to sing in accord with the beats of a metronome, so that they could keep a constant tempo.

Actually, these constraints guarantee ad hoc the best performance of the system and hide the limitations of it. The question to be asked is if any constraints should be imposed at all? They might enhance the system performance but certainly not the user performance. I believe that it is important to look at the abilities and preferences of the users. Assuming that the best performance occurs when users can act in a natural way, what then is the vocal query method that users would prefer? Will they hum, sing in syllables, or use lyrics and whistling instead?

In short, experimenters often assume that users may somehow be willing to accept certain constraints such as using one specific syllable. It is argued that constraints on vocal querying are compensated by a higher reliability of the system. However, several of these assumptions have not been justified. We simply do not know which vocal querying method

[80] See 1.2.1: Digital music databases, and http://www.musipedia.org/

users would prefer, and we do not know how frustrated users may become by having such constraints imposed. Furthermore, flexible systems might be more difficult to develop, but as they do not exist yet, it has not been proven that their performance would be worse.

4.4.2 Music stimuli

The limitations of existing systems thus motivate a more elaborate study of spontaneous user behaviour. In view of such a study a set of target audio files was collected. It contained 30 pieces of music from the MAMI target database that consists of 160 pieces of music[81] ❸. The MAMI target database differs from existing databases because it: (1) consists of entire pieces of music stored in .wav format, (2) includes polyphonic music (multiple voices) and (3) has a representative genre structure.

A The MAMI database

The MAMI database consists of two sets of music files: the *target* files and the *query* files. Target files are the entire music pieces to be retrieved from a database. Query files are vocal reproductions of fragments of target files. These are a kind of audio files that users could give as input to a music information retrieval system that allows vocal querying. In such a system query files should be matched with target files. The MAMI database is conceived as a test bed and is not intended to be an operational site model database. The design of the MAMI database, however, has enhanced know how applicable to more extensive site model databases.

Prior to the putting together of the MAMI target database two questions had to be answered: what kind of pieces should be included and how much examples were needed. The subdivision of music into genres and/or styles is most subjective and complex, especially nowadays because artists are crossing so many genres. For that reason it was decided to use a criterion based on the statistics provided by the International Federation of the Phonographic Industry[82] (IFPI). This statistics (see table 6) is according to the genre sales of music in Belgium in 2000.

For creating the MAMI database one example per percentage in the list was chosen. This would imply, however that for most genres, the database would contain only few examples. Ten examples per category are considered as the absolute minimum for relevant testing. Therefore, the amount of music pieces for genres with lower percentages has been complemented. This resulted in a database containing 160 pieces. The query files in the MAMI database resulted from the experiment on spontaneous vocal query behaviour.

[81] A complete list of the 160 musical pieces in the MAMI project database is included in the electronic appendix. Permission to use this collection for research purposes has been given by SABAM, the Belgian author rights association.

[82] http://www.ifpi.org

Sales %	Genre
46	Pop/MOR/easy listening
24	Rock/heavy metal
7	R&B/urban (incl. disco, funk, fusion, motown, reggae and soul
5	Oldies
4	Soundtracks
3	Dance (incl. techno, house, jungle
3	Classical
3	Rap/hip hop
2	World/ethnic
2	Children's music/spoken word/comedy
2	New age

Table 6: Genre sales in Belgium in 2000 according to IFPI.

B The test set

The selection of 30 pieces for the query-by-voice experiment was further based on the following criteria: (1) a high percentage of subjects should be familiar with the piece, (2) the pieces should be relatively easy to remember and to imitate (3) the selection should reflect the heterogeneous musical landscape. The list contains different genres: besides popular music songs, ranging from chanson to heavy metal, well-known Flemish children songs and classical music are included as well.

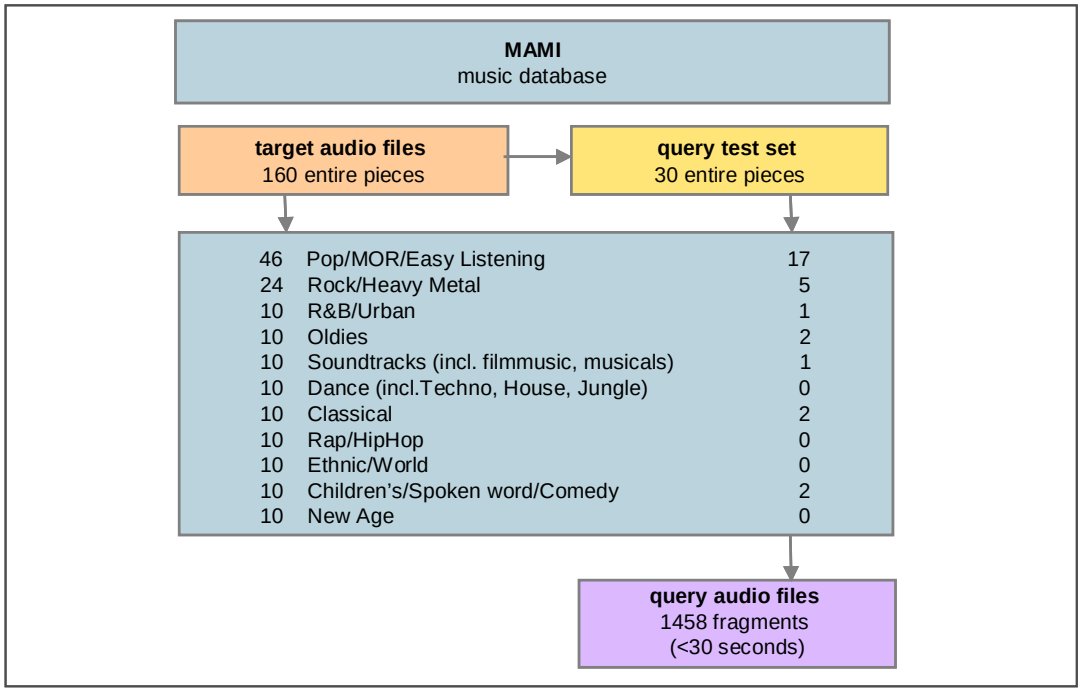


Figure 11: The MAMI music database.

Figure 11 shows the composition of the MAMI music database: a set with 160 target files, a query test set that consists of 30 pieces selected from the target files and a set with 1458

vocal queries generated by the participants in the experiment. The amount of pieces per genre category is indicated as well.

4.4.3 Set up, method and procedure

In this part, a description is given of the participants, software, physical set up and introductory phase.

A People

The 72 subjects were recruited from students and staff members of Ghent University. The experiment was anonymous. Participants did not know in advance that they were participating in a study that involved the imitation of music.

B Software

The application used for conducting the experiment was implemented by K. Tanghe. It is a Win32 console application using MSVC++ 6.0 and was tested on Windows98SE and Windows2000 systems[83].

C Physical set up

The participants were accompanied to a separate room where they could perform the experiment in isolation. The experimenter gave the subjects an outline of the general procedure and handed them a text with a detailed description. They were seated at a desk with a PC under normal office conditions. The PC running Windows98SE was installed on a table in a room with no special sound isolation. The door was closed but environmental noise was still slightly audible. The idea was finding a compromise between recording in a natural environment and giving the user privacy. A Labtec Verse 514 low-budget microphone was connected to the microphone input of the computer's sound card (a Yamaha DS1 PCI standard sound card). Music stimuli were presented through headphones that were connected to the sound card output. The subjects were sitting in front of the computer monitor and the microphone, which was fixed on top of the monitor keeping a distance of about 30 centimetres from the subject's head. A text on the screen, similar to the printed text that was handed out, guided them through the procedure. The experiment took about 35 minutes.

D Info on the subject

Before starting the experiment, subjects were asked to fill out a questionnaire, first gathering information about age, gender and musical training and second on their familiarity with the music. To start with, each subject was automatically given a unique ID that was used for labelling the generated information files. A general profile of the user was obtained by asking following questions: "What is your age?", "Are you male or female?", "How many hours a week do you actively listen to music?", "Do you play a musical instrument?" and if

[83] Standard C++ was used as much as possible. The used libraries PortAudio (Bencina et al. 2001-) and libsndfile (de Castro Lopo 1999-) are cross platform.

so ”How many hours a week?" and "What is your highest level of musical education?". This information was written to a profile file, together with the name of this file and the name of the file that is used as a log of the experiment for this subject.

E Familiarity with the music

Then, information about the subject's knowledge of the various pieces used in the experiment was gathered. All pieces were specified in the configuration file and gathered in a set called Set1. For Set1, thirty pieces were used and for each subject this same set is reorganized at random. Starting with the first piece of Set1, its title was shown textually together with (between brackets) its composer, performer or other brief specification and the subject was asked whether he/she would be able to imitate a fragment of that piece. If the answer was "yes", the piece was added to a set of "known, imitable" pieces (Set3)[84]. If the answer was "no", the subject was offered multiple possibilities to choose from that are: “I do not know it" (piece is added to Set4K), "(I think) I know it, but I cannot remember how it sounds" (piece is added to Set4R) or "I really do know it, but I cannot imitate it" (piece is added to Set6). Then the second piece was presented and so on. This iteration over the pieces in Set1 stopped when all pieces were handled, or as soon as Set3, Set4R and Set4K all contained enough pieces. The results of this categorization were also written to a log file where each of the above four sets shows the ID's of the pieces that were added to it. In order to obtain enough diversity in the subject's knowledge about the pieces, thirty pieces were used for Set1 and the aim was to get ten pieces for Set3 (the final number depended on the subject's knowledge of the pieces). This was a trade-off between getting enough recordings of the same piece by different subjects and getting recordings for enough different pieces. To reduce the total time of the experiment, only two pieces from Set4R and two from Set4K were used. An overview of the different sets of pieces is given in Table 7.

	Set1	Fixed set of pieces from MAMI target database	
	Set3	Known and imitable	
	Set4K	Not known	
	Set4R	Thought to be known, but not remembered	
	Set5	Fixed fragment to be imitated in different ways	
	Set6	Known, but not imitable	

Table 7: Sets with musical stimuli in the query-by-voice experiment.

4.4.4 Experiment part one

The first experimental task focused on the reproduction of musical pieces from *long-term memory*. It gave information on how participants prefer to imitate musical pieces and which parts are imitated without having heard the piece immediately prior to reproduction.

[84] Set2 is not included in the final set up of the experiment. In a preliminary stage, the experiment was set up in order to offer different test sets distinguishing between people who indicated that they could imitate the melody for less of more than ten titles out of thirty. After testing it was found that this distinction would not add value to the experiment.

Subjects were asked to perform a vocal query for ten titles previously indicated as known and imitable. The pieces (Set3, known and imitable pieces) were presented one by one. Participants only saw the title, composer, performer or other brief specification, so there was no sound. On the screen and in the accompanying text, four vocal query methods were suggested and described: humming (making sound with the lips closed), singing lyrics (singing the text of the song), singing syllables (singing any suitable nonsense syllable) and whistling (to utter a shrill clear sound by blowing or drawing air through the puckered lips). It was made clear to them that these four methods could be combined and that the choice of fragment and voice was totally free, but limited to 30 seconds. Subjects could choose when to start and stop the recording by pressing the enter key.

After each performance, the subjects were offered a second chance to produce another query for the same piece. This could be done when the subject was not satisfied with the first recording, if he/she wanted to try another method, or wanted to perform another fragment from the same piece. After the vocal imitation(s) and before moving on to the next title, the subject was invited to describe the piece in another way. The choices were the following: make a recording (using a method other than the previous ones), provide a textual description of the piece or describe an alternative query method using typed text. Each of these choices could be made at most once. The iteration over the pieces stopped if the predefined number of ten pieces was reached (or if all pieces were gone through if there were not enough pieces in Set3 to reach this number). All recordings were stored in WAV files (44.1 kHz, 16-bit mono) and the textual inputs were stored in the log file. A list of the pieces that were presented, the names of the files that were recorded for each piece and the choices made by the user are also stored in the log file. In the experiment, a number of pieces that were supposed to be familiar to a large majority of the participants had been included. This had to guarantee Set3 to contain a sufficient number of queries to obtain relevant results.

4.4.5 Experiment part two

The second task focused on imitations from *short-term memory*, after listening to the piece shortly prior to reproduction. It was designed to investigate differences with imitations from long-term memory, and to get an idea of which parts of a piece tend to "stick" after having heard the entire piece. The subject had to listen to four entire pieces of music, previously indicated as unknown or not possible to recall. They were asked whether they knew the piece and if so to perform a vocal query following the same instructions of the first task. Again, the subject had the option to make a second recording. The pieces for this task were selected from Set4K (not known) pieces. If there were less than four songs in the list, pieces indicated as imitable in the preparatory stage, but not included in the first task were used. Again, recordings were written to .wav files and all used pieces and the responses were logged in the log file.

4.4.6 Experiment part three

The third task was meant to gather information on *differences in performances* of the same melody by various subjects using different query methods (male/female differences, pitch fluctuations, use of vibrato, accuracy of imitation, ability to whistle a melody…). The same short musical fragment (Set5) was played back for all subjects. The subject could listen to it up to three times and the first time he/she was asked whether he/she knew the fragment. Then the subject was asked to imitate it in four different ways that are singing with words (the text being shown on the screen), singing using the syllable “ta", humming and whistling. Table 8 gives an overview of the major steps in the query-by-voice experiment.

	PREPARATORY STAGE
	Collecting info on the subject Collecting info on familiarity with the music
	EXPERIMENT PARTS
	Imitating known pieces without hearing them first Imitating pieces afther hearing them entirely Imitating a fixed fragment in four different ways

Table 8: Overview of the query-by-voice experimental procedure.

4.4.7 Output

For each participant the experiment generated a profile file, a log file and a set of sound files.

A Profile file

For each subject a profile file was generated that contains the following information: a unique ID for the subject, age and gender, the number of hours a week the subject actively listens to music, whether the subject plays an instrument or not (if so, the number of hours a week is added), the highest level of musical education (no musical education, music academy or music conservatory) and paths to the log file and profile file.

B Log file

A log file was generated that keeps track of the course of an experiment for a specific subject (the subject ID is stored in the file name). Each log file consists of four parts, corresponding to the flow of the experiment as described above.

C Query sound files

These are the files that have been analysed to investigate user preferences and that can be used for the evaluation of audio feature extraction algorithms. All generated query sound files are named consistently in a way that allows identification of that stage of the experiment at which they were generated. There are sound files for the different query-by-voice trials in part one of the experiment (reproducing known pieces from long-term memory), the first recording and possibly one or two extra recordings (see supra) and similarly for the second part (producing queries after listening to the piece). In part three, all

subjects were asked to imitate a fragment heard using several methods: by singing words, singing syllables, humming and whistling (the latter not being applicable if the subject indicated that she/he could not whistle). This gives an extra three or four sound files per subject. A last sound file contains some spontaneous comments (if any) by the subject, recorded after the experiment has ended (mainly noise or laughter).

4.4.8 Analysis strategy

This section regards the analysis of data gathered in part one and two of the experiment that focused on spontaneous vocal querying behaviour. Part three of the experiment regards fixed query methods and its queries are therefore not included in this analysis.

Figure 12 gives a general overview of the structure of the experiment on spontaneous vocal query behaviour.

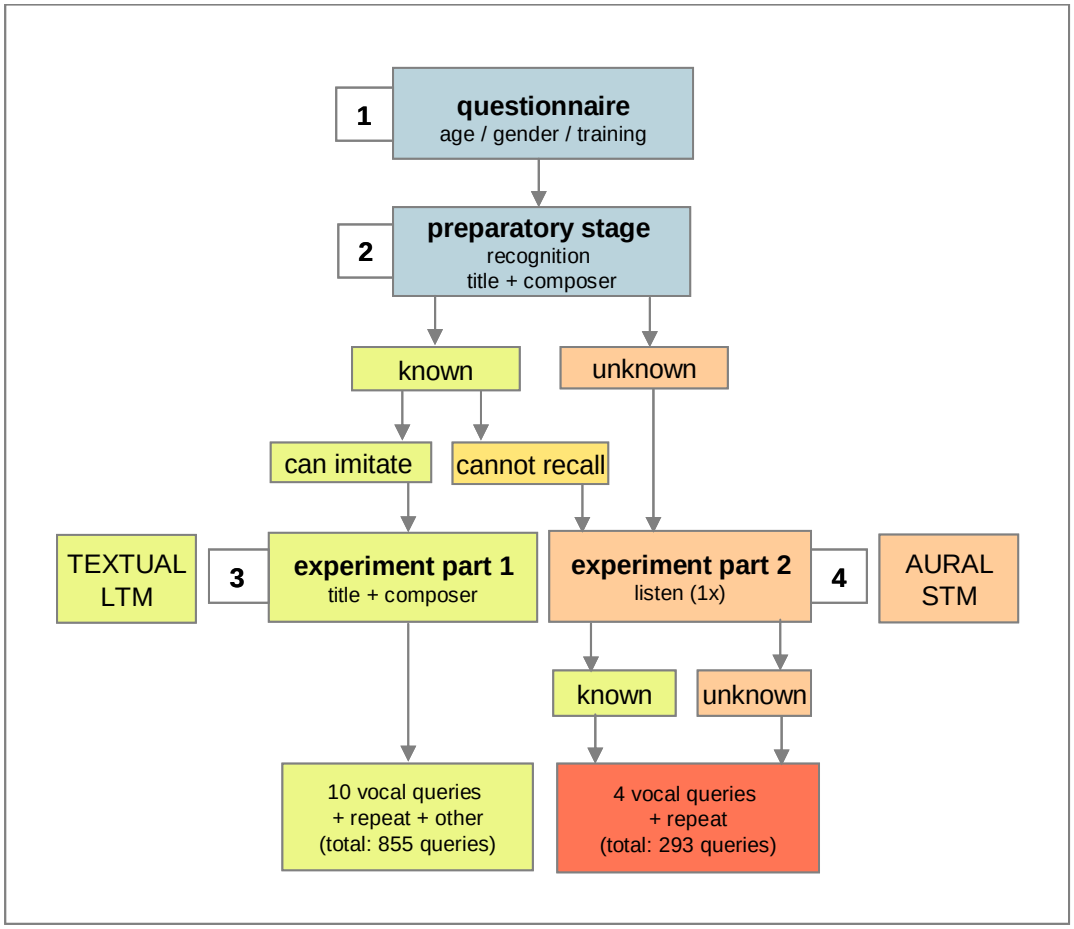


Figure 12: General overview of the experiment on spontaneous vocal query behaviour.

The experiment includes (1) a short questionnaire asking background information on the subjects, (2) the preparatory stage that collects information on participants' familiarity with the 30 pieces of music in the stimuli set, (3) experiment part one that for each participant

collects around ten vocal queries produced from long-term memory and (4) experiment part two that for each participant collects a set of four vocal queries produced from short-term memory. Participants were also offered the possibility to repeat a vocal query or to suggest other query methods. This generated a set of 855 queries for experiment part one and a set of 293 queries for experiment part two. (Examples of queries in ❸).

Two conditions were investigated: reproducing music purely from long-term memory, and imitation after listening to the piece. To begin with, a description is given of the general and segment specific aspects of the queries. Then attention is paid to syllabic queries and syllable structure. Subsequently, analysis by subjects is considered. After that aspects related to the stimuli are looked at. Finally, the effects of memory are discussed.

Part one (queries from long-term memory) and part two (queries from short-term memory) of the experiment together generated 1458 query sound files, 73 user profile files and 73 log files that are freely available at the MAMI site[85] for research purposes. The output data in the profile files were stored in an Excel file together with manual annotations for the sound files (i.e. vocal queries). SPSS statistical analysis was used for calculating distributions, correlations and ANOVA. The analysis deals with a large data set of queries, and the strategy therefore was oriented onto finding more general trends and characteristics rather than carrying out a detailed analysis of, for example issues of pitch and rhythm.

The features investigated are related to the beginning and ending time of the query, the vocal query method (lyrics, humming, whistling…), performance style (whether the query focused on melody, rhythm…), target similarity (whether the query was similar to part of the original), and syllabic structure (what syllables were used). Some of these aspects are rather subjective, such as the assessment of the degree of similarity between query and the original, and results are therefore rather tentative. But other aspects, such as temporal properties and vocal query method allow a much more precise annotation.

At the first level of analysis, the queries where segmented according to the methods used. Vocal queries may contain a mix of different query methods such as humming, singing syllables, singing lyrics and whistling, as stipulated in the guidelines for the subject. In addition to those methods percussion (such as tapping along with the drum) and comments (spoken comments made by the subjects while performing a query) was found. Studying these methods in more detail requires segmentation into homogeneous parts, and this resulted in a set of 2114 segments.

In the next sections, some further methodological issues, associated with this analysis strategy, will be addressed in more detail. In what follows, first a global view on the characteristics of the music queries is given. General features are addressed using the original set of 1148 queries in their entirety, and the features of the segment-based 2114

[85] http://www.ipem.UGent.be/mami/public/Data/QbVExperiment

query-set are discussed as well. Finally the effects of long- and short-term memory are considered.

4.4.9 General aspects of the queries

General features are addressed using the original set of 1148 non-segmented queries.

A Beginning and length of the queries

Seventy-eight queries (6.79% of the total) exceeded the 30-second limit of the recording time. In these cases the recording was stopped and the computer warned the users. Twenty-six recorded fragments (2.26%) contained no query at all, as the subjects accidentally or deliberately stopped the recording before producing any kind of query. These files were not taken into account in the statistical analysis.

The average starting time of the query occurred 634 ms after the subjects started the recording, with 99.3% starting within 2 seconds. The mean length is 14.04 seconds, but the distribution is asymmetric, peaking towards 6 seconds and then slowly diminishing towards the maximum allowed length of 30 seconds, with a peak roughly between 5 and 15 seconds. However, large differences occur between subjects, with personal averages ranging from less than 5 seconds to close to the maximum recording time of 30 seconds. Also the average starting time shows huge differences between subjects, with personal averages varying roughly between 150 and 1500 ms. (❷ p.50).

B Number of segments, based on query methods

About 60% of the queries consisted of one homogenous query method and have been treated as one single segment. Other queries contained different methods, as well as changes from one method to the other and back (e.g. lyrics, whistling, and back to lyrics). The maximum number of segments observed in one single query was 12, but 97.8% of the queries contained a maximum of six segments. The distribution of the queries according to the number of segments is shown in (❷ p.50).

4.4.10 Segment specific aspects of the queries

For segment-specific aspects the set of 2114 segments has been used.

A Length of the segments

As with the analysis of the queries as a whole, the timing of the segments has been investigated as well. As for the duration of the full queries, a similar asymmetric distribution was found. 50% of the segments are shorter than 8 seconds and 75% is shorter than 15 seconds. The distribution of the segment lengths is shown in (❷ p.51).

B Query method

Table 9 gives an overview of the different query methods in terms of the number of segments and total time within the output (see pie charts in ❷ p.52). It is clear that the two methods standing out are singing on text and singing on syllables. Together they account

for 83.4% of all queries, and 78.2% of the total time. The frequency of text segments is higher, but the syllabic method is more prominent within the total time, which shows that syllabic segments in general are longer. Among the other methods, whistling is prominent, particularly because the average duration of whistled segments is rather long (with a mean length of 14.63 seconds, almost double the average). Thus, although only 8.6% of the segments are whistled, they take up 17.1% of the total time. By contrast, 8.1% of the segments contain humming, percussion, or comments, but these methods together take up only 4.5% of the total time.

Query method	N segments	Segments %	Total time ms.	Total time %
Text	926	45.6	5558959	37.4
Syllabic	766	37.8	6056644	40.8
Whistle	174	8.6	2544864	17.1
Hum	101	5.0	541815	3.6
Comment	42	2.1	65108	0.4
Percussion	20	1.0	77394	0.5

Table 9: QbV: Occurrence of different query methods.
The occurrence is hierarchically ordered following the number of segments.

C Performance style

The annotation of performance style was aimed at distinguishing between melodic, rhythmic and intermediate performances. A performance style was considered to be melodic when a clear succession of different pitches, or melodic intervals, could be observed. A segment was labelled as rhythmic when no clear pitch intervals were noticed (such as in a spoken text or a percussive sequence). An intermediate category was necessary in order to classify segments in which a sense of pitch is present, but without a clear melody (e.g. using a reciting tone). The distribution of these performance styles is given in (❷ p.53). Clearly, segments with a melodic content are dominant, accounting for 72.6% of the total number, and 76.7% of the total time. Intermediate and particularly rhythmic performance styles are much scarcer and in general also shorter than melodic segments. Table 10 summarizes the occurrence of performance styles.

Performance	N segments	Segments %	Total time ms.	% total time
Melodic	1502	72.6	11969478	76.7
Intermediate	469	22.7	3264400	20.9
Rhythmic	98	4.7	373249	2.4

Table 10: Occurrence of performance styles.
The occurrence is hierarchically ordered following the number of segments.

D Similarity between segments and targets

To each of the segments, a relative similarity rating was given on a six-point scale, ranging from 0 (not recognizable) to 5 (sounds similar). The estimation of the degree of similarity between query segment and target is subjective to a degree. Similarity, moreover, was

focused on melodic and rhythmic properties and aspects of timbre or use of lyrics was neglected. Due to the fact that queries may contain more than one segment, an overall similarity rating for the queries, had to be defined. Each of the segment similarities was multiplied by the segment duration; these values were then summed and finally divided by the total query length. As a matter of fact, providing a scientifically more reliable similarity rating would necessitate an additional rating-experiment, which would be extremely time-consuming in view of the large amount of data. Therefore the similarities mentioned in this study must be seen as a rough judgment carried out by musical experts. Due to the subjectivity of the numbers similarity measures are only used to compare large sets of data: to compare the efficiency of the different methods, the performance of different groups of users and the effects of differences in memory recall.

4.4.11 Syllabic queries

Syllabic queries form an important part of the query dataset. Therefore, an essential part of the manual annotation work carried out was related to the analysis of syllables. Syllables are considered to be non-semantic vocal events, containing a vowel, which is preceded and/or followed by a consonant or a complex of consonants. After segmentation of the queries following query method it was found that 766 segments out of 2114 had syllabic content. These contain a total number of 14748 syllabic units, and additionally some 104 events such as “tr”, “pf”, and tongue clicks that were not considered as syllables.

In spontaneous production of syllables, the sounds are not strictly determined. They can slightly change between two successive syllables of the same type. This is most prominent for the vowels, but it also applies to certain consonants (e.g. t-d). For further processing a categorization was introduced that reflects the normal practice in Dutch, this being the native language of the participants. This analysis yields 179 different syllables, of which 45 occurred only once. In what follows, syllables and syllable parts are transcribed using the Speech Assessment Methods Phonetic Alphabet (SAMPA)[86], which is a machine-readable variant on the International Phonetic Alphabet. To facilitate reading an example of each of the vowel categories used is given, illustrated with an English word that reflects the ‘typical’ sound of the category (table 11).

Vowel category	[a]	[@]	[E]	[e:]	[i]	[o]	[u]	[y]
Sounds like	l**a**ger	th**e**	s**e**x	c**a**re	t**i**psy	r**oa**d	t**oo**	t**u** (Fr)[87]

Table 11: Vowel categories for syllable annotation.

Syllables were analysed according to their structural components: the onset (initial consonant or complex of consonants), the nucleus (vowel) and the coda (final consonant).

[86] http://www.phon.ucl.ac.uk/home/ampa/home.htm
[87] The sound [y] is foreign to the English language, but common in French (e.g. tu) and German (e.g. hüpsch).

Nucleus and coda together form the rhyme. In total 23 different onsets and 37 rhymes were found. Table 12 shows a condensed representation of the count of syllables following the components onset and rhyme. It summarizes the results for 14642 syllables[88] (99.3% of the total), organized in 11 onsets and 19 rhymes. The most commonly used syllables are [na], [n@], [la], [t@] and [da], together they make up 52% of the total number of syllables within the results. Four onsets [n, d, t & l] initiate 86.9% of the syllables, as for the rhymes only two elements [a] and [@] stand out, together ending 67.7% of the syllables. The rhyme can further be divided into nucleus and coda. An overview according to these structure components is given in table 13. Here, 77.7% of the syllables has or an [a] or an [@] as vowel, while [i] and [u] still occupy a significant share, but the importance of the other vowels is marginal. Almost 87% of the sung syllables ends on a vowel, and thus has no coda. Of the remaining 13%, about 75% has an [m] as coda, 6 other codas, together used in 14 syllables are not listed.

	a	aj	am	an	aw	@	@j	@m	@n	e	E	i	im	o	om	u	um	y	ym		%
-	50	2				73						4		8		20	1	3		161	1,1
b	16	1	4			3				1	1			2	1	5	31			65	0,4
d	998	166	189	18	9	627	11	115	23	27	8	681	42	22	4	450	89	144	20	3643	24,9
nd	30	2				2	4					1				1		2		42	0,3
h	26		1			39				5	1	3		5		4			1	85	0,6
j	20	5				5			1	4										35	0,2
l	1455	7	91			157		2				5				1				1718	11,7
n	2247	24	19		11	1962	1	16	5	24	52	10		86		66	2			4525	30,9
p	245	1	250			51	2	12				10		1	49	12	32		1	666	4,5
r	194	7	34	3		298		33		13	18	132	7	2		36	3	16		796	5,4
t	454	7	54	16	3	1006	2	222	60	44	43	335	23	2	46	321	47	170	16	2871	19,6
w	2				18	6					7			2						35	0,2
	5737	222	642	37	41	4229	20	400	89	118	130	1181	72	130	100	916	205	335	38	14642	
%	39,2	1,5	4,4	0,3	0,3	28,9	0,1	2,7	0,6	0,8	0,9	8,1	0,5	0,9	0,7	6,3	1,4	2,3	0,3		

Table 12: Analysis of syllable structure following onset and rhytme.

Nucleus	a	@	e	E	i	o	u	y	Total
N	6697	4765	135	142	1259	247	1129	374	14748
%	45,4	32,3	0,9	1,0	8,5	1,7	7,7	2,5	10,0

Coda	-	j	m	n	w	Total
N	12800	247	1466	152	69	14734
%	86,8	1,7	9,9	1,0	0,5	99,9

Table 13: Analysis of syllable structure following rhyme and coda.

In the next section syllable structure is further analysed according to the way in which different subjects treat syllables in their queries.

[88] Onsets and rhymes that occur less than 10 times were not taken into account.

4.4.12 Analysis by subjects

This part of the analysis addresses subjective aspects related to *age, musicianship* and *gender*. Individual findings were summarized for each subject taking part in this experiment. Averages were calculated to give a representative view on the subject's behavior. In this way it was possible to see if different subjects use different strategies, for general aspects of queries and segments (timing characteristics, segmentation behavior, methods used) as well as for the use of syllables in the syllabic query segments. One subject (ID 16) did not succeed in producing any valid queries and was omitted from the dataset. Analysis, then, shows a large variance in the timing and segmentation behavior of the different subjects.

A Beginning, length and segmentation

The subject's average query length is between 4.49 and 26.84 seconds, with a mean of 13.94 s. The average start of the queries varies between 155 and 1510 ms, with a mean of 642.6 ms. The average number of segments within the queries lies between 1 and 3.5 (mean =1.87 segments), which gives an average segment length varying between 3.33 and 23.54 seconds, with a mean of 7.98 seconds. (❷ p.54).

B Query methods

The syllabic and textual query methods are the most widely used query methods among subjects. Of the 71 subjects producing valid queries, 68 produced syllabic and 67 textual segments. Humming is used at least once by 39 of the subjects, whistling by 31 subjects. Twenty subjects gave comments, and 11 subjects produce percussive queries. The overall tendency is remarkable because 31 subjects seem to prefer the textual method, 30 choose the syllabic method, 9 the whistling method and only 1 subject prefers humming. Looking at the distributions of the different methods for all subjects, different strategies could be distinguished (see table 14).

N methods	N subjects	N subjects : method
1	38	18 : text 16 : syllable 04 : whistle
2	17	15 : text +syllable 01 : text + whistle 01 : syllable + whistle
More	16	

Table 14: Use of query methods.

A small majority of the subjects (38) concentrates on one method (at least 60% of the query time). Eighteen subjects concentrate on text and 16 on syllabic queries while a small group of 4 concentrates on whistling (2 of them solely produce whistled queries). A quarter of the subjects (17), divides its query time between two methods (each taking between 30 and 60 %, and together at least 80% of the total query time of the subjects). In 15 of these 17 cases textual and syllabic queries are combined, the combinations of text with whistling and

of syllables with whistling each occurring once. The remaining 16 participants use several methods (with three or four methods covering at least one eighth of the total query time), most common is a combination of textual, syllabic and whistled queries, in the other cases humming is added or replaces one of the three other methods.

From these results five user groups were distinguished. Four apply to about a quarter of the population and a small but rather distinctive group consists of “whistlers”. The typical whistler is a young (mean = 24 y.), male musician. This group produces the longest queries (mean = 19.3 s.), with the highest overall degree of similarity (mean = 4.12). They hardly switch method within queries and if they do, they choose the syllabic method as the only alternative.

To summarize, according to query methods the five following user groups have been distinguished:

- textual (25%)
- syllabic (23%)
- textual and syllabic (21%)
- more than two methods (23%)
- whistling (6%)

An ANOVA on the timing and segmentation aspects of the queries summarized by subject with the four remaining method types as categories reveals significant effects of method type on the mean number of segments ($F(3.63) = 1.418$, $p < 0.01$) and the mean segment length ($F(3.63) = 5.320$, $p < 0.01$). As shown in (❷ p.55) subjects divide their attention over two methods (average segment length = 6.60 s., average number of segments = 2.14). Subjects who use text as dominant method (length = 5.92 s., number = 2.19) stand against subjects mixing more than two methods (length = 7.55, number = 1.59) and those who use syllabic queries as dominant method (length = 9.29, number = 1.74).

C Syllable structure by subject

Since many existing MIR systems start from syllabic input, it is of interest to take a closer look at the way in which different subjects treat syllables in their queries. This includes the spread of the syllables (how many different syllables are used by each subject, and which syllables are used by many subjects) and the clustering of syllables.

The 68 subjects using syllables produce 17.9 different syllables on the average. Nine subjects use more than 30 different syllables, with a maximum of 48, while 16 subjects use less than 10 different syllables. Besides, large differences exist between the subjects and the differences in spread between the syllables are also considerable, and they do not always reflect the total count of the syllables in the experiment as a whole.

The syllables used by the largest number of subjects are [t@] (N=53), [na] (N=50) and [d@] (N=49). Some less common syllables are still used (occasionally) by quite many subjects, e.g. [d@m] (N=115) occurs in 25 subjects, [ram] (N=34) and [r@m] (N=33) each by 15

subjects. Other syllables still are used by fewer subjects, e.g. the 2nd most common syllable [n@] is only used by 40 subjects. For less common syllables this can be much more extreme: [daj] (N=166) by 6 subjects, [no] (N=86) by 3 subjects and the 31 occurrences of [bum] are all attributed to the one subject.

Correlations between the spread of the different syllables and syllable parts show which types of syllables are often combined in the output of one subject and which usually appear more exclusively. To begin with, a correlation analysis was carried out on all of the 56 syllables that occurred at least 20 times in the whole data set. The effects are illustrated with the most common syllables. The analysis shows two types of clustering. First, almost in every case very high positive correlations are found between syllables without coda and their analogues with coda (e.g. [di-dim] r = .910, [r@-r@m] r = .608 and [ta-tam] r = .764). Second, clear relations are seen between syllables with the same vowel but starting with either [d], [t] and [r] (e.g. [da-ra] r = .817, [da-ta] r = .649, [ra-ta] r = .576; [d@-r@] r = .477, [d@-t@] r = .477, [r@-t@] r = .640; all correlations are significant at a $p < 0.001$ level).

On the other hand, syllables with [n] and especially [l] as the onset consonant appear largely unrelated to other syllables. This aspect was noted after calculating the correlations between the distributions of the different syllable parts within the queries of each subject. The findings of this analysis complete our view on the clustering of syllables: significant negative correlations are found with [l] and [n] onsets ([l-d] r = -.354, [l-n] r = -.320, [l-t] r = -.345; [n-d] r = -.432, [n-l] r = -.320, [n-t] r = -.411, all significant at a $p < 0.01$ level), which confirms their relative isolation. Also remarkable is the very strong negative correlation between [a] and [@] nuclei ($r = -.611$, $p < 0.001$), which points to a distinction between subjects that either use syllables with an [a] nucleus syllables with an [@] nucleus.

These results show that clear differences exist between different subjects, both in the number of different syllables they use and in their preferences for certain types of syllables. An analysis of onsets and nuclei by subject equally reveals the existence of distinct groups characterized by a different behavior. Based both on the onsets and the nuclei, four profiles may be distinguished in the subject population. One small (N=5) but distinct group (1) can be distinguished for onsets, with a dominance (>50%) of the onset [l] (average share = 86.2%). Three other, larger groups can be observed: the largest groups (2) of subjects (N = 23) uses syllables with a clear dominance (>60%) of the [d-t-r] complex (average share = 37.6 + 35.2 + 7.5 = 80.3%), and a somewhat smaller (N=18) group (3) prefers (with a dominance of >50%) the onset [n] (average share = 68.6%). Finally, there is a group (4) of subjects (N=20), using a mixture of the three main onsets, often combined with many other onsets (e.g. [p], [h]).

As mentioned in the general overview of the results, a large majority of the nuclei contains one of two vowels: [a] or [@]. The four user profiles distinguished by their use of syllables are: (1) a group (N = 21) with a clear (>60%) dominance of [a] nuclei (average share = 79.7 %), (2) a group (N = 13) with a clear (>60%) dominance of [@] nuclei (average share =

72.6 %), (3) a group (N = 13) that uses both [a] and [@] (each < 60%, but together > 80%; averages [a] = 44.9 %, [@] = 48.7 %) and finally a group (4) of subjects (N = 21) that uses [a] and/or [@] together with other syllables ([a] and [@] each < 60% and other syllables > 20% (e.g. the average shares of nuclei [u] = 24.0%, [i] = 12.6% and [y] = 7.1%).

Some interactions become clear between the group using onsets and the group using nucleuses arise (as can be seen in ❷ p.55). The most prominent interactions are (1) the clear relation between the [l] onset and [a] nuclei, pointing to a strong preference for the syllable [la]; (2) the over-representation of [d-t-r] onsets in the "mixed" nucleus group, and the smallest amount of coda-less syllables (only 80.5% against a 87.0% average) indicates the use of a large number of different syllables (the average number of syllables used in the [d-t-r] group is 22.2 against 4.2 for the [l] group); and (3) a relation between [n] onsets and the group switching between [a] and [@]. No significant effects of either of the two syllable based divisions were found in the timing and segmentation characteristics of the subjects.

D Effects of age, gender and musical experience

Analyses showed a significant negative correlation between age and query-target similarity ($r = -.293$, $p < 0.05$) and a significant positive correlation between age and the average starting time of the query ($r = .279$, $p < 0.05$). Correlation between age and syllable parts showed relations between the relative use of nuclei [a] ($r = .426$, $p < 0.001$) and [@] ($r = -.348$, $p < 0.01$) and onset [l] ($r = .390$, $p < 0.01$). As for the methods used, an increase of comment with age ($r = .307$, $p < 0.01$) can be observed.

ANOVA also reveals some significant effects of gender on the choice of the syllables. For the onset, men tend to use [t] significantly more often than women do, with a 24.9% average against a 11.9% average ($F(1.66) = 7.74$, $p < 0.01$). For the nuclei, a significant effect of gender on the use of the [a] is found, being a very dominant vowel in women's syllabic queries (56.1% average, against 37.5% for men; $F(1.66) = 7.33$, $p < 0.01$). Men also tend to use a larger variety of syllables, using an average of 20.19 different syllables, against 14.15 for women ($F(1.66) = 7.33$, $p < 0.01$). Finally, women, tend to start their query later than men (after 716 ms. against 595 ms.; $F(1.69) = 5.97$, $p < 0.05$).

Comparison between musicians and non-musicians also yields some significant effects, mainly on the methods used. Within the query method of musicians, text is of significantly minor importance, with an average of 35.3% of the time spent using text as compared to 46.2% for non-musicians ($F(1.66) = 4.57$, $p < 0.05$). This is compensated for by a larger share of syllabic and whistled time in the musicians' queries (though the effects of each of these separately do not reach significance). Finally, there is also an effect of musicianship on the average length of the segments: musicians' segments are longer than non-musicians' (means 9.18 and 7.12 seconds; $F(1.66) = 4.90$, $p < 0.05$), this being a result of longer queries and less segmentation by musicians.

4.4.13 Stimuli related aspects

In this study, there are 26 pieces of music involved and the amount of queries by song varies considerably, ranging from 13 to 87. Moreover, since the choice of the songs should reflect the variety of the musical landscape, they are not equally distributed between variables like language and style. All this makes reliable analyses difficult. Nevertheless, a few general results are presented that can serve as a starting point for further research. Table 15 gives an overview of the titles in the experiment and the number of queries generated in each experiment part.

Title	Experiment part 1			Experiment part 2		Total recordings
	Recording 1	Recording 2	Recording 3	Recording 1	Recording 2	
Blowin' in the wind	47	2	4	7	2	62
All I really want	0	0	0	23	1	24
Mooi, 't leven is mooi	68	12	7	0	0	87
De Marie-Louise	61	9	6	3	0	79
How do you do	0	0	0	19	1	20
Paloma blanca	0	0	0	11	3	14
YMCA	67	10	4	2	0	83
When a man loves a woman	57	7	3	3	0	70
Don't worry, be happy	65	9	6	1	0	81
Walk on the wild side	44	7	1	6	1	59
Waterloo	60	3	3	5	0	71
Sunday, bloody sunday	48	8	5	7	1	69
It's raining men	60	5	5	1	0	71
Whithout you	0	0	0	17	3	20
Ad mortem festinamus	0	0	0	22	2	24
My way	40	7	1	12	3	63
Rosa	17	2	2	16	4	41
Smells like teen spirit	10	1	1	20	3	35
Only happy when it rains	6	1	0	19	2	28
Temple of love	6	0	1	16	3	26
Highway to hell	0	0	0	13	0	13
Don't cry for me Argentina	27	2	1	4	3	37
Klein klein kleuterke	17	1	2	6	2	28
'k zag twee beren	10	1	0	2	0	13
O Fortuna	4	0	0	11	1	16
Fur Elise	2	0	0	12	0	14
Total	716	87	52	258	35	1148

Table 15: Overview of the number of queries created per piece of music.

There is some variability in the timing of the queries between songs which is much less than between subjects. The average length of the queries by song varies between 9.90 and 20.96 seconds and the start time between 514 and 752 ms. Differences in segmentation, although also smaller than between subjects, are more important. The average number of segments varies between 1 and 2.94 and the average segment length between 4.55 and

18.49 seconds, which indicates that some songs, more than others, give rise to a production of homogenous queries.

Analogous to the subject-related variables like gender, age and musicianship, song-related variables can be defined such as genre and language[89]. Within the stimuli set, 18 songs were in English (one mixed with Spanish), 4 were in Dutch and 3 in Latin (one mixed with French), and one piece was purely instrumental. Stylistically, the dataset includes pop (14), rock (8), children's song (2) and classical music (2). The limited number of songs and the unequal distribution within both categories prevents us from making far-reaching conclusions. Nevertheless, some interesting results may serve as a starting-point for further research. For example, there seems to be a link between the query method and the language. Depending on their knowledge of the source language, it could be expected that subjects produce more text and fewer syllables. In this study the use of text is the highest in the songs with a Dutch text (the native language of the subjects; 47.3% average), is slightly less for the English texts (40.0%) and much lower for the Latin texts (12.4%). This is mirrored in the relative use of syllables (33.0 - 38.7 - 70.4%). Song style, on the other hand seems to have an influence on the syllable structure, i.e. the predominance of the syllable [la] in the imitations of the children's songs. (❷ p.56).

4.4.14 Effects of memory use

For this analysis, queries produced from both long-term and short-term memory are kept separate. Queries generated in the first part of the experiment rely entirely on long-term memory. They are produced from memory for songs indicated as "known, remembered and possible to imitate". The queries from the second part of the experiment, however, were produced after hearing the complete song first. Within the last category a distinction is made between songs indicated as "known" after listening and those indicated as "unknown" after listening. For "unknown" songs, the subjects of course had to rely entirely on their short-term memory. For "known" songs, there was interaction between long and short-term memory.

Thus, the experiment provided three classes of queries, characterized by a different use of memory:

- long-term memory (LTM) (833 queries from part 1)
- short-term memory (STM) (81 queries from part 2)
- mixed (LTM+STM) (208 queries from part 2)

Memory use has a significant effect on both starting time and the length of queries (❷ p.57), as well as on the query-target similarity, and there is a very noted effect on the query length ($F(2.1119) = 30.343$, $p < 0.001$), with following results:

- a mean of 13.21 s. when using LTM

[89] Time period would be another possibility, but it is often difficult to specify the age of songs since they can be traditional, covers, recent interpretations of older composed material.

- a mean of 12.78 s. when using STM
- a mean of 17.87 s. when using LTM and STM

These findings are reflected in a highly significant effect of memory category on the segment length ($F(2.2111) = 46.500$, $p < 0.001$). The segments within the queries using LTM and STM are significantly longer (mean: 10.6 s.) than those in which only one type of memory is used (means: 6.8 s. for LTM and 6.3 s. for STM). There is also a significant effect on the starting time of the queries ($F(2.1119) = 3.136$, $p < 0.05$). When only relying on LTM (mean: 643 ms.) subjects tend to start earlier than when using STM (mean: 685 ms), but when using LTM+STM they still start sooner (mean: 579 ms.).

Memory use had also a highly significant effect on query-target similarity, both at the level of the query ($F(2.1119) = 6.668$, $p < 0.01$) and at the level of the segment ($F(2.2111 = 10.571$, $p < 0.001$). The distribution of the different similarity levels at the segment level within the three categories, shown in (❷ p.57), indicates only a very small difference between LTM and LTM+STM but a drastic fall in similarity for the queries based on STM only.

Besides effects on timing and similarity, there is a clear influence of memory use on query method and performance style. There is a change from textual dominance to syllabic dominance with a growing importance of short-term memory: for LTM, 48.7 % of the queries are textual, taking 41.7 % of the total time, for LTM+STM this becomes 39.7/33.3 % and for STM 34.4/26.6 %. (❷ p.57).

The importance of syllabic queries moves in the other direction: LTM 34.9/36.0 %, LTM+STM 43.1/47.2%, STM 49.1/58.3%. Likewise, the importance of whistling also decreases: LTM 8.6/18.0 %, LTM+STM 9.5/15.3%, STM 4.3/8.0%. Remarkable, finally, is also the sudden increase in percussive queries when long-term memory is no longer present.

The influence of memory type on the performance style (melodic, rhythmic or intermediate) is shown in (❷ p.57). Just as with textual queries, the amount of queries, characterized by a melodic performance style, diminishes with an increasing importance of short-term memory (LTM: 73.9/79.6%, LTM+STM: 69.0/73.7%, STM: 47.2/51.7%), this in favour of intermediate (LTM: 19.1/18.2%, LTM+STM: 25.6/22.8%, STM: 45.5/41.9%), and, to a lesser extent, rhythmic style (LTM: 4.7/1.8%, LTM+STM: 3.7/3.2%, STM: 5.5/5.8%). The latter is only visible in the share of total query time, which indicates that queries in a rhythmic style get relatively longer.

4.4.15 Discussion

Analysis of the timing characteristics of the queries shows a mean query length around 14 seconds. Considerable differences, however, exist between subjects, with personal averages ranging from less than 5 seconds to close to the maximum allowed recording time

of 30 seconds. The start of the actual query occurs on average 634 ms. after the start of the recording, but here inter-individual differences are extensive as well, with personal averages varying roughly between 150 and 1500 ms. Thus, a user-friendly music information retrieval system should be flexible in timing, expecting some people to start up to 2 seconds after the start of the recording and expecting from a few seconds to over 30 seconds of information input.

In classifying and segmenting the queries a distinction was made between six methods: singing lyrics, singing syllables, whistling, humming, percussion and comments. In about 60% of the queries, subjects used only one method. In the other cases subjects used at least two methods, sometimes alternating between two or more methods, but queries containing more than six segments were scarce. The amount of segmentation is depending on both the subjects and the target song.

The most common query methods are singing lyrics and singing syllables. Syllabic segments are generally longer and therefore more prominent than textual segments. Whistling is the third most popular method, while humming, percussion and comments amount to only a small share.

Just as with the timing characteristics, the use of certain methods is user dependent. Three types of users could thus be differentiated: (1) subjects concentrating on one method (lyrics, syllables, whistling), (2) subjects dividing their query time between two methods (mostly a combination of textual and syllabic) and (3) subjects mixing several methods. The choice for certain methods also seems to depend on the familiarity of the subjects with the language of the target song: the less familiar people are with the language, the more they tend to use syllables and whistle instead of using lyrics. These findings indicate that the ideal MIR system should also be able to cope with changes in method. Alternatively one could offer users the choice of one specific method: lyrics, syllables or whistling.

Concentrating on one single method does not seem so attractive since some classes of users largely rely on one single method that is not necessarily the same method as a system may impose. This is most clearly illustrated by whistling. Less than half of the people used whistling as a query method, but a small group of those did provide long, high-quality queries. Thus, allowing only whistling (e.g. Prechelt and Typke, 2001) does not seem to be a good choice, but excluding whistling would exclude an interesting body of queries.

Around 75% of the queries were performed melodically, which supports the idea that melodic content is to be regarded as a salient feature of vocal queries (e.g. Chai, 2001).

Since the use of syllables is one of the most important query methods, and most existing MIR systems require syllabic input, the nature and structure of the syllables occurring in spontaneous queries was further investigated. The ten most common syllables are (in order of decreasing importance) [na], [n@], [la], [t@], [da], [di], [d@], [ta], [tu] and [ti], of which

[t@], [na] and [d@] belong to the syllable repertoire of the largest number of subjects. If a system requires specific syllables for ease of processing, it is desirable to choose some of these most common types. Some of the systems, however, ask the users to use syllables that hardly ever occur in spontaneous queries, e.g. [ba] (Brøndsted et al., 2001), occurring only 16 times and [fa] (Pauws, 2002), which was used only 5 times. These syllables might be less comfortable to users and might so yield a larger amount of errors, or diminish the attractiveness of the system to the public. Additionally, just as with the methods, different categories of users were distinguished, some was sticking to particular onsets and/or nuclei, others were mixing them. In the view of segmentation issues, the onset of the syllables is probably the most important feature. It was found that some users stick to [l] or [n] as a onset, but that [d], [r] and [t] are often mixed. As for the nuclei, the choice for [a] in most systems is in agreement with the findings, as it is by far the most common nucleus (45%). On the other hand, I am not aware of experiments in the literature that give participants the choice to use syllables with an [@] nucleus, although it is found in almost a third of the syllables in this study. Allowing both vowels would increase the user-friendliness of a system. The fact that most of the sung syllables end on a vowel and have no coda is in agreement with the syllable constraints found in the experimental literature.

Significant effects of age, gender and musical background were found as well as differences between the reproduction of unfamiliar melodies (STM) and the recall of known melodies (LTM). Younger people tend to start their queries sooner and get better query-target similarity scores. The use of [a] nuclei and [l] onsets and of additional comment as a query method increases with age. Men start their queries later than women and use a larger variety of syllables with a larger share of onset [t] and a smaller amount of [a] nuclei. Musicians make longer queries than non-musicians and use less text in favour of syllables and whistling. The timing of the queries as well as their similarity with the target is dependent on the type of memory used by the subjects. When known songs are “refreshed” by playing them to the subjects, the queries start sooner and last longer. The reproduction of unfamiliar melodies from STM has less quality than recall of known melodies even if the query only relies on LTM. The lesser degree of acquaintance with the language of the target song is also reflected in the larger share of syllabic segments and the increased importance of rhythmic and intermediate performances. The present findings support the notion that people are quite good at reproducing tones of familiar songs from long-term memory (Levitin, 1994). Thus far, however, very few researchers have investigated the effect of long-term memory. Most experiments within the field of MIR research (see supra) were set up in such a way that the participants were asked to sing (e.g. McNab, 1997) or repeat a sequence of notes played to them (e.g. Lindsay, 1996).

4.4.16 Conclusion

When participants are allowed maximum freedom in the formation of queries, different types of behaviour have been found. Although a large variety in vocal querying strategies

was observed, some basic characteristics could be established and some categories of users could be distinguished. The findings of the query-by-voice experiment generate important guidelines for the development of user-friendly music information retrieval systems. The application of these guidelines will enhance the interactive process and open the way to make music information retrieval systems more efficient and more attractive to the user.

5 User Context

The study reported in chapter four provided an idea into the music search behaviour of university students, but this study was limited. Above all it was indicative of the need for extended investigation of the user in the broad sense. This chapter (based on original articles IV and VI) reports the set up and outcome of a large-scale user study in the form of a survey. The primary goal of this study was to enable user context (e.g. background, habits, preferences, expertise and interests) to be taken fully into consideration in the development of music information retrieval tools. The survey was aimed at (1) attracting potential MIR users on a large scale, (2) identifying real users' musical background, habits and interests and (3) recruiting within this population those people that would be willing to participate in future music annotation experiments.

5.1 Background in user context

In chapter four, a review of the literature involving users performing vocal querying was given. It has been remarked that most vocal query experiments have been carried out in view of processing queries to symbolic transcription, without any concern for the user's query preferences. The same is true for the user context. Although more people discover vocal query methods, genre classification, notation languages and emotion related keywords as new possibilities for looking for music, there are but a few studies that analyse the profile of the so-called "MIR user". Such profile would include a music information retrieval user's musical background, habits and interests.

In view of the planning for a music information retrieval test bed, Cunningham (2002) emphasizes the need to ascertain who the potential users of a music collection are. However, studying large, amorphous groups of people looking for specific types of music is not an easy task. Who are they and what are their music desires? Cunningham suggests that a good starting point for coming to an understanding of people looking for popular music, is the small but significant research that examines the relationship between everyday people and music, including examinations of the ways in which people use music in their daily lives. It is not said, however, how a selected group of participants from the population of "everyday people" in this small research should be defined. In William's study (2001) of young people's relationships to music, for example, it is reported that "young people found music to be significant in terms of its routine practical uses relating to rather meaningless and mundane engagements" (p. 236). Although this study is directly engaged with those who consume popular music, it also involved unstructured group discussions that were not conducted in view of music information retrieval research. Again, we are confronted with a lack of knowledge about a core issue related to music information retrieval.

While the data for this study was being gathered another similar project was started. The Human Use of Music Information Retrieval Systems (HUMIRS) project (Downie, 2004) aims at collecting and analyzing data concerning "real-world users". Thus far, however, results

have been reported only from a study population group that comprises the University of Illinois at Urbana-Champaign community (Lee, 2004). For that reason, the study reported here can be considered as the largest study which currently exists about users in relation to music information retrieval.

To summarize, within the scope of music information retrieval investigation, no user studies have been conducted that collected context data from a population that is representative of potential music information retrieval users. As music information retrieval systems should be created with a variety of different potential users in mind, they should accommodate lay people presumably seeking popular music as well as musical experts looking for detailed information and all categories in between those extremes. This vague population shares at least two things: they are music lovers and they are open-minded to new technology. From this assumption the survey on user context was designed.

5.2 Global set up of the study

The survey on user context is the initial step of a large-scale investigation that looks at both definition of user groups and exploration of multiple annotation forms. After the collection of context data, experimental research has been conducted with participants recruited from this survey. The background idea of the global set up is both methodological and practical. First, it is interesting to have at one's disposal the background information of subjects involved in several different annotation experiments. This background information supports comparison of the effect of user context within diverse annotation tasks. Performing experiments with representative samples taken from the same population, moreover, offers opportunities for statistical generalization. Another practical thing is that each time a participant takes part in an experiment it is no longer needed to present a questionnaire on personal data, as all of this is already included in the survey output. The empirical data for this study was collected in several stages using different methodologies. Figure 13 shows the global set up of the experimental procedures.

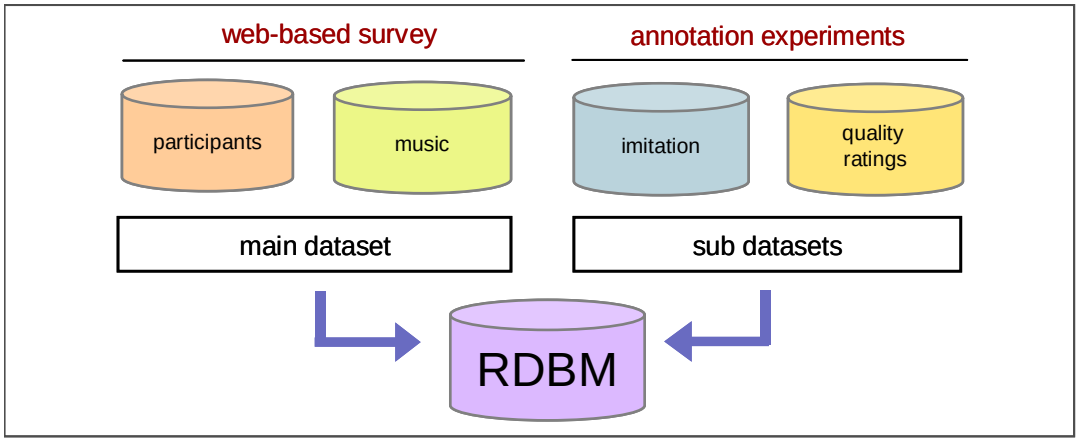


Figure 13: Global set up of the study of user context.

The *main dataset* consists of data generated by the survey (a web-based questionnaire distributed online). Apart from contextual information it produced a list of subjects who agreed to take part in additional experiments (*participants*). The main dataset also contains titles of participants' favourite music. These titles were the starting point for the construction of a *music audio database* with participants' favourite music. Various subsets of this database have then been used in different experiments.

Sub datasets contain data from diverse annotation experiments based on *imitation* and *quality ratings*. Imitation experiments have been set up primarily to create data (i.e. audio files as imitations of real music) as ground truth for algorithm development. Annotation by imitation (Lesaffre, 2004) has been tested for rhythm, melody and tonality and some of these studies (Heylen, 2004) are still going on.

Next sections are dedicated to the survey and in chapter six, full analyses are given for an experiment that focused on the annotation of perceived semantic qualities. Finally, a relational database model was set up to store the large quantities of data from the multiple parts of the study. This database is presented in chapter seven.

5.3 Survey method

In this part the set up, design and procedure of the survey are described.

5.3.1 Survey set up

A self administering survey[90] into people's global and musical background was set up. It is presented in the form of a questionnaire available online. For participation there was no restriction on age, gender or background. However, the survey was designed to reach a specified target population: potential users of music search and retrieval technology.

A web-based methodology has been chosen because in many respects it is a beneficial approach. Using the Internet made it for example possible to address the targeted population more accurately. This method also allowed starting a study on a much larger scale than the usual experiments in this research area. Other advantages to this method are that no personal guidance[91] is required and that it is low cost.

Recruitment of respondents was done in numerous ways. To begin with, the survey was advertised on the IPEM website; mailings and postings to newsgroups (e.g. music.be) were done; flyers and posters[92] were distributed and interviews[93] were given, all encouraging people to participate. Furthermore, a special stand with a PC - where the questionnaire could be completed - was built at the music library of the city of Ghent for a period of three

[90] PHP language is used. The survey was hosted by http://www.visionR.be
[91] II am aware that in some respect this aspect might also be a disadvantage, but the questionnaire was designed in order the overcome as much as possible negative effects.
[92] Designed by Tom Desmet, Public Relations section of Ghent University.
[93] Live interviews on 20040331 (Jongens en Wetenschap, Radio 1) and 20040402 (Ochtendpost, Radio 2) and an interview with Marc Coppens published in "Klankbord" issue 11, 2004.

months. This approach was assumed to guarantee that the response would represent a valid cross-section of potential end-users.

A potential risk was the length of the questionnaire[94]: respondents might become unwilling to complete it. This could especially have been the case because people were not personally guided during the process by a research member either. The number of individuals responding in full to the questionnaire was nevertheless satisfactory. The music library employees, who distributed the flyer with basic information, even said that people were very enthusiastic about the initiative and enjoyed participating.

5.3.2 Survey design

The survey questionnaire had 29 questions plus personal data, the latter being disconnected from the rest in the dataset, so that any association with individual persons would be impossible. A 32 bit key identification number was automatically generated. Personal data was saved in a separate table that was only kept to contact the people who were willing to participate in the annotation experiments. In total, ten forms were presented with questions that probed the global and musical background of the respondents. Table 16 shows the global structure and keywords referring to the questions asked. More details on question phrasing, question types, and measurement are given in (❷ pp.61-66)

GLOBAL BACKGROUND		MUSICAL BACKGROUND	
Demographic	Internet	Study	Behaviour
Name	Use	Education	Way
Address	Activities	Level	Context
Age	Place		Instrument
Birthplace	Time		Singing skills
Gender			Dancing skills
Job function			Movement
Language			Listen medium
Culture			Radio station
			Genre
			Taste evolution
			Favourites

Table 16: General concept of the survey design.

5.3.3 Survey procedure

The questionnaire was first tested for validity and reliability both by music experts working at IPEM and by novice listeners recruited among friends. After quality testing the survey was made publicly available and was maintained until January 2005. The dataset used in this thesis is based on the data that was gathered in the period from April 2004 until January 2005.

[94] 29 questions were asked of which some require more work than just checking a yes/no box (see 5.4.5: Favourites).

Filling out the questionnaire was rather demanding and took about an hour on average. It was hoped, however, that the time consuming aspect of the survey finally would not be a hindrance and that only motivated people would feel encouraged to fulfill the task. Analysis of the number of forms that have been filled in has shown that the information was gathered from subjects who took the survey seriously and who were willing to provide a lot of information.

In an introductory text on the screen, the goal of the study was explained and privacy of the participants was guaranteed. People who filled in the questionnaire could not see the whole set of questions in advance. Respondents could only proceed to a next form when the previous one was fully completed. The survey includes open-ended as well as closed-ended questions. In view of creating a large analysis potential it contains different types of questions that are: true/false radio questions, pull down menus, open answer text fields, multiple choice checkboxes and matrix radio buttons. The questionnaire is included in digital format in ❸.

First, information was gathered about global background, followed by music education, musical skills and musicianship. To begin with, a number of questions was asked that verify *socio-demographic* and *cultural background* of the respondents. Their *familiarity with the Internet* was examined as well. Then, information was gathered about the kind of *music education* they had and what their *music level* is, whether they play an *instrument* and if so, which type. They also had to report on their *singing* and *dancing skills*. They were also asked how they *interact with music* and what their *preferred medium* is for listening to music. Music *genre preferences* were addressed in multiple ways. They were asked which genres they mostly listen to, the *evolution of personal taste* was checked and participants were requested to provide up to ten titles of their *favourite music*.

The genre taxonomy presents a set of classes that is not too extensive and that covers genres and categories that are assumed to be familiar to the subjects. Participants could make multiple choices from a checklist with fourteen predetermined music genres and categories. Additionally, they had the possibility of naming a genre that was not in the list. The evolution of musical taste was tested by linking the genres with six phases of age. Finally, an open-ended questionnaire was included in the survey with the aim of generating of a music audio database that is representative of people's favourite music: music that reflects their genre and title preferences as well.

5.4 Questions and findings

The number of individuals responding to the questionnaire was 774 (up to January 2005). After quality-control checks and data cleaning (e.g. deletion of double entries), 711 ID's (92%) were considered valid for analysis. Three quarter (N=523) of the respondents filled in the complete questionnaire, showing that the contributors were considerably dedicated to

the task. Table 17 shows the number and percentage of participants according to the forms that were filled in by the 711 ID's. (❷ p.67).

Form	Entitling	Frequency	Percentage
F1	About you	17	2.4
F2	General information	6	0.8
F3	Culture	2	0.3
F4	Internet	0	0.0
F5	Music education	4	0.6
F6	Music practice	6	0.8
F7	Music listening	13	1.8
F8	Music genres	7	1.0
F9	Music taste	133	18.7
F10	Favourites	523	73.6
Total		711	100

Table 17: Distribution of participants in the survey according to their response.

Explorative analysis and statistical inferences have been performed on a dataset containing information from the 663 participants who filled in eight forms of the survey. Eight forms were considered an optimal threshold for statistical analysis because they contain enough data for relevancy and cover background information for 93% of the valid participants. Comparison of the output with datasets for six forms (N682, 96%) and for seven forms (N676, 95%) showed no noticeable differences.

In what follows, a summary of the descriptive statistics is given. The focus is especially on participants' music background and on the results that are important in view of the study described in chapter seven that deals with the perception of qualities of music. Tables and figures with detailed statistical output are comprised in (❷ pp.68-78).

5.4.1 Global background

Participants' global background information is about socio-demographic issues including language and culture and gives an idea of their Internet behaviour.

A Socio-demographic

Q # 1: In what year were you born?

The age of the participants ranges from 15 to 75 year with a mean of 29.4 year. The age category between 15 and 35 year is most strongly represented (75,1 %). (❷ p.68).

Q # 2: Where were you born?

97,6 % was born in Belgium.

Q # 3: What is your occupation?

This was an open-ended question. The answers were manually checked and grouped into eight categories that represent most occurring professions. An occupation taxonomy was built that includes following categories: student, administrative and secretarial, technical, educational, scientific (i.e. academic, researcher, PhD. student), management and sales, medical and not specific (e.g. unemployed, pensioned, no profession, housewife). Table 18 summarizes the results of this categorization. Occupational information shows an interesting spread of professions. (❷ p.70).

Occupation	Frequency	%
Student	261	39,4
Administrative	114	17,2
Technical	52	7,8
Educational	61	9,2
Scientific	45	6,8
Management	40	6,0
Medical	16	2,4
Not specific	74	11,2
Total	663	100

Table 18: Distribution of occupational categories.

Q # 4: What is your gender?

Male respondents (55.8 %) slightly outnumbered females (44.2 %). (❷ p.69).

Q # 5: What is your native language?

The native language of 98.8% of the subjects is Dutch

Q # 6: What culture do you belong to?

The cultural background of 99.5% is Western.

Q # 7: Which country determines your background?

94,1% say that Belgium determines their cultural background; the other countries can be seen as part from the western culture, except for Cambodja (1), Bolivia (1) and Uganda (1).

B Internet activities

Q # 8: Are you used to working with the Internet?

As could be expected from the survey design, targeting a population that is open minded to new technology, 92.6% of the respondents said they are used to working with the Internet. (❷ p.71).

Q # 9: What kind of Internet activities do you do?

Participants were asked to indicate the type of activities they do. A selection of seven activities (emailing, chatting, playing games, downloading music, watching movies, searching information and other was offered). Their most popular activity is sending emails (30.4%) and information search (31.1%). With regard to music activities, 12% said that they download music from the Internet. (❷ p.71).

Q #10: Where do you use the Internet?

In 50% of the cases the Internet is used at home. (❷ p.71).

Q #11: How many hours on average a week do you spend using the Internet and how many of these do you spend on music?

Participants in the survey spend 9.6 hours per week on the average using the Internet for general activities (❷ p.72) and they spend 3.1 hours per week on the average for activities related to music (❷ p.73). People were requested to select the time from a list with numbers from 1 to 30 and "more". The output has been organized into seven time categories. The time spent actually varies greatly among people (see table 19). A small group (4.8%) spend 30 hours or more per week using the Internet and a fifth of the respondents (21.1%) spend 10-14 hours per week, while about a third (32,1%) spend only 1-4 hours per week. The question is what do these different groups of people spend their time doing. This is not an easy question to ask. Previous tests performed with students at Ghent University have shown that it is possible to get a reasonable accurate estimate if you ask people about the total amount of time that they spend on the Internet per week on average. However, it is nearly impossible for people to divide this time into mutually exclusive categories. Therefore, participants were just asked about the time they spend on the Internet in general and in particular for activities related to music.

Time	General %	Music %
1-4 h	32,1	78,9
5-9 h	24,6	13,0
10-14 h	21,1	3,9
15-19 h	7,4	1,1
20-24 h	7,4	1,7
25-29 h	2,6	0,0
30+ h	4,8	1,5
Total	100,0	100,0

Table 19: Internet time spent for general activities and for music related activities.

5.4.2 Music background

A Music education

Q #12: What kind of music education did you get?

Here multiple choices could be made from 6 possibilities, including “no music education” (36,8%). From those participants having had music education 26.6% report self-study, 10.1% private education, 52.3% music school[95], 7.1% conservatory and 3.8% university. 44.6% selected one education type, 14,0% selected two types, 4,1% three types and 0,5% all four types in the list. (❷ p.74).

Q #13: What level have you reached?

A choice was offered from no level (36,5%), over beginning (19,9%) and medium level (23,2%) up to high-level (20,4%).(❷ p.74).

B Interaction with music

Q #14: In which way are you occupied with music?

The way people interact with music was equally distributed: 49.3% said they just listen to music and 50.7% answered they listen as well as play. (❷ p.75).

Q #15: What is the context of your music activities?

Most participants (91.4%) call themselves amateur. (❷ p.75).

C Musical instruments

Q #16: Which instrument do you play?

The survey statistics show that 57.2% percent of the participants answered that they play one ore more musical instruments. Keyboard was the most popular instrument group (25.6%) followed by plucked stings (21.1%) and woodwind (19%) The seven types involved are woodwind, brass, percussion, electronic, keyboard, string bowed and string plucked. Brass instruments did not do well (5.4%) in this survey. From the instrument players, 54% play one instrument, 32% play two instrument types and 10.1% three types. (❷ p.76).

D Singing and dancing skills

Q #17: How do you judge your singing skills?

Q #18: How do you judge your dancing skills?

Participants in this survey evaluate themselves as better dancers than singers: 40.9% said they cannot sing or are bad singers and 30% cannot dance or are very bad dancers while 21.3% are good or very good singers and 25.3% are good or very good dancers. (❷ p.77).

[95] The use of music education categories is similar as in the survey of music search behaviour (see 4.3.3).

Q #19: Do you move along with the music you hear?

Nearly all participants (94.9%) said they spontaneously move along with the music they hear. (❷ p.77)

E Music activity

Q #20: Do you listen to music actively or passively?

A question was included to make a distinction between active and passive music listeners. To make sure that participants would understand what was meant, a description was given on the screen. Active listening happens in the conscious mind when we listen more closely to individual musical elements. Which elements actually capture our attention may change within the duration of the composition or tune. We listen passively when music is not necessarily what we are concentrating on at the time. We may be doing homework or driving in our car while music is playing: i.e. we are not really paying attention to it. A distinction is made between three categories: active listeners (32%), a group that listens both actively as well as passively (43.3%) and passive listeners (24.7%).

F Medium for listening to music

Q #21: What is the medium of your first choice for listening to music?

Q #22: What is the medium of your second choice for listening to music?

Participants were asked to give information on the medium of their first and of their second choice. Four possibilities were offered: CD/minidisk, radio, television and Internet. Participants' preferred first medium is the CD (62.9%) and the medium of their second choice is the radio (47%). (❷ p.78).

G Radio

Q #23: What are your favourite broadcast stations?

With regard to the broadcast stations a selection of the ten most known Belgian radio stations was presented. The highest scores went to Studio Brussels (critical trendsetter, mainly pop/rock), Radio 1 (mainly informative, many genres) and Klara (classical) respectively checked by 54%, 42.9% and 34.3% of the participants. (❷ p.78).

5.4.3 Genre

Q #24: What kind of music do you mostly listen to?

Participants had to choose from a checklist with fourteen predetermined music genres and types. Additionally they had the possibility of naming a genre that was not in the list. Subjective data was solicited by asking extra questions to which the participant was free to answer. Allowing an explicit unstructured response enabled us to collect more detailed information on the terminology people use for describing music genres and as checks of the reliability of the information provided. For each selected category participants could give

an example plus a description of the preferred subgenre. This opportunity was designed to collect a wider variety of responses that more truly reflected the listening habits of the respondents and to check participants' knowledge of genres.

A Genre taxonomy

The genre taxonomy that was presented to the participants was based on a set constructed for a previously created audio database[96] for the MAMI research project. The MAMI database was constructed to have a test collection for the development of a system for content-based search of music. This aim implied the consideration of different types of music. It also required characterization of a wide range of sound characteristics, necessary for multiple ways of querying on data (i.e. by means of text queries and by music-based query techniques as well). The adaptation of the MAMI genre taxonomy for the survey still presented a general set of classes that was not too extensive and covered well-known genres and categories that were assumed to be familiar to the subjects. The following fourteen categories of genres and styles were available for choice: classical, world/ethnic, light/oldies, pop, rock/metal/noise, jazz, blues/soul/reggae, folk/country, rap/hip hop, new age, dance/house/techno, soundtracks, children's songs, other. The genre taxonomy is composed of a wide range of music styles that are rather global but that have an immediate appeal. Participants could make multiple choices from this fixed genre taxonomy, and as they had the possibility of adding (naming) a genre that was not in the list, the most frequent responses reflected the general trends in the answers.

B Preferred genres

Table 20 shows the fourteen classes involved sorted by the percentage of responses. (❷ p.79).

Genres	Checked %	Response %
Pop	60,6	13,0
Classical	56,1	12,0
Rock	50,2	10,8
Jazz	40,4	8,7
World	36,5	7,8
Dance	33,3	7,1
Light	32,9	7,0
Soundtrack	32,4	6,9
Blues	31,4	6,7
Other	27,8	5,9
Folk	24,9	5,3
Rap	24,3	5,2
New Age	9,4	2,0
Children	6,8	1,5

Table 20: Preferred music genres sorted by percentage of responses.

[96] See 4.4.2 The MAMI database.

The 663 participants who filled out the genre questionnaire together provided 3096 selections of genres out of the genre taxonomy presented. There were 4.7 genres categories selected on average. About half of the people (53.5%) have chosen between one and four genres, the other half have chosen between five and fourteen genre categories (❷ p.80). Only 13.1% selected more than seven genre categories. The top-ranked genre categories among the respondents are pop (13.0%), classical (12.0%) and rock (10.8%).

The open-ended "other" responses included "religious", "kleinkunst" (cabaret), "hafabra" (brassband), "chanson", "lounge", and music from specific cultural regions such as Japanese, Indian, Flemish, etc..

5.4.4 Taste

Q #25: Did your musical taste change? What kind of music did you listen to when you were younger?

The evolution of musical taste was tested by linking the same genre set used as in previous question with six phases of age. In research, a variety of age categories is used. Here, a rather broad age categorization as often used in social psychology was chosen. The age classes are: 1-12, 12-18, 18-25, 25-35, 35-45 and 45+[97]. Given the targeted population, it was assumed that introducing categories for still older age groups would needlessly overload the list. This hypothesis is in agreement with the age distribution of the participants, because only 4,3% is older than 55 years.

A Evolution of taste

According to the sum of counts for all age categories, it was found that pop (19%), classical (12,9%) and rock/metal/noise (11,2%) are the most popular genres. The evolution of taste over the six age categories gives a clearer insight into the influence of age on the appreciation of the genre classes included in the genre taxonomy. The importance of classical music is remarkable high and increases with the age. Most extremes appear in the youngest age category (1-12), where popular music gets the highest score (29,5%), followed by children's music (28,7%). The interest in pop music remains rather stable between 18 and 35 years and then gradually decreases. Interest in rock/metal/noise music is most prominent around the age of 12-18, after which it slowly decreases.

Table 21 and figure 14 summarize the percentages of genre selections per age category. (❷ pp.81-82).

[97] Maybe it was not the best choice to introduce categories with overlapping boundaries. However, among the participants who replied to the form "taste evolution" there were but few persons aged 12/18/25 and 35 (8%). The majority of them consequently chose the highest category.

Genre	1-12	12-18	18-25	25-35	35-45	45+
Classical	10,4	12,1	12,2	14,5	18,0	19,3
World/ethnic	1,9	5,8	8,5	9,3	10,9	12,2
Light/oldies	11,5	6,8	6,1	6,9	8,7	8,7
Pop	29,5	21,8	15,0	15,1	13,2	12,2
Rock	4,8	17,6	12,1	9,5	6,8	5,5
Jazz	1,4	4,1	8,7	9,2	10,0	10,3
Blues	1,6	5,6	8,4	9,2	10,0	10,3
Folk/country	1,4	4,0	6,3	6,6	7,3	8,7
Rap/hip hop	1,3	5,1	4,9	3,9	3,5	2,6
New Age	0,8	2,2	2,4	2,8	2,5	2,6
Dance	2,6	7,6	7,4	5,6	2,8	1,3
Soundtrack	4,3	5,7	6,9	6,0	4,8	4,8
Children	28,7	1,4	1,1	1,3	1,3	1,6

Table 21: Evolution of taste.
Overview of genre distribution per age category (in percentages).

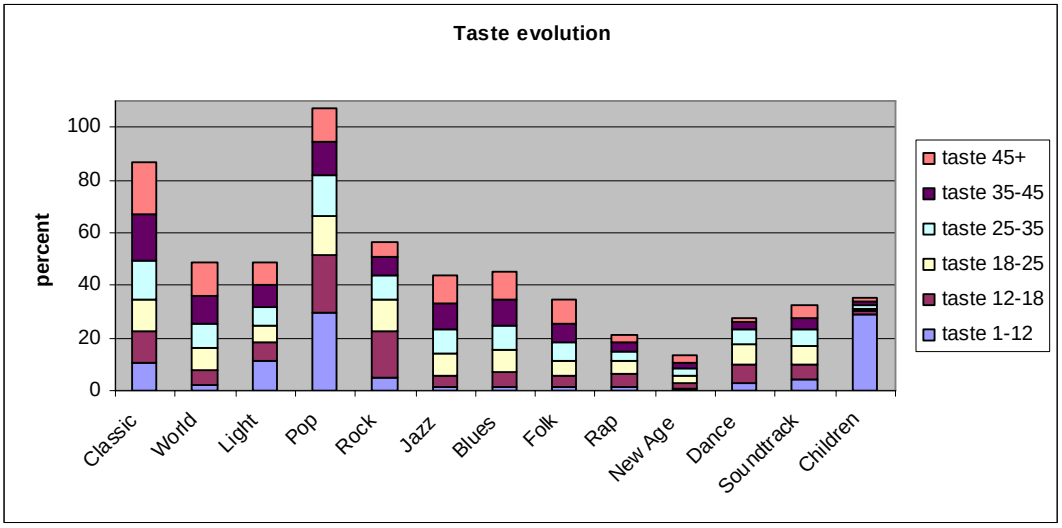


Figure 14: Contribution of genre per age category.

B Broadness of taste

People's musical taste is the broadest between 12 and 25 years and then progressively decreases. Table 22 summarizes the counts and the number of genre selections on average per person per age category.

Age category	Count	N participants	Selections
Age 1-12	1402	678	2.14
Age 12-18	2094	678	3.19
Age 18-25	2664	643	4.06
Age 25-35	1619	377	2.47
Age 35-45	984	174	1.50
Age 45+	736	93	1.12

Table 22: Number and mean of genre selections per age category.

C Comparison between genre distributions

Although the genre distribution for participants' currently preferred genres and for the sum of counts for the six age categories is quite similar, one remarkable difference was found. Popular music was more often selected when annotating taste (19%) than when annotating favoured genre categories (13%). This divergence is mainly due to the high score for children's songs in the youngest age category and the absence of "other" in the genre set presented for taste evolution. The global similarity of genre structure is in agreement with the assumption that respondents to the survey took the questionnaire serious and it assures that the data is reliable. Table 23 illustrates this comparison (❷ p.83).

Genre	Form 8	Form 9
Classical	12,0	12,9
World/ethnic	7,8	7,1
Light/oldies	7,0	7,6
Pop	13,0	19,0
Rock/metal/noise	10,8	11,2
Jazz	8,7	6,5
Blues/soul/reggae	6,7	6,9
Folk/country	5,3	5,1
Rap/hip hop	5,2	4,0
New age	2,0	2,2
Dance/house/techno	7,1	5,8
Soundtracks	6,9	5,8
Childrens	1,5	5,9
Other	5,9	NA

Table 23: Comparison of distributions for "genre" (F8) and "taste" (F9).

5.4.5 Favourites

Given the aim that the survey would recruit participants for annotation experiments with real music, this called for extra information on people's favourite music. At the end of the survey, an open-ended questionnaire was included in view of creating a music database for the experiments involving said subjects. The purpose of composing such a database was twofold. Firstly, it would be representative of the preferred music of the envisaged subject population (i.e. reflecting their genre as well as their title preferences). Secondly, it was expected that subjects annotating real music would perform the task more accurately when dealing with familiar music. In experimental psychology - where music is used as one of the more effective mood induction procedures - stronger effects have been obtained if one used music selected by the subjects themselves rather than selected by the experimenters (Carter et al., 1995, Thaut and Davis, 1993). The following question was asked hoping participants would feel encouraged to provide many titles.

Q #26: Give titles of pieces of music and names of composers, songwriters or performers you very often listen to. Let us also know the genre you think this music belongs to. Also rate the descriptions that fit with the music. There is space for ten pieces of music you are free to name less than that amount. However, the more titles we get, the better you help us with our research.

523 (74%) participants in the survey did indeed provide from one up to ten titles of their preferred music together with the composer or performer of the piece in question. As a result, we arrived at a list of 3021 titles representing music with long-term familiarity to the participants. A quarter of the people (N=132) who presented titles (N=523) gave the maximum of 10 examples. The 3021 titles were taken as a starting point for creating the music database used in the annotation experiment reported in chapter six. Table 24 shows the number of subjects according to the number of titles supplied. (❷ p.84).

N title	N subjects	%
1	523	17,3
2	467	15,5
3	419	13,9
4	361	11,9
5	309	10,2
6	258	8,5
7	216	7,1
8	181	6
9	155	5,1
10	132	4,4
Total titles	3021	100

Table 24: Distribution of favourite titles.

They were also asked to give an indication of the genre they thought the piece belongs to (open-ended). Here, the focus was first on collecting information on people's familiarity with styles for music they have known for a long time and second, on evaluating the categories in the fixed genre taxonomy that had been presented.

A Preferred genres

Genre descriptions were checked and labelled manually by music experts. For pieces the experts did not know, aural examination was done. Where possible, the same labels as used in the fixed genre taxonomy were applied. Labelling of the genres generated two new variables for each of the 3021 titles. The first variable (see table 25 column "favourites") contains labels for the fourteen classes as used in the genre taxonomy plus two new classes that are "variable" and "not known". "Variable" was used when subjects had given more than one genre and "not known" was used for cases where subjects wrote that they could not name the genre or gave a wrong indication. As in the survey, the class "other" applied to genres that were not included in the fixed genre taxonomy. When the participant named a sub-category instead of a general genre class, then the label for the latter was

assigned. For example “opera” or “renaissance” was marked as “classical”. When the relation with a particular genre class was not clear, it was assigned to the “other” class, for example, “piano music”, “instrumental”, “rhythmical” or “romantic” are terms that can refer to different styles. The second variable (see table 25 column “favourites RE”) implies the manual classification of the classes “variable”, “not known” and “other” into the global genre categories. Re-labelling was based on auditory control of the title. In some cases, for example, where the music was not found or the title was vague, it was impossible to perform re-labelling. Finally, a new genre class “electronic” was introduced. It was the only term that occurred frequently in the list that was not included in the fixed genre taxonomy. In most of the cases it appeared to refer to the use of electronic instruments in one of the fixed genre classes (e.g. rock guitar), but in other cases (1,4%) is referred to a particular composition style (i.e. electronic music).

Genre	Genre %	Favourites %	Favourites RE %
Classical	12,0	16,3	17,7
World/ethnic	7,8	3,5	5,3
Light/oldies	7,0	2,3	6,3
Pop	13,0	16,7	20,4
Rock/metal/noise	10,8	19,5	23,5
Jazz	8,7	4,1	4,6
Blues/soul/reggae	6,7	3,1	3,8
Folk/country	5,3	2.0	2,4
Rap/hip hop	5,2	3,1	3,6
New age	2,0	1,5	1,5
Dance/house/techno	7,1	4,4	5,4
Soundtracks	6,9	3,1	3,2
Children	1,5	0,4	0,4
Other	5,9	10,6	0,0
Variable		3,7	0,1
Not known		5,8	0,3
Electronic			1,4

Table 25: Comparison of genre distributions in “genre” and “favourite titles”.

It was found that there is a clear preference for pop and rock styles, but classical music is strongly represented as well. Pop, rock and classical thus remain the top three genre categories. Compared with the genre questionnaire, the titles of people’s favourite music include remarkably higher percentages: pop (+7,4%) , rock (+12,6) and classical (+5,2%). This picture suggests that the types of music most of interest to a music information retrieval system user will be pop, rock and classical genre categories. Tables and figures with detailed overviews of genre distributions according to the number of titles are included in (❷ pp.85-100).

B Greatest hits

Although three genre categories clearly stood out, a great variety of titles was found. Only 8,5% of the favourite titles was given by more than one respondent with a maximum of fourteen times. The top three composers in the favourite list are Bach (56), Mozart (45) and Beethoven (26). The top three titles are Mozart's "Requiem" (14) Yann Tiersen's "La valse d'Amélie" soundtrack for the movie "Le fabuleux destin d'Amélie Poulin" (10) and Bach's "Matthäuspassion" (8). This can be explained from the popularity those composers and titles have in general. Mozart's "Requiem Mass" (KV 626), for example, is often considered as one of the best compositions ever. In popular imagination this requiem has frequently been connected to the myth surrounding the composer's death. The film "Le Fabuleux Destin d'Amélie Poulin" (2001) has met with incredible approval of almost every festival audience. Tiersen's waltz for the movie has moved many nostalgic and romantic souls. Finally, in 2000, the musical world commemorated the 250th anniversary of the death of J. S. Bach, with international Bach year events and the release of many recordings of his music.

5.4.6 Qualities of favourite titles

Additionally, for each title that was given, two sets of five bipolar adjectives were presented to be rated on a scale. Of course, we have to bear in mind that people will evaluate rather positively when their opinion is asked about preferred music. The goal of including quality annotation in the survey was double. Firstly, it was meant to give a global impression of the way participants judge music characteristics of favourite music in terms of expression and structure. Secondly, it would serve as a test case in view of defining requirements for the experiment about quality annotation of music.

Q #27: Which mood best fits this music?

Q #28: How does this music sound?

Annotation of expression and structure by means of ratings is a difficult task for people who are not music experts. As a consequence the results from this part of the survey might yield some ambiguity. Given the set up of the investigation as an online activity, it was impossible to have fully control over the situation. A text on the screen explained what people were expected to do:

- "Select for each pair of adjectives the point on the scale that best fits the global mood of this piece of music. For example for the pair "cheerful - sad" this means that when the music is "very cheerful" you should select the point on the extreme left of the scale and for "very sad" you should select the point on the extreme right. The button in the middle means that you think that none of both adjectives fits. When the music both contains "cheerful" and sad "parts", you are expected to select "variable".

The choice of the adjective pairs was based on previous research on semantic differential studies of music (Leman, 2003) and a pilot study that was performed at IPEM, Ghent

University (December 2003) within the context of a course “Music Psychology”. In view of constructing the sets of adjectives for the survey and the experiment that was going to follow, thirteen musicology students were requested each to provide ten musical excerpts of 30 seconds. Multiple genres were allowed. The students had to label the most typical emotions/feelings that they thought the music fragments they had chosen evoked (a list with the titles of the pieces chosen by the students and the annotations they made is included in ❸. The adjectives were analysed and went back to the group for classical reevaluation of their relevance. In a separate task the students had to choose six pairs of adjectives from the list of fifteen that was used in the previous experiment (Leman et al., 2003) and argue their choice. Finally, ten adjective pairs were chosen for the questionnaire. This set was then tested and discussed by music experts working at IPEM. In the survey, the adjectives were presented in two sets, related to expression and style. Table 26 shows the bipolar adjectives for *expression* (set I). and *style* (set II) presented for quality annotation of favourite titles (in Dutch and English).

Set I	Dutch	English	Dutch	English
I.1	Opgewekt	Cheerful	Triestig	Sad
I.2	Zorgeloos	Carefree	Angstig	Anxious
I.3	Teder	Tender	Brutaal	Brutal
I.4	Kalm	Calm	Rusteloos	Restless
I.5	Passioneel	Passionate	Koel	Restrained
Set II	**Dutch**	**English**	**Dutch**	**English**
II.1	Zacht	Soft	Hard	Hard
II.2	Langzaam	Slow	Snel	Quick
II.3	Helder	Bright	Donker	Dull
II.4	Harmonieus	Harmonious	Wringend	Rough
II.5	Dynamisch	Dynamic	Statisch	Static

Table 26: Bipolar adjectives used for the annotation of favourite titles.

The adjectives had to be judged on a 7-point scale (i.e. very, rather, little, not, little, rather, very). The quality ratings are responses based on long-term memory and regard the entire piece of music. Although it was requested to give titles of music that respondents hear very often, it cannot be measured how frequently “very often” actually is. Besides, people were not requested to listen to the music before ratings were made and we do not know if some spontaneously did so. Therefore, it has to be taken into account that the judgments might be limited by the vagaries of memory. As a result of this constraint, statistical significant inferences cannot be made. However, adding more questions to this final part of the survey was assumed to produce a negative effect on the feedback of titles.

Bearing in mind that some people might find it hard to make a selection from a specific pair of adjectives, the possibility was offered to choose “variable”. Indeed, non-experienced subjects might have doubts on the meaning of the adjectives when connected to music. Music experts on the other hand, who are assumed to concentrate more on the music,

might find it difficult to decide on rather vague terms that are supposed to be valid for an entire piece of music. A table was constructed, including all quality descriptions of respondents' favourite titles, organized per genre category. It was impossible to perform inferential statistical analysis on the set of titles because the occurrence of alike titles was too small (i.e. highest occurrence = 14 for Mozart's "Requiem"). Explorative analysis was performed on the frequencis (counts) for all the adjective ratings. The titles for which the annotations were incomplete[98] were omitted. After data cleaning quality annotations for 2937 titles (out of 3021) (97,8%) were kept for analysis.

A Global annotation of affect

Over the whole set of titles, looking at the sum of percentages for the ratings "very much" and "rather", the highest rating for affect qualities was assigned to "passionate" (47,6%). The next important affect feature was "cheerful" (31,5%) followed by carefree (29,3%) and tender (27,5%). Table 27 and figure 15 show the results (in percent) for all rating levels for each pair of affect adjectives. Detailed tables and figures are included in (❷ pp.101-105).

Affect	Very	Rather	Little	Not	little	Rather	Very	Variable
Cheerful-Sad	16,9	14,6	10	9,4	13	13,2	9	13,9
Carefree-Anxious	15,3	14	12,2	17	16,1	9,5	5,5	10,4
Aggressive-Tender	7,2	9,5	11	16,3	13	15,6	16,1	11,4
Calm-Restless	9,8	10,8	11,2	13,1	14,8	16,4	11,1	12,8
Passionate-Restrained	24,5	23,1	21,8	14	5	2,8	2	6,9

Table 27: Ratings in percentage points for the adjective pairs related to affect.

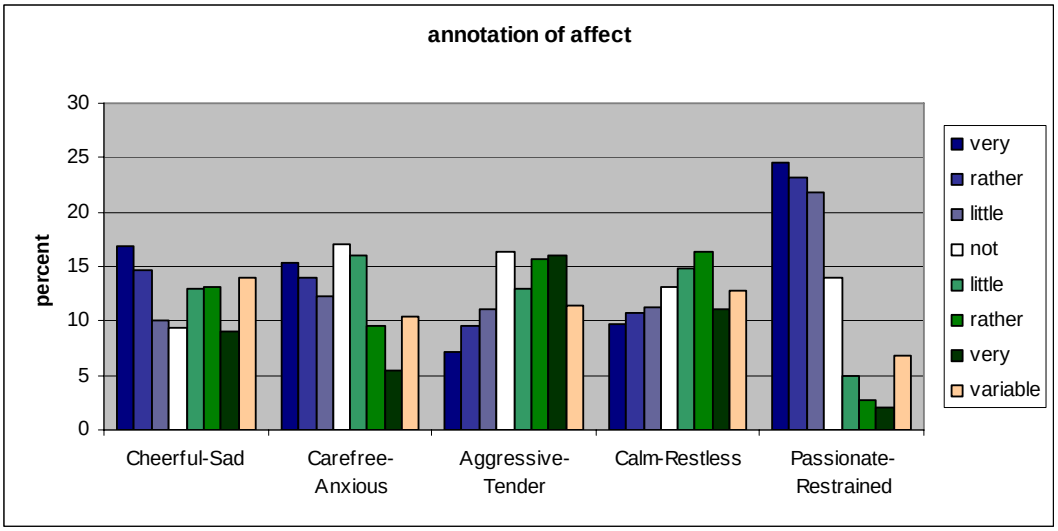


Figure 15: Annotation of affect.

[98] This happened when a respondent stopped in the middle of an annotation process or when the time limit of 10 minutes between two judgments was exceeded.

B Global annotation of structure

In hierarchical order, the highest ratings for structural features were assigned to “bright” (34,9%), dynamic (32,4%), harmonious (31,9%) and soft (29,6%). Table 28 and Figure 16 show the results (in percentage points) for all rating levels for each pair of adjectives related to structure. Detailed tables and figures are included in (❷ pp.106-110).

Structure	Very	Rather	Little	Not	Little	Rather	Very	Variable
Soft-Hard	14,3	15,3	15,4	10	13	9,6	7	16
Slow-Quick	8,8	12,1	13,7	10,1	16	12,2	7	20
Bright-Dull	15,4	19,4	16,7	11,2	12	9,8	5	11
Rough-Harmonious	2,3	4,7	8,3	12	15	25,6	21	11
Static-Dynamic	1,5	3,4	6,3	14,9	18	23,7	21	12

Table 28: Ratings in percentage points for the adjective pairs related to structure.

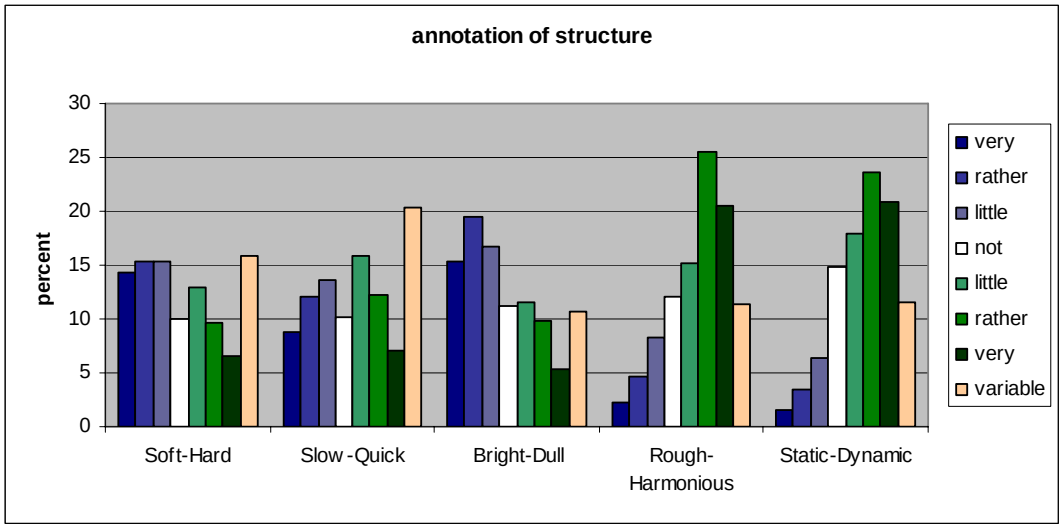


Figure 16: Annotation of structure.

C Assessment of genre

Given the finding (see supra) that pop, rock and classical are the most occurring styles for global genre selection, taste and favourite titles as well, these genres clearly include participants most preferred music. To have an idea of the spread of these categories, the genre taxonomy was grouped into four broad categories including subclasses that can easily be classified under one of the main categories. The new genre taxonomy comprises classic, pop, roots and other.

Table 29 shows the structure of the grouped genre categories. This configuration was the starting point for the genre classification in the database that was created in view of the experiment on annotation of music qualities reported in chapter six.

Genre	Grouping	%
Classical	Classical	17,7
New Age Soundtracks Light music Pop Rock/Metal/Noise Electronic music Dance/House/Techno Rap/Hip hop Soul/Reggae	Pop	63,9
Jazz & Blues World/Ethnic Folk/Country	Roots	16,1
Children Not known Variable	Other	2,3

Table 29: Taxonomy of broad genre categories.

Looking at the affect and structure ratings (❷ pp.111-120) for the different genres categories, it was found that there are tendencies in the way participants assign emotions and "sound like" characteristics to music styles. In what follows a summary according to the highest rankings (rather and very) in "classical", "pop" and "roots" is given. The class "other" was not considered because it represents only a small percentage (2,3%) of the titles.

Affect ratings for genre category "classical"

In all but one adjective pair, the highest percentage was found for "variable". The exception was "passionate-restrained" that ranks "very passionate". Second-highest rankings were "little anxious", "rather tender" and "little restless".

Affect ratings for genre category "pop"

Pop music was evaluated by the extremes "very cheerful", "very passionate" and "rather restless". The adjective pairs "aggressive-tender" and "carefree-anxious" reveal highest percentages for "not". The second-highest ratings for these pairs were "very careless" and "very tender".

Affect ratings for genre category "roots"

Root music was judged by the extremes as "very cheerful", "very carefree" and "very passionate". The adjective pairs "aggressive-tender" and "calm-restless" reveal highest percentages for "not". The second-highest rating for "aggressive-tender" was "very tender". Ratings were equally assigned to "little calm" and "little restless".

Structure ratings for genre category "classical"

Similar to affect ratings the option "variable" was chosen frequently. This is the case for the adjective pairs "soft-hard", "slow-quick" and "bright-dull". Apart from this, classical music has been evaluated as "rather harmonious", and "rather dynamic".

Structure ratings for genre category “pop”

Pop music was estimated restrained: “little quick”, “rather bright”, “rather harmonious” and “rather dynamic”. There was some vagueness in the rating of “soft-hard” that hold highest percentages of ratings on both the positive and negative side of the adjective pair, namely “rather soft” and “little hard”.

Structure ratings for genre category “roots”

Music titles in the class “roots” were rated moderate: “little soft”, “little quick”, “rather bright” and “rather harmonious”. “Dynamic” on the other hand was judged “very”.

Variable and undecided

High ratings for “variable” and “not” require further analysis. A comparison between the attribution of “variable” and of “undecided” for all adjective pairs in the three genre classes was made (❷ pp.121-122).

Table 30 summarizes per set (i.e. variable, undecided) the two adjective pairs with highest percentages and the genre class (CL=classical, PO=pop, RO=roots) in hierarchical order.

		Affect	Genre	Structure	Genre	
	Variable	Cheerful-sad	RO-PO-CL	Slow-quick	RO-CL-PO	
		Calm-restless	CL-RO-PO	Soft-hard	CL-PO-RO	
	Undecided	Aggressive-tender	CL-RO-PO	Static-dynamic	CL-PO-RO	
		Carefree-anxious	CL-PO-RO	Soft-hard	CL-RO-PO	

Table 30: Adjective pairs with highest scores for “variable” and “undecided”.

To summarize, affect ratings according to broad genres categories for participants’ favourite music show global agreement on “very passionate”. Within the other adjective pairs there is a similar trend in the ratings for pop and roots. For the pair “cheerful-sad”, classical music contrasts with pop and roots as it is judged “rather and very sad”, while pop and roots are both judged “very cheerful”.

With regard to structure, among the three genre classes the ratings for “rough-harmonious”, “bright-dull” and “static-dynamic” fit in rather well with one another. All genres have been high-rated for harmonious and bright timbre and for dynamic articulation. Tempo in classical music was judged slower than in pop music and roots. Assessments of loudness vary. Although all genre classes show low ratings for “rather hard” or “very hard”, rating levels for “soft” differ among genres. The highest score for classic is “little soft”, pop is “rather soft” and roots is “very soft”.

Affect features get more extreme rating levels (very), while structure features globally get moderate rating levels (little, rather).

5.4.7 Relations

In the next step, the effect of age groups, gender and music background was investigated. From the dataset (N=663) four new binary grouping variables *music expertise, age*

categories ≤25/>25 and age categories ≤35/>35 and *broadness of taste* were created. For music expertise first a variable expertise level was calculated as being the sum of values attributed to the variables music way (listen/listen and play), music context (amateur/professional), music level (none/beginning/medium/high), music instrument (yes/no) and way of listening (active/active and passive/passive). The label “novice” was assigned to scores 1 to 6 for music level and the label “expert” was assigned to scores between 7and 10. The variable broadness of taste was based on the number of genre selections. Scores from 1 to 4 were labelled as “narrow” and from 5 to14 as “broad”.

Table 31 shows the distribution of the binary variables used. For each binary variable two-way contingency table (r x c tables) analysis using crosstabs was conducted to evaluate statistical significance of relationships between the categorical variables. The Pearson Chi-square statistic was used to estimate the likelihood that the differences between the observed and expected values would occur under the null hypothesis that there is no difference between these values.

gender		female	male
	%	44,2	56,8
musical expertise		novice	expert
	%	60	40
age category 25		below 25	above 25
	%	49,6	50,4
age category 35		below 35	above 35
	%	76,6	23,4
music education		yes	no
	%	63,2	36,8
active musicianship		play	do not play
	%	57,2	42,8
broadness of taste		< 5 genres	> 5 genres
	%	53,6	46,4
familiarity with classic		yes	no
	%	56,1	43,9

Table 31: Distribution of binary variables (N=663).

Multiple effects of gender, age, music background, and music taste were found. In what follows a summary is given of the main significant statistical relationships that were found between the categorical variables. Crosstabs for genre are included in (❸).

A Effects of gender

It is likely that:

- within the age category older than 25, two out of three (63%) are men;
- of those who listen to music and play music, two out of three are men (60%);
- of those who play an instrument, two out of three are men (59%);

- of those who cannot sing, three out of four (74%) are men;
- of the very good dancers, the majority (93%) are women ;
- of active listeners, two out of three (68%) are men;
- of experts, two out of three (62%) are men.

B Effects of age

It is likely that:

- of professionals, two out of three (65%) are older than 25;
- of the very good dancers, more than three quarter (79%) are younger than 25;
- of those who do not listen to classical music, two out of three (66%) are younger than 25;
- of those who listen to music and play music, more than three quarter (80%) are younger than 35;
- of those musically educated, almost three quarter (70%) are younger than 35;
- of those who play an instrument, more than three quarter (81%) are younger than 35;
- of the good singers, more than three quarter (80%) is younger than 35;
- of those who do not listen to classical music, the majority (86%) is younger than 35.

C Effects of expertise

It is likely that:

- within gender, two third of the women are novice (65%);
- within the way people deal with music, most of those (96%) who just listen to music are novice;
- within professionals, three out of four (77%) are experts;
- most (98%) people who had no music education are novice;
- most people with no (99%) or beginning (86%) music level are novice;
- most (99%) people who do not play an instrument are novice;
- most (88%) of the very good singers are experts;
- of classical music listeners, more than two third (70%) are expert;
- of active listeners, two out of three (66%) are expert.

D Effects of music education

It is likely that:

- within the age category younger than 35, two out of three are musically educated (66%);
- most people who listen to music and play music are musically educated (95%);
- of the amateurs, almost two third (62%) are musically educated;
- of professionals, more than three quarter (81%) are musically educated (95%);
- most people with medium (97%) or high (99%) music level are musically educated;
- most (95%) people who play an instrument are musically educated;
- most (93%) of the very good singers are musically educated;
- of classical music listeners, three out of four (75%) are musically educated.

E Effects of playing an instrument

It is likely that:

- within the age category younger than 25, almost two third (62%) play an instrument;
- within gender, almost two third (61%) of the men play an instrument;
- of professionals, more than three quarter (77%) play an instrument;
- of musically educated people, the majority (86%) plays an instrument;
- of the very good singers, the majority (84%) plays an instrument;
- of those who listen to classical music, two out of three (63%) plays an instrument.

F Effects of broadness of taste

It is likely that:

- of people who cannot dance, two out of three (67%) have a narrow taste;
- of those who do not listen to classical music, two out of three (63%) have a narrow taste.

G Effects of listening to classical music

It is likely that:

- within the age category older than 35, three out of four (74%) listen to classical music;
- of people who listen to music and play music, two out of three (64%) listen to classical music;
- of professionals, three out of four (74%) listen to classical music;
- of musically educated people, two out of three (64%) listen to classical music;
- of people with a high music level, three out of four (76%) listen to classical music;
- of people who play an instrument, two out of three (62%) listen to classical music;
- of good singers, more than three quarter (79%) listen to classical music;
- of people with a broad taste, two out of three (65%) listen to c classical music.

5.5 Conclusion

The aim of the survey that was reported in this chapter was to define the global profile of the so-called music information retrieval population and to find out what are the styles and qualities of the preferred music. With 774 participants representing a broad distribution of “music lovers and technology minded people” the sample size was significant. The number of respondents was sufficient to make quantitative and qualitative deductions.

Trying to ascertain real users’ musical background, habits and interests lead to results that support the assumption that music plays an active role in the lives of the people that will use interactive music systems. This is in agreement with the hypothesis that the targeted population consist of dynamic music lovers. According to the findings in the survey, a global profile of the envisaged users can be outlined.

The average music information retrieval system users:

- are young people (three quarter is younger than 35 years);

- are used to working with the Internet (92,6%);
- spend one third of their Internet time on music related activities;
- do not earn their living with music (more than 90 % is amateur);
- are actively involved with music: about two out of three had a music education and one out of five has a high-level of music expertise;
- have the broadest music taste in the age category between 12 and 25 years;
- have as their preferred genre categories pop, rock and classical music;
- are good at genre description;
- have difficulties assigning qualities to classical music;
- assign most variability to classical music.

Effects were found of:

- gender;
- age;
- expertise;
- music education;
- playing an instrument ;
- broadness of taste;
- listening to classical music.

6 Description of High-level Features

This chapter (based on original article VI) deals with the description of high-level features of music. High-level concepts contribute to the definition of meaning in music. But can meaning and emotion in music be measured? High-level features are hard to describe because they are based on subjective feelings and, therefore, user-dependent. In order to figure out whether music can be described by means of high-level concepts, an experiment which focuses on the annotation of high-level features was set up. The subject group consisted of some of the participants from the survey reported on in the previous chapter. This annotation experiment explores how real users perceive affects in music, and what structural descriptions of music best characterize their understanding of music expression. The experiment on perceived musical qualities regards the annotation of real music stimuli. These stimuli comprise a dataset with excerpts of music that are well-known to most of the participants. It was assumed that their familiarity with the music they hear would lead to high quality responses, and consequently to reliable data. The aim was to reveal relationships that could support making the link between musical structure and musical expressivity. Inter-subjective similarities and subjective differences were also investigated.

6.1 Background in annotation of high-level features

It has been shown that most people agree that music is able to evoke emotions (Lavy, 2000). Quite some psychological and social behavioural studies over the past few years have explored this issue (Juslin and Sloboda, 2001). An overview of music stimuli, subjects, response types and descriptive attributes in empirical studies of expression in music is given in (Gabrielson and Juslin, 2003). These studies have clearly shown that many features of music are perceptual and user-dependent. The literature scan below draws attention to the description of high-level features, subsequently from the viewpoint of *emotion-based music information retrieval*, commercial applications and *user studies* about the *attribution of affect and structural content* respectively.

6.1.1 Emotion-based music information retrieval

Researches in several disciplines underline the importance of affect and emotion in system development (Zhang and Li, 2005). Although emotion has been shown to be useful for music information retrieval, investigation analysing music qualities based on annotation of real music is scarce. Yang and Lee (2004), for example, claim that machine learning techniques can be embedded in annotation tools using software agents, in support of high-level design goals. They evaluated a structured emotion rating model for integration into software agents. No studies, however, have thus far been reported that gather annotations from potential users of content-based music information retrieval systems. Because of this user-dependency, emotional features in particular remain hard to describe. As an alternative, Chai and Vercoe (2000) suggested bringing user models into play in music information retrieval. However, in order to develop such models, investigation is needed

that supports the definition of music information retrieval user groups (e.g. experienced-novice) and their perception of emotion in music. Above all, there is lack of information on MIR system user's interests and their definition of high-level features of music. Music retrieval systems, on the other hand, scarcely regard query-by-emotion because there are no databases available with annotated music that would support the development of tools for emotion-based content extraction of music.

6.1.2 Commercial applications

Kalbach (2002) reviews three commercial applications (i.e. MoodLogic Music Management Client[99], Barnes and Noble Jazz Discover[100] and Glass Engine[101]) that allow people to search for songs based on musical characteristics and moods. Although Kalbach praises the innovative character of these programs in using emotions for music organization and retrieval, there are quite some shortcomings. MoodLogic, for example, is a song organizer powered by a proprietary network of song data provided by the MoodLogic user community and professionals employed by MoodLogic. However, the system only supports MPEG-1 audio files (e.g. MP3) and requires specific filenames.

Existing systems (e.g. MoodLogic, All Music Guide MSN, the Glass engine and Musicat) that allow people to search for songs based on musical characteristics and moods can be confusing to the user. The reason is that in commercial systems, annotation of music emotion relies on ad hoc taxonomies which have not been based on empirical studies. The MoodLogic mood classification, for example, is limited to six adjectives which seem rather trendy (i.e. aggressive, upbeat, happy, romantic, mellow and sad). The underlying correlation between terms like "romantic" and "mellow" is not clear.

Another example, The Glass Engine, offers a mix of five musical features that pertain to expression and structure categories (i.e. joy, sorrow, intensity, density and velocity). One could ask for example what images or concepts thinking of "music density" or "music velocity" evokes in an average music information retrieval user. More investigation is needed to give a clear view of the user's perception of these concepts and of their usability in music information retrieval system development.

6.1.3 User studies

Kim and Belkin (2002) report on a study in which eleven students were asked to write words describing seven pieces of classical music. All the words used by the subjects were classified into seven categories. Remarkable in this study is that the highest ranked category is the one that comprises words that explicitly indicate emotions (31%). However, this study does not allow generalization as the sample was very small, restricted to

99 http://www.moodlogic.com
100 http://music.barnesandnoble.com/jazz developed by Savage Beast Technologies.
101 http://philipglass.com

students who were non-music experts and the stimuli were limited to classical music. A more rigorous methodology is needed to categorize perception of musical features.

6.1.4 Attribution of affect and structural content

The study reported here expands on experiments previously conducted at IPEM, Ghent University (Leman et al., 2004 and 2005), both in terms of scale and approach. In the studies conducted by Leman et al., affect attribution was done by university students while structural content was judged by music experts (i.e. musicologists). However, in order to define relations between users' background and their description of expressive qualities and structural features of music, annotations provided by those same subjects are needed.

The focus in this study was on gathering reliable and representative data as a basis for fine tuning of the methodology for modelling music affect, as presented by Leman et al. Unlike previous research where the selected stimuli were assumed to be unknown to the subjects, the idea now is to work with participants having a high degree of familiarity with the music which they are requested to annotate.

To summarize, from a literature review and previous experiments conducted at IPEM, it is clear that in music information retrieval research opinion and attitude questions are most challenging and still need a lot of investigation. So far, there exist no music databases annotated by means of characteristics of expression and structure in music according to the perception of music information retrieval users. Furthermore, user dependency with regard to description of music has been acknowledged but also largely neglected, as becomes clear from the fact that experimental research recruited subjects from selected populations of university students or musical experts.

6.2 Description of music qualities

The perception of qualities in music is highly subjective, and therefore, it can be expected that it is highly user-dependent. The study of qualitative description of music handles subjective features related to cognitive and emotional or affective issues and is situated at the high-level of the conceptual framework that forms the theoretical framework for this study[102].

The description of perceived music qualities involves four aspects: (1) the attribution of affect and emotion, (2) the description of structural features, (3) the description of involved activity and (4) the impact of memory. In what follows an explanation of these four description aspects is given.

[102] In contrast with this, research on music quantities regards objective and measurable concepts that are situated at the low- and mid- levels of this conceptual framework (e.g. tempo, pitch, loudness). These features are less complicated to describe, because they can be extracted from the music itself. However, little is known about the interaction between those levels and there is no agreement on a methodology that connects bottom-up and top-down approaches.

6.2.1 Attribution of affect and emotion

Emotion is an expressive category that reflects aspects of a person's mental state, normally tied to the person's internal (physical) and external (social) feelings. Studies in social psychology indicate that emotion occurs before cognition but also intervenes with cognition (Russell, 2003).

Affect is considered a type of emotion or *subjectively experienced* feeling. According to Norman (2002), affect and cognition have in common that they both are information processing systems, but with different functions and operating parameters. The affective system is judgmental, assigning positive and negative valence to the environment rapidly and efficiently. The cognitive system interprets and makes sense of the world.

Russell (2003) defines a "*core affect"* as a "neuro-physiological state consciously accessible as simply feeling good or bad, energized or enervated" (p.147). It is considered as a combination of valence (pleasure-displeasure) and arousal or activation (sleepy-activated, the degree of feeling energized). From this viewpoint, core affect is non-reflective.

Affect quality is related to the ability of a stimulus to cause or change core affect. Whereas core affect exists within a person, affect qualities exist in the stimulus. Emotion qualities enter consciousness being affectively interpreted.

Perception of affective qualities of music is the individual perception of the ability of a music stimulus to change an individual's core affect. Perception of music qualities influences the subsequent reactions a person has to music.

6.2.2 Description of structural qualities

Structural music information involves many features and representation forms such as the hierarchical structure of form (e.g. movements, sections, subsections) and of harmony (e.g., key, cadence, harmonic function). In the perception of polyphonic music, several features appear simultaneously and the perceived structure that comes to the fore may vary from person to person. In order to minimize this perceptive variability the stimuli in the quality annotation experiment are not just taken from the beginning of the music pieces. The excerpts have been selected by musical experts by the criterion of maximum homogeneity. The focus is thus on the musical **sound qualities** that constitute the perceived essence for each fragment. As a consequence, the chosen structural components relate to the description of the *most typical* musical features in the fragments. Structural qualities include sound qualities such as loudness, timbre, melody and harmony, tempo and rhythm, and **articulation**. These are constituent elements of the overall musical structure.

6.2.3 Description of involved activity

In contrast with the static approach to structural qualities, **activity qualities** relate to the physical experience of movement. Activity qualities are reflected by the physical response of humans to music, such as tapping with the toes, waving one's arms or swinging around. Activity qualities rely on concepts such as **gesture energy** and **catchiness**. Here, the idea of gesture is regarded as an expressive concept related to body movement of a listener while hearing music. Gesture in this sense refers to the notion of *expression of feelings*. The concept of catchiness is defined by features that are *easy to recognize* and *easy to imitate* such as a singing along with an attractive melody or tapping a repeating rhythmic pattern.

6.2.4 Impact of familiarity

In addition, the concept of **familiarity** with the music examples or **styles** has been linked to the annotation of music qualities. The degree of acquaintance with music in general and the knowledge of a specific **piece** in particular are assumed to affect the attribution of emotion, the description of structural qualities and the involved activity as well.

In figure 17 a global overview of the model for the description of qualities at the high-level of the conceptual framework is given.

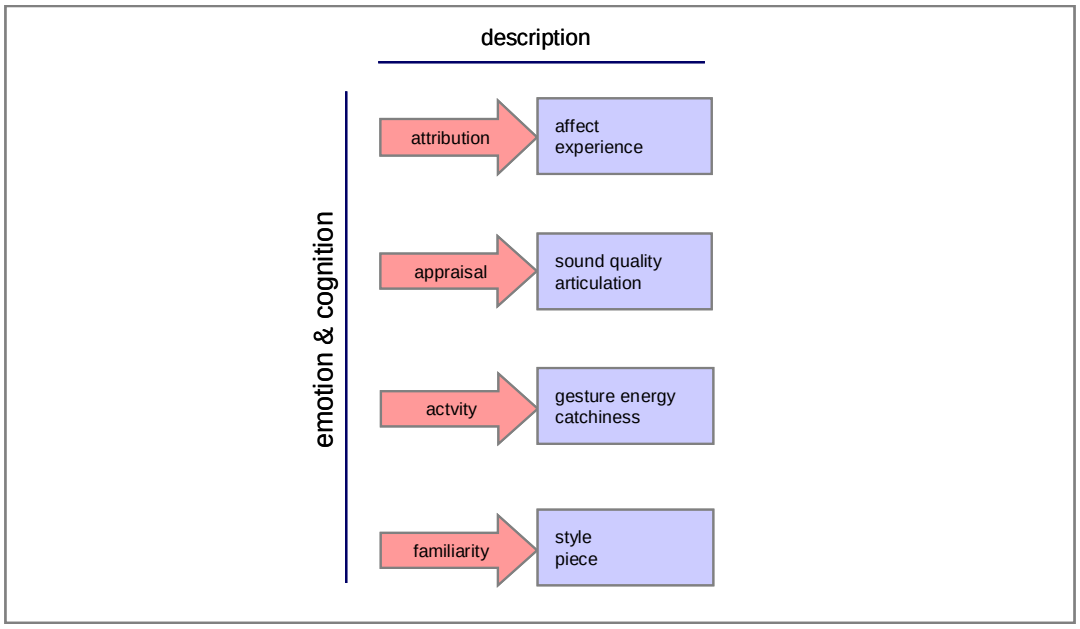


Figure 17: Model for quality description.

6.3 Method

6.3.1 Subjects

An invitation was sent to 490 respondents of the survey on user context (see previous chapter) who at a certain point in time (2004.07.15) had filled out all 10 forms of the

questionnaire and said that they were willing to participate in future annotation experiments. Out of 490 invited persons, 115 responded of whom 94 finally participated in the experiment.

6.3.2 Stimuli

Out of the 3021 favourite titles provided by the participants in the survey, 160 titles were selected for the creation of a test database with music fragments (see 5.4.5). This database was divided into three categories: “classical”, “pop” (with a balanced distribution of pop music, rock/metal/noise, electronic music, dance/house/techno, rap/hip hop, soul/reggae, light music, new age and soundtracks) and “roots” (jazz, blues, world/ethnic, and folk/country). Table 32 shows the distribution of the 160 music fragments over the three broad genre categories and the subclasses involved.

CATEGORY	Num	%	SUBCLASS	Num
Classical	49	30,6		49
Pop	69	43,1	New Age	2
			Soundtracks	4
			Light music	4
			Pop	15
			Rock/Metal/Noise	24
			Electronic music	5
			Dance/House/Techno	6
			Rap/Hip hop	5
			Soul/Reggae	4
Roots	42	26,3	Jazz & Blues	16
			World/Ethnic	21
			Folk/Country	5
TOTAL	160	100		160

Table 32: Distribution of genre categories and sub-classes.

They contain respectively 49, 69 and 42 musical fragments. The number of titles in the three categories was based on the statistics of the top three classes in the genre questionnaire. The choice was made for three more or less distinguishable style categories and a large representation of audio within each in order to have enough material for statistical tests, in which the effect of style on musical perception would be analysed. It is assumed that familiarity with music increases the competence for annotating. Therefore, only titles with a high popularity, measured by their repeated occurrence in the survey study of chapter five, were retained for selection. The criterion was that the same title should have been mentioned by at least ten different participants of the group that provided favourites (N=523). This choice resulted in a set of music examples displaying a broad diversity of music. Then, for each title that was selected, one excerpt of exactly 30 seconds was recorded. This method generated a dataset on which most subjects could probably

agree. In addition, the dataset should reflect the musical preference of an average user of a music information retrieval system. The main criteria for the selection were *homogeneity* within the fragments[103] and *absence of lyrics*[104]. The set of 160 music excerpts were digitally extracted from various commercial music CD's. The selection from the original piece was done by musical experts on the basis of global musical homogeneity. The excerpts were carefully selected so that the beginning should sound natural (e.g. the beginning of a melodic phrase). The end of the excerpt was faded out. The 160 excerpts were converted to a 16 bit 44100 kHz PCM .wav format. The intensity remained the same as in the original; only in case of saturation the amplitude was normalized. A table with the references for all music excerpts and onset-offset times of the excerpts used is included in ❸.

6.3.3 Design

The experiment consisted of seven forms gathering ratings of adjectives representing expressive and structural qualities of music. Respondents were also asked to give more information on the music excerpts with regard to familiarity, appreciation, degree of difficulty and physical response. In table 33 an overview is given of the design of the annotation experiment showing the main parts, the adjectives involved and between brackets the abbreviations used in the analysis.

	I. EXPRESSION	II. STRUCTURE	III. ACTIVITY	
	I.1 AFFECT (aff)	II.1 SOUND		
	Cheerful (cheer)	Soft/hard (SoHa)	gesture (mov)	
	Sad (sad)	Clear/dull (ClDu)	melody imitation (imi)	
	Carefree (care)	Rough/harmonious (RoHa)	IV. MEMORY	
	Anxious (anx)	Void/compact (VoCo)	No recognition (no)	
	Tender (tend)	Slow/quick (SloQu)	Style recognition (style)	
	Aggressive (aggr)	Flowing/stuttering (FloStu)	Vaguely known (vague)	
	Passionate (pass)	Dynamic/static (DySta)	Well known (well)	
	Restrained (restr)			
	Most typical (mt)			
	I.2 EXPERIENCE	II.2 TYPICAL	V. JUDGMENT	
	Annoying (anno)	Timbre (tim)	Beautiful/awful (BeAw)	
	Pleasing (plea)	Rhythm (rhy)	Difficult/easy (DiEa)	
	Touching (touch)	Melody (mel)		
	Indifferent (indif)	None (no)		

Table 33: Overview of the design of the annotation experiment.

[103] See 6.2.2: Description of music structural qualities.
[104] This option was based on the idea that participants should respond to the music. With sung pieces there is a risk that the annotation of affect would rely more on the lyrics than on the music. For the same reason titles were not shown either. Of course, given the criterion of familiarity, recognition could not be excluded.

To avoid ambiguity and make sure that participants focus on the music, questions that assess qualitative measures were phrased. For example, with regard to the question "*what feeling emanates from this music*?" the adjectives were incorporated in short phrases (e.g. this music is cheerful). A detailed overview comprising the questions and answers is included in (❷ pp.125-126).

A Annotation of expressive features

The annotation of "expressive features" (I) was subdivided in two sets of adjectives related to *affect* (I.1) and *experience* (I.2). Expression was viewed as based on the well known concepts of attribution and induction (Scherer and Zentner, 2001). For *affect* the feeling that emanates from the music was judged. Subjects had to make ratings on a 5-point scale (ranking from "not" to "very") for a set of eight adjectives that appeared in random order on a list. The list of adjectives contains opposite terms but a unipolar scale was used to avoid weak choices. By offering the possibility of having "no opinion" on a particular affect, people were not forced to make a rating. Additionally, respondents had to indicate which of the eight attributes they found the most important. The *experience* (I.2) section focused on the mental and physical state of the subjects, i.e. on the effect of the music on the subject. Participants rated four adjectives that were presented in the same way as for *affect*.

B Annotation of structural features

The section dedicated to "structural features" (II) had two parts assessing participant's perception of how the music *sounds* (II.1) and what *characteristic* features (II.2) are. Unlike for expressive features, ratings of sounding qualities (II.1) were done on a 9-point scale (bipolar version of the 5-point scale used for affect). Indeed, it is obvious that when music is soft, it is not hard. For affect features, on the contrary, it is difficult to maintain bipolarity. One can easily imagine a piece of music that at the same time expresses some degree of cheerfulness and some degree of sadness. As an indication of the most typical structural feature(s) (II.2) multiple choices had to be made between timbre, rhythm and melody.

C Annotation of involved activity

In the form called "activity" (III) the concept of physical response was examined. The two following phrases: "*to this music I start moving spontaneously*" and "*I could imitate the melody*", were rated on a 5-point scale (ranking from not to very).

D Annotation of familiarity with the music

The extent of *familiarity* with the excerpts was checked in the part dedicated to "memory" (IV). A forced choice had to be made between four options: (1) not knowing the piece of music or the genre, (2) not knowing the piece of music but being familiar with the genre, (3) having heard the piece before and (4) having heard the piece a lot.

E Annotation of judgment

The last “judgment” (V) task was a highly subjective one. It was meant to learn what kind of music the participants personally like or dislike. The bipolar adjectives “beautiful-awful” and “difficult- easy” were rated on a 9-point scale.

6.3.4 Procedure

The experiment for manual annotation of perceived qualities of music was very extensive and time consuming. It took place in four sessions each presenting 40 excerpts of music, 30 seconds in length each. As each fragment had to be annotated for all the features involved (see previous section) 40 excerpts per session was considered the maximum. The experiment was conducted in groups of maximum ten participants who performed the test under guidance. The sessions took place in a computer classroom, the participants sat in front of a PC, while the music was being played through headphones. The order of the music excerpts was randomized. Only the number of the excerpts was presented on the screen. The test was conducted in Dutch (in this thesis the adjectives have been translated into English).

Before participants could begin with the actual experiment, the experimenter explained its aim and procedure of the study. Examples clarifying the meaning of the concepts were given. A “dull” sound, for example, was illustrated by a clap with the palm of the hand on a wooden table. After that, participants received a written guideline (included in ❸) and were asked for each fragment to fill out seven web-based forms containing scaling matrixes.

While giving their ratings, the participants could listen to the music excerpts as much as they wanted to. They rated each music excerpt using the phrased adjectives for qualities of music on the response sheet.

Figure 18: Web-interface for the annotation of affect features.

As an example, figure 18 shows the first of seven forms as it was presented to the participants. The question is “What feeling emanates from this music?” and the adjectives involved are: anxious, passionate, sad, aggressive, cheerful, restrained, carefree and tender. These adjectives had to be rated on a scale from o (not) to 4 (very) which was also

reflected in the colour scale. The neutral grey column next to the rating scale could be used when participants had “no opinion”. Per adjective one out of five ratings or “no opinion” had to be checked. In the red column, additional information had to be given by indicating which one of the eight adjectives best described the music excerpt they heard.

Whilst the total duration of the 40 music excerpts was 20 minutes, since the subjects could listen several times to each excerpt and since they had to fill out 280 forms, the total time of each session took about three hours on average. Although this was a demanding task, most participants enjoyed doing the tests and some of them even asked for more sessions than the four they had already done.

Out of 94 participants, 79 (84 %) participants judged the whole set of 160 musical fragments, one person stopped after three sessions, three after two sessions and eleven after one session.

6.4 Statistics

6.4.1 Data exploration

Prior to analysis, the raw data was controlled and imported in a relational database management system that was set up for this investigation (see chapter seven). Statistics were run using SPSS 12.0 to examine the means and standard deviations of the bipolar pairs as well as to check for possible outliers or entry errors. No outliers were found but some incompleteness in the data entry was corrected. The annotations of two persons who did not finish one set were omitted. Although they were motivated enough, they found the task too complicated. This leaves us with a dataset of 13360 entries provided by 92 subjects. Analysis was done for the 12640 responses generated by the 79 subjects who judged 160 fragments. Tables and figures representing the global distributions for the annotated features are included in (❷ pp.127-145).

6.4.2 Analysis

First, a comparison was made between the background information provided by the participants in the survey (N=663) and the participants who took part in the four sessions from the annotation experiment (N=79). Recall that the latter group of participants (the sample) forms a subset of the former group (the whole group). The aim was to investigate whether the outcome of the sample was typical of the whole group. Then, influence of subject-related factors when describing music features and the effect of familiarity with the music excerpts was investigated. Factor analysis was run in order to investigate underlying dimensions of the assessments. Standard deviations were used as a measure of unanimity in the perception of structural features. Finally, correlation analysis was done in order to establish relationships between variables.

6.5 Results

6.5.1 Comparison of datasets

Comparison between the main survey dataset (see chapter five) and the survey subset (the data provided by participants in the experiment) has shown that descriptive statistics for both cases are quite similar. Frequencies for “downloading music”, “being musical educated”, “being an amateur” and “being an active listener” for example, are approximately equal (less than 2% difference). The differences that have been found are rather small and support the idea of having a sample representing the targeted population. The only noticeable difference is that there is a better spread of age categories with a smaller group younger than 35 in the subset (58% instead of 75%). Figure 19 represents the difference in spread of age categories for the main dataset and the subset. Table 34 is a brief summary of the main differences and similarities found.

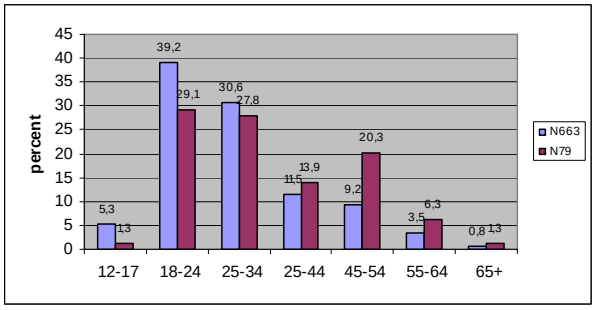


Figure 19: Comparison between age categories in the main dataset and the subset.

Comparison	Variable	Subset (N=79) %	Main (N=663) %
LESS	Younger than 35	58,2	75,1
	Self-taught in music	19,7	26,6
	Playing an instrument	53,2	57,2
	Playing plucked strings	14,3	19,9
	Playing electronic instruments	5,7	10,6
	Very good dancer	1,3	6,5
MORE	High music level	29,1	20,4
	Music education at university	8,5	3,3
	Very good singer	11,4	8,4
	Playing keyboard	31,4	25,6
	Listening to the radio	40,5	31,7
	Listening to classical music	16	12
SIMILAR	Downloading music	10,4	12
	Had music education	63,3	63,2
	Amateur	91,1	91,4
	Active listener	31,6	32
	Aged between 25-44	41,8	42,1

Table 34: Comparison between the subset and the main dataset.

6.5.2 Influence of subject related factors

Using the profile information of the 79 subjects, the data were divided into six binary categories, according to: (a) *gender* (46% female, 54 % male); (b) *musical expertise*[105] (based on answers about music education and music skills 58% of the participants were classified as expert, 42 % as novice); (c) *age category* (58 % is between 1 and 35 years old, 42 % older than 35); (d) *broadness of taste*[106] (50 % indicated listening to at least 5 different musical genres); (e) *familiarity with classical music* (75 % reported listening to classical music) and (f) *active musicianship* (53.2% plays an instrument). The table below shows the distribution of the binary variables.

BINARY VARIABLES			
(a) gender		female	male
	%	46,6	53,4
(b) musical expertise		novice	expert
	%	58,2	41,8
(c)age category 35		below 35	above 35
	%	58,2	41,8
(d) broadness of taste		< 5 genres	> 5 genres
	%	50,6	49,4
(e) familiarity.with classic		yes	no
	%	74,7	25,3
(f) active musicianship		play	do not play
	%	53,2	46,8

Table 35: Distribution of binary variables (N=79).

For each of these six categories, non-parametric Mann-Whitney tests were performed on data of each of the judged adjectives and adjective pairs describing the expressive (I), structural (II), and activity (III) qualities of the music examples. Calculations were performed with summed ratings of variables per participant (79 cases). Split up for the different parameters, an overview of the significant results is given in (❷ pp.146-147).

A Influence of gender

Although they are not numerous (from the 164 tests, 18 appear as significant), some effects of subject-related factors revealed by the Mann-Whitney tests warrant discussion. Finding the maximum of five significant values for the category “gender”, it is likely that gender has an influence on the perception of music. Men rated the music excerpts significantly more “restrained”, more “harmonious” and more “static” than women did. The latter then judged

[105] For the variable “music expertise” first a variable expertise level was calculated as the sum of values attributed to the variables music way (listen/listen and play), music context (amateur/professional), music level (none, beginning, medium, high), music instrument (yes/no) and way of listening (active, active and passive, passive). The label “novice” was assigned to scores 1-6 for music level and the label “expert” was assigned to scores between 7-10.

[106] The variable “broadness of taste” was based on the number of genre selections. Scores 1-4 have been labelled as “narrow” and 5-14 as “broad”.

the music to be more “beautiful” and more “difficult” (-2,87, p<0,01) than did their male counterparts.

B Influence of age

For the variable concerning age categories some significant values were found as well and this supports the notion that the age of the listeners is a contributing factor with regard to differences in perception of affect features: people older than 35 found the music more “passionate” (-2,33, p<0,01) and less “static” (-2, 80, p<0,01) than did younger listeners.

C Influence of music expertise

Listeners with no (advanced) music education judged the pieces as being more “cheerful”, more “passionate” and more “dull” than did experts. As could be expected, the experts have obviously less difficulty (-3,87, p<0,001) with imitating the melodies they heard in the musical pieces. The tests also taught that people with a broad musical taste judged the music as being more “pleasing” and more “beautiful” than did subjects with a narrow musical taste.

D Influence of familiarity with classical music

There is a significant influence (-2,47, p<0.01) of being a listener to classical music and the familiarity with the music excerpts in the stimuli set.

E Influence of musicianship

These tests reveal almost identical results as did the test on musical expertise, an extra significant effect being that a musician often considered the music as being more “static” than did a non-musician.

6.5.3 Effect of familiarity with the music excerpts

When the participants in the experiment indicated their familiarity with the music excerpts, the most occurring answer was “*I do not know this piece but I am familiar with the genre*” (34,3%), followed by “*I have heard this piece before*” (24,8%), “*I do not know this piece and I am not familiar with its genre*” (21,8%) and finally “I have heard this piece a lot” (19,1%). From these percentages it can be concluded that in 78,2 % of the cases people were familiar with the style of the excerpts and in 43,9 % of all the cases they even know the specific piece of music. These findings are in agreement with the aim to minimize the amount of unfamiliar pieces in the stimuli set.

In addition to the above described Mann-Whitney tests, the familiarity of the subjects with the excerpts was used as a binary categorization factor that was added to the six subject related factors. The music excerpts were divided into two groups (known and unknown fragments, respectively 43,9 % and 56,1 %). Because there are no 100% unknown excerpts it is obvious that working with summarized data was not an option for this test. As a consequence all 12640 cases (79*160) were retained for analysis. The results are displayed in (❷ pp.146-147), category g.

Familiarity with the music excerpts shows a highly significant effect for all adjectives describing perceived (the affect emanating from the music) (I.1) and felt (the experience of the effect of the music) (I.2) emotions.

In descending order of significance, known music is perceived as being more “tender”, more “passionate”, less “restrained”, more “cheerful”, less “aggressive”, less “anxious”, more “sad” and more “carefree” than unknown music.

Familiarity also very strongly affects the felt emotions which are more subjective ratings of music: known excerpts were judged as being more “pleasing”, less “annoying”, less “indifferent” and more” touching”.

Even for the annotation of structural adjective pairs, except for two features (void-compact and slow-quick), there is a very significant effect of familiarity.

And finally, highly subjective adjectives about appreciation judgments (beautiful-awful) and the difficulty level (easy-difficult) of the music excerpts are influenced in the same way: known music is judged as being more “beautiful” and “easier” than unknown music.

6.5.4 Factor analysis of expressive features

A non-parametric correlation analysis of all “expression” adjectives (the twelve adjectives described under I.1 and I.2 in table 33 over all subjects (79) and all audio excerpts (160) showed that several adjectives are correlated. High significant correlation coefficients are found for touching and pleasing (,631**, p<0,01), cheerful and carefree (,565**), touching and tender (,467**) and for pleasing and annoying (-,476**).

Factor analysis was used to better understand the nature of these correlations in terms of underlying factors. The question investigated here is whether the 12-dimensional space used for the description of the perceived expression qualities in music can be reduced to a lower dimensional space and how much of the inter-subjective variance can be explained. Using maximum likelihood estimation and varimax rotation, three factors are revealed (see figure 20) which, together, explain 51% of the variance in the data.

The first dimension (first factor) shows a contrast in judging the excerpts between the positive adjectives “pleasing”, “touching”, “passionate” and “tender” and the negative adjectives ”indifferent” and “annoying”. This dimension explains 25 % of the total variance in the data and includes terms that express rather *high intense experiences* of music.

The second dimension (second factor) refers to the adjectives “cheerful”, “carefree” and “pleasing” (positive loadings) versus the adjectives “sad”, “anxious” and “touching” (negative loadings). This dimension explains 14% of the differences in the data and account for *diffuse affective state*.

The third dimension (third factor) refers to adjectives “aggressive”, “anxious”, “annoying” and “restrained” (positive loadings) versus “tender” and “carefree” (negative loadings). This dimension explains 12% of the total variance. The adjectives now denote more *physical involvement*.

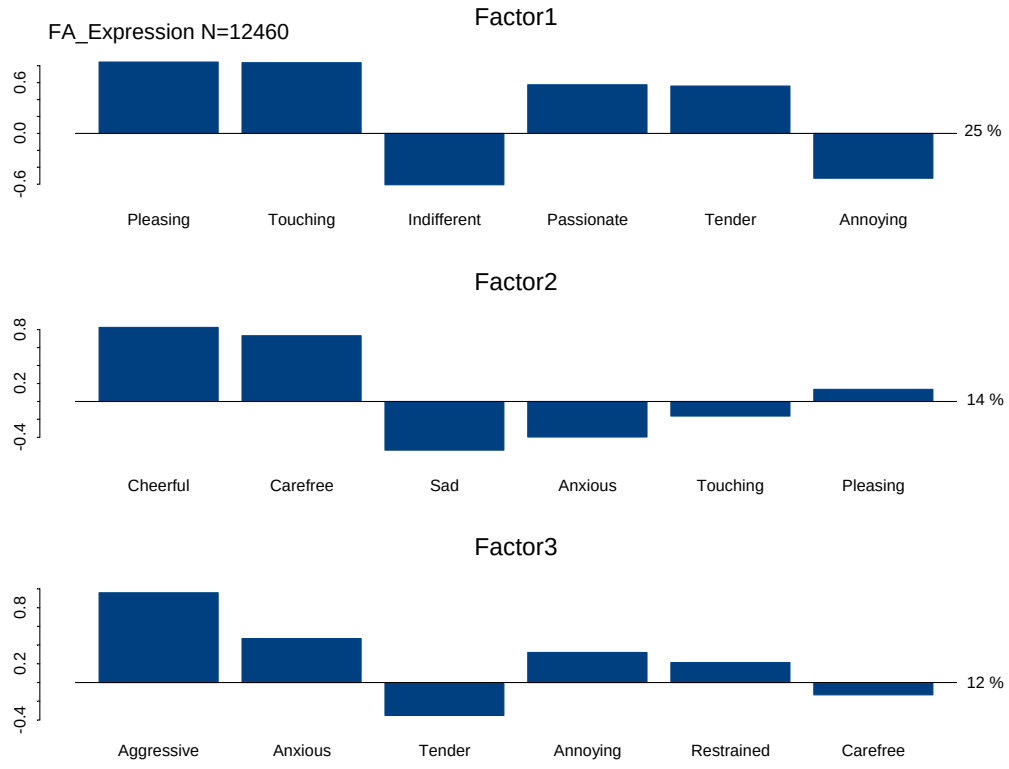


Figure 20: Factor loadings of the expression space.

6.5.5 Factor analysis of structural features

A non-parametric correlation analysis of all “structure” adjectives (the seven adjectives described under II) over all subjects (79) and all audio excerpts (160) showed that several adjective pairs are correlated. High significant correlation coefficients are found between “soft-hard” and “void-compact” (,459**, p<0,01), “soft-hard” and “slow-quick” (,529**), “soft-hard” and “flowing-stuttering”(,458**) and “slow-quick” and “dynamic-static” (-,429**). Factor analysis was performed in a similar way as for expressive features. Manual annotation of structural features reveals three dimensions.

The first dimension accounts for ”soft-hard”, “void-compact”, “flowing-stuttering”, “bright-dull” and “slow-quick” (positive loadings) versus “rough-harmonious” (negative loading). This dimension explains 26 % of the total variance. Loudness, spectral density, articulation, timbre and tempo are located on the positive pole of the axis. The negative pole only has one adjective pair that is related to timbre.

The second dimension refers to “slow-quick” and “soft-hard” (positive loadings) versus “dynamic-static” (negative loading). This dimension explains 19% of the differences in the data. It juxtaposes tempo and loudness versus articulation.

The third dimension refers to “rough-harmonious” (positive loading) versus “flowing-stuttering” and “clear-dull” (negative loadings). Timbre gets a very high loading (95%) and is not clearly correlated with other features. This dimension explains 16% of the total variance.

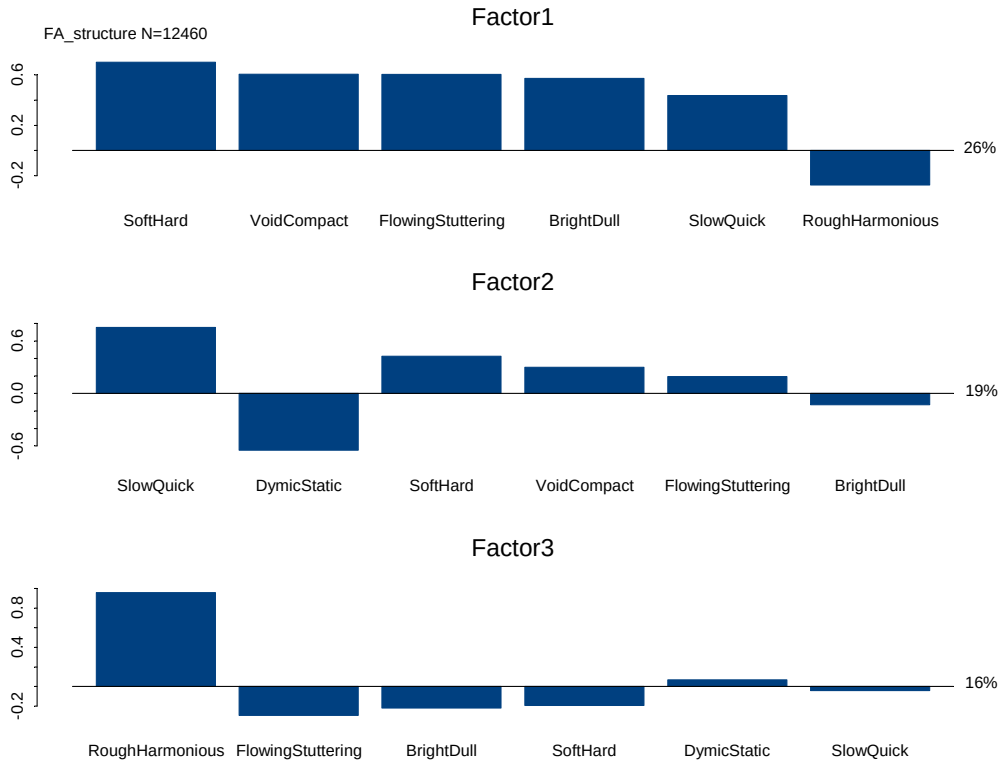


Figure 21: Factor loadings of the structure space.

6.5.6 Unanimity among subjects about structural features

To what extent do people have the same opinion about, for example, the appearance of tempo, loudness, and brightness in a particular piece of music? Moreover, do some musical features reach more unanimity in described perception than other musical features? In other words, it is interesting to know to what degree the listener’s ratings of perceived structural qualities of music claim objectivity.

As participants in the experiment were asked to rate seven structural features (the seven adjective pairs described under II.1 in table 33 for all 160 excerpts, the experiment provided enough data to investigate this issue. With 79 responses for each music excerpt, standard deviations were calculated for the 160 excerpts for the structural features. The mean

standard deviation then calculated for each structural feature gives a rough idea of the objectivity level or the degree of unanimity by which the concerning structural feature of the music was judged (see table 36).

	Mean StDev	Min.StDev	Max.StDev
Soft-Hard	1,4	0,68	2,05
Slow-Quick	1,4	0,65	1,98
Void-Compact	1,8	1	2,5
Flowing-Stuttering	1,8	0,75	2,42
Dynamic-Static	1,8	0,95	2,81
Clear-Dull	1,9	1,12	2,46
Rough-Harmonious	1,9	0,9	2,52

Table 36: Standard deviations for the structural features.

Brightness (clear-dull) and roughness (rough-harmonious) count for the least unanimity among responses (stdev=1,9), closely followed (stdev=1,8) by density (void-compact), articulation (flowering-stuttering) and movement (dynamic-static). With a standard deviation of 1,4 for both loudness (soft-hard) and tempo (slow-quick), these features are the most unanimously perceived.

6.5.7 Relationships between perceived qualities

As a useful guide for further statistical exploration of the relationship between perceived expressive (I.1 and I.2) and perceived structural qualities (II.1), several correlation analyses were carried out on the data of the relevant variables. Two possible approximations of analysis were explored: Kendalll's tau_b and Pearson's correlations. The first considers analyses including the data of all 12640 (79*160) cases independently and the second considers analyses on the same data but summarized for the 160 music excerpts.

A Non-parametric correlations: Kendall's tau_b

Given the non-parametric character of the full dataset (the data consists out of points on an ordinal rating scale), the chosen correlation coefficient for the first analysis method was Kendall's tau_b. The output of the calculation of these non-parametric correlations is summarized in table 37. The full table is included in (❷ pp.148-149).

Although most relationships between the variables are highly significant (p<0, 01), high correlation coefficients in the table are rare (most coefficients stay under ±,3. Correlations between ±,3 and ±,5 are highlighted with light yellow and those higher than ±,5 are marked dark yellow.

High correlation coefficients are found between "tender" and "soft" (-,507**). The affect "tender" also correlates with "harmonious" (,326**), "void"(-,347**), "slow" (-,371**) and flowing (,392**).

High correlation coefficient are also found between “aggressive” and “hard” (,527**). The affect “aggressive” also correlates with ”rough” (-,298**), “compact” (,352**), “quick” (,354**) and “stuttering” (,384**).
Furthermore a correlation is found between the experience adjective “touching” and “soft” (-,300**).

The few relationships in the table found to be not significant (highlighted with blue) are those between “anxious” and “slow-quick” (p=,695) and between “carefree” and “flowing-stuttering” (p=,828)

		Non Parametric Correlations Kendall's tau						
		SoftHard	ClearDull	Rough-Harmonious	Void-Compact	Slow-Quick	Flowing-Stuttering	Dynamic-Static
Cheerful	CC	,101(**)	-,136(**)	,076(**)	,133(**)	,286(**)	,049(**)	-,318(**)
	Sig	,000	,000	,000	,000	,000	,000	,000
Sad	CC	-,309(**)	-,048(**)	,102(**)	-,248(**)	-,366(**)	-,206(**)	,216(**)
	Sig	,000	,000	,000	,000	,000	,000	,000
Carefree	CC	-,012	-,121(**)	,125(**)	,050(**)	,150(**)	-,002	-,161(**)
	Sig	,085	,000	,000	,000	,000	,828	,000
Anxious	CC	,131(**)	,177(**)	-,216(**)	,046(**)	,003	,146(**)	,031(**)
	Sig	,000	,000	,000	,000	,695	,000	,000
Tender	CC	-,507(**)	-,283(**)	,326(**)	-,347(**)	-,371(**)	-,392(**)	,116(**)
	Sig	,000	,000	,000	,000	,000	,000	,000
Aggressive	CC	,527(**)	,269(**)	-,298(**)	,352(**)	,354(**)	,384(**)	-,175(**)
	Sig	,000	,000	,000	,000	,000	,000	,000
Passionate	CC	-,087(**)	-,103(**)	,088(**)	-,043(**)	-,063(**)	-,128(**)	-,127(**)
	Sig	,000	,000	,000	,000	,000	,000	,000
Restrained	CC	,133(**)	,122(**)	-,148(**)	,065(**)	,055(**)	,185(**)	,093(**)
	Sig	,000	,000	,000	,000	,000	,000	,000
Annoying	CC	,264(**)	,214(**)	-,243(**)	,161(**)	,126(**)	,258(**)	,057(**)
	Sig	,000	,000	,000	,000	,000	,000	,000
Pleasing	CC	-,256(**)	-,239(**)	,257(**)	-,121(**)	-,092(**)	-,253(**)	-,120(**)
	Sig	,000	,000	,000	,000	,000	,000	,000
Touching	CC	-,300(**)	-,220(**)	,218(**)	-,191(**)	-,207(**)	-,286(**)	-,026(**)
	Sig	,000	,000	,000	,000	,000	,000	,000
Indifferent	CC	,130(**)	,156(**)	-,135(**)	,100(**)	,060(**)	,153(**)	,116(**)
	Sig	,000	,000	,000	,000	,000	,000	,000

**Correlation is significant at the 0.01 level (2-tailed) *Correlation is significant at the 0.05 level (2-tailed).

Table 37: Non-parametric Kendall tau-b correlations: expression*structure (N=12640)

B Pearson’s correlations

As a second possible option to correlate the data, Pearson’s correlation analyses were done using the mean ratings of all variables assigned to each excerpt (thus, the data of 12640 cases were summarized for the 160 excerpts). The results of the analyses are presented in table 38. The full table is included in (❷ pp.150-151). It is remarkable that many variables are related in a very significant way while others show no significant relationships at all.

		Pearson's Correlations						
		SoftHard	ClearDull	Rough-Harmonious	Void-Compact	SlowQuick	Flowing-Stuttering	Dynamic-Static
Cheerful	PC	,298(**)	-,196(*)	,093	,426(**)	,621(**)	,238(**)	-,699(**)
	Sig	,000	,013	,241	,000	,000	,002	,000
Sad	PC	-,718(**)	-,331(**)	,400(**)	-,753(**)	-,838(**)	-,694(**)	,710(**)
	Sig	,000	,000	,000	,000	,000	,000	,000
Carefree	PC	,076	-,266(**)	,257(**)	,219(**)	,403(**)	,061	-,488(**)
	Sig	,342	,001	,001	,005	,000	,443	,000
Anxious	PC	,338(**)	,521(**)	-,598(**)	,173(*)	,047	,331(**)	,078
	Sig	,000	,000	,000	,029	,553	,000	,327
Tender	PC	-,942(**)	-,695(**)	,770(**)	-,858(**)	-,795(**)	-,885(**)	,542(**)
	Sig	,000	,000	,000	,000	,000	,000	,000
Aggressive	PC	,887(**)	,740(**)	-,757(**)	,766(**)	,648(**)	,776(**)	-,430(**)
	Sig	,000	,000	,000	,000	,000	,000	,000
Passionate	PC	-,256(**)	-,261(**)	,176(*)	-,251(**)	-,246(**)	-,364(**)	-,056
	Sig	,001	,001	,026	,001	,002	,000	,484
Restrained	PC	,466(**)	,512(**)	-,497(**)	,330(**)	,297(**)	,555(**)	,033
	Sig	,000	,000	,000	,000	,000	,000	,680
Annoying	PC	,756(**)	,699(**)	-,704(**)	,604(**)	,509(**)	,722(**)	-,215(**)
	Sig	,000	,000	,000	,000	,000	,000	,006
Pleasing	PC	-,701(**)	-,707(**)	,703(**)	-,561(**)	-,447(**)	-,708(**)	,128
	Sig	,000	,000	,000	,000	,000	,000	,108
Touching	PC	-,737(**)	-,593(**)	,574(**)	-,696(**)	-,665(**)	-,742(**)	,407(**)
	Sig	,000	,000	,000	,000	,000	,000	,000
Indifferent	PC	,548(**)	,526(**)	-,413(**)	,485(**)	,413(**)	,549(**)	-,127
	Sig	,000	,000	,000	,000	,000	,000	,110

**Correlation is significant at the 0.01 level (2-tailed) *Correlation is significant at the 0.05 level (2-tailed).

Table 38: Pearson's correlations: expression*structure (N=160).

The expressive adjectives "sad", "tender", "aggressive", "annoying" "pleasing" and "touching" correlate most strongly with almost all structural features.

The affect adjective "sad" strongly correlates with:

- "soft" (-,718**), "void" (-,753**), "slow" (,838**), "flowing" (-,694**) and "static" (,710**).

The affect adjective "tender" strongly correlates with:

- "soft" (-,942**), "clear" (-,695**), "harmonious" (,770**), "void" (-,858**), "slow" (-,795**), "flowing" (-,885**) and "static" (,542**).

The affect adjective "aggressive" strongly correlates with:

- "hard" (,887**), "dull" (,740**), "rough" (-,757**), "compact" (,766**), "quick" (,648**), "stuttering" (,776**).

The experience adjective "annoying" strongly correlates with:

- "hard" (,756**), "dull" (,699**), "rough" (-,704**), "compact" (,604**), "quick" (,509***), "stuttering" (,722**).

The experience adjective "pleasing" strongly correlates with:

- "soft" (-,701**), "clear" (-,707**), "harmonious" (,703**), "void" (-,561**) and "flowing" (,722**).

The experience adjective “touching” strongly correlates with:

- “soft” (-,737**), “clear” (-,593**), “harmonious” (,574**), “void” (-,696**), “slow” (-,665**), “flowing” (-,742**).

Other strong correlations are found between:

- “cheerful” and “quick” (,621**)
- “cheerful” and “dynamic” (-,699**)
- “anxious” and “dull” (,521**)
- “anxious” and “rough” (-,598**)
- “restrained” and” dull” (,512**)
- “restrained and “stuttering” (,555**)
- “indifferent” and “hard” (,548**)
- “indifferent and “dull” (,526**)
- “indifferent” and “stuttering” (,549**)

The expressive adjective with the lowest significance is “carefree”. Most relationships in the table that are not significant (highlighted with blue) are between the majority of expressive adjectives and the structural adjective pair “dynamic-static”.

To summarize, positive attributes such as “tender” and “pleasing” are negatively correlated with “bright-dull”, “void-compact”, “flowering-stuttering”, and “soft-hard” and positively with “rough-harmonious”. In other words, these positive attributes are correlated with “bright”, “void”, “flowing”, “soft”, and “harmonious”. Negative attributes such as “aggressive” and “annoying” are correlated with “dull”, “compact”, “stuttering”, “hard”, and “dull”. The correlations between the experience adjective “indifferent” and structural adjectives correspond with the kind of relationships found for negative attributes. These findings contradict with what one would expect from an adjective that expresses an uninterested experience. This discrepancy is due to the fact that participants had difficulties in rating the phrase “this music leaves me indifferent”. Many subjects rated “not indifferent” (0) whilst they meant “indifferent”. The adjectives “carefree” and “passionate” on the contrary are not or very slightly correlated with the rated structural features.

6.6 Consistency within annotations

Can objectivity be claimed for these test results? The subjectivity inherent in listener evaluations is a well known problem. For quality judgments we need to be concerned about the consistency of measurement. A study on the reliability of the annotations was done by means of consistency tests that were carried out over time.

6.6.1 Set up

About two months after participants (N=79) had done the experiment they were requested to perform the same annotation task again. A set of six music excerpts was selected by

means of a randomizer[107]. The randomizing was based on the arbitrary choice of a number between one and the amount number in the genre subsets (classical, pop, roots) according to a distribution by which each number from the series has an equal probability. The random selector picked two classical, two pop and two roots excerpts. Table 39 shows the six titles involved in the consistency according to the genre classes.

Genre	Nr	Composer	Piece title
CLASSICAL	10	Stravinsky	Le sacre du printems
	125	Shubert	Stringquartett No. 14 in D minor "Der Tod und das Madchen"
POP	140	Underworld	Jumbo
	20	Hendrix	Voodoo Chile
ROOTS	150	Anonym	Raga Kirvani - tânam (traditional Indian)
	71	Anonym	Merai lassu csardas es szapora (traditional Hungarian)

Tabel 39: Stimuli used in the consistency test.

People were asked to perform the test several times (four times was suggested) with an interval of at least a week between each. Response was quite satisfactory in that 63% (N=50) did one consistency test, 47% (N=37) did 2 tests, 35% did 3 tests (N=28) and 28% (N=22) did four tests. Analysis of repeated responses on the same categorical scale was performed on the set with data from the original annotations plus two consistency tests (N=37) which is considered an adequate sample size.

6.6.2 Analysis

The data of the two restrained consistency tests (further abbreviated as CT1 and CT2) together with the corresponding data extracted from the original experiment (CTo), made it possible to compare judgments of identical musical excerpts made at different points of time. Thirty-seven subjects evaluating six music excerpts each gives 222 observations per judged variable. To compare these three different samples (CTo, CT1 and CT2) with each other, the Marginal Homogeneity Test was used. This test is recommended for analyzing repeated measurements on individuals when the variable outcome is ordinal. The objective is to test the null hypothesis that two paired samples come from the same distribution.

6.6.3 Results

Consistency has been checked for affect (I.1) and structural (II) features. First, the results for repeated measures for judged affect are discussed and secondly the results for judgments of structural features.

[107] Excel add-in for statistics was used.

A Consistency of affect qualities

The results of the Marginal Homogeneity (MH) Test executed on the judged *affect qualities* are given in table 40.

	CT's	Num. obs.	ties%	Obs. MH Stat.	Mean MH Stat.	Std.Dev. MH Stat.	Std. MH Stat.	Asymp. Sig. (2-tailed)
Anx_o*Anx_CT1	CTo*CT1	219	51,1	176	154	7,000	3,143	0,002
Cheer_o*Cheer_CT2	CTo*CT2	219	56,6	177	158	7,762	2,512	0,012
Aggr_o*Aggr_CT2	CTo*CT2	221	50,2	194	175	7,714	2,463	0,014
Aggr_o*Aggr_CT1	CTo*CT1	221	56,6	179	163	6,519	2,454	0,014
Pass_o*Pass_CT1	CTo*CT1	221	47,1	237	221	8,201	2,012	0,044
Caref_o*Caref_CT2	CTo*CT2	219	47,9	164	179	8,775	-1,709	0,087
Sad_o_Sad_CT2	CTo*CT2	221	45,2	151	165	7,921	-1,704	0,088
Cheer_CT1*CheerCT2	CT1*CT2	222	57,2	170	161	6,265	1,516	0,129
Sad_o*Sad_CT1	CTo*CT1	220	48,2	139	151	8,093	-1,483	0,138
Anx_o*Anx_CT2	CTo*CT2	222	27,5	259	241	13,565	1,327	0,185
Caref_CT1*Caref_CT2	CT1*CT2	220	48,2	185	195	7,746	-1,291	0,197
Pass_o*Pass_CT2	CTo*CT2	222	50,5	250	240	8,441	1,244	0,214
Cheer_o*Cheer_CT1	CTo*CT1	219	48,4	203	194	7,348	1,225	0,221
Pass_CT1*Pass_CT2	CT1*CT2	221	50,2	175	182	6,576	-0,988	0,323
Tender_CT1*Tender_CT2	CT1*CT2	221	64,3	107	111	5,244	-0,763	0,446
Caref_o*Caref_CT1	CTo*CT1	219	49,8	172	178	8,559	-0,643	0,520
Tender_o*Tender_CT2	CTo*CT2	221	54,3	144	148	6,442	-0,621	0,535
Aggr_CT1*Aggr_CT2	CT1*CT2	222	60,4	157	154	6,205	0,483	0,629
Sad_CT1*Sad_CT2	CT1*CT2	221	57,9	133	136	6,305	-0,397	0,692
Restr_CT1*Restr_CT2	CT1*CT2	221	61,1	107	109	7,159	-0,210	0,834
Restr_o*Restr_CT1	CTo*CT1	221	50,2	149	148	7,921	0,189	0,850
Anx_CT1*Anx_CT2	CT1*CT2	219	30,6	216	218	12,787	-0,156	0,876
Restr_o*Restr_CT2	CTo*CT2	220	49,5	145	145	8,078	0,062	0,951
Tender_o*Tender_CT1	CTo*CT1	221	59,7	113	113	6,205	0,000	1,000

Table 40: Marginal Homogeneity test results for affect qualities.

The MH test was performed on the comparison between judged affect qualities in consistency tests at different points of time. The values are sorted according to the significance of difference between two measurements.

The first column in the table shows the two related samples, which are compared on consistency. For example, in case of the first row, the judgments for “anxious” from CTo are compared with the judgments for “anxious” from CT1. The amount of observations is given in the column labelled as “Num Obs.” (total observations)[108]. In the column labelled ”ties”, the output of a contingency table of the two concerning variables is summarized for the percentage of participants who gave the same answer in two tests. For the comparison in the first row, for example, 51,1% of the participants gave the same rating when judging a specific musical excerpt for the adjective “anxious” both in the original and in the first

[108] The fifth point on the rating scale ‘no opinion’ is excluded because not being part of an ordered scale. Hence, the number of observations is not always 222.

consistency test. Logically, 48,9% gave two different ratings for the same musical excerpt. Below, an example of the contingency table (table 41) for the comparison between the ratings for “anxious” in CTo (the original experiment) and CT1 (the first consistency test) is given.

			Anxoius_CT1					
			0	1	2	3	4	Total
	Anxious_CTo	0	51	23	1	0	0	75
		1	18	20	5	2	1	46
		2	10	12	19	6	0	47
		3	2	5	13	11	2	33
		4	0	1	1	5	11	18
	Total		81	61	39	24	14	219

Table 41: Contingency table for the ranking of “anxious”.

The next columns in table 40 contain the output of the Marginal Homogeneity Test, which is based on the counts in the above-mentioned contingency tables for each variable. The p-values in the last column indicate if paired samples are different or not. For the comparison in the first row, for example, the p-value is 0,002, which means that the judgments of “anxious” in CTo are significantly different from the ratings of “anxious” for the same music excerpts in CT1. Out of twenty-four comparisons five turn out to be significant. It is remarkable that p-values lower than 0.05 all refer to comparison with the original annotations. This could be due to the longer time interval between the start of the consistency test and the annotation experiment (two months) and between the tests (approximately a week). It seems that the adjectives “anxious”, “cheerful”, “aggressive” and “passionate” are ambiguous terms when measurements are repeated.

B Consistency of structural qualities

The results of the Marginal Homogeneity Test performed on the judgments for *structural features* are given in table 42. Out of twenty-one comparisons, eight are significantly different. Similar as for repeated measures of judged affect qualities significant p-values are the outcome of comparison with the original annotations, except for the adjective pair “clear-dull” that shows different judgments for CT1 and CT2. Furthermore this adjective pair also had varying judgments for CTo and CT1. “Rough-harmonious” and “slow-quick” seems adjective pairs that are ambiguous in the whole consistency test. Previous tests (see section on unanimity) already had shown that there is the least agreement among subjects for brightness (clear-dull) and roughness (rough-harmonious).

	CT's	Num. obs.	ties%	Obs. MH Stat.	Mean MH Stat.	Std. Dev.MH Stat.	Std. MH Stat.	Asymp. Sig. (2-tailed)
RoHa_CTo*RoHa_CT1	Cto*CT1	222	34,23	-59	-7,5	14,620	-3,523	0,000
RoHa_CTo*RoHa_CT2	CTo*CT2	222	35,14	-73	-27,5	15,612	-2,914	0,004
FloStu_CTo*FloStu_CT1	Cto*CT1	222	27,93	138	92	16,643	2,764	0,006
ClDu_CTo*ClDu_CT1	Cto*CT1	222	38,74	37	-2	14,177	2,751	0,006
SloQu_CTo*SloQ_CT2	CTo*CT2	222	44,14	144	117	12,000	2,250	0,024
ClDu_CT1*ClDuCT2	CT1*CT2	222	39,64	-54	-29	11,790	-2,120	0,034
SloQu_CTo*SloQu_CT1	Cto*CT1	222	41,44	106	82	11,576	2,073	0,038
DySta_CTo*DySta_CT2	CTo*CT2	222	36,49	-200	-167	16,416	-2,010	0,044
SoHa_CTo*SotHa_CT2	CTo*CT2	222	41,44	146	125,5	10,524	1,948	0,051
DySta_CTo*DySta_CT1	Cto*CT1	222	39,19	-179	-150	16,271	-1,813	0,070
FloStu_CT1*FloStu_CT2	CT1*CT2	222	39,19	55	78,5	12,990	-1,809	0,070
VoCo_CTo*VoCo_CT2	CTo*CT2	222	27,93	91	61,5	16,785	1,757	0,079
VoCo_CTo*VoCo_CT1	Cto*CT1	222	28,83	67	42,5	15,548	1,576	0,115
FloStu_CTo*FloStu_CT2	CTo*CT2	222	27,93	125	102,5	16,317	1,379	0,168
SoHa_CTo*SoHa_CT1	Cto*CT1	222	41,44	127	112	11,225	1,336	0,181
ClDu_CTo*ClDu_CT2	CTo*CT2	222	36,49	45	31	14,230	0,984	0,325
SoHa_CT1*SoHa_CT2	CT1*CT2	222	50,45	64	58,5	9,097	0,605	0,545
RoHa_CT1*RoHa_CT2	CT1*CT2	222	44,59	-6	-12	14,509	0,414	0,679
VoCo_CT1*VoCo_CT2	CT1*CT2	222	44,14	4	-1	12,590	0,397	0,691
SloQu_CT1*SloQu_CT2	CT1*CT2	222	44,59	88	85	10,050	0,299	0,765
DySta_CT1*DySta_CT2	CT1*CT2	222	43,69	-115	-112	14,221	-0,246	0,806

Table 42: Marginal Homogeneity test results for structural features.

The MH test was performed on the comparison between judgments for structural features in consistency tests at different points of time. The values are sorted according to the significance of difference between two measurements.

6.7 Discussion

The results from the comparative investigation on the profile of the participants in the survey and those in the experiment confirmed the assumption that there would be no major differences between these datasets. This finding indicates that it is likely that the results from the annotation experiment may account for the target population, namely the potential users of music information retrieval systems.

The choice of the stimuli for the annotation experiment is in contrast with previous studies that often focused on classical music (e.g. Wedin, 1972) or on presenting music that has been systematically manipulated (e.g. Juslin, 1997; Peretz, 1998). This study drew attention on the users of music information retrieval systems. The stimuli used, therefore respond to the global profile of the music taste of the average user of music information retrieval systems. The excerpts were explicitly selected out of the favourite music titles provided by the participants in the survey on user context (chapter five). Given the fact that users probably will search for music that corresponds to a certain style they like or to a specific mood, it may be assumed that the stimuli set covers the type of music that the

targeted population would like to retrieve. Even though, the list contains a broad diversity of music, from the titles in users' favourite list it is clear that this type of music can be described as being rather light and broadly popular.

The number of excerpts in the dataset (N=160) largely exceeds the quantity of stimuli used in previous research. With a few exemptions (e.g. Leman et al., 2004) the number of stimuli generally used in research regarding musical structure and expression is rather small, varying from two to fifty music pieces (Gabrielson and Juslin, 2003). Furthermore, the main criteria for the selection of the excerpts were the homogeneous character of distinguishing music features (such as a slow tempo, a repeating rhythm pattern, a staccato performance) and absence of lyrics, whereas in previous research stimuli have often been selected on the basis of emotional expressivity.

As the annotation experiment relied on verbal descriptions, it was necessary to deal with the well-known problem that there is no agreement about the precise meaning of the adjectives and on the classification of emotions (Gabrielson, 2001; Juslin, 2001). In contrast with studies that very often present just a list of terms, phrased adjectives were presented in order to give the subjects a better understanding of how to interpret the adjectives. In addition, the annotation experiment also involved two response formats that were (1) a choice between descriptive terms and (2) ratings using bipolar and unipolar scales. As a consequence, the data that was collected by user annotation of expression, structure and movement have been analysed by means of a variety of statistical methods.

A novel approach to the investigation of the relationship between expressive and structural features is that in current study the influence of subject related factors according to the global profile of the users of music information retrieval systems was brought into account. More specifically, the profile information on the subjects in the experiment that was established by means of their participation in the survey on user context was used. This approach revealed significant subject dependencies for gender, age, music expertise, musicianship, broadness of taste and familiarity with classical music. In other words, music information retrieval system developers should take into account that perceived emotion qualities are affected by the profile of the user.

Two approaches to the expression descriptors have been used: a categorical approach and a dimensional approach. The categorical approach divided the descriptors in two sets: affect descriptors (e.g. cheerful, sad) and experience descriptors (e.g. pleasing, boring). The dimensional approach was based on previous research reported by Leman et al. (2004). In their investigation on correlation of gestural musical audio cues and perceived expressive qualities the dimensions "Valence", "Activity" and "Interest" were defined. In the experiment on annotation of music qualities, for each of these three dimensions two adjective pairs were included.

Although some differences were found, these dimensions are closely related to the three factors found in this study. The “Interest” dimension (moving, exiting, pleasing and passionate, versus indifferent, boring, annoying and restrained) has much similarity with the factor denoting intense experience of music (pleasing, touching, passionate and tender versus indifferent and annoying). The “intense experience” factor, however, explains 25% of the differences in the data whereas the “Interest” dimension in the Leman data explains 19,5%. The “Valence dimension” (carefree, gay, hopeful and positive versus anxious, sad, desperate and negative) corresponds with the factor denoting a “diffuse affective state” (cheerful, carefree and pleasing versus sad, anxious and touching). The latter explains 14% of the variability whereas the former explains 19%. The “Activity” dimension (bold, restless and powerful versus tender, calm and fragile) resembles the factor denoting “physical involvement” (aggressive, anxious, annoying and restrained versus tender and carefree). The latter explains 12% of the variability whereas the former explains 18%. Note that in the experiment on quality annotation the semantic space was reduced from fifteen to eight dimensions. Unlike in the study by Leman et al., where the subjects involved were university students the sample represented a much broader population. This provides an interesting addition to the previous study, namely the finding that this semantics is likely to be the same for different groups of users.

The finding that participants in the experiment most unanimously agree on tempo and loudness supports results from previous research. Indeed, tempo is considered the most important among structure features affecting emotional response to music (e.g. Scherer and Oshinsky, 1997; Juslin, 2000). In the annotation experiment, slow tempo is associated with various affect descriptions such as “sad” and “tender” and with the experience descriptions “pleasing” and “touching”. Quick tempo is associated with the affect description “aggressive” and the experience description “annoying”. The findings that hard music may be determinant for the perception of aggression and the experience of annoyance, and soft music for the perception of tenderness and a pleasing experience, is also confirmative of known research (e.g. Baroni and Finarelli, 1994; Juslin, 1997).

6.8 Conclusion

Comparison of the user context defined by descriptive statistics of the main dataset and the sample of participants in the experiment showed no such differences that they would warrant the questioning of the reliability of the results presented in this chapter.

Explorative analysis reveals *subject dependencies* for the six binary variables that were investigated: gender, age, expertise, musicianship, broadness of taste and familiarity with classical music. Familiarity with the music stimuli set showed a highly significant effect for all the expressive features.

Factor analysis of expressive features revealed three dimensions that were described as “high intense experience”, “diffuse affective state” and “physical involvement”. These

factors are closely related to the dimensions “Interest”, “Valence” and “Activity” found in previous research (Leman, 2004). With regard to unanimity in describing structural features, seven adjective pairs were tested that relate to loudness, timbre, tempo and articulation. Participants most unanimously agree on loudness and tempo, whilst less unanimity was found for timbre and articulation. Interesting relationships were found between expressive and structural features and a strong correlation was found between affect (tender-aggressive) and loudness (soft-hard).

In order to check the reliability of the music annotations, the outcome of two repeated measurements for affect qualities and structural features were compared with the original annotations. Although there were a few inconsistencies, evaluations for the majority of the variables in the annotation experiment seem to be rather reliable. It was found that the ambiguous terms already showed a rather high standard deviation when consistency among subjects was checked.

7 Three applications

In this chapter, the conceptual framework and the results from empirical investigation are used to build three practical music information retrieval applications. The first application uses the taxonomy of chapter three for the design of a music information retrieval user interface. The taxonomy structure suggests possible ways of connecting taxonomy concepts with query statements. The second and third applications draw on the study of semantic description and content-based retrieval of music which have been described in chapters five and six. The second application has a methodological relevance in that it focuses on data handling in view of future scientific research and statistical modelling of empirical data. The application provides an annotated data model from a user perspective. The third application, a semantic recommender system prototype, uses the output data from the subject sampling of potential music information retrieval system users.

7.1 Application 1: User interface taxonomy

In chapter three, the case studies for manual annotation of music presented, focused on user-oriented and model-oriented annotation. These annotations were made by musical experts and non-musical experts, using linguistic labelling as well as action labelling. Linguistic labelling relies on the use of keywords such as an open-ended text box, the selection from a predefined list or the rating of adjectives. This applies to both the musical syntax and the musical semantics. Concerning musical syntax, linguistic annotation is related to the description of perceptual and structural features (e.g. loudness and timbre). Concerning musical semantics, linguistic annotation is related to the description of expressive qualities such as emotion and affect. Action labelling refers to a physical action that has to be undertaken such as the imitation of a melody or a rhythm, editing graphs, tapping on a key and moving a slider.

The manual annotation methodologies developed so far provide guidelines for the development of a user-oriented interface, which can act as a bridge between user preferences and needs on the one hand and system development on the other. In what follows, a description is given of a prototype user interface, which can serve as a model for a linguistic and action-based querying system. The prototype system implements a user interface taxonomy which has been designed as a referential framework for software development within the MAMI project. An example of such tool, based on the user interface taxonomy has been developed by K. Tanghe. The MAMI-prototype is build as a dialog window occupied by tab pages for each of the categories of the user interface taxonomy[109]. Screenshots of this prototype are included in (❷ pp.155-164), an example of a preliminary version is included in ❸.

[109] The user interface has been developed using wxWindows: http://www.wxwindows.org

In the next sections, the music categories and query methods included in the user interface are discussed. After that, a detailed description is given of each music category and matching query methods according to the MAMI-prototype dialog windows.

7.1.1 Music categories

The user interface classifies query methods into six categories that are similar to the constituent music categories of the conceptual framework plus a category called “standard info”. Table 43 shows the categories and subcategories of the user interface taxonomy and related query methods. These are facilities for vocal input, verbal specifications in terms of low to high-level structural descriptions and qualitative descriptions.

	CATEGORY	SUBCATEGORY	QUERY METHOD	
	STANDARD INFO		textual list selection	
	MELODY & HARMONY	MELODY	notated melody recording (audio, MIDI) musical excerpt	
		CHORD PROGRESSION	notated chord progression recording (audio, MIDI) musical excerpt	
		TONALITY	notated tonality recording (audio, MIDI) musical excerpt	
	TIMING & RHYTHM	TEMPO	BPM, tempo tapping recording (audio, MIDI) musical excerpt, drawing	
		TIMING / STRUCTURE	textual list selection	
		DRUM PATTERNS	notated drum pattern recording (audio, MIDI) musical excerpt, drawing	
	LOUDNESS		list selection drawing	
	TIMBRE		list selection	
	EXPRESSION	AFFECT / EXPERIENCE	selection of qualities moving a slider	

Table 43: User interface taxonomy.

The category set thus comprises standard info, melody and harmony, timing and rhythm, loudness, timbre and expression. Standard information only requires linguistic input about standard meta-data such as the title of the piece, name of the composer, orchestra

conductor, or record company. Apart from that, content specific information can be given about melody, harmony, rhythm, loudness and timbre. Inputs may be based on specifications of structural descriptors (e.g. by selecting pre-defined verbal descriptors related to pitch and rhythm), audio queries and audio examples (e.g. recording of a voice or an instrument, prerecorded examples), graphics (e.g. by making a drawing of tempo evolution) or actions (e.g. by tapping the beat or handling a slider) according to physical, sensorial, perceptual and structural descriptors.

7.1.2 Query methods

In what follows, a non exhaustive list of conceivable query methods which are likely to be important for “searching inside the music” is presented. Each query method can be seen as corresponding with one or more properties of music for which a user might want to specify some kind of criterion. Most characteristics of music vary over time, whereas only few properties apply to a musical piece as a whole. The different query methods are categorized according to the type of properties they correspond with. Since query methods potentially deal with multiple properties that would conceptually be placed in a different category, the categorization should not be interpreted in a strict sense. It is attempted to place these methods in the category where a user would most intuitively look for it. Next to traditional text-based queries the following musical queries were considered: live audio input, live MIDI input, audio input from file and MIDI input from file.

A Live audio input

Singing a melody, playing a sequence of guitar chords are query methods that require live audio input through a soundcard. This could be done using a microphone or using a line-in level connection between the instrument and the sound card. Recording circumstances are usually rather noisy in real-life situations and this should be taken into account by the modules processing the audio. As an alternative, a "calibration stage" could be included that assures that the sound card mixer settings are set to appropriate recording levels. A software tool should offer buttons for recording, playback, pause and stop. Playback is important, because users might want to listen to what they entered to decide if the recording needs to be redone or not.

B Live MIDI input

This method allows users to enter musical data via a MIDI device (e.g. keyboard, MIDI guitar controller and drum pads). This is especially useful for musicians that are used to working with MIDI devices. More and more people have a MIDI device attached to their computer, although the average user may not. The tool should offer buttons for recording, playback, pause and stop.

C Audio input from file

This is a general input method for queries where a user wants to specify a sound file that should be used as a reference (e.g. ”same tempo as this sound file", "same melody as in this sound file”). The path to the sound file could be entered via file browsing. Specification of a time range in the sound file, either by typing in start and end time, or by clicking a selection button might be added. For example, by opening a new dialog box in which the user could select the relevant region, while having the possibility to play back that selection.

D MIDI input from file

A MIDI file might also contain musical information that a user might want to specify in a query (e.g. a section of melody track or bass line, a drum pattern). Similar with audio file input, the path to the MIDI file must be specified and the opportunity to select a specific region could be offered. In this case, for example, a piano roll view of the MIDI file could be displayed. MIDI track selection is interesting as well, because it is usually preferable to only consider one track in the MIDI file for a query[110]. For tonality extraction though, all tracks with tonal content might be needed.

7.1.3 Use modes

Because of the wide range of potential music information retrieval system users, and also because of the diversity of music itself, it is necessary to provide users with flexibility in the way they want to perform a search within a music database. Music experts will want to have full control over all possible properties, whereas non-expert users might feel intimidated when being confronted with music concepts they have never heard of before or which they do not know the meaning of. The latter would rather prefer a minimal interface without minimal control options. For people who often use a search engine though, the possibility to customize the interface would be useful as well. As a consequence, multiple search modes are considered: *advanced search*, *basic search*, *guided search* and *custom search*. In what follows, a short description of these use modes is given.

A Advanced search

The advanced search mode contains all possible query methods gathered both from reflection about music descriptions, and from the feedback from users involved in the study on search and query behaviour (see chapter four). An advanced search mode does not necessarily oblige users to use all of the query fields. In the prototype many search fields are included in order for the user to keep an overall idea of how music can be described and approached. All the music categories and query methods from the user interface taxonomy are reflected in the advanced user interface. Basic search, guided search and custom search are special versions of the advanced search mode that contain a reduced set of features from this mode.

[110] MIDI channel 10 for instance is most often used for drum tracks.

B Basic search

The average music information retrieval users will probably be satisfied having a simplified search method at their disposal.

The basic search mode is designed in order to offer the opportunity to perform a search using the following query methods:

- specify an artist or group name;
- specify the title of a song;
- imitate a simple melody line by voice;
- specify a genre or style.

C Guided search

This is an extra search mode that could be added to improve usability. That some people will prefer to use a guidance tool solicits necessary information from them by asking a few well-chosen questions. Guided search could for example be implemented as a wizard that asks questions. This method helps the system narrow down the search as much as possible without bothering the user with useless search fields. Each time the user makes a choice, and based on all previous choices, the wizard asks more directed questions until the system has gathered enough information.

Beginning and average users might find this very helpful, whereas advanced users would probably find this too time-consuming or too limited. They are likely to prefer more direct control over what they want to specify.

D Custom search

Advanced users might be interested in using an extra search mode such as a custom search. The idea is that the advanced user selects the types of search fields that will be presented when performing a search. A customizable search interface could be useful as some users probably never use certain search fields at all. Instead of making a different interface for each type of user, the system could keep track of how many times people have used each search field and only show the ones that have been used regularly, while hiding the others.

7.1.4 Feedback

A search result is usually reported as a list of titles, ranked in order of decreasing overall relevance. A significance threshold could be introduced in order to quickly skip the pieces with low relevance. For each piece of music in a database, for example, a set of relevant fragments could be shown depending on whether the user was searching for a whole piece or for a particular excerpt. Sorting the results according to each of the specified properties would be another possibility. One could, for example, offer the option to sort the results by overall relevance or by one of the specified properties. Audio feedback is of course important as well. In order to check whether the search engine really returned the desired

piece, users must be able to listen to the results, either the whole piece or a relevant excerpt, depending on what the user was looking for.

7.1.5 Detailed overview

In what follows, tables are presented according to the application of each music category in the user interface taxonomy in the prototype. For each item, a description is given of the kind of information a search field could include and the method by which the information could be entered.

Standard information

PIECE INFORMATION	
Piece title	
Description	The title of a piece of music, a track on a compact disc, a part of a composition.
Method	Linguistic input.
Performer(s)	
Description	Name of the person(s) or group who exhibits musical or acting skills.
Method	Linguistic input + List Selection, alphabetical.
Composer(s)	
Description	Name of the person(s) who created (composed) the musical work.
Method	Linguistic input + List Selection, alphabetical.
Author(s)	
Description	Name of the writer(s) of the text or lyrics.
Method	Linguistic input + List selection, alphabetical.
Example	Text writer of the libretto of an opera, songwriter etc.
Lyrics	
Description	Parts of the words of a piece of music.
Method	Linguistic input.
Example	Words of a song, libretto of an opera, text of an oratorio or a cantata
Language	
Description	Language of the lyrics of the piece.
Method	List Selection, alphabetical.
Genre	
Description	A general term describing the type or kind of music and overall characteristics of a work. As supplied by the record label (usually specified as "file under").
	Groupings: classical / folk - ethnic - world / blues - jazz / pop - rock - country - electronic / musical - opera - theatre / …
Method	List Selection, hierarchical.
Performance type	
Description	Type of performance used for the piece of music.
Method	List Selection, alphabetical.
Example	Band, choir, orchestra, ensemble, soloist…

Year of composition	
Description	Year in which the piece of music was composed.
Method	Numerical input (4-digit year) with allowed deviation before and after (in years).
Year of publication	
Description	Year in which the piece of music was offered (for sale or for free) to the public
Method	Numerical input (4-digit year) with allowed deviation before and after (in years).
Piece duration	
Description	Duration of the piece of music.
Method	Numerical input in hours, minutes and seconds (h:m:s)
Example	1:22:35
Track number	
Description	Number of the track on the album containing the piece.
Method	Numerical input + List Selection (for "first" and "last").

ALBUM INFORMATION	
Album title	
Description	Title of the album, where album refers to any long-playing record (LP), cassette tape, CD, DVD, MD etc…
Method	Linguistic input.
Catalog number	
Description	Identification number assigned to each issue of an album by a record company.
Method	Linguistic input.
Record label	
Description	Company that recorded or compiled the album.
Method	Linguistic input + List Selection, alphabetical.

Melody

NOTATED MELODY	
Melody	
Description	A textual representation of a melody (a sequence of non-overlapping notes performed by a single instrument).
Method	Linguistic input.
Notation type	
Description	The way in which the melody is notated.
Method	List Selection.
	Possible methods with their corresponding representations:
	Notes: text field contains note names in the format NmO, where
	N = note symbol (A to G),
	m = accidentals (# or b),
	O = octave number (-1 to 9)
	Note intervals:

Method	Text field contains note intervals in semitones (positive or negative numbers)
	Contour (fine):
	Text field contains sequence of D, d, R, u, U, where
	D = pitch goes down with at least 3 semitones
	d = pitch goes down with 1 or 2 semitones
	R = pitch is repeated
	u = pitch goes up with 1 or 2 semitones
	U = pitch goes up with at least 3 semitones
	Contour (rough):
	Text field contains sequence of D, R, U, where
	D = pitch goes down
	R = pitch is repeated
	U = pitch goes up
Example	Notes: C4 G3 G3 F#4 Bb4 A4
	Note intervals: -5 0 +11 +4 -1
	Contour (fine): DRUUd
	Contour (rough): DRUUD

MELODY: MONOPHONIC AUDIO RECORDING (vocal query)	
Recording / playback control	
Description	Allows the user to perform a melody using his/her voice and a microphone.
Method	Live Audio Input.
Type of query	
Description	The type of voice query that was performed.
Method	Selection from a list. Possible types:
	Humming
	Singing using syllables
	Singing using text
	Whistling
	Mixed (same as not specified)

MELODY: MONOPHONIC INSTRUMENT RECORDING	
Recording / playback control	
Description	Allows the user to play a melody using an instrument (using a microphone or the line-in input of the sound card). This should really be a monophonic melody and no chords.
Method	Live Audio Input.

MELODY: MONOPHONIC INSTRUMENT MIDI RECORDING	
Recording / playback control	
Description	Allows the user to play a melody on a MIDI keyboard (or other MIDI device). This should be a monophonic melody and no chords. Any MIDI channel is accepted.
Method	Live Midi Input

MELODY: FROM AUDIO FILE	
Audio file selection control	
Description	Allows the user to supply a (fragment of a) sound file and use that as a reference melody in the query ("same melody as in this audio fragment"). This works best when there is only one melody line in the fragment. When there is more than one melody line, a selection of the pitch range can be specified to help in choosing the relevant one.
Method	Audio Input from File.
Pitch range selection	
Description	Specifies the pitch range of interest for extracting the melody from the supplied audio file (especially useful when the audio file contains multiple simultaneous melody lines).
Method	List Selection. Possible ranges: low - middle - high - edit manually
	When edit manually is selected, a graphical view of the melodic content in the audio fragment is shown (melody lines) and the user can "grab" the ones he would like to use. Also allows playback for feedback purposes.
Example	Low: bass line
	Middle: leading vocals
	High: flute, violin

MELODY: FROM MIDI FILE	
MIDI file selection control	
Description	Allows the user to supply a (fragment of a) MIDI file and use that as a reference melody in the query ("same melody as this one").
	The relevant MIDI track must be selected. If a name for a track is contained in the MIDI file, it is shown next to the track number, which aids in selecting the relevant one.
Method	MIDI Input from File.
Example	MIDI tracks can have names like: voice, bass guitar, drums, lead synth, …

Harmony

NOTATED CHORD PROGRESSION	
Progression	
Description	Notated sequence of chords. Used representation depends on notation type (see below).
Method	Linguistic input.
Notation type	
Description	The way in which the chord sequence is notated.
Method	List selection. Possible types:
	Chord names:
	Uses absolute chord names
	Chord functions:
	Tonality followed by colon and then Roman number notation for the chords
Example	Chord names: Em G A C
	Chord functions: C: iii V iv I

CHORD PROGRESSION: POLYPHONIC INSTRUMENT AUDIO RECORDING	
Audio recording control	
Description	Allows the user to record a chord sequence performed on an instrument (using microphone or line-in).
Method	Live Audio Input.
Example	Guitarists might play a chord sequence to find all songs containing that particular sequence.

CHORD PROGRESSION: POLYPHONIC INSTRUMENT MIDI RECORDING	
MIDI recording control	
Description	Allows user to record a chord sequence performed on a MIDI device (usually a MIDI keyboard).
Method	Live Midi Input.

CHORD PROGRESSION: FROM AUDIO FILE	
Audio file selection control	
Description	Allows the user to specify a (fragment of a) sound file containing a chord sequence to be used as reference in the query ("same chord sequence as this one").
Method	Audio Input from File.
Example	Users might have a small fragment of a song containing a typical chord sequence and want to find out from which song it is.

CHORD PROGRESSION: FROM MIDI FILE	
MIDI file selection control	
Description	Same as for audio file selection, but from a MIDI file.
	The MIDI track containing the chords one is interested in can be selected from a list.
Method	Midi Input from File.

NOTATED TONALITY	
Tonality	
Description:	Depending on the type (see below), this can be a single tonality, or a sequence of tonalities.
Method:	Linguistic input.
Type	
Description	Specifies the type of text input in the tonality field.
Method	List Selection. Possible types of input:
	Main tonality: a single constant tonality for the whole piece or fragment
	Tonality sequence: varying tonality, specified as a space separated list of tonalities

TONALITY: POLYPHONIC INSTRUMENT AUDIO RECORDING	
Audio recording control	
Description	Allows the user to record a performance on an instrument and use the tonality extracted from this audio fragment as a reference for the query ("same tonality as this fragment").
Method	Live Audio Input.

TONALITY:POLYPHONIC INSTRUMENT MIDI RECORDING	
MIDI recording control	
Description	Same as for audio recording above, but with live MIDI input.
Method	Live Midi Input

TONALITY: FROM AUDIO FILE	
Audio file selection control	
Description	Allows the user to specify a (fragment of a) sound file to be used as reference in the query ("same tonality as in this file").
Method	Audio Input from File.

TONALITY: FROM MIDI FILE	
MIDI file selection control	
Description	Same as for audio file above, but with a MIDI file. The MIDI track to be used can also be set if needed by default, all tracks (except track 10) are used, because there are no two tonalities at the same time, so each track with notes contributes to the tonality.
Method	Midi Input from File.

Timing and rhythm

AVERAGE TEMPO	
Description	Allows searching for pieces having the specified tempo. Tempo is specified in BPM (beats per minute).
Method	Linguistic input or Move a Slider (from slower to faster tempo).
	A "Listen" button allows hearing a click played back at the specified tempo.
Example	120 BPM

TEMPO TAPPING	
Description	Allows the tempo to be specified by tapping on a button. This is coupled to the BPM field and slider of the average tempo above.
Method	Tap a button or a key.
	First, "Start tapping" is clicked, and then the user taps on the "TAP" button at the desired tempo. At least 5 taps are needed to measure the tempo.

TEMPO: MONOPHONIC INSTRUMENT AUDIO RECORDING	
Audio recording control	
Description	Allows the user to record a performance on an instrument (or voice), and use the tempo extracted from this audio recording to search the database. This could be a constant tempo or a tempo curve.
Method	Live Audio Input.
Example	One could just tap on the microphone, record the sound of ticking with a pencil on a table or simulate the tempo by voice.

TEMPO: MONOPHONIC INSTRUMENT MIDI RECORDING	
MIDI recording control	
Description	Allows the user to record a performance on a MIDI keyboard or MIDI drum pads, and use the tempo extracted from this MIDI recording to search the database. This could be a constant tempo or a tempo curve.
Method	Live Midi Input
Example	Hitting a note "on the beat".

TEMPO: FROM AUDIO FILE	
Audio file selection control	
Description	Allows the user to specify a (fragment of a) sound file, and use the tempo extracted from that file to search a database. Again, constant tempo or tempo curve.
Method	Audio Input from File.
Example	"Search pieces with the same tempo as in this piece"

TEMPO: FROM MIDI FILE	
MIDI file selection control	
Description	Same as for audio, but from a MIDI file.
Method	Midi Input from File.

TEMPO EVOLUTION	
Whole piece	
Description	Sets the tempo evolution of the piece to search for.
Method	List Selection. Possible types of tempo evolution (names say what they mean):
	Rather constant
	Increasing tempo
	Decreasing tempo
	Lots of clear tempo changes
	Draw tempo evolution by hand
	Drawing the tempo by hand allows the user to draw the tempo evolution as a curve over time. Tempo is on the y-axis (BPM) and time on the x-axis (in % of the length of the piece, which is not known in advance of course).
Fragment	
Description	Sets the tempo evolution of the fragment to search for.
Method	List Selection. Same as for the whole piece but on a local time scale, and specified for the fragment to search for, not the whole piece.
Example	Whole piece could have a rather constant tempo, except for a small fragment where the tempo accelerates heavily.

TIMING AND STRUCTURE: TIME SIGNATURE	
Description	The time signature tells how many and what kind of notes per measure there are. The number on top is the number of notes per measure, and the bottom number is what kind of note.
Method	Linguistic input.
Example	5 / 2 (= 5 half notes per measure)

TIMING AND STRUCTURE: SEARCH LOCATION RESTRICTIONS	
No restrictions	
Description	Considers entire pieces.
Method	Exclusive Selection.
Fuzzy location	
Description	Considers only sections of the pieces as specified in several "fuzzy" ways.
Method	List selection. Possible "fuzzy" sections:
	Beginning of the piece (probably first 10% will be fine)
	End of the piece (probably last 10% will be fine)
	Somewhere in the main part
	First half
	Second half
Structural part	
Description	Considers only a specific structural part of a piece. This may depend on the genre of the piece.
Method	List selection.
	Possible structural parts:
	Chorus, verse, bridge, intro, outtro, theme, motif, solo
Exact time range	
Description	Considers only the exact time range of the pieces.
Method	Linguistic input.
	Start time and end time (in seconds). If a time is negative, it is considered to be from the end of the piece. If end time is left blank or zero, it is considered to specify the end of the piece.
Example	-20 0 = last 20 seconds of a piece

TIMING AND STRUCTURE: STRUCTURE OF THE PIECE	
Abstract	
Description	Allows using an abstract representation of the structure of a piece (i.e. without using any context or genre dependent musical terms). Sometimes this is better, because music terms are not always generic enough to search all types of music.
Method	Linguistic input.
	Structural parts are notated using textual labels, and the structure to be searched for is specified as a sequence of these labels separated by spaces.
Example	A B B C B C D C E
Concrete	
Description	Allows using concrete musical terms related to the structure of the piece (genre dependent).
Method	List selection
Example	The same as for the abstract example, but with typical terms for a pop song: e.g., intro, verse, chorus, solo, outtro

Take into account	
Description	Allows the user to specify which types of musical information should be taken into account when speaking about a structural part (structuring can be done according to different criteria, so we need a way to specify what should be considered and what not).
Method	Multiple Selection from a set of items.
	Possible musical info:
	Timbre
	Loudness
	Tempo
	Rhythm
	Harmony

NOTATED DRUM PATTERN	
Edit… / Play	
Description	Allows the user to search for pieces containing a specific drum pattern. A drum pattern is defined here as a time pattern of bass and snare drums (maybe later also hi hats and cymbals).
Method	Graphical Editing.
	The user can specify the pattern by graphically editing the relative positions and amplitudes of the drum events on a time line (possibly using a grid). He/she can listen to the pattern while editing it. If a grid is used, the length (number of subdivisions) can be specified.
	Possible extension: hitting specific keys on the computer keyboard to enter the drum pattern.
Example	A typical drum and bass pattern might look like this (B = bass drum, S = snare drum):
	B----B--
	--S---S-

DRUM PATTERNS: AUDIO RECORDING	
Audio recording control	
Description	Allows the user to record a drum pattern performance using a microphone or line in audio signal.
Method	Live Audio Input.
Example	A user might imitate a pattern of bass and snare drums using his/her voice.
	A recording might be done from a real drum set or from hitting different objects.

DRUM PATTERNS: MIDI RECORDING	
MIDI recording control	
Description	Allows the user to record drum patterns by using a MIDI keyboard or a MIDI drum pad.
Method	Live Midi Input
	Only MIDI channel 10 will be used (drum channel in General MIDI standard).

DRUM PATTERNS: FROM AUDIO FILE	
Audio file selection control	
Description	Allows the user to specify a (fragment of a) sound file to be used as reference drum pattern ("same drum pattern as in this fragment"). The fragment can be from a real piece, so it may contain other instruments as well (drum detection needed).
Method	Audio Input from File.

DRUM PATTERNS: FROM MIDI FILE	
MIDI file selection control	
Description	Allows the user to specify a (fragment of a) MIDI file to be used as reference drum pattern ("same drum pattern as in this fragment"). Extraction of the drums is easier since they are always located in track 10 for standard MIDI files.
Method	Midi Input from File.
	Only MIDI channel 10 (= drum channel in General MIDI standard).

Loudness

OVERALL LOUDNESS IMPRESSION	
Whole piece	
Description	Allows the user to specify the overall loudness he/she perceives for the piece as a whole.
Method	List selection.
	Possible items: 5 levels from "very silent" to "very loud".
Fragment	
Description	Allows the user to specify the overall loudness he/she perceives for the fragment he/she is looking for.
Method	List selection.
	Possible items: 5 levels from "very silent" to "very loud".
Example	Some acoustically performed songs are overall quite silent, but also contain loud parts due to the use of some electrical amplification at some points.

LOUDNESS EVOLUTION	
Whole piece	
Description	Specifies the evolution of the loudness over the entire piece.
Method	List selection.
	Possible items:
	Rather constant
	Increasing loudness
	Decreasing loudness
	Lots of alterations between loud and silent
	Draw loudness evolution by hand
	When "Draw…" is selected, the user can specify the evolution of the loudness over time by drawing a loudness level on a 2D graph with on the y-axis loudness and on the x-axis time.
Example	The "Bolero" from Ravel is becoming louder and louder over time.

Fragment	
Description	Specifies the evolution of the loudness for the fragment being searched for.
Method	List selection.
	Same as above, but for the fragment instead of for the whole piece.

Timbre

INSTRUMENTATION	
Vocal	
Description	Sets the type of vocal content.
Method	Multiple Selection from a set of items. (any combination, except with "No voice", selecting nothing will just ignore these search fields, as usual). Possible vocal content types are:
	Male
	Female
	Child
	No voice
Example	A duet might be performed by a male and a female singer.
Non-vocal	
Description	Sets the instrument types that are being used. Main instrument groups are used, and also a few specific instruments that are easy to recognize.
Method	List Selection, hierarchical.
	Possible groups / instruments:
	String instruments
	Wind instruments
	Percussion instruments
	Electronic instruments

EVENT DYNAMICS EXTRAS	
Pitch	
Description	Extra information about noticeable pitch dynamics.
Method	List selection
	Possible items:
	A lot of vibrato: the music contains a lot of events with noticeable vibrato
	(low frequency pitch oscillations)
	A lot of glissando: the music contains a lot of events with noticeable glissando ("sliding notes"; pitches seem to slide from one note to another without being stable in between, as in a trombone)
Amplitude	
Description	Extra information about noticeable amplitude dynamics.
Method	List selection
	Possible items:
	A lot of tremolo (low frequency amplitude oscillations)
	A lot of notes with sharp attacks
	Slowly decaying notes

Expression

BIPOLAR EMOTION ADJECTIVES	
Description	Limited set of bipolar adjectives that relate to subjective qualities one perceives by listening to the piece ("the kind of feeling it generates").
Method	List selection
	Selection of the qualities that should be used in the search.
	Move a Slider.
	Moving a slider to indicate the way in which each selected quality is perceived. Possible subjective qualities and their extreme perceptions:
	Happiness: from "sad" to "gay"
	Assertivity: from "tender" to "brutal"
	Interest: from "boring" to "exciting".

ASSOCIATED KEYWORDS	
Description	Free lists of keywords that the user thinks are related to the piece he/she is searching for. The database might have a set of keywords associated with each piece.
Method	Linguistic input.

7.1.6 Conclusion

The user interface presented above consists of music categories and query methods to consider when developing a music information retrieval software tool that connects taxonomy concepts with query statements. The MAMI-prototype interface is based on the constituent music categories of the conceptual framework (chapter two), the annotation taxonomy (chapter three) and query methods (chapter four). It illustrates how these taxonomies could be reflected in a tool that handles multiple query techniques. Such a tool could be a starting point for easily hooking up algorithms to a working system. At the same time it would constitute a kind of reference of different query methods that are desirable in a content-based music information retrieval system. Mid and low-level algorithms need to be connected to the high-level concepts in order to interface between the internal data representations, calculations and user input.

7.2 Application 2: Data management

An important aspect that will contribute to the enhancement of music information retrieval research is the storing of experimental data. These datasets should be stored in such a way that they are easily accessible to many users (i.e. researchers) and that new records can be easily added. In this section, a study is presented that comprises the setting up and the normalization of a relational database containing user data. The objective for maintaining such a database with user data is twofold. First, a relational database makes the data gathered by experimental investigation as well as the interpretation process

transparent to others than the researchers that collected and processed them. Second, multiple users can work with it and append their own research data. To begin with, the choice for a relational database is motivated. Then, the conceptual design, the practical set up of the MAMI-relational database model (MAMI-RDBM), the process of normalization and the query interface are described.

7.2.1 Approach

Given the scarcity of investigational data on music information retrieval users' background and their perception of music qualities[111], the primary goal of setting up a database was to provide meaningful storage and retrieval of less-structured information collected through multiple investigations. There are many database options available such as flat-file, relational, object-oriented and device-independent structured document approach. Flat-file database management systems (e.g. Filemaker Pro) comprise straight tables with no connection between them. A relational database management system (RDBMS) allows cross-referencing information between multiple files that share a common field and establishes a variety of relations among sets of data (e.g. MS Access, MS SQL server). Object-oriented database management systems (ODBMS) manipulate objects which encapsulate complex data structures and processes (e.g. ObjectDB, Versant).

The starting point here was to make a music information retrieval research tool, using the collected contextual data (i.e. high-level descriptions). For that purpose, a relational database was chosen. The MAMI-relational database model (MAMI-RDBM) was designed to store the large quantities of data gathered by the study on user context (chapter five) and by the experiment on annotation of musical qualities and structure (chapter six). A relational database has the advantage of using matching values in multiple tables to relate the information in one table with the information in another table. All operations on data are done on the tables themselves or produce other tables as a result. Operations between tables are performed by treating them as sets. It is possible to navigate through records and indexes, and operations can be carried out both through a relational view of the database and by using SQL statements.

Given the different types of data, collected via the online survey and through the experiment, and the aim of getting a grip on associations between these datasets, the use of a relational database was considered the best option. Such a database is able to support the needs of different approaches to look at the data from the perspectives of diverse research goals.

The MAMI-RDBM has been developed in different phases. First, a top-down approach to database design was used to define the global database construction. After identification of the datasets, the preliminary set up was created. This involved the definition of tables and

[111] See 1.2.4 and 5.1 Background in user context studies.

relationships. In the next phase the database was normalized. Finally, a query interface was designed.

7.2.2 Top-down approach: conceptual design

The output for both the survey on user context and the experiment on annotation of music qualities was initially stored in plain text format. In the earliest stage of the investigation the raw data was imported in MS-Excel spreadsheets for control, corrections and basic data mining. For setting up the MAMI-RDBM database, a standard relational database management system (RDBMS) software was used[112].

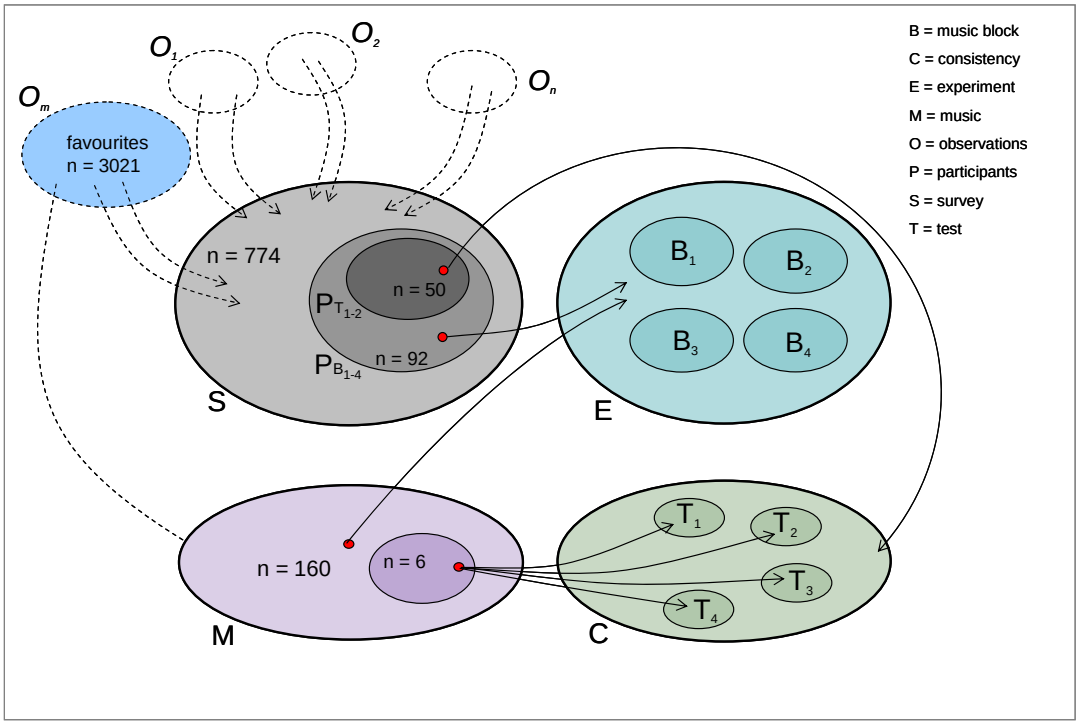


Figure 22: Conceptual design of the MAMI-RDBM.

Figure 22 shows the conceptual design of the database. The datasets included in the relational database are represented as Venn diagrams. The internal structure of the MAMI-RDBM consists of four data sets: the data output from the survey on user context (S), the annotations from the experiment on musical qualities (E), the annotations collected from the consistency tests (C) and the musical stimuli (M).

Diagram S represents the set with observations (O) for 774 participants in the survey. The subset $P_{B_{1-4}}$ contains all the observations of participants who took part in the experiment (n=92). The subset $P_{T_{1-4}}$, a subset of $P_{B_{1-4}}$, contains all the observations of participants (n=50) who took part in the consistency test.

[112] MS Access RDBMS 2003 was used.

Diagram E represents the set with annotations of musical qualities made by the participants in subset P_{B1-4}. It consists of four subsets B1, B2, B3 and B4, each representing the annotations for 40 musical excerpts.

Diagram C represents the annotations of musical qualities provided by the participants who took part in the consistency tests. Set C consist of four subsets T1, T2, T3, and T4 each representing the same annotations for the six musical excerpts performed at different moments in time.

Diagram M represents the set with the musical stimuli used for making the annotations in set E and C. The musical audio dataset, a set of 160 audio files in .wav format, was based on the observations (3021 titles) in the "favourites" part of the survey. A subset of six excerpts was used in the consistency test.

A practical reason for setting up this database was the large amount of manual work caused by the concept of the survey. Hence, set S with user context data, is the result of ten sequentially collected survey items online presented as individual forms. It was a logical set up: only if survey item one was finished was it possible to move on to survey-item two and so on. Each participant, however, could interrupt the survey at any of the survey item levels. As a consequence, the sample size decreases for every subsequent output[113]. For all survey items a separate text file was generated. The importation of these text files in Excel-spreadsheets generated columns of differing lengths. Although for each participant a unique key was created and repeated in each text file for each participant, it sometimes happened that another key was generated[114]. Importing the data in a user-friendly management system overcame the problem of mistakes, inherent in manual correction of key mismatch.

7.2.3 Preliminary set up: tables and relationships

The MAMI-RDBM database is a collection of tables that track data about the categories (e.g. musical background, genre, expression, structure) that form the basis of the global set up of the experimental procedures. In other words, the data for the four datasets is contained in tables that are taxonomy based. They were created according to the design of the survey and experiment taxonomy. This means that for each category investigated, a table was generated containing all the descriptors for the concerning category. Descriptors are represented in table fields or columns. A table row holds all the information about one subject. Table 44 shows the structure of the tables in the MAMI-RDBM according to the four datasets generated by the empirical investigation.

[113] See 5.4: Questions and findings.
[114] This problem already occurred during the test phase of the survey and although several computer scientists were asked for advice, this inconvenience could not be overcome and necessitated manual correction.

	DATASET	TABLE	
	SURVEY (S)	tbl_Enq_01_PERSONAL tbl_Enq_02_GENERAL_BACKGROUND tbl_Enq_03_CULTURE tbl_Enq_04_INTERNET tbl_Enq_05_MUSICAL_BACKGROUND tbl_Enq_06_MUSIC_INSTRUMENT_ACTIVITY tbl_Enq_07_LISTEN tbl_Enq_08_GENRES tbl_Enq_09_TASTE tbl_Enq_10_FAVOURITES	
	EXPERIMENT (E)	tbl_EXP_ALL tbl_EXP_EXPRESSION tbl_EXP_EXPERIENCE tbl_EXP_STRUCTURE tbl_EXP_VARIABILITY tbl_EXP_ACTIVITY tbl_EXP_MEMORY	
	CONSISTENCY TEST (C)	tbl_CT_ALL tbl_CT_EXPRESSION tbl_CT_EXPERIENCE tbl_CT_STRUCTURE tbl_CT_VARIABILITY tbl_CT_ACTIVITY tbl_CT_MEMORY	
	MUSIC (M)	tbl_AUDIODATA	

Table 44: Table structure in the MAMI relational database model.

The relational part of the MAMI-RDBM deals with how these multiple tables relate to each other, allowing retrieval of information from a combination of tables. Relationships have been defined between the four datasets (i.e. S, E, C and M) in the database. Figure 23 shows a screenshot of the basic relationships in the MAMI-RDBM database, displaying the attributes for the survey data (tbl_SURVEY_DATA, for the data collected by means of the annotation experiment (tbl_EXP_ALL), for the data collected by means of the consistency test (tbl_CT_ALL) and for the music data (tbl_AUDIODATA).

Structured Query language (SQL) and MS-Visual Basic for Applications was used to manipulate the data in the MAMI-RDBM[115] database. The top-down approach to the database design provided the preliminary datasets. In view of improving query handling and using the MAMI-RDB for the development of a recommender system (see further), the datasets were normalized.

[115] Implementation of the relationships and the Visual Basic modules and normalization of the datasets carried out by F. Desmet, administrator of the MAMI-RDBM database and member of the IPEM staff.

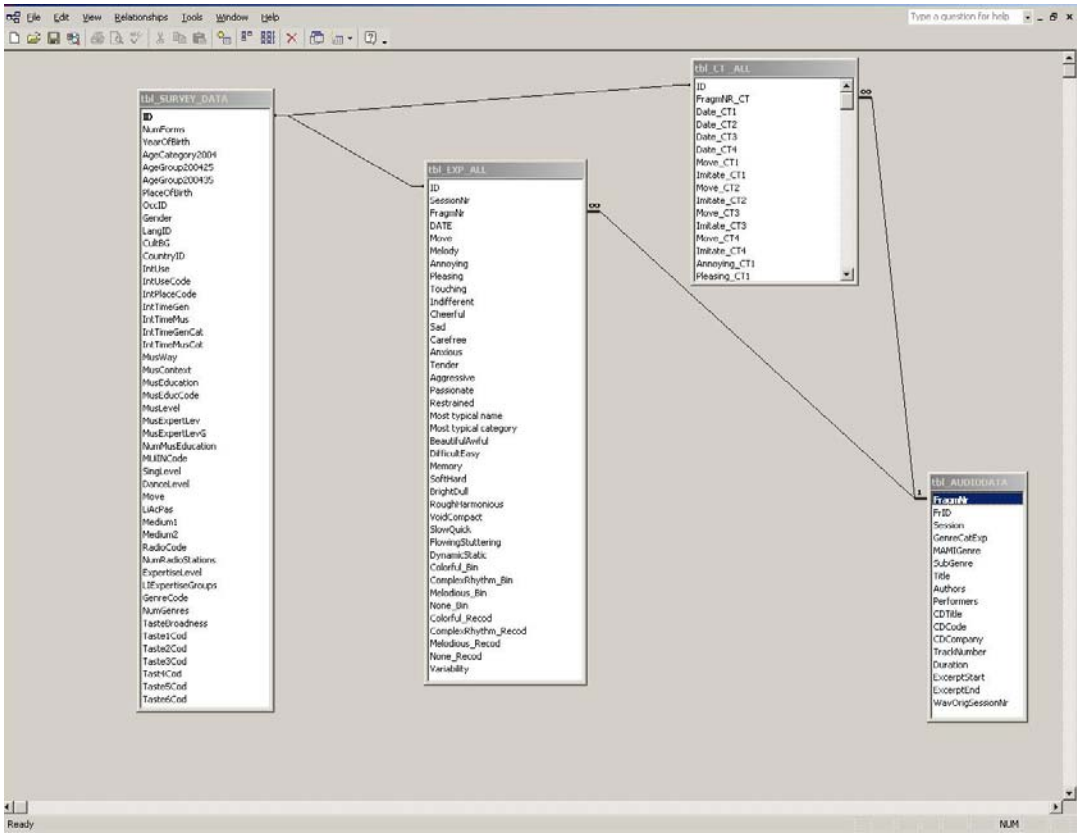


Figure 23: Screenshot of the basic relationships of the MAMI-RDBM.

7.2.4 Bottom-up approach: database normalization

Normalization is the process of using a set of rules in order to design a set of related tables. The normalization process has two goals: elimination of redundant data (e.g. identical data stored in more than one table) and ensuring that data dependencies make sense. Although the principles of normalization are well known they are often neglected. The guidelines for normalizing a database are referred to as normal forms (NF). For the MAMI-RDBM, rules for the first normal form (1NF), the second normal form (2NF), the third normal form (3NF) and the fourth (4NF) were applied. The zero normal form or un-normalized form (0NF) basically consists of the preliminary datasets according to the top-down database design. The normalization rules for the normal forms 1NF to 4NF are cumulative, which means that a database must meet all the criteria of a previous normal form before it can move to the next one. Criteria are for example the removing of repeating groups (e.g. duplicate columns in the same table), of partial dependencies (e.g. subsets of data that apply to multiple rows of a table) and of transitive dependencies (e.g. columns that are not dependent on the primary key).

7.2.5 Query builder

Although predefined queries are incorporated, any person familiar with relational databases can easily make new queries. Tables can be consulted directly from the database in MS-

Access format. As the internal structure of the database might be rather complex to a foreign researcher, a data query builder has been implemented in order to provide a user-friendly database query interface. This allows easy retrieval of the records in the datasets without having to know the internal structure of the MAMI-RDBM. A demo of this query builder is included in ❸. Figure 24 shows a screenshot of the interface of the query builder for the following query: retrieve all male participants in the survey, older than 35 who said they download music from the Internet. The system retrieved 9 records (out of 774), the output can be viewed in the form of a web page.

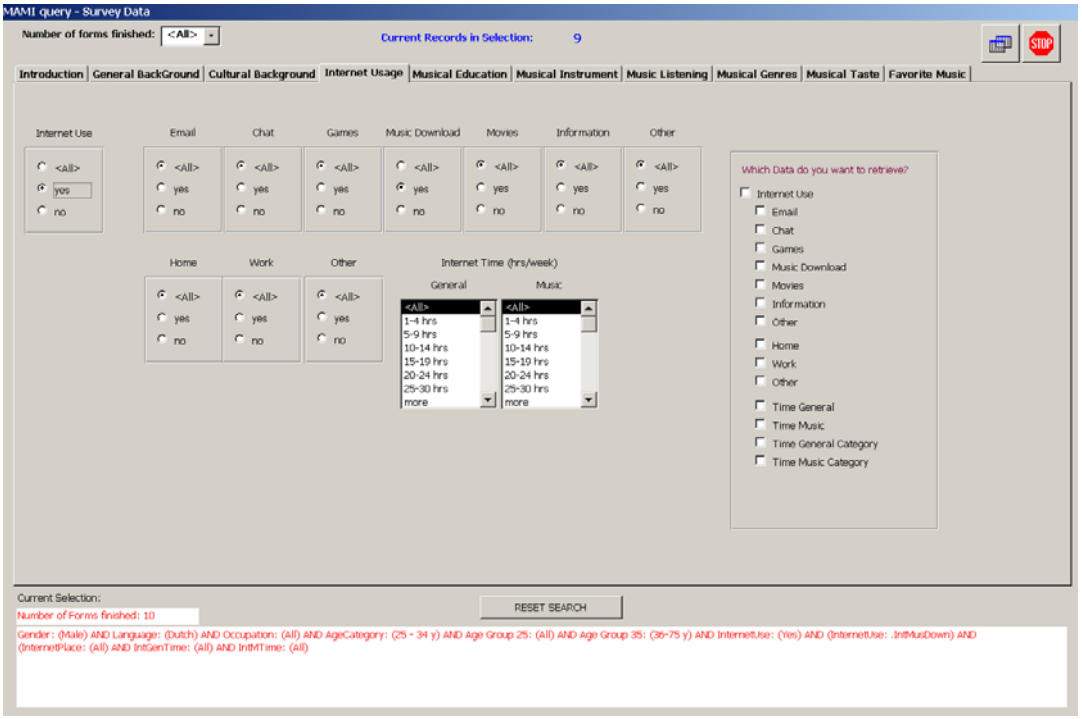


Figure 24: Screenshot of the interface of the MAMI-RDBM query builder.

7.2.6 Conclusion

In this section an application is described for the data management of user data collected via experimental research and the MAMI-relational database model (MAMI-RDBM) is presented. Given the aims of easy accessibility and the possibility of appending new data to empirically collected test sets, there are many reasons for choosing a relational database. To sum up but a few:

- the data are located in a secure central place (i.e. the IPEM server);
- the data are within the reach of multiple MIR researchers;
- the execution of the normalization process reduces the space the database occupies and ensures that the data is logically stored whilst maintaining data integrity;
- multidimensional querying is possible (SQL);
- data storage and retrieval are transparent (to the programmer and the end user);

- the database allows Open Database Connectivity (ODBC) with statistical programs such as SPSS and S-Plus;
- relations are established between contextual data, annotations of perceived qualities of music and music data.

The MAMI-RDBM database allows easy tracking of the high-level descriptions of musical data that underpin the interpretations and findings reported in previous chapters. Furthermore, it allows other music information retrieval researchers who conduct user studies, to review the user data and to generate alternative interpretations by combining it with new data that they have collected. It was also pointed out that the set up of a relational database model for storing and handling experimental data has multiple advantages in view of music information search and retrieval research.

7.3 Application 3: Semantic music recommender system

In this section a music information retrieval application is presented that supports the querying of a music database by the semantic descriptors[116] “affect”, ”structure” and “motion”. This application is based on the empirical data from the subject sampling of potential music information retrieval system users described in chapters five and six. The application was conceived as a semantic music recommender system (SeMuReS) that is based upon agreement among users in the experiment on the annotation of music qualities. The goal of the application is twofold. First, it aims at offering researchers a simple tool that allows testing the quality of the gathered data[117]. Secondly, it is a means for collecting users’ evaluations of system usability. In what follows, the background in music recommender systems is sketched. Then the approach to the setting up of the prototype system is explained, after which the design and procedure are described. Finally, details are given about the implemented fuzzy logic functions.

7.3.1 Background

The application on “music and emotion” presented here is an example of a user-centred approach to a music recommender system. The system recommends music from a relational database containing user background and user annotations of qualities of music. An overview of the general background on emotion-based music information retrieval was given in chapter six.

In general, a recommender system is defined as an application that acts as a personalized decision guide aiding users in making choices about matters related to personal taste (Swearingen and Sinha, 2002). Recommender systems are based on a synthesis of ideas from human-computer interaction, sociology, information retrieval and the technology of the

[116] See 2.5.1 Global or contextual auditory descriptors.
[117] The SeMuReS prototype is included in the electronic appendix. It runs on Windows 2000 and Windows XP platforms (the Visual Basic .NET framework is needed).

Internet. There are different types of recommendations and techniques for generating recommendations such as personalized or non-personalized recommendations and social filtering.

Existing music recommender systems such as Amazon, MediaUnbound, MoodLogic, or Songexplorer[118], differ a lot in the kind of input users must provide. What these systems have in common is that they haven been developed from a commercial viewpoint (i.e. selling music). From that perspective these systems might be deemed successful, but if you define success in terms of how much users are helped in discovering of music, usability may not be very satisfactory. According to a study by Swearingen and Sinha (2002), MediaUnbound was rated best as far as those features are concerned that lead a user to trust a system's recommendations. In this study it was found that trust was affected first and foremost by the accuracy of the recommendations. Apart from that, trust also seemed to depend on the transparency of the system logic, familiarity with the items recommended and the complexity of the process for retrieving recommendations.

Usability to the user is also a topic of interest in the musical digital library community. Although in the literature the importance of interface and system usability is acknowledged (e.g. Arms, 2000), it has only recently been suggested that users themselves should be consulted. Previous studies rather focus on trying to find out what people do and would like to do with music. These studies involve, for example, analyzing music queries posted to Usenet News (Downie, 2002) and to the Google ask-an-expert service (Bainbridge, 2003) or watching people's behaviour in CD stores (Cunningham, 2003). The usability of existing systems and prototypes, however, has not been tested with real music information retrieval users. Indeed, the most common method used for studying usability is laboratory-based testing (Covey 2002). Notess (2004) on the other hand, describes studies that explore the use of the digital musical library projects Variations and Variations2[119] in a more natural setting. These projects, however, are being developed for use by university students and staff as opposed to by typical music information retrieval users.

To summarize, most existing music recommender systems are designed from a commercial viewpoint. They widely vary in the type of input they require and the number of recommendations they generate. As a consequence they are difficult to compare. From a user perspective, the main unsolved problem is how to make use of different representations and techniques to meet information needs. Usability needs to be studied by involving real music information retrieval system users. I am not aware of examples that are ready for testing in the real world.

[118] Amazon: http://www.amazon.com,
MediaUnbound: http://www.mediaunbound.com,
MoodLogic: http://www.moodlogic.com,
Songexplorer: http://songexplorer.com

[119] Variations and Variations2 are digital musical library projects under development at Indiana University.

7.3.2 Approach

The SeMuReS prototype is a music information retrieval system able to deal with queries that are related to emotion, structure and movement. It operates on the limited dataset of 160 audio excerpts, the data used in the experiment on the description of high-level features[120]. Fuzzy logic is used to incorporate response distributions over the empirical data in the MAMI-RDBM database.

In order to reshape it for the benefit of the global user, the recommender system exploits ratings provided by a sample of music information retrieval system users. The rating values are regarded from the viewpoint of the consistency of semantic descriptors among users. The purpose is to make user-dependent music recommendations. Subject dependencies, such as gender and expertise, which have been studied by investigation of relationships between users' background and their perception of music qualities[121] are brought into account.

Fuzzy logic is used in order to account for the fuzzy or subjective character of the semantic descriptions of music qualities. Indeed, musical quality has no clearly defined boundary. One subject, for example, might have rated the excerpt from Strawinsky's "Le sacre du printemps" as "very aggressive", "rather rough" and "rather stuttering", while another subject could have rated it as "not aggressive", "very passionate" and "rather harmonious". The difficulty is that, when attempting to define what quality features constitutes a piece of music, individual perceptions and backgrounds have to be taken into account. From that perspective, the incorporation of fuzzy logic is an interesting option, because it bypasses the sharp-edged true-false logic of semantic descriptors. The truth statement thus becomes a matter of degree. In practice this means that fuzzy logic allows a semantic description to have any value between the numerical value of 1 (true) and the value of 0 (false).

The interface for the SeMuReS prototype is designed for use at exhibitions and for multiple testing possibilities by a broad range of users. In fact, SeMuRes has thus been demonstrated at ACCENTA 2005[122] at the stand of Ghent University. At that exhibition, the system was used by 648 visitors. Ratings of the system's performance have been gathered and will be used in a follow-up study[123].

7.3.3 Design and procedure

The SeMuReS application basically consists of four parts: the definition of the user profile, the presentation of the input options, the recommendations and evaluation tasks. After the choice of the language (Dutch or English), a short introduction is given on the screen. It is

[120] See chapter six.
[121] See 6.5.2.: Influence of subject related factors.
[122] ACCENTA is Flanders' international annual fair in Ghent that celebrated its 60th anniversary in 2005 (September 17-25). The prototype on music and emotion was one of the demonstrations illustrating the research activities at the department of musicology (IPEM) of Ghent University.
[123] Given the short time interval of three weeks between the presentation at ACCENTA and the submission of this thesis it was impossible to include results from that study.

explained that it is possible to not only look for music but ALSO to listen to it. Apart from that, it is also made clear that database searching is based on input adjectives expressing emotion, music characteristics and movement. Finally, it is clarified that some personal data are required before the actual search can begin as the system adapts itself to the user. The interaction paradigm is the following: a user provides the input and the system processes that information to generate a ranked list of recommendations. A demo of the prototype is included in ❸.

A Profile

The user must enter three profile specifications: gender, year of birth and level of musical interest. The choice of the subject dependencies gender and expertise is based on the findings in the study on quality ratings, i.e. that these factors explain differences in perception of high-level features[124]. Age is included in order to investigate the distribution of age categories of different test samples. Five levels of musical interest are presented to the user. These are a simple translation of the factor music expertise. After having provided the system with this information, users get access to the search screen. Figure 25 shows the profile screen.

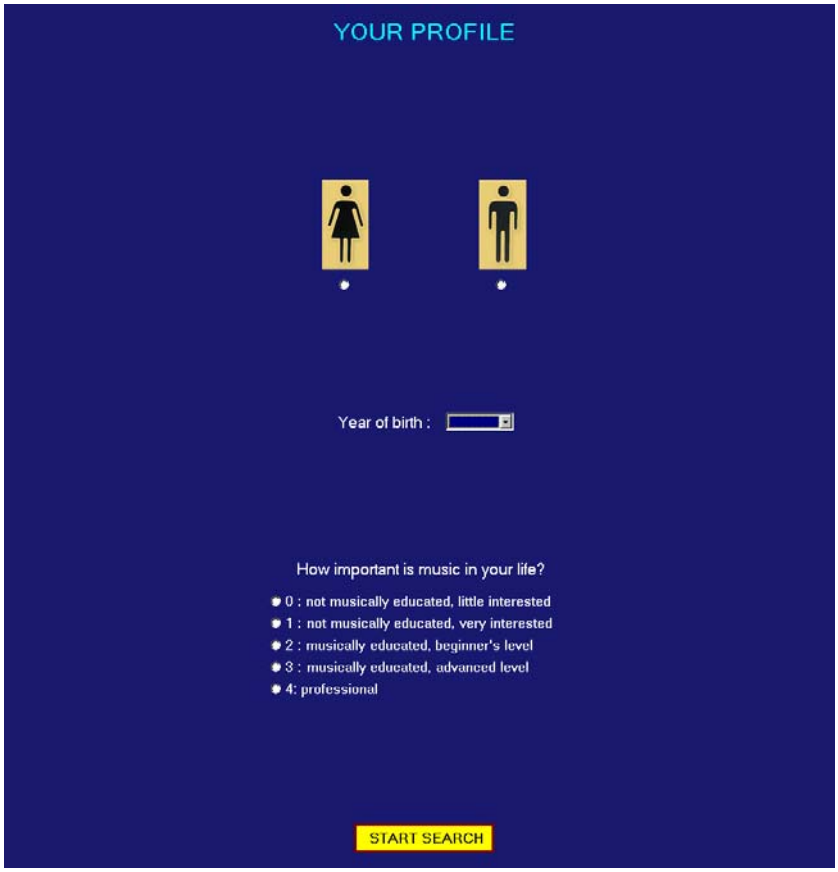


Figure 25: Profile screen of the semantic recommender system prototype.

[124] See 6.5.2.: Influence of subject related factors.

B Input options

Before entering a search a simple pop up window informs the user on what to do. That is to say, a user can make as many selections as desired from four selection fields (i.e. genre, emotion, sound and movement) at the top of the search screen (figure 23). The user can also make a choice between five genre categories, eight emotion labels, four adjective pairs referring to the characteristics of how the music sounds and three adjective pairs reflecting movement.Concerning genre categories, we learned that a user's taste may cover several genres[125]. Given the limited amount of music excerpts in the test set, a reduced set of five genre filters is offered: classical, pop/rock, jazz, folk/country and world/ethnic. Users are given the choice of searching one, more or all genres.

All the "affect" and "sound" adjectives that were used in the annotation experiment[126] are also used in the user interface. The adjectives are presented exactly in the same way as in the experiment: unipolar for "affect" adjectives and bipolar for "sound" and "movement" characteristics. The rationale for including all these adjectives is based on the fact that, although in the quality annotation experiment there was less unanimity on some structural features such as brightness and roughness, there was in general a rather high consistency among subjects[127] about the other adjectives. The application is constructed in such a way that users can query the database as much as they like. They can modify previous searches or reset the system and make another query without having to start all over again.

C Recommendations

The output is a hierarchically ordered list displaying entry number of the excerpt in the database, genre, title, author(s) and a degree of confidence score. The ranked list is shown underneath the four search fields. The score (in percentage points) reflects the global degree[128] of agreement among subjects on the features entered in the query. All titles are shown up to a threshold of 20% of inter-subjective agreement. While the user selects search options by means of check boxes, the list adapts itself according to the query input.

The user can browse the list, listen to the musical excerpts and retrieve more detailed information about the items. Figure 26 shows a screenshot of the search screen and a recommendation list according to a query by a female expert searching the database for classical music and using the emotion descriptor "sad", the sound descriptor "soft", and the movement descriptors "dynamic" and "slow".The user can listen to any recommended[129] music excerpt. By means of clicking a button, a popup screen appears on top of the search screen. Using a media player, the user can listen to the musical excerpt as often as he/she wants.

[125] See 5.4.3: Genre.
[126] See 6.3.3: Design.
[127] See 6.5.6: Unanimity among subjects.
[128] The calculation of the scores is described in 7.4.4.
[129] The 30 second excerpts are embedded in the application in MP3 format..

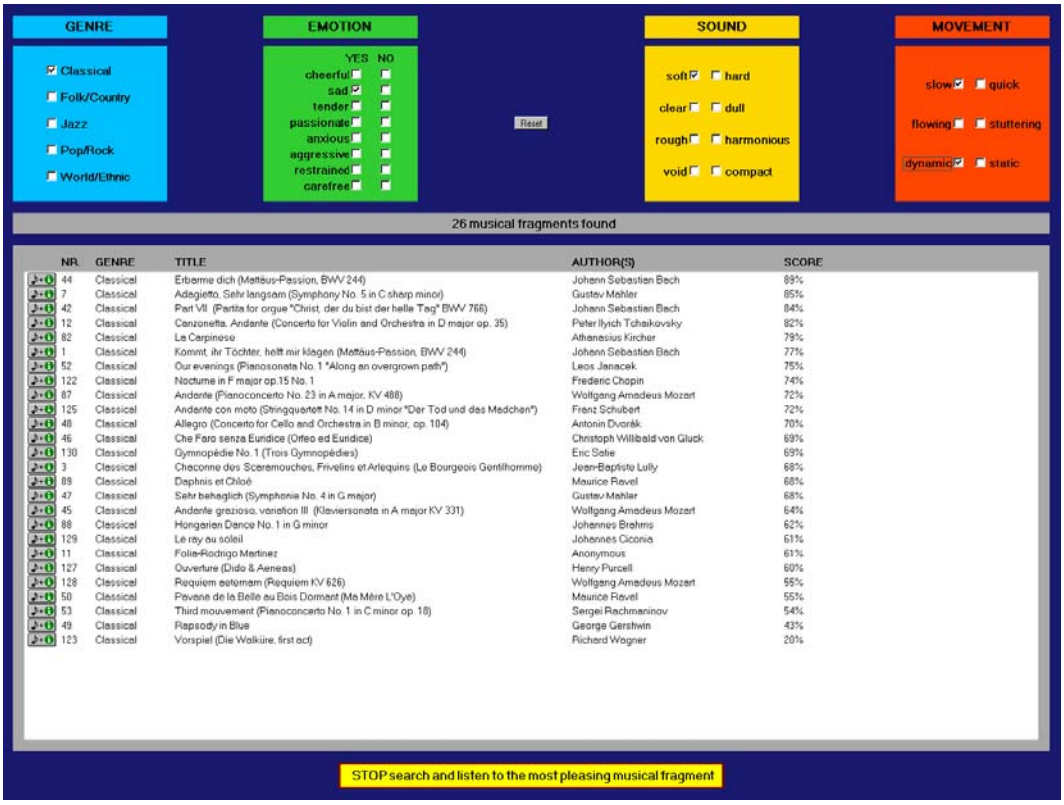


Figure 26: Search screen of the semantic user recommender system.

In their study on recommender systems, Swearingen and Sinha (2002) found that users like to have more information about the recommended item. On this basis, the popup window also provides standard information on the selected piece such as performer(s), CD title, publisher and subgenre. When a user, for example, requests pop/rock, the recommended piece may be classified in the database under “new age” or “soundtrack”. In that case, apart from the basic genre category, the name of the subclass is displayed as well.

Furthermore, for each feature in the query an individual score is given that reflects the agreement among subjects on each feature entered. For example, figure 27 shows the metadata of the musical excerpt “Erbarme dich” (Matthäus-Passion, BWV 244) by J.S. Bach. There is 100% agreement among female experts that this piece is sad, 95% agreement that it is soft, 93% agreement that it is slow and 72% agreement that it is dynamic. Before users can close the popup window and start another search, they are tasked with evaluating the recommendation they listened to. Figure 27 shows the output window for the query as shown in figure 26, for music excerpt number 44 that was ranked first in the recommendation list.

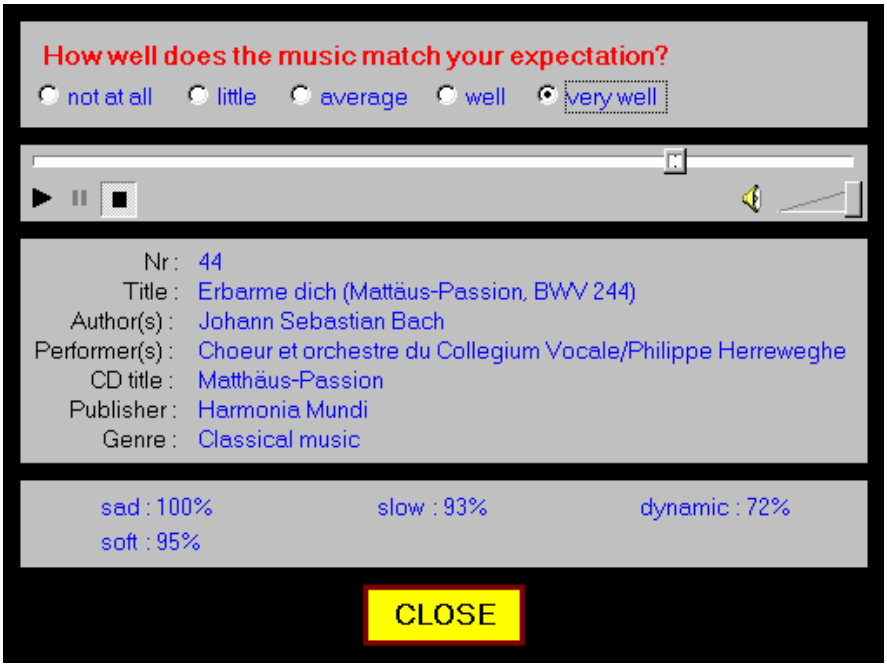


Figure 27: Window with media player, query output and satisfaction rating.

D Evaluation tasks

Two evaluation tasks are included in the application: a repeated task that involves the assignment of the degree of satisfaction after listening to a recommendation and a final evaluation of the usability of both the system and the feature sets provided to make a query to the system. An evaluation is required after listening to a recommendation before the next search can be entered. In order to prohibit it from becoming boring to the test user, the task is simple to do. An indication is asked of the degree of satisfaction ("How well does the music match your expectation?") by means of a rating on a five point scale rating (from "not at all" to "very well"). At the bottom of the search screen a button with the text "Stop the search and listen to the most pleasing excerpt" leads the users to the last screen. Arrived there, they hear the music excerpt[130] that was judged as pleasing by 92% of the participants in the experiment. A text on the screen explains that the application is based on empirical data. The users are invited to evaluate the demo regarding user-friendliness by simply clicking relevant boxes. They are also asked whether they find it interesting to look for music using adjectives that express emotion, structural music characteristics and movement.

E Data logging

For each session, entries are added to tables with user data in the Access database described in a previous section. These tables contain the user profiles (i.e. a numeric ID, gender, year of birth, level of music interest), the search queries for each excerpt that the users listen to (i.e. checked genres and adjectives) and the evaluations (i.e. satisfaction rating and usability of the interface and of the search options).

[130] Yann Tiersen's "La valse d'Amélie" from the movie soundtrack "Le fabuleux destin d'Amélie Poulin".

7.3.4 Fuzzy logic functions

In what follows, a description is given of the fuzzy logic functions of the SeMuReS application. This component of the application was built by K. Vermeulen[131] using Visual Basic .NET for Microsoft Access Databases.

A Calculation of fuzzy functions per adjective and profile

Recommendations are based on a user profile that accounts for two different types of user information: gender and expertise. As a consequence, for each adjective, four fuzzy functions were calculated, characterized by three numbers: the 25th, 50th and 75th percentile values, representing the cumulative value of a semantic descriptor. First, for each adjective, the rating values attributed by all the subjects who fit a specific profile (i.e. female novice, female expert, male novice and male expert) were sorted in ascending order. After that, the values according to the cumulative percentages of 25% (v1), 50% (v2) and 75% (v3) were calculated.

The three values v1, v2 and v3 define the following fuzzy function score:

IF $x \leq v1$ THEN score(x) = 0
IF $v1 < x \leq v2$ THEN score(x)=0,5*[(x – v1)/(v2 – v1)] (a number between 0 and 0,5)
IF $v2 < x \leq v3$ THEN score(x)=0,5+0,5*[(x – v2)/(v3 –v2)] (a number between 0,5 and 1)

v1, v2 and v3 are calculated as follows:

IF “rating x” = 0 : x / frequencyRatings(0)
ELSE
“rating x”+(x - frequencyRatings (0.. ”rating value x” - 1))/frequencyRatings (”rating value x”)
to which frequencyRatings (0..y) = the number of the rating values $\leq$ y

In what follows an example of the calculation of v1, v2 and v3 for the profile “male expert” and the descriptor “cheerful” is given. Figure 28 shows a plot of the fuzzy function.

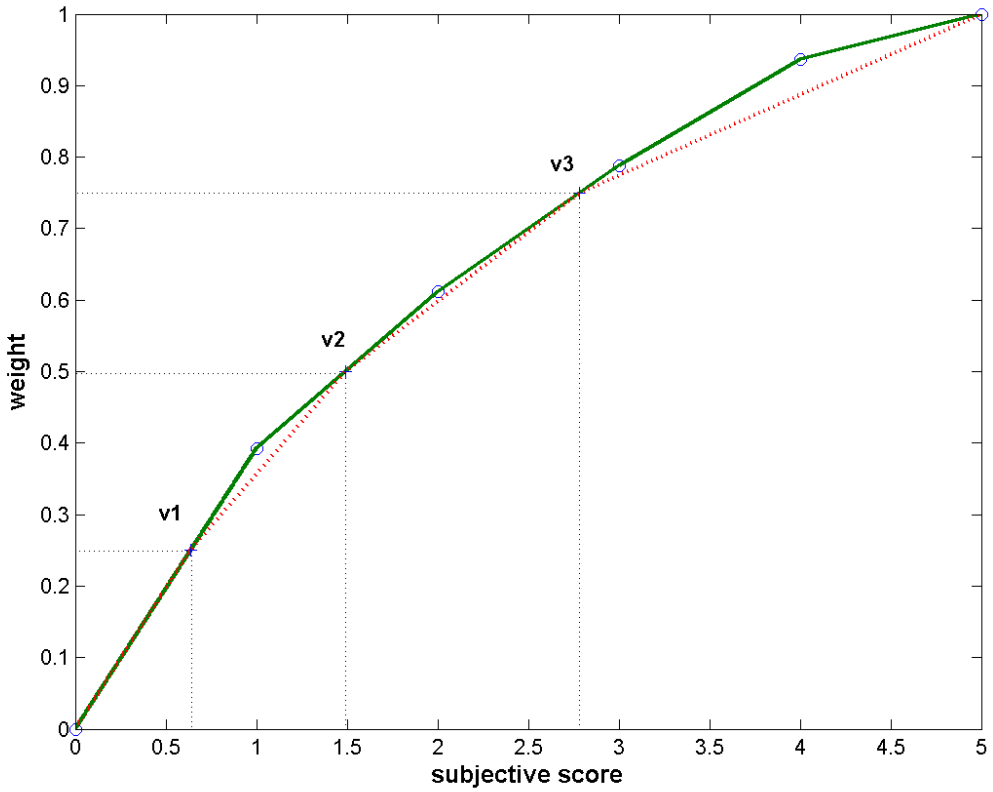
The cumulative distribution function (see figure 28 green line) is built on the number of ratings given by male experts for “cheerful”. The number of ratings for:

- “not” [frequencyRatings(0)] = 1252
- “little” [frequencyRatings(1)] = 698
- “moderate” [frequencyRatings(2)] = 564
- “rather” [frequencyRatings(3)] = 474
- “very” [frequencyRatings(4)] = 202

The total number of evaluations = 3190

From this discrete set of five known data points (1252, 698, 564, 474 and 202) a new, fuzzy function is built (see figure 28 dotted red line) on a set of three data points (v1, v2, and v3).

[131] A task fulfilled by K. Vermeulen, graduate student in Inforatics.



Figuur 28: Cumulative function and fuzzy function.

Calculation of v1 (25%):

x = 3190 / 4 = 797

rating value 797 = 0 (<1252)

v1 = 797 / frequencyRatings(0) = 797 / 1252 = 0,64

Calculation of v2 (50%):

x = 3190 / 2 = 1595

rating value 1595 = 1 (<1252 + 698)

v2 = 1 + (1595 – frequencyRatings(0..0)) / frequencyRatings(1)

v2 = 1 + (1595 – 1252) / 698 = 1,49

Calculation of v3 (75%):

x = 3190 * 3 / 4 = 2392

rating value = 2 (< 1252 + 698 + 564)

v3 = 2 + (2392 – frequencyRatings (0..1)) / frequencyRatings (2)

v3 = 2 + (2392 – (1252 + 698)) / 564 = 2,78

B Calculation of scores per music excerpt, adjective and profile

In order to calculate the scores per music excerpt, to begin with, the rating values attributed by all subjects who fit a specific profile, were sorted in ascending order as to adjective and excerpt number respectively. After that, the cumulative median value (v2) was calculated.

The score for each adjective, profile and excerpt concerned results in the following function value:

score(median), with “score” being the fuzzy function corresponding to the adjective and profile concerned.

In what follows an example of the calculation of the score for music excerpt one, the profile “male expert” and the descriptor “cheerful” is given. The number of ratings for:

- “not” [frequencyRatings(0)] = 12
- “little” [frequencyRatings(1)] = 6
- “moderate” [frequencyRatings(2)] = 0
- “rather” [frequencyRatings(3)] = 2
- “very” [frequencyRatings(4)] = 0

The total number of evaluations =20

Calculation for 1/2(x = 20 / 2 = 10) :

“rating value 10” = 0 (< 12), dus :

median = 10 / frequencyRatings (0) = 10 / 12 = 0,83

score(median) = score(0,83) = 0.5 * (0,83 – v1) / (v2 – v1),

then v1 = 0,64 < 0,83 < v2 = 1,491...

=> score(0,83) = 0.5 * (0,83– 0,64) / (1,49– 0,64) = 0,11

C Calculation of combined scores

IF no adjectives are selected	THEN score = 1
IF 1 adjective is selected	THEN score = score for the adjective concerned
IF n adjectives are selected	THEN score = the nth power root of the product of the adjective scores.

In what follows an example of the calculation of combined scores for music excerpt one, the profile “male expert” and the descriptors “cheerful”, “sad” and “passionate” is given.

cheerful = 0,11

sad = 1

passionate = 0,97

case 1: “cheerful” and “sad” are selected:

score = square root(0,115 * 1) = 0,34

case 2: “cheerful”, “sad” and “passionate” are selected:

score = 3th power root(0,11 * 1 * 0,97) = 0,48

7.3.5 Conclusion

A prototype database system has been developed that to test and evaluate the usability of a contextual, emotion-based music recommender system. As evaluation of the demo was done just prior to the submission of this thesis, the log files could not be analyzed. A possibility would be to set up a new evaluation test by inviting those music information retrieval system users who took part in the survey but who were not involved in the experiment on quality annotation. It would be interesting to compare the evaluations from that sample with the sample from the broad population.

Although both the satisfaction ratings and usability evaluations have not been statistically analyzed yet, the overall reactions of the users at ACCENTA were very positive. IPEM collaborators, who observed people using the application, could see that they enjoyed discovering music by entering emotion-based queries. An advantage of a recommender system over social recommendations is that a system can pull from a large database. No friend or shop keeper knows about all the music a person might like.

From conversations with the test users we learned they found the recommendations made by the system very helpful and relevant. Using fuzzy logic and membership functions as to quality ratings is an important alternative to overcome the vagueness of subjective concepts. It can be assumed that the hypothesis that the agreement among participants in the experiment reflects agreement for a broad population will probably be confirmed by the satisfaction ratings.

Furthermore, test users have shown interest in the applicability of the system for educational purposes. Indeed, the application easily generates new sets of recommendations without a lot of effort from the user. One of the advantages of the audio feedback then is that it could for instance be used to give children an idea of how music that is passionate, void, static, or else sounds.

Discussion and conclusion

In this thesis I put forward the assertion that the development of a content-based music information retrieval system will benefit from a user-centred approach to music description. With this in mind, a conceptual framework was defined, which in turn formed the basis for an integration of bottom-up and top-down approaches to music information retrieval. Starting from this framework, the research focused on empirical validation and support of a content-based music description model. This model forms the core component in computational applications for music information retrieval. In what follows I shall focus on the following aspects of my study (1) defining a user-centred approach, (2) methodological and theoretical implications of the study, (3) summary of the contributions and findings and (4) suggestions for future research. The final conclusion follows thereafter.

Defining a user-centred approach

Throughout this thesis, techniques have been developed for the manual annotation of music. Although, some of these techniques were developed in different projects at IPEM, the scale at which these techniques have been applied, as well as the different levels of annotation which have been addressed, are entirely new and untried before in musicology.
In addition, it is likely that within the area of music information retrieval, this is the first study that presents a conceptual framework which supports the integration of research input from different research communities, in particular: musicology, psychology and engineering. It is also the first study that established a conceptual framework built on methodological considerations for manual annotation. This framework is an intermediate stage between the automatic description of low-level music content and user description of high-level content. Furthermore, it is the first research that makes serious attempts at defining both a profile of the potential music information retrieval users as well as the semantic concepts at the basis of their interpretation of the music of their interest.

When I started this study, about four years ago, music information retrieval research was very much in its infancy. It appeared to me that within the interdisciplinary music information retrieval community, the disciplines involved were not equally represented. Especially contribution by musicology and extensive experimental research was rather scarce. Meanwhile, research on music information retrieval has become an active interdisciplinary research field, involving disciplines such as computer science, musicology, psychology, and neurosciences. Research in these areas proliferates, and as a consequence there arose the need to link these research activities. I believe that the musicological approach, put forward in this thesis, can offer a model for unification.

In the domain of music information retrieval, music is often approached from a “scientific” viewpoint. Much to my regret, science is still seen as something belonging solely to scientists and music is considered the domain of musicians and musicologists. From my viewpoint as a musicologist, I consider music both as an integral part of the arts as well as

encompassing everything which has to do with the phenomenon called “sound”. Therefore, it was my aim to arrive at an integration of the scientific and musical approaches.
In this thesis, I did not want to go into the discussion between science and the arts. On the contrary, I envisaged a framework that would aid exploration of the relationship between them. As music exists by virtue of the human response to it, the characterization of musical content cannot be separated from that very response. Furthermore, it may be assumed that the future success of music information retrieval system applications will be strongly influenced by the user’s sensibility to this technology. Given this conviction, I reject the idea that music can be described without reference to the human subject who perceives it and interprets it. Hence the idea to focus on a user-centred approach to music information retrieval. This approach can thus fill the gap between the human way of dealing with content and meaning, and the machine encoding of physical energy.

Methodology and theoretical implications

Advantage of the approaches

In this thesis, my research was based on the following three approaches to music information retrieval: user-centred, phenomenological and application-oriented. By means of elaborate experimental research I have demonstrated that the user-centred and phenomenological approaches are important viewpoints, both for theories of music perception and cognition and for the practical construction of computer systems destined for search and retrieval of music content. A characteristic of the user studies conducted in this thesis is that they are motivated by content-based system design. In other words, all the music stimuli used in the investigated were real music. Although a user-centred approach was the starting point, this intention places my research in between user-oriented and system-oriented research. The significance of this approach is that it takes into account the double nature of music as an acoustic and as an expressive phenomenon. This objective led to the third approach, which is application-oriented. This approach involved the design of three applications that validate the conceptual framework.

To recapitulate, the research challenges of my study were: (1) constructing a conceptual framework, (2) defining a general framework for the manual annotation of music content, (3) providing a framework for the evaluation of music information retrieval tools, (4) providing user-dependent knowledge on high-level music content, and (5) providing a foundation for new music analysis.

Basic observations of the conceptual framework

The model presented in this thesis conceptualizes the multiple approaches to the domain of music information search and retrieval by distinct research groups. The term “model” herein should of course be put in the right context. As the subtitle suggests, the aim of this thesis was not to present a model in the sense of a theoretical construct that represents physical processes and sets of logical and quantitative relationships between them. Rather, it sought

to provide an overall framework for placing music information retrieval models and empirical research within the complex whole of approaches to music that exist within this research community. In other words, its aim was to present a skeletal structure, within which empirical observations and algorithm development can be understood as a part of a coherent whole, and upon which further music information retrieval research can be based.

This conceptual framework for content-based music information retrieval, indeed runs parallel with the aim of supporting the global picture of the variety of music information search and retrieval approaches. The argument is founded upon five basic observations concerning the nature of the model: (1) it is methodologically well-founded, (2) it has an empirical basis, (3) it is coordinating, (4) it is a dynamic structure and (5) it is fully integrated.

The conceptual framework is **methodologically well-founded** in that its tenets are based on previous research in which I was involved. It was developed in a series of steps, going from the global internal representational framework of the IPEM toolbox, via the MAMI-taxonomy to the MIR-conceptual framework.

The conceptual framework has an **empirical basis** in that taxonomies based on experimental observations were the foundation for defining constituent categories and concepts.

The conceptual framework is **coordinating** in that it is a multi-leveled structure of the phenomenon “music”, including acoustical, perceptual and semantic levels. It comprises physical properties of sound and perceptual, affective and cognitive processes as well. As a consequence it ties in with existing research within the music information retrieval domain.

The conceptual framework is a **dynamic structure**. Although it is built on rigid structures, the elaborated framework is open-ended, in that more rigid structures and unique relations can be added. It is open to new concepts in music that may emerge from new ways of description and perception, which are related to modern technology. It is an intermediate structure toward a taxonomy driven music information retrieval system.

The conceptual framework is **fully integrated**, in that it starts from real audio and the main processes on which it is based are assumed to apply to any listener and to any music culture. Although the experimental part of the study has mainly concentrated on responses to western music by western listeners (i.e. Belgian), the conceptual framework is assumed to be independent of these. Further studies are needed to confirm this assumption.

Accomplishment of the research processes

As outlined in chapter one, categorization, annotation and contextualization were delineated as the three research processes required for achieving the goal of this thesis. In order to show how these processes were reflected in this thesis, in what follows I shall give

a brief overview of what was achieved, by summarizing the objectives dealt with in each chapter.
Chapter one preceded the implementation of the research processes. It situated the investigation within the domain of music information retrieval search and retrieval. Apart from this, the music information retrieval problems addressed in this thesis were defined and the approach was argued.

Chapter two was theoretical and dealt with the process of **categorization**.
Following steps were undertaken:

- the MAMI-taxonomy as conceptual framework was presented and elaborated;
- the components of the model were discussed in detail;
- an explanation was given of the ontology of the conceptual framework in terms of ordinary language;
- the spatial-temporal structure class was put forward as the major concept class for the structuring and grouping of sound;
- a model of descriptors defining the spatial-temporal structure class was provided;
- the requirements for system development were defined.

Chapter three was theoretical and methodological and dealt with the process of **annotation.**
Following steps were undertaken:

- the concept and problems of music annotation were specified;
- a three-level framework based on methodological considerations concerning manual annotation of real audio was provided;
- a manual annotation framework definition was given, supporting context dependencies, computer modelling and multiple representation levels;
- the framework was validated by means of two case studies: one on the manual annotation of vocal queries and one on the manual annotation of drums;
- the case study on the manual annotation of vocal queries provided methodology for user-oriented and modelling-oriented annotation of melodies;
- the case study on the manual annotation of drums provided ground truth for algorithm development.

Chapters four to seven focused on empirical investigation and applications and dealt with the process of **contextualization**. Depending on the experimental aims, various research methods were used.

In chapter four, spontaneous user behaviour was investigated by means of two approaches. The first approach focused on music search behaviour, whereas the second approach drew attention on vocal querying methods. Following steps were undertaken:

- a literature based survey of query methods was given;
- investigation from the field of music distribution was reported;
- a survey on music search behaviour was discussed in detail;

- an elaborate experiment on vocal spontaneous query behaviour was discussed in detail.

The empirical investigation discussed in chapter five and six involved two aspects of a large-scale user study. Chapter five focused on user context, whereas in chapter six the attention was drawn on the description of high-level qualities of music. Following steps were undertaken:

- a background in user context was provided;
- a survey on users' musical background, habits and interests was discussed in detail;
- the global profile of a potential music information retrieval user was defined;
- user dependent effects on user context were investigated;
- background on the description of music qualities was provided;
- an experiment on the annotation of music qualities was discussed in detail;
- influences of subject related factors were discussed;
- relationships between perceived music qualities were uncovered;
- consistency within user annotations was tested.

In chapter seven the set up of three applications that evaluate the experimental investigation and validate the conceptual framework was discussed. Following applications were created:

- a prototype of a user interface based on a taxonomy;
- an annotated database of experimental findings;
- a prototype of a semantic music recommender system.

Although a conceptual framework for content-based music information retrieval was inexistent and user-centred research focusing on users' perception of music qualities is scarce, for each issue dealt with in this thesis, the investigation was situated within existing research. Throughout this thesis, the significance of each investigation was shown by referring to key studies that place the presented research within context.

Summary of research contributions and findings

Apart from the development of the conceptual framework, which was the main goal of this thesis, a variety of music information retrieval contributions were made by means of the methodologies worked out for the case studies, the experimental investigation and the evaluation. This thesis presented findings from various research methods but not all of them showed significant results. However, they illustrated the steps in the research process. The findings that were highly significant will undoubtedly be of benefit to the improvement of music information retrieval systems. In this section the results obtained through a user-centred and phenomenological approach to study music information retrieval issues are summarized. An overview is given of the major research contributions and findings according to the research questions that were postulated in chapter one.

RQ #1: What is a suitable taxonomic scheme?

In view of a taxonomy-driven content-based music information retrieval system, the taxonomies, the conceptual framework and the manual annotation methodologies presented in this thesis, offer a multi-dimensional skeleton within which user-centred music information retrieval issues can be discussed and compared.

Rather than a taxonomy comprising finite categories with defined boundaries, I suggested a dynamic conceptual framework modelling content-based music information retrieval as a multi-layered space.

The global model resulting from this study is a comprehensive framework, encompassing music from its physical properties up to the human perception and cognition of it. I suggested a five-leveled conceptual framework that involves acoustical, sensorial, perceptual, structural and expressive description levels. The model is a high-level theory for a fully integrated approach to music description.

Basic observations of that conceptual framework have been discussed in a previous section. Although this concluding discussion is concerned with highlighting the potential of the conceptual framework, the significance of this thesis lies not only within the presented model itself but also within the approach and methodologies applied to formulate and explore it.

RQ #2: What is a suitable reference database?

Content-based music information retrieval deals with music as it is heard. The observations made during my study show that apart from metadata, the ideal reference database would be one that contains raw audio and that can match high-level descriptions of music with low-level descriptions automatically extracted form music content. Such a database would allow users to find any fact from any entry point into the database and new facts would be automatically added. This is in contrast with most existing information systems that claim to be music databases but are nothing more than collections of music excerpts in MIDI format. Overly simplified stimuli, however, are not sufficient to the construction of fully integrated systems. Moreover, they might become an easy way to escape from making efforts to understand the implications of real music.

The study that was presented on the kind of music that a reference database for music information retrieval research should include and the amount of examples needed to have a representative collection resulted in a the MAMI-database. This database contains target audio files as entire pieces of music in CD format. The use of the MAMI-database as a test bed throughout the experimental investigation contributed to the definition of a suitable reference database.

A reference database should meet two fundamental requirements: it should contain entire pieces of real music and the music should be representative of the global music distribution. In contrast with digital music collections, consisting of short excerpts that

mostly are only the beginning of a piece, my study has shown that entire pieces are needed for multiple reasons. A spontaneously performed vocal query, for example, is not necessarily the beginning of a piece or a recurring theme. Entire pieces are also needed for the application of analysis methods that look at the music piece as a whole, such as statistical analysis of occurrences of sequences. The requirement for music that is representative of the global music distribution was evidenced by the list containing over 3000 titles provided by potential users of a music information retrieval system who participated in the survey on user context.

Apart from these requirements and in order to generate statistically meaningful results and to evaluate music information retrieval algorithms, a suitable reference database should contain large sets of music. It was assumed that at least ten examples per global genre category are needed for relevant testing. The MAMI-database which responds to this assumption may be small in order to cover the needs of algorithm testing. However, further research is needed on this issue.

A suitable reference database has user-oriented and a modelling-oriented implications. From a user-oriented perspective it has the advantage that test databases can be created containing excerpts from the music pieces in the reference database (i.e. the target files to be matched). From a modelling-oriented viewpoint it has the advantage of enabling easy testing of algorithms for similarity between aural queries and the target files.

RQ #3: What are appropriate annotation methodologies?

In this thesis annotation of music has been approached from a variety of perspectives. The definition of a general framework for manual annotation of musical audio that was provided is an intermediate but necessary step towards the fully automated annotation of music. Although manual annotation is very demanding, the case studies that were carried out have shown the great potential of manual annotation.

The strategy for user-oriented annotation of vocal queries, for example, provided useful insights into the query-target similarity, a task that algorithms still cannot do. It also provided new insights in the structure of a query-by-voice. Apart from that, further understanding was gained on query length, query method and query performance style.
The methodology used for modelling-oriented annotation of vocal queries has shown its efficiency in that it improved a new auditory model-based transcriber of vocal melodic queries. This transcriber clearly outperforms other state-of-the art systems with which it was compared.

The case study on manual annotation of drums using real musical stimuli has provided ground truth for a drum detection algorithm that was ranked second in the first annual Music Information Retrieval Evaluation EXchange (MIREX) contest of 2005. Besides, the user annotations collected by means of this case study were also used as ground truth data for the MIREX audio drum detection contest.

RQ #4: What are suitable ways for users to interact with music information retrieval systems?

The study on spontaneous music search behaviour consisted of fieldwork and a survey, both resulting in some remarkable findings. Interviews in the field of music distribution revealed that the impact of having personal contact with a seller to whom questions about desired music can be asked may not be underestimated. Although the sample in the survey was not entirely representative of the typical music information retrieval population, important indications for further music information retrieval research were supplied. The finding that people who know the music they are looking for, prefer to go to a music shop is in agreement with the assumption that users of music information retrieval systems most likely are searching for unknown music.

Other interesting findings were the following:

- when searching for known music, title and genre are the preferred look-up methods;
- when searching for unknown music, genre and rhythm are the most suitable features;
- providing the system with an example would be the preferred new look-up method;
- users' retrieval preferences are title, composer/performer and musical audio.

An essential contribution to the development of music information retrieval systems based on vocal querying was made. Research into how users behave spontaneously when they perform a vocal query has resulted in new approaches to the vocal query paradigm. In contrast with the many constraints that vocal query systems impose on users, the experiment on spontaneous user behaviour showed that, when users are allowed maximum freedom, different types of vocal querying occur. In descending order of significance, the preferred query methods are: singing the lyrics, singing with syllables, whistling, humming and vocal imitation of percussive sounds. Producing vocal queries of 14 seconds on the average, about two third of the users make use of only one of these methods, the others alternating between two or more methods. Another finding was that syllabic query segments are longer and thus more prominent than queries based on lyrics. Detailed analysis of the use of syllabic queries provided new insights that in turn might inspire new applications.

RQ #5: Who are the potential users of music information retrieval systems?

In contrast with most user studies that take a small sample of university students as their subjects, the survey on user context was designed in order to recruit a broad range of participants. The recruitment methodology that was set out was successful as it clearly appealed to a representative sample of potential music information retrieval system users.
The survey on user context provided a major contribution because it made it possible to define a global profile of the potential user of music information retrieval systems. It also confirmed the assumption that users of music information retrieval systems are music lovers and are used to working with the Internet. The global profile of the average user of

music information retrieval systems that could be established is defined by the user's musical background, habits and interests, preferred genres, taste and favourite titles.

To recapitulate the essential points, the average music information retrieval users:

- are younger than 35;
- spend one third of their Internet time on music activities;
- are actively involved in music information retrieval;
- have a broad musical taste;
- cite pop, rock and classical music as their preferred genres.

Furthermore, analysis of relationships has provided insight into of the effect of gender, age, expertise, music background, broadness of taste and interest in classical music. These findings are a useful guide in deciding which relations might be of use in music information retrieval. From these observations it can also be suggested that including different target user groups in the design process might enhance the functionality of an interactive system.

RQ #6: What is the effect of the users' background on quality description?

Using the profile information collected in the survey on user context, analysis of the influence of subject related factors revealed subject dependencies for gender, age, expertise, musicianship broadness of taste and familiarity with classical music.

To recapitulate, the significant subject dependencies found were:

- male users judged the music excerpts as being more restrained, harmonious and static than did female users;
- female users judged the music excerpts as being more beautiful and more difficult than did male users;
- users aged over 35 found the music excerpts to be more passionate and less static than did the majority of the users, who are younger than 35;
- users with a broad musical taste judged the music excerpts as being more beautiful and more pleasing than did users with a narrow taste;
- users who said they were familiar with the music in the test, often listened to classical music.

Apart from these findings, it was shown that familiarity with the music excerpts had a highly significant effect for all adjectives involved in the experiment describing perceived and felt emotions.

To recapitulate:

- known excerpts were perceived as being more tender, more passionate, less restrained, more cheerful, less aggressive, less anxious, more sad and more carefree (in descending order of significance) than unknown music;
- known excerpts were experienced as being more pleasing, less annoying, less indifferent and more touching (in descending order of significance).

RQ #7: What are fitting music qualities?

Search behaviour also depends on highly developed abilities to perceive and interpret musical information. A listener must call to mind a great deal of analogies, metaphors and memories in order to make coherent sense out of the music content. The knowledge provided on how a user of music information retrieval systems perceives music qualities will be beneficial to future retrieval systems that deal with semantic description of high-level content.

Factor analysis of user ratings for expressive features of real music revealed three dimensions that are in agreement with the dimensions interest, valence and activity found in previous research. With regard to unanimity in describing structural music features, adjective pairs were tested that describe loudness, timbre, tempo and articulation. Participants most unanimously agree on loudness and tempo, whilst less unanimity was found for timbre and articulation.

It was shown that the complexity of musically induced emotions resides in one or several different structural categories of music. Interesting relationships for example were found between expressive and structural features and a strong correlation was found between affect (tender-aggressive) and loudness (soft-hard).

Analysis that checked the unanimity among users on ratings of perceived structural qualities have shown that loudness and tempo are the structural features that are the most unanimously perceived.

The semantic music recommender system provides direct evidence of the degree to which users agree about music qualities.

RQ #8: Are users capable of making reliable judgments?

A consistency test of quality judgments carried out over time and using different time intervals revealed that, for the majority of the adjectives describing affect qualities and structural features, user evaluations are quite reliable.

But, it is of interest to the music information retrieval community that the study on user reliability also provided knowledge on ambiguous terms. Marginal Homogeneity tests revealed that when measurements are repeated and when the time interval is greater than a week, the affect adjectives anxious, cheerful, aggressive and passionate appear to be ambiguous terms.

As far as structure is concerned, it was found that there was the least agreement among subjects on adjective pairs describing brightness and roughness. These findings were also confirmed by the high standard deviation of ratings when consistency among subjects was analysed.

From these findings music information retrieval system developers can assume that if one user really likes a given piece of music, there will almost certainly be a large number of

other users who also like that piece of music (even though not everyone may like it). It is certainly not trivial to discover new music that has this property.

The knowledge gathered about the reliability of queries based on semantic description of high-level features of music and on subject related factors could enhance the development of user-dependent music recommender systems. This is evidenced by the prototype of a semantics based music recommender system that was created on the basis of user ratings of music qualities.

RQ #9: What is an appropriate user interface?

To make sure that the development of music information retrieval systems can look forward to a successful future, it is important to build user–friendly systems that prevent users from getting bored or frustrated. The conceptual framework was used to build a practical music information retrieval application as an example of a user interface taxonomy. This taxonomy reflects the objectives of a user interface design.

An appropriate user interface should include *multiple use modes*, offer *different query methods*, provide *musical audio feedback*, be *consistent*, *intuitive and comprehensible*.

An appropriate user interface should;

- be **multi-modal** in that it can then account for distinct user groups. A multi-modal interface is suggested because it was shown that the effect of gender, age, expertise, music background, broadness of taste and interest in classical music defines different user groups of music information retrieval systems. To accommodate, for example, both the expert and the novice, a flexibility of use modes is needed. Those who have a passion for music may have advanced searching skills in music due to experience and practice. Moreover, integrating target user groups into the system design process may enhance the functionality of a system. Someone who is fond of music may often look up unfamiliar music that could become a new favourite;
- offer **multiple query methods**. The survey on music search behaviour showed that users found several search methods suitable, preferring especially providing linguistic information, a prerecorded example and singing into a microphone. It was also shown that when performing vocal queries people prefer not a single query method but a combination of them. Therefore it is suggested that including different query methods in the system may enhance the functionality of an interactive system;
- provide **musical audio feedback**. In view of music recommender systems (such as the example that was discussed in this thesis), it is important that the users can listen to the recommended music;
- be **consistent**, so that the user will be able to interact with the system in a uniform way, without having to change between different input devices;

- be **intuitive**, so in that it is easy to use and visually appealing;
- be **comprehensible** in that it will provide the user with guidance and feedback.

Future research

Although this concluding discussion is mainly concerned with highlighting the conceptual framework for content-based music information retrieval and its inferences, the significance of this thesis is also within its potential to suggest, outline and support further research. The conceptual framework could thus inspire the formulation of useful research questions which may lead to new empirical investigation and to the development of new systems. In what follows some suggestions for future research are given.

One possible approach to further research would be to refine the methodologies used in the experiments presented here. Several improvements and extensions could be made ranging from the choice of stimuli (e.g. not solely western music) to the types of measurement (e.g. move along with the music). Alternatively, new experimental techniques could be designed with the aim of observing the impact of extra musical features (e.g. pictorial). Such a multimedia approach, involving the manipulation of music and other contexts may yields useful results.

The implications being that if music information retrieval is to make the leap from physical observation and feature extraction of sound to the coherent understanding of emotional response to music, then, future research must embrace the multimedia context in which music is experienced.

Another future research path could involve further investigation of user behaviour. In order to make music information retrieval appealing to the average user, more investigation will be needed, incorporating possibly predicting user objectives and desires to retrieve music, images and video. The aim of such investigation would be to maximize natural interaction with the system and to minimize thought and manual involvement on the part of the user.

Normal users may just look up music and not information about it, but those users who probably will make use of most of all the possibilities offered by music information retrieval systems are researchers, musicologists, composers and producers. They will typically not only look up a recording, but be equally as interested or more in the higher-order features of music. Investigation into specific domains of interest of music professionals of various kinds may be of future interest.

Based on the findings from this study, three applications were created that may be of benefit to the future development of commercial applications and research tools such as searchable databases. Widening this research could include music audio archives, for example in museums, radio broadcasting, specialized archives and private collections.

In conclusion

The major conclusion this thesis makes is that by assessing the interpretative consequences of high-level music description, an analytical approach can be developed which focuses on the interplay between the “real” environment (the user) and the “virtual” environment (the music as physical object). It has been shown that research into users’ music background, interest and perception of music has the potential of improving music information retrieval system development. By considering a broad range of musical stimuli (real music) and a broad population (real users), I believe that my work may be of relevance to the field of music information retrieval research. A conceptual framework for content-based music information retrieval is defined within which other content-based music information retrieval work can be considered. This is where I hope this research may be of particular interest for those working in the music information retrieval community.

List of figures

Figure 1: Generalized representation of a MIR system architecture. 12
Figure 2: Components of a Music Information Retrieval (MIR) framework. 34
Figure 3: Description levels of he conceptual framework. 37
Figure 4: Concept classes specifying spatial-temporal structure features. 42
Figure 5: Conceptual framework for music information retrieval. 45
Figure 6: Constituent music categories of the expressive and structural layers. 46
Figure 7: General framework for annotation of musical audio. 62
Figure 8: Representation levels and associated annotation methods. 63
Figure 9: Vocal query annotation using PRAAT. 69
Figure 10: Multi-track sequencer and beat annotation set up. 74
Figure 11: The MAMI music database. 95
Figure 12: General overview of the experiment on spontaneous vocal query behaviour. 100
Figure 13: Global set up of the study of user context. 120
Figure 14: Contribution of genre per age category. 131
Figure 15: Annotation of affect. 137
Figure 16: Annotation of structure. 138
Figure 17: Model for quality description. 151
Figure 18: Web-interface for the annotation of affect features. 155
Figure 19: Comparison between age categories in the main dataset and the subset. 157
Figure 20: Factor loadings of the expression space. 161
Figure 21: Factor loadings of the structure space. 162
Figure 22: Conceptual design of the MAMI-RDBM. 195
Figure 23: Screenshot of the basic relationships of the MAMI-RDBM. 198
Figure 24: Screenshot of the interface of the MAMI-RDBM query builder. 199
Figure 25: Profile screen of the semantic recommender system prototype. 203
Figure 26: Search screen of the semantic user recommender system. 205
Figure 27: Window with media player, query output and satisfaction rating. 206
Figuur 28: Cumulative function and fuzzy function. 208

List of tables

Table 1: General model of the research processes. 27
Table 2: Taxonomy for user-oriented annotation of vocal queries. 66
Table 3: Taxonomy for model-oriented annotation of vocal queries. 67
Table 4: Taxonomy of drum types, labels and MIDI notes. 73
Table 5: Overview of query methods under research. 82
Table 6: Genre sales in Belgium in 2000 according to IFPI. 95
Table 7: Sets with musical stimuli in the query-by-voice experiment. 97
Table 8: Overview of the query-by-voice experimental procedure. 99
Table 9: QbV: Occurrence of different query methods. 103
Table 10: Occurrence of performance styles. 103
Table 11: Vowel categories for syllable annotation. 104
Table 12: Analysis of syllable structure following onset and rhytme. 105
Table 13: Analysis of syllable structure following rhyme and coda. 105
Table 14: Use of query methods. 106
Table 15: Overview of the number of queries created per piece of music. 110
Table 16: General concept of the survey design. 122
Table 17: Distribution of participants in the survey according to their response. 124
Table 18: Distribution of occupational categories. 125
Table 19: Internet time spent for general activities and for music related activities. 126
Table 20: Preferred music genres sorted by percentage of responses. 129
Table 21: Evolution of taste. 131
Table 22: Number and mean of genre selections per age category. 131
Table 23: Comparison of distributions for “genre” (F8) and “taste” (F9). 132
Table 24: Distribution of favourite titles. 133
Table 25: Comparison of genre distributions in “genre” and “favourite titles”. 134
Table 26: Bipolar adjectives used for the annotation of favourite titles. 136
Tabel 27: Ratings in percentage points for the adjective pairs related to affect. 137
Table 28: Ratings in percentage points for the adjective pairs related to structure. 138
Table 29: Taxonomy of broad genre categories. 139
Table 30: Adjective pairs with highest scores for “variable” and “undecided”. 140
Table 31: Distribution of binary variables (N=663). 141
Table 32: Distribution of genre categories and sub-classes. 152
Table 33: Overview of the design of the annotation experiment. 153
Table 34: Comparison between the subset and the main dataset. 157
Table 35: Distribution of binary variables (N=79). 158
Table 36: Standard deviations for the structural features. 163
Table 37: Non-parametric Kendall tau-b correlations: expression*structure (N=12640). 164
Table 38: Pearson’s correlations: expression*structure (N=160). 165

Tabel 39: Stimuli used in the consistency test. 167
Table 40: Marginal Homogeneity test results for affect qualities. 168
Table 41: Contingency table for the ranking of “anxious”. 169
Table 42: Marginal Homogeneity test results for structural features. 170
Table 43: User interface taxonomy. 178
Table 44: Table structure in the MAMI relational database model. 197

Content of the electronic appendix

- _Original Articles
  - ART_I_UserBehaviour_ESCOM2003.pdf
  - ART_II_UserDependentTaxonomy_SMAC2003.pdf
  - ART_III_QueryByVoiceExperiment_ISMIR2003.pdf
  - ART_IV_ ManualAnnotation_ISMIR2004.pdf
  - ART_V_SpontaneousUserBehaviour_CIM2004.pdf
- Chapter3_ Annotation
  - CaseStudy1_VocalQueries_ModelOriented
    - QbVExperiment_ModelOrientedAnnotation_Description.pdf
    - .wav and .TextGrid files (examples of model-oriented annotation)
  - CaseStudy1_VocalQueries_UserOriented
    - QbVExperiment_UserOrientedAnnotation.xls
    - QbVExperiment_UserOrientedAnnotation_Description.pdf
  - CaseStudy2_AnnotationOfDrums
    - AnnotationOfDrums_MusicStimuli.pdf
    - AnnotationOfDrums_Guidelines.pdf
- Chapter 4_Spontaneous User Behaviour
  - SurveyMusicSearchBehaviour_Questionnaire.pdf
  - MAMIMusicCollection_MusicStimuli.pdf
  - QbV_Experiment_QueryExamples
    - .wav and files (examples of vocal queries)
- Chapter 5_Survey on User Context
  - SurveyUserContext_Questionnaire.pdf
  - ExpressiveAdjectives_PilotStudy.xls
  - UserContext_Relations_CrosstabsGender.rtf
- Chapter 6_Description of High Level Features
  - AnnotationExperiment_Guidelines.pdf
  - AnnotationExperiment_WebInterface.pdf
  - AnnotationExperiment_MusicStimuli.xls
- Chapter 7_Three Applications
  - Application 1: User interface taxonomy (MAMI-USIT)
    - MAMIPrototypeUserInterface.exe
  - Application 2: Data management (MAMI-RDBM)
    - MAMIQueryBuilderDemo.exe
  - Application 3: Semantic Music Recommender System (SeMuReS)
    - MAMISeMuReSDemo.exe

References

Arms, W. Y. (2000). *Digital Libraries*. Cambridge, MIT Press.

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

Bainbridge, D., Cunningham, S. J., & Downie, J. S. (2003). How people describe their music information needs: a grounded theory analysis of music queries. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR03)*, Baltimore, 221-222.

Baroni, M., & Finarelli, L. (1994). Emotions in spoken language and vocal music. In *Proceedings of the International Conference for Music Perception and Cognition*, Liège, 343-345.

Baumann, S., & Halloran, J. (2004). An ecological approach to multimodal subjective music similarity perception. In *Proceedings of the Conference in Interdisciplinary Music (CIM04)*, Graz, CD-Rom.

Baumann, S., Klüter, A., & Noriel, M. (2002). Using natural language input and audio analysis for a human-oriented MIR system. In Proceedings of the International Conference on Web Delivering of Music (Wedelmusic), Darmstadt. Retrieved, February 8, 2003, from

http://www.dfki.uni-kl.de/~baumann/pdfs/WEDEL2002.pdf

Bigand, E., & Pineau, M. (1996). Context effects on melody recognition: a dynamic interpretation. *Current Psychology of Cognition, 15*, 121-134.

Bigbee, T., Loehr, D., & Harper, L. (2001). Emerging requirements for multi-modal annotation and analysis tools. In *Proceedings of Eurospeech Conference*, Aalborg.1533-1536.

Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, *32*(1). 23-60

Birmingham, W., Pardo, B., Meek, C., & Shifrin, J. (2002). The MusArt Music-Retrieval System: an overview. *D-Lib Magazine*, *8*(2). Retrieved, March 22, 2003, from

http://www.dlib.org/dlib/february02/birmingham/02birmingham.html

Boersma, P., & Weenink, D. (1996). *PRAAT, a system for doing phonetics by computer*. Amsterdam: Institute of phonetic sciences of the University of Amsterdam. Retrieved March 26, 2002, from http://www.fon.hum.uva.nl/praat/

Bonardi, A. (2000). IR for contemporary music: what the musicologist needs. In *Proceedings of the International Symposium on 1st Music Information Retrieval (ISMIR00)*, Plymouth. Retrieved January 17, 2002, from http://ciir.cs.umass.edu/music2000

Brøndsted, T., Augustensen, S., Fisker, B., Hansen, C., Klitgaard, J., Nielsen, L., & Rasmussen, T. (2001). A system for recognition of hummed tunes. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFX-01)*, Limerick, 203-208.

Byrd, D., & Crawford, T. (2002). Problems of information retrieval in the real world. *Information Processing and Management, 38*, 249-272.

Byrd, D. (2003). *Candidate music IR Test Collections*. Retrieved July 31, 2003, from

http://php.indiana.edu/~donbyrd/MusicTestCollections.html

Carré, M., Pierrick, P., & Apélain, C. (2001). New query-by-humming music retrieval system conception and evaluation. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFX-01), Limerick*, 227-232.

Carter, F. A., Wilson, J. S., Lawson, R. H., & Bulik, C. M. (1995). Mood induction procedure: importance of individualizing music. *Behavior Change, 12*, 159-161.

Casey, M. (2002). Generalized sound classification and similarity in MPEG-7. *Organised Sound, 6*(2), 153-164.

Chai, W. (2001). *Melody retrieval on the web*. Unpublished thesis, Master of Science in Media Arts and Sciences, MIT, Massachusetts.

Chai, W., & Vercoe, B. (2000). Using user models in music information systems. In *Proceedings of the 1st International Symposium on Music Information Retrieval (ISMIR00)*, Plymouth. Retrieved January 17, 2002, from http://ciir.cs.umass.edu/music2000

Clarisse, L. P., Martens, J.-P., Lesaffre, M., De Baets, B., De Meyer, H., & Leman, M. (2002). An auditory model based transcriber of singing sequences. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR02)*, Paris, 116-123.

Cope, D. (1991). *Computers and musical style*. Oxford, UK: Oxford University Press

Covey, D. T. (2002). *Usage and usability assessment: library practices and concerns*. Washington, DC: Digital Library Federation and Council on Library and Information Resources. (CLIR Report105). Retrieved October 15, 2003, from http://www.clir.org/pubs/reports/pub105/contents.html

Cumming, N. (2000). *The Sonic Self: Musical Subjectivity and Signification*. Bloomington and Indianapolis: Indiana University Press.

Cunningham, S. J. (2002). User studies: a first step in designing a MIR testbed. In *The MIR/MDL Evaluation Project White Paper Collection*, *2nd ed*., 17-19.

De Mulder, T., Martens, J.-P., Lesaffre, M., Leman, M., De Baets, B., & De Meyer, H. (2003). An auditory model based transcriber of vocal queries. In *Proceedings of the 4t International Conference on Music Information Retrieval (ISMIR03)*, Baltimore, 245-248.

De Mulder, T., Martens, J.-P., Lesaffre, M., Leman, M., De Baets, B., & De Meyer, H. (2004). Recent improvements of an auditory model based front-end for the transcription of vocal queries. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP04)*, Montreal. Vol. IV, 257-260.

Doraisamy, S., & Rüger, S. (2002). A Comparative and fault-tolerance study of the use of N-grams with polyphonic music. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR02)*, Paris, 101-106.

Downie, J. S. (2002). Toward a theory of music information queries: system design implications. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR02)*, Paris, 299-300.

Downie, J. S. (2003). Music Information Retrieval. In B. Cronin (Ed.), *Annual Review of Information Science and Technology*, 295-340.

Downie, J. S. (2003). Toward the scientific evaluation of Music Information Retrieval Systems. In *Proceedings of the 4 th International Conference on Music Information Retrieval (ISMIR03)*, Baltimore, 25-32.

Downie, J. S. (2004). The creation of music query documents: framework and implications of the HUMIRS project. In Proceedings of the Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH), Göteborg. Available at http://www.hum.gu.se/allcach2004/AP/html/prop134.html

Foote, J. (1999). *Methods for the automatic analysis of music and audio*. Xerox Park Technical Report FXPAL-TR-99-038. Retrieved December 14, 2002, from http://citeseer.ist.psu.edu/foote99methods.html

Foote, J. (2000). ARTHUR: Retrieving orchestral music by long-term structure. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR00)*, Plymouth. Retrieved January 17, 2002, from http://ciir.cs.umass.edu/music2000

Foote, J., Cooper, M., Unjung Nam. (2002). Audio retrieval by rhythmic similarity. In *Proceedings of the 3rd International Symposium on Music Information Retrieval (ISMIR02)*, Paris, 265-266.

Futrelle, J., & Downie, J. S. (2002). Interdisciplinary communities and research issues in Music Information Retrieval. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR02)*, Paris, 215-221.

Gabrielsson, A., & Juslin, P. N. (2003). Emotional expression in music. In R. J. Davidson & H. H. Goldsmith & K. H. Scherer (Eds.), *Handbook of affective sciences*. New York: Oxford University press, 503-534.

Gabrielsson, A., & Lindström, S. (2001). The Influence of Musical Expression on Emotion in Music. In P. Juslin & J. Sloboda (Eds.), *Music and Emotion: Theory and Research*. New York: Oxford University Press, 223-248.

Gardner, H. (1993). Multiple Intelligences: The Theory in Practice. New York: Basic Books.

Gødoy, R., & Jorgensen, H. (2001). *Musical Imagery*. Lisse: Swets & Zeitlinger.

Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2003). RWC music database: music genre database and musical instrument sound database. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR03)*, Baltimore, 287-288.

Haitsma, J., & Kalker, T. (2002). A highly robust audio fingerprinting system. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR02)*, Paris, 107-115.

Handschuh, S., Staab, S., & Volz, R. (2003). On deep annotation. In *Proceedings of the International World Wide Web Conference (WWW03)*, Budapest. Retrieved February 12, 2004, from http://www2003.org/cdrom/papers/refereed/p273/p273_handschuh.html

Harb, H., & Chen, L. (2003). A query by example music retrieval algorithm. In Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS03), London, 122-128.

Haus, G., & Pollastri, E. (2001). An audio front end for query-by-humming systems. In *Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR01)*, Bloomington, 65-73.

Heylen, E. (2004). Een systematisch muziekwetenschappelijke, methodische en empirische studie naar tonaliteit. Unpublished MA thesis, Ghent University, Ghent.

Huron, D. (1999). *Music Research Using Humdrum: A user's guide*. Menlo Park, CA: Center for Computer Assisted Research in the Humanities.

Huron, D., & Aarden, B. (2002). *Cognitive issues and approaches in Music Information Retrieval*. Unpublished writing. Retrieved January 9, 2003, from http://dactyl.som.ohio-state.edu/Huron/Publications/huron.aarden.MIR.html

Jang, J.-S., & Lee, H.-R. (2001). Super MBox: an efficient/effective content-based music information retrieval system. In *Proceedings of the 9th ACM International Conference on Multimedia*, Ottawa.

Juslin, P. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. Journal of *Experimental Psychology: Human Perception and Performance, 26*, 1797-1813.

Juslin, P. N. (1997). Perceived emotional expression in synthesized performances of a short melody: capturing the listener's judgment policy. *Musicae Scientiae, 1*, 225-256.

Juslin, P. N., & Sloboda, J. A. (2001). *Music and Emotion: Theory and Research*. New York, Oxford: Oxford University Press.

Kalbach, J. (2002). Classifying emotion for information retrieval: three websites. *Notes, 59*(2), 408-411.

Kamenetsky, S., Hill, D., & Trehub, S. (1997). Effect of tempo and dynamics on the perception of emotion in music. *Psychology of Music, 25*(2), 149-160.

Kapur, A., Benning, M., & Tzanetakis, G. (2004). Query by BeatBoxing: Music information retrieval for the DJ. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR04)*, Barcelona, 170-177.

Kassler, M. (1966). Toward Musical Information Retrieval. *Perspectives of New Music, 4*(2), 59-67.

Kessler, A. (2004). Subjectivity, emotion and meaning in music perception. In *Proceedings of the Conference in Interdisciplinary Music (CIM04)*, Graz, CD-Rom.

Kim, H.-G., Burred, J., & Sikora, T. (2004). How efficient is MPEG-7 for general sound recognition. In *Proceedings of the Audio Engineering Society (AES04) International Conference*, London. Retrieved, March 16, 2005, from http://www.nue.tu-berlin.de/publications/papers/AES_MPEG-7.HGK.pdf

Kim, J.-Y., & Belkin, N. (2002). Categories of music description and search terms and phrases used by non-music experts. *In Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR02)*, Paris, 164-171.

Klapuri, A. (1997). *Automatic transcription of music*. Unpublished MA thesis, Tampere University of Technology, Tampere.

Kosugi, N., Nishihara, Y., Sakata, T. Yamamuro, M. and Kusima, K. (2000). A practical Query-By-Humming System for a Large Music Database. In *Proceedings of the 8 th ACM International Conference on Multimedia*, Marina del Rey, CA, US, 333-342.

Laban, R. (1963). *Modern Educational Dance*. London: MacDonald and Evans Ltd.

Lavy, M. (2000). Emotion and the experience of listening to music: components of a model. In *Proceedings of the Society for Education, Music and Psychology Research (SEMPRE) Conference on The Effects of Music*, Leicester. Retrieved March 22, 2002, from http://www.sempre.org.uk/papers.html?confID=19

Lee, J. H., & Downie, J. S. (2004). Survey of music information needs, uses and seeking behaviours: preliminary findings. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR04)*, Barcelona, 441-448.

Leman, M. (2002). Musical Audio Mining. In J. Meij (Ed.), *Dealing with the Data Flood: Mining Data, text and multimedia.* Rotterdam: SST Netherlands Study centre for Technology Trends.

Leman, M. (in press). *Being involved with music: A theory of embodied music cognition and mediation technology*. (Provisional book title).

Leman, M., & Camurri, A. (2004). Musical content processing for Interactive multimedia. In: *Proceedings of the Conference in Interdisciplinary Music (CIM04)*, Graz, CD-Rom

Leman, M., & Camurri, A. (in press). Understanding musical expressiveness using interactive multimedia platforms. *Musicæ Scientiæ, 10*.

Leman, M., Clarisse, L. P., De Baets, B., De Meyer, H., Lesaffre, M., Martens, G., Martens, J.-P., & Van Steelant, D. (2002). Tendencies, perspectives, and opportunities of musical audio-mining. In A. Calvo-Manzano, A.Pérez-Lopez, J. Salvador Santiago (Eds.), *Forum Acusticum Sevilla (FAS), Special Issue of the Journal Revista de Acústica XXXIII (3-4)*, Sevilla, CD-Rom.

Leman, M., Lesaffre, M., & Tanghe, K. (2001). Introduction to the IPEM Toolbox. In *Proceedings of the XIII Meeting of the FWO Research Society on Foundations of Music Research*, Ghent. Available at http://www.ipem.ugent.be/IPEMToolbox

Leman, M., Lesaffre, M., & Tanghe, K. (2001). An introduction to the IPEM Toolbox for Perception-Based Music Analysis. *Mikropolyphonie - The Online Contemporary Music Journal, 7*.

Leman, M., Lesaffre, M., & Tanghe, K. (2001). *IPEM Toolbox - An auditory toolbox for perception based musical analysis*. Manuscript, Ghent University. Available at http://www.ipem.ugent.be/IPEMtoolbox.

Leman, M., Vermeulen, V., De Voogdt, L., Moelants, D., & Lesaffre, M. (2005). Prediction of Musical Affect Attribution Using a Combination of Structural Cues Extracted from Musical Audio. *Journal of New Music Research, 34*(1), 39-67.

Leman, M., Vermeulen, V., De Voogdt, L., Taelman, J., Moelants, D., & Lesaffre, M. (2004). Correlation of gestural musical audio cues and perceived expressive qualities. In A. Camurri & G. Volpe (Eds.), *Gesture-based communication in human-computer interaction*. Berlin Heidelberg: Springer-Verlag, 40-54

Lesaffre, M., Leman, M., De Baets, B., & Martens, J.-P. (2004). Methodological Considerations Concerning Manual Annotation of Musical Audio in Function of Algorithm Development. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR04)*, Barcelona, 64-71.

Lesaffre, M., Leman, M., Tanghe, K., De Baets, B., De Meyer, H., & Martens, J.-P. (2003). User-dependent taxonomy of musical features as a conceptual framework for musical audio-mining technology. In *Proceedings of the Stockholm Music Acoustics Conference (SMAC03)*, Stockholm, 635-638.

Lesaffre, M., Moelants, D., & Leman, M. (2004). Spontaneous User Behavior in Vocal Queries for Audio-Mining. In W. B. Hewlett & E. Selfridge-Field (Eds.), *Music Query: Methods, Models, and User Studies. Computing in Musicology, 13*, 129-146.

Lesaffre, M., Moelants, D., Leman, M., De Baets, B., De Meyer, H., & Martens, J.-P. (2003). User Behavior in the Spontaneous Reproduction of Musical Pieces by Vocal Query. In *Proceedings of the 5th triennial European Society for the Cognitive sciences of Music conference (ESCOM)*, Hanover, 208-211.

Lesaffre, M., Tanghe, K., Martens, G., Moelants, D., Leman, M., De Baets, B., De Meyer, H., & Martens, J.-P. (2003). The MAMI Query-By-Voice Experiment: Collecting and annotating vocal queries for music information retrieval. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR03)*, Baltimore, 65-71.

Levitin, D. (1994). Absolute memory for musical pitch: evidence from the production of learned melodies. *Perception & Psychophysics, 56*, 414-423.

Lindsey, A. (1996). *Using contour as a mid-level representation of melody*. Unpublished M.S. thesis, Massachusetts Institute of Technology, Massachusetts.

Lu, L., You, H., & Zhang, H. J. (2001). A new approach to query by humming in music retrieval. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Tokyo. 1080-1083.

Martens, J.-P., & Van Immerseel, L. (1990). An auditory model based on the analysis of envelope patterns. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Albuquerque, New Mexico, 401-404.

Martin, K. D., & Kim, Y. (1998). Musical instrument identification: A pattern-recognition approach. In *Proceedings of the 136 th Meeting of the Acoustical Society of America (ASA)*, Seattle. Retrieved March 16, 2003, from http://sound.media.mit.edu/Papers/kdm-asa.pdf

Martin, K. D., Scheirer, E., & Vercoe, B. (1998). Music content analysis through models of audition. In *Proceedings of the 6 th ACM International Conference on Multimedia Publications*, Bristol. Retrieved March 16, 2003, from http://sound.media.mit.edu/Papers/ACMMM98.pdf

McNab, R., Smith, L. A., Witten, I. H., & Henderson, C. L. (2000). Tune Retrieval in the Multimedia Library. *Multimedia Tools and Applications, 10* (2/3), 113-132.

McNab, R., Smith, L. A., Witten, I. H., Henderson, C. L., & Cunningham, S. J. (1996). Towards the digital music library: Tune Retrieval from Acoustical Input. In *Proceedings of the 1st ACM International Conference on Digital Libraries*, Bethesda, US, 11-18.

Meek, C., & Birmingham, W. (2002). Johnny Can't Sing. In Proceedings of the International Conference on Music Information Retrieval (ISMIR02), Paris, 125-132.

Norman, D. (2002). Emotion and Design: attractive things work better. *Interactions, 9*(4), 36-42.

Notess, M. (2004). Three looks at users: a comparison of methods for studying digital library use. *Information Research, 9*(3). Available at http://InformationR.net/ir/9-3/paper177.

Notess, M., & Swan, M. (2004). Timeliner: Building a learning tool into a digital music library. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA)*, Lugano, 603-609.

Pachet, F. (2005). Knowledge management and musical metadata. In D. Schwartz (Ed.), *Encyclopedia of Knowledge Management*: Idea Group.

Pachet, F., & Cazaly, D. (2000). A Classification of Musical Genre. In *Proceedings of the International Conference Recherche d'Information Assistée par Ordinateur (RIAO)*, Paris. Available at http://www.csl.sony.fr/downloads/papers/2000/pachet-riao2000.pdf

Pachet, F., & Zils, A. (2003). Evolving automatically high-level music descriptors from acoustic signals. In *Proceedings of the International Symposium on Computer Music Modeling and Retrieval (CMMR)*, Montpellier, 42-53.

Pampalk, E., Dixon, S., & Widmer, G. (2003). Exploring music collections by browsing different views. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR03)*, Baltimore, 201-208.

Pardo, B., Meek, C., & Birmingham, W. (2003). Comparing aural music information retrieval systems. In *The MIR/MDL Evaluation Project White Paper Collection, 2nd ed*., 39-41.

Pauws, S. (2002). CubyHum: A fully operational query by humming system. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR02)*, Paris, 187-195.

Peretz, I. (1998). Music and emotion: Perceptual determinants, immediacy, and isolation after brain damage. *Cognition, 68*(2), 111-141.

Pickens, J. (2000). A comparison of language modeling and probabilistic text information retrieval approaches to monophonic music retrieval. In *Proceedings of the International Symposium on 1st Music Information Retrieval (ISMIR00)*, Plymouth. Retrieved January 17, 2002, from http://ciir.cs.umass.edu/music2000.

Pickens, J. (2001). *A survey of feature selection techniques for Music Information Retrieval*. Technical Report, Center for Intelligent Information Retrieval, Department of Computer Science, University of Massachusetts.

Prechelt, L., & Typke, R. (2001). An interface for melody input. *ACM Transactions on Computer-Human Interaction, 8*(2), 133-149.

Puckette, M. S. (no date). *Pure Data.* Retrieved July 31, 2003, from http://www.pure-data.org

Reiss, J., & Sandler, M. (2004). Audio issues in MIR evaluation. In Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR04), Barcelona, 28-33.

Russell, J. (2003). Core affect and the psychological construction of emotion. *Psychology Review, 110*(1), 145-172.

Scherer, K. H., & Zentner, M. R. (2001). Emotional Affects of Music: Production Rules. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and Emotion: Theory and Research*. New York: Oxford University Press. 361-392.

Scherer, K. R., & Oshinsky, J. S. (1977). Cue utilization in emotion attribution from auditory stimuli. *Motivation and Emotion, 1*, 331-346.

Selfridge-Field, E. (2000). What motivates a musical query. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR00)*, Plymouth. Retrieved January 17, 2002, from http://ciir.cs.umass.edu/music2000.

Shalev-Shwartz, S., Dubnov, S., Friedman, N., & Singer, Y. (2002). Robust temporal and spectral modeling for query by melody. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, 331-338.

Shepherd, J., & Wicke, P. (1997). *Music and Cultural Theory*. Camebridge: Camebridge Polity Press.

Sorsa, T., & Halonen, K. (2002). Mobile melody recognition system with voice-only user interface. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR02)*, Paris, 279-280.

Swearingen, K., & Sinha, R. (2002). Interaction Design for Recommender Systems. In *Proceedings of the Designing Interactive Systems (DIS2002) conference*, London. 321-329.

Tanghe, K., Lesaffre, M., Degroeve, S., Leman, M., De Baets, B., & Martens, J.-P. (2005). Collecting ground truth annotations for drum detection in polyphonic music. *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR05)*, London, 50-57.

Tanzi, D. (2003). Musical experience and online communication. *Crossings, 3*(1). Available at http://crossings.tcd.ie/issues/3.1/Tanzi.

Thaut, M., & Davis, W. (1993). The influence of subject-selected versus experimenter-chosen music on affect, anxiety, and relaxation. *Journal of Music Therapy, 30*, 210-223.

Toiviainen, P., & Krumhansl, C. (2003). Measuring and modeling real-time responses to music: the dynamics of tonality induction. *Perception, 32*(6), 741-766.

Turnbull, D. (2005). *Automatic Music Annotation*. Research Exam, Department of Computer Science and Engineering, UC San Diego. Retrieved June 16, 2005, from http://www.cse.ucsd.edu/~elkan/254/Turnbull.pdf.

Typke, R., Wiering, F., & Veltcamp, R. C. (2005). A survey of music information retrieval systems. *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR05)*, London, 153-160.

Tzanetakis, G., & Cook, P. (2000). Experiments in computer-assisted annotation of audio. In *Proceedings of the International Conference on Auditory Display (ICAD)*, Atlanta, US, 111-116.

Tzanetakis, G., Ermolynskyi, A., & Cook, P. (2002). Beyond the query-by-example paradigm: new query interfaces for Music Information Retrieval. In *Proceedings of International Computer Music Conference,* Göteborg. Retrieved, March 12, 2003, from http://www.cs.cmu.edu/~gtzan/work/pubs/icmc02gtzan.pdf

Uitdenbogerd, A., & R., v. S. (2002). A review of ractors affecting music recommender success. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR02)*, Paris, 204-208.

Vinet, H. (2003). The representation levels of Music Information. In *Proceedings of the International Symposium on Computer Music Modeling and Retrieval (CMMR)*, Montpellier, 193-209.

Wedin, L. (1972). Multidemensional study of perceptual-emotional qualities in music. *Swedish Journal of Musicology, 13*, 241-257.

Willemze, T. (1975). *Muziek Lexicon*. Utrecht: Spectrum.

Williams, C. (2001). Does it really matter? Young people and popular music. *Popular Music, 20*(2), 232-242.

Yang, D., & Lee, W. (2004). Disambiguating Music Emotion Using Software Agents. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR04)*, Barcelona, 52-58.

Zannos, I. (1999). Music and signs - semiotic and cognitive studies in music. Bratislava: ASKO Art & Science.

Zhang, P., & Li, N. (2005). The importance of affective quality. *Communications of the ACM (CACM), 48(9)*, 105-108.