From music similarity to music recommendation: Computational approaches based on audio features and metadata

Dmitry Bogdanov

TESI DOCTORAL UPF / 2013

Director de la tesi:

Dr. Xavier Serra i Casals
Dept. of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona, Spain

Dissertation submitted to the Deptartment of Information and Communication Technologies of Universitat Pompeu Fabra in partial fulfillment of the requirements for the degree of

DOCTOR PER LA UNIVERSITAT POMPEU FABRA

Music Technology Group (http://mtg.upf.edu), Dept. of Information and Communication Technologies (http://www.upf.edu/dtic), Universitat Pompeu Fabra (http://www.upf.edu), Barcelona, Spain.

To my Family and Friends.

Acknowledgements

This research was conducted from October 2008 to June 2013 at the Music Technology Group in Barcelona, supervised by Xavier Serra and Perfecto Herrera.

I would like to thank a lot of people who helped me on my way. First of all, I thank Xavier Serra for giving me the opportunity to work and learn at MTG, and, in particular, to participate in the research projects which I feel passionate about. Secondly, I am very grateful to Perfecto Herrera. His invaluable and numerous ideas and suggestions helped me immensely to progress in my research work, develop my methodology, and provided me a lot of inspiration.

I am very grateful to my colleges who made inestimable contribution to the part of the present research: Joan Serrà for our fruitful collaboration on music similarity measures, Martín Haro, Emilia Gómez, Ferdinand Haro, and Anna Xambó for their original ideas, inspiration, and work on realization of the Music Avatar project. Their help and advices provided significant contribution on my way to complete this thesis.

Finally, I send best regards to other MTG and ex-MTG members whose company always meant for me high motivation, energy, great fun and extra knowledge to learn practically every day: Oscar Celma, Cyril Laurier, Nicolas Wack, Mohamed Sordo, Owen Meyers, Piotr Holonowicz, Agustín Martorell, Waldo Nogueira, Justin Salamon, Gerard Roma, Enric Guaus, Ricard Marxer, José Zapata, and all others. It was a big pleasure to work with all of you!

This thesis has been carried out at the Music Technology Group of Universtitat Pompeu Fabra (UPF) in Barcelona, Spain from Oct. 2008 to Jun. 2012. This work has been supported by the FI Grant of Generalitat de Catalunya (AGAUR) and the PHAROS (IST-2006-045035) Buscamedia (CEN-20091026), Classical Planet (TSI- 070100-2009-407, MITYC), and DRIMS (TIN2009-14247-C02-01, MICINN) projects, and the R+D+I scholarship from UPF.

Abstract

The amount of music available digitally is overwhelmingly increasing. Vast amounts of music are available for listeners, and require automatic organization and filtering. In this context, user modeling, which consists in customization and adaptation of systems to the user's specific needs, is a challenging fundamental problem. A number of music applications are grounded on user modeling to provide users a personalized experience. In the present work we focus on user modeling for music recommendation, and propose a preference elicitation technique in conjunction with different recommendation approaches. We develop algorithms for computational understanding and visualization of music preferences. Our approaches employ algorithms from the fields of signal processing, information retrieval, machine learning, and are grounded in cross-disciplinary research on user behavior and music.

Firstly, we consider a number of preference elicitation strategies, and propose a user model starting from an explicit set of music tracks provided by this user as evidence of his/her preferences. The proposed strategy provides a noise-free representation of music preferences. Secondly, we study approaches to music similarity, working solely on audio content. We propose a novel semantic measure which benefits from automatically inferred high-level description of music. Moreover, we complement it with low-level timbral, temporal, and tonal information and propose a hybrid measure. The proposed measures show significant improvement, compared to common music similarity measures, in objective and subjective evaluations.

Thirdly, we propose distance-based and probabilistic recommendation approaches working with explicitly given preference examples. Both content-based and metadata-based approaches are considered. The proposed methods employ semantic and hybrid similarity measures as well as they build semantic probabilistic model of music preference. Further filtering by metadata is proposed to improve results of purely content-based recommenders. Moreover, we propose a lightweight approach working exclusively on editorial metadata. Human evaluations show that our approaches are well-suited for music discovery in the long tail, and are competitive with metadata-based industrial systems.

Fourthly, to provide insights on the nature of music preferences, we create regression models explaining music preferences of our participants and demonstrate important predictors of their preference from both acoustical and semantic perspectives. The obtained results correlate with existing research on music cognition. Finally, we demonstrate a preference visualization approach which allows to enhance user experience in recommender systems.

Resum

La quantitat de música disponible en format digital creix de forma aclaparadora. Hi ha grans quantitats de música disponibles per als oients, i per tractar-les es requereix organització i filtratge automàtic. En aquest context, el modelatge d'usuari, que consisteix en la personalització i adaptació dels sistemes a les necessitats específiques de l'usuari, és un problema fonamental i complex. Vàries aplicacions musicals es basen en aquest modelatge per proporcionar als usuaris una experiència personalitzada. En el present treball ens centrem en el modelatge d'usuari per la tasca de recomanació musical, proposem una tècnica d'inferència de preferències així com diferents mètodes de recomanació. A més, desenvolupem algoritmes per la comprensió automàtica i visualització de preferències musicals. Els nostres mètodes fan servir algoritmes de processament de senyals, recuperació d'informació, aprenentatge automàtic, i es basen en la investigació interdisciplinària en comportament de l'usuari i música.

En primer lloc, es consideren diverses tècniques d'inferència de preferències i es proposa un model d'usuari. El model és construït a partir d'un conjunt explícit de peces musicals proporcionades per l'usuari com a evidència de les seves preferències. Aquesta estratègia permet una representació fiable de les preferències musicals.

En segon lloc, s'estudien mètodes d'estimació de similitud musical, treballant exclusivament en el contingut d'àudio. Per una banda, es proposa una nova mètrica semàntica que es beneficia de la descripció musical d'alt nivell (etiquetes de gènere, emoció, instrumentació) extreta automàticament. A més, es complementa amb informació de baix nivell (tímbrica, temporal i tonal) per proposar una mètrica híbrida. Les mètriques proposades mostren una millora significativa en avaluacions objectives i subjectives, comparades amb els mètodes més comuns de similitud musical.

En tercer lloc, es proposen diversos mètodes de recomanació musical que funcionen a partir d'exemples explícits de preferència. Aquests mètodes estan basats en l'anàlisi del contingut d'àudio i metadades. Els recomenadors proposats utilitzen les mètriques de similitud proposades o models probabilístics. Per una banda, es fan servir mesures de similitud semàntica i híbrida, i alternativament es construeix un model probabilístic semàntic de preferència musical. Es proposa un filtratge addicional basat en metadades per millorar els resultats dels recomanadors basats en contingut d'àudio. D'altra banda, es proposa un mètode senzill, basat exclusivament en metadades editorials. Les avaluacions amb usuaris demostren que els nostres mètodes són molt adequats per al descobriment de la música, i són competitius amb relació als estàndards actuals.

En quart lloc, per proporcionar informació sobre la naturalesa de les preferències musicals, es creen models de regressió que relacionen les preferències musicals dels nostres participants amb els descriptors utilitzats. Per cada usuari, es presenten els predictors de preferència rellevants a nivell acústic i semàntic. Els resultats obtinguts estan molt relacionats amb els de la investigació existent en cognició musical. Finalment, es presenta un mètode de visualització de preferències que permet millorar l'experiència d'usuari en sistemes de recomanació.

Preface

I am always fascinated by music and its power to affect us, human beings. Being a music lover, a collector, an occasional disc-jockey, and musician, I always seek for new music which can impress me profoundly. Developing my own preferences, I always wondered how music can have an impact on our emotions, behavior, and way of life. Having this personal interest in music discovery, I entered MTG to work on the theme of music recommendation and user modeling. Back in 2008 this topic was only starting to gather interest in the research community. While user models and recommender systems for other domains were considerably developed, there was a lack of research in the music field. Attacking a problem of music recommendation, one has to deal with a lot of factors which are located on the intersection of different disciplines: information theory, signal processing, data mining and information retrieval together with cognitive sciences and sociology. Working with human factors implies additional difficulties and a desperate need for subjective evaluations.

This thesis deals with the problems of music recommendation, music similarity, and music preference elicitation, and challenges the goal to facilitate music discovery for the listeners. In particular, it focuses on bridging the current performance gap between approaches working with audio content information and the state-of-the-art approaches working with metadata. Furthermore, it focuses on how music preferences of listeners can be explored by audio analysis, and how recommender systems of the future can be enhanced by this knowledge. The outcomes of the present research have been published in a number of international peer-reviewed conferences and journals. The proposed approaches have been proved to be competitive to the best state-of-the-art approaches, in particular, in several international evaluation campaigns. Moreover, part of this research has been incorporated into a commercial music recommendation service and has been featured in the media. Till this date, a number of music recommender systems exists, but we are still on the long way to truly understand the underlying structure of music preferences. I hope that the findings of this thesis will help to pave the way for further research on the topic of music recommendation.

Contents

List of figures

List of tables

CHAPTER 1

Introduction

1.1 Motivation

We are living in the age of overwhelming information. The growth of digital technologies, the multimedia industry, and the Internet over the past decade facilitated physical access to information. We are now privileged to have more possibilities than ever to create, share, explore, and discover content on topics of our interest. However, the ease of access to information is burdened by the infinity of choices, large part of which is irrelevant for our particular needs and desires. Narrowing the scope to music, we encounter this problem in the present world of digital music distribution. We are already faced with the new paradigm of music consumption: listeners now have instantaneous access to digital music collections of an unprecedented size. The majority of music recordings are available online, and the amount of digital music counts tens of millions of tracks and grows extensively. Major Internet stores such as the iTunes Store contain up to 28 million tracks,[1] adding thousands of new tracks every month. Such amount of music is not surprising, as music takes an important part of people's everyday life, and more and more people express and share their music creativity by owning to the modern technology. A recent study found that British adults, when randomly probed via their mobile phones (North, 2004), are in presence of music in 39% of the cases. Listening to music have become the top leisure-time activity for most people (Rentfrow & Gosling, 2003, 2006). Music enthusiasts may now have a wider access to music than ever without any doubt. However they might be lost while searching for the "cream of the crop" in the non-stopping flow of new music content due to the infinity of choices.

With current technology we may expect better tools for search and discovery. There are numerous possibilities to be challenged in order to facilitate the access to music, and to bring discovery and interaction with music collections

[1] http://en.wikipedia.org/wiki/ITunesStore, retrieved on December 19, 2012.

on a whole new level. These challenges agitated the rise of music recommendation systems, which are yet far from being perfect, being limited in the type of information they are able to process and followed approaches, and in explanations given for specific recommendation. In particular, a very popular standard in music recommendation is collaborative filtering: *"if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue x than to have the opinion on x of a person chosen randomly."*[2] While this approach can be applied to music, it poses certain technical limitations in terms of maintaining a solid user base, gathering costly collaborative data, dealing with sparsity of such data and popularity bias. More importantly, it brings a number of conceptual limitations. Firstly, the listener's choices might be significantly reduced to an artificial "filter bubble" defined by social behavior patterns (Morozov, 2012; Pariser, 2012). Secondly, such a system does not require an understanding of the music preference itself, nor the music, but only captures the behavioral patterns of consumption for a certain population of users, meanwhile the tracked behavior could have been generated by different transient motives that would not reflect the inherent preferences of the customer: buying stuff for a relative, buying because of being enrolled in a course, etc.

Creating music recommenders which are able to provide listeners with more profound recommendations is a challenge. A recent survey of 500 US adults on their music habits (OMR, 2011) revealed that 54% of respondents have used music recommendation tools. Among those, 40% felt that recommendations are accurate only about half of the times or less. Among the respondents who actively searched for new music, 22% found it difficult or nearly impossible to discover new, unheard-of music that they like.

We believe that alternative approaches, which are able to generate musically informed decisions, should by explored in order to provide listeners with more profound recommendations. We foresee intelligent recommender systems of the future to have a deeper understanding of the underlying factors of music preference, i.e., to understand the listener. We want such systems to solve a number of complex problems: to understand the music itself (i.e., "to listen" to music audio) and the associated cultural context (i.e., to mine contextual information from the Internet), and to be able to compare music based on this knowledge. A lot of questions are to be resolved on the way to such systems, which requires a cross-disciplinary research on the edge of signal processing, data mining, and information retrieval together with music theory, cognitive sciences, and sociology. Music information retrieval (MIR) is a recently emergent field of research which addresses these questions to a great extent.

[2] http://en.wikipedia.org/wiki/Collaborative_filtering

1.2 Problems of music recommendation

At present, the majority of music discovery industrial systems provide means for manual search of music (Nanopoulos et al., 2009). This type of search is based on textual information about artist names, album or track titles, and additional semantic[3] properties which are mostly limited to genres. Music collections can be queried by tags or textual input using this information. However such search systems do not provide enough experience for the listener willing to explore and discover music.

From this perspective, a challenging task is to build a music recommender system, which is able to gather preference information about the users, predict their preference for music items (e.g., tracks, albums, or artists) with which they have not yet interacted within the system (i.e., expectedly novel items), and provide recommendations based on these predictions. Typically, the system should solve two problems formulated by Celma (2008):

- *Prediction.* Let $U = \{u_1, u_2, ..., u_m\}$ be the set of all users and $I = \{i_1, i_2, ..., i_n\}$ be the set of all possible items that can be recommended (a *target music collection*). Each user u_i expressed his/her interest in a list of items $I_{u_i} \subseteq I$. Compute function Pu_a, i_j which represents the predicted preference of item i_j for the active user u_a, such as $i_j \notin I_{u_i}$.

- *Recommendation.* Create a list of N items, $I_r \subset I, I_r \cap I_{u_i} = \varnothing$, that the user will like the most, i.e., with high Pu_a, i_j values.

However, it is necessary to solve another instrumental problem before prediction can be done. This problem is to obtain a user profile representing user interest in I_{u_i}, the henceforth referred as the problem of *preference elicitation.*[4] An open question is how to infer knowledge about the user and create a proxy representation of her/his music preferences in a way which is sufficient to provide successful recommendations. Getting a thorough and reliable user profile can imply considerable amount of user effort, including up-to-date explicit surveying and gathering user feedback. Alternatively, it is possible to address this problem with incomplete or partially reliable information, minimizing user effort by the cost of possible decrease in the quality of recommendations. Existing systems can obtain a user profile by monitoring user behavior, e.g., music consumption, listening statistics, or user ratings. The simplest user profile can be formed as a vector of ratings or playback counts for different artists, albums, and tracks, however, other representations can be considered. Starting from

[3] We use the term "semantic" to refer to the concepts that music listeners use to describe items within music collections, such as genres, moods, musical culture, instrumentation, etc., throughout this work following Amatriain (2005) and Aucouturier (2009).

[4] This procedure is commonly called as preference elicitation in literature on recommender systems. For consistency, in the rest of this thesis we will follow this terminology instead of other possible terms, such as "preference inference".

this data, different approaches can be taken to match the user profile to the target music collection and make the prediction.

Current systems provide basic means for music recommendation, which are not related to the audio content, i.e., using *metadata*[5] (Baltrunas & Amatriain, 2009; Celma, 2008; Firan et al., 2007; Jawaheer et al., 2010; Levy & Bosteels, 2010; Shardanand & Maes, 1995). However, it is not yet clear which type of metadata or their combination is the most beneficial. Addressing large-scale recommendation problems is of primary importance, and further research studies on comparison of metadata sources suitable for that will be of great interest for both academic and industrial communities. User ratings, listening behavior, manual annotations, social tags, or keywords extracted from the Internet, have their advantages and disadvantages. Manual expert annotations might be very accurate, but costly and even infeasible on large collections (Szymanski, 2009). In opposite, information extracted from the web is noisy. User ratings or listening statistics are hard to gather as they require a large user base, and, therefore, are expensive and undisclosable sources of information. This implies a *cold start problem*, i.e., lacking the metadata when new items are being added to music collection.

Many recommenders are biased towards popular items and often they leave long-tail items out of reach as long as they work with metadata. This is the common problem of *popularity bias* for the majority of metadata sources: essentially, popular music items will gather much more user-driven metadata (for example, user ratings) than unpopular items, which does not have enough exposure among the users of the system, and in general among listeners, to be able to generate that data. In practice, this implies the lack of user ratings, playback counts, social tags, or web-pages to extract keywords from, or even their total absence for music items in the so-called *"long tail"* (Celma, 2008). As a result, recommendations are biased towards popular items often leaving long-tail items out of scope. This might lead to another problem, specific for collaborative filtering systems: constantly incorporating user feedback, the system is at risk of perpetuating the situation known as "the rich gets richer". It may be thoroughly expected that the highlighted problems significantly limit music discovery experience.

Surely, solving problems of popularity bias and the long tail is a hot-topic of the research for the recommender systems community. *Audio content* analysis is advocated by MIR researchers as an alternative or a complement to metadata (Barrington et al., 2009; Casey et al., 2008; Celma, 2008). Recommender systems based on audio content are expected to reveal the long tail to listeners and, therefore, democratize and revolutionize music search breaking the popularity bias and enabling novel ways for querying and interacting with music collections. This challenge has been addressed by a number of researchers

[5] We pragmatically use the term "metadata" to refer to any information about or related to the music that is not extracted from the audio signal itself.

focused on content-based music recommendation for the last ten years. Until recently, the vast majority of academic efforts within MIR have ignored user-centric music recommendation, and instead have focused on the related tasks of automatic semantic annotation and music similarity, which were more easily addressable with the existing methodologies. The advances of MIR in other fields are to be properly integrated into the context of recommendation problem, and we expect that the existing content-based approaches have a large room for improvement in terms of user satisfaction. Currently, research studies evidence a detrimental performance of content-based approaches compared to recommenders working with metadata (Slaney, 2011). This might be explained by the fact that approaches typically extract low-level acoustic features from music audio, i.e., the features which are close to the signal, but far from the way listeners do conceptualize or think about music). In contrast, it might be desirable to work on a higher level of automatic description and utilize semantic concepts used by humans. It is challenging to try to bridge the so-called *semantic gap* (Aucouturier, 2009; Celma et al., 2006), which arises from the weak linking between human concepts related to musical aspects and the low-level features.

In general, data representations in both metadata approaches (for example, operating on very large vectors of user ratings) and content-based approaches (operating on vectors of acoustical features) might be relatively complex and incomprehensible for humans. However, transparency of recommendations, or the possibility of their justification for the user, is very important (Cramer et al., 2008; Sinha & Swearingen, 2002; Swearingen & Sinha, 2001). The employment of semantic concepts extracted from metadata or audio content can be an effective solution to this problem.

Before closing this section, let us recapitulate the main challenges in the design of music recommenders. Ideally, the system should be able to understand listeners, being adaptive to their music preferences, help them in music discovery by reducing popularity bias, facilitate the access to the long-tail items in music collections, and improve serendipity of recommendations, that is to increase the amount of novel and relevant of recommendations for a given user (Celma, 2008). Moreover, the independence from large datasets of user-generated data (ratings or tags), which are generally proprietary, is a great advantage to exploit in order to avoid the cold-start problem. We believe that the exploration of the factors which determine music preferences (such as acoustical or cultural properties) rather than the factors which are effects of music preferences (collaborative filtering data) is fundamental. Research related to this task will not only benefit the design of better recommender systems, but it will also contribute to the general understanding of music cognition. Currently, there is a lack of comprehensive research on these topics and those related or involving “users” or “listeners”. Evaluation methodologies followed by researchers are generally limited: they employ small music collections and simulate user-based evaluations due to the absence of real participants or due to their limited amount. In addition, a large part of research studies focus on evaluation of music similarity instead of music recommendation itself.

1.3 Our goal

In the present thesis, we consider a number of research questions related to the problems of music recommendation:

- How much metadata is relevant and what is the user satisfaction with the metadata-based approaches in terms of liking and novelty of recommendations?
- Can the content information, automatically retrieved from the raw audio signal, be effectively used for recommendation?
- What is the performance gap between content-based and metadata-based approaches in terms of user satisfaction?
- Can we bridge the gap between low-level features and human-level judgments about music, and how can the audio content provide valuable insights on the music preferences in an understandable form for humans?

To answer these questions, we aim for certain goals to address in the thesis:

1. Propose a noise-free preference elicitation strategy which is suitable to employ audio content information as well as metadata and which can be used as a ground for building an effective evaluation methodology. Explore how audio content-based information can innovate and improve approaches to modeling music listeners. In particular, we want to reduce the semantic gap between human judgments of music preferences and the low-level audio characteristics of music. To this end, we want to incorporate high-level semantic information to our content-based preference model.
2. Improve content-based approaches to music recommendation. Specifically, propose novel approaches to music similarity, and evaluate them in the context of music recommendation. Propose novel metadata-based and hybrid approaches, which are more prone to the cold-start problem and less costly in terms of required data. To this end, starting from the proposed preference elicitation strategy, we need to conduct a comprehensive user-based subjective evaluation of content-based and metadata-based approaches to music recommendation, focusing on liking and novelty of recommendations, i.e., music discovery.
3. Provide computational insights on the important factors of music preferences, analyzing audio content, and correlating the conclusions with the existing research on music cognition.

4. Explore additional applications of the content-based user models, in particular, music preference visualization.

1.4 Outline

This thesis is structured as follows: Chapter 2 reviews the foundations of music preferences by music cognition, psychology, and sociology, and systematizes the existing applied research on the topic of music recommendation and music similarity within the MIR community. Chapter 3 considers approaches to preference elicitation and proposes an explicit noise-free strategy to build a semantic user model by automatic inference of high-level concepts from the audio content. In Chapter 4 we focus on the problem of content-based music similarity. We propose and evaluate a novel semantic similarity measure together with a hybrid low-level/semantic approach. These measures allow for better music similarity estimation, according to the conducted objective and subjective evaluations. Chapter 5 considers different approaches to music recommendation, including content-based, metadata-based, and hybrid methods. We employ the proposed similarity measures in the context of recommendation, and study how their simple filtering by genre metadata can improve the performance. As our baselines, we use the state-of-the-art approaches, working by means of collaborative filtering and social tags. In addition, we propose our own approach working with editorial metadata. Chapter 6 studies how audio content information can be exploited to provide quantitative insights on the factors of music preferences from both acoustical and semantic perspectives. In Chapter 7 we demonstrate an approach to music preference visualization which takes advantage of the proposed semantic user model. Finally, Chapter 8 discusses open issues and concludes this thesis.

CHAPTER 2

Literature Review

2.1 Introduction

Recommender systems are active information filtering systems that attempt to present to the user information items (film, television, music, books, news, web pages) the user is interested in. In particular, they seek to predict the "rating" or "preference" that user would give to an item they had not yet considered. The term "preference" can be interpreted as an evaluative judgment in the sense of liking or disliking an object (Scherer, 2005) which is the most typical definition employed in psychology.

Music recommender systems are specifically focused on music items, and their main task is to propose to the user interesting music to discover, including unknown artists or particular tracks, based on the user's musical preferences (Celma, 2008). Different research disciplines are to be followed in order to create effective approaches to music recommendation. In particular, it is important to develop a solid understanding of the factors that influence music preferences. This is a complex problem, which can be be considered from the points of view of music perception, psychology, and sociology. Practical applications, such as music recommendation and automatic preference inference, will furthermore require knowledge from the fields of signal processing, information retrieval and machine learning. In general, we are absolutely sure of the necessity of an interdisciplinary approach to these problems. In this section we will review the existing studies on the foundations of music preferences, and approaches to user modeling for music recommendation. We will also recapitulate the existing approaches to the problem of measuring music similarity, using metadata and audio content, as these methods can serve as the basis for music recommendation. As well, we will highlight methodological problems of evaluation, currently faced by the researchers in the field of music recommendation.

2.2 Foundations of music preferences

Music is a highly rewarding stimulus for a typical listener (Menon & Levitin, 2005). A variety of studies have been conducted during the last three decades in attempt to develop theoretical models able to explain music preferences or at least provide insights on their driving factors. These studies highlighted a number of auditory, perceptual, psychological, and sociological factors influencing music choices by individuals.

2.2.1 Theoretical models

Leblanc (1982) developed the first integral model of music preferences in his "interactive theory of music preference". The structure of his multi-level model is presented in Figure 2.1. The lowest level includes physical properties and complexity of music stimulus, referential meaning of the stimulus, the quality of music performance, type of media that presents the music, social ties, and influence of authoritative figures. These factors, responsible both for musical environment and cultural environment, are interacting between themselves being the input information for the listener. Current context of the listener, including physiological condition, attention focus and affective state, will condition whether the listener will actually listen to the music or not. If the context conditions are fulfilled, the musical input is then "filtered" by the characteristics of the listener such as auditory sensitivity, musical ability and training, personality, sex, ethnic group and socio-economic status, maturation, and memory. All of these factors are processed in the listener's brain and contribute to a preference decision together with the fact of previous exposure (i.e., familiarity of the listener with the music). Remarkably, this model suggests interactions and relations between factors that can be investigated empirically, although the interaction pattern is too complicated to isolate influence of specific factors. Furthermore, the model does not address the reasons why people listen to music nor describes their selection process when confronted with a large amount of music pieces. These reasons impede a practical application of the model for music recommender systems.

Another model, presented in Figure 2.2, was proposed by Hargreaves et al. (2005). Its focus is not exactly on music preference, but on different kinds of responses that can be provoked in the listener due to interaction with the music qualities, the listener's characteristics and listening context. The responses to music can affect variables in the listener, the influence of listening context, and the perception of music over time. Music preference appears as a part of the effective response. The interaction between different variables and causal relations between them are poorly formulated in this model, which also makes it of lower value for practical application, apart from providing a general idea on the determinants of response to a specific musical stimulus at a given point in time.

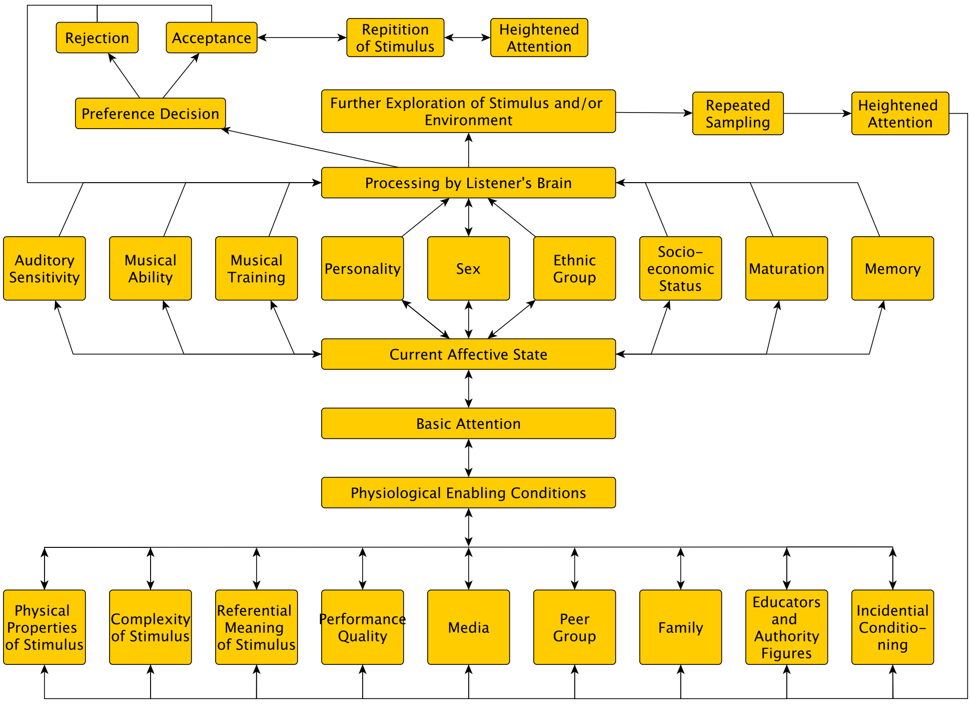


Figure 2.1: Interactive theory of music preference (Leblanc, 1982).

In addition to these two proposed theoretical models, a number of research works focus on studying particular factors of music preference. A comprehensive literature overview is provided in the dissertation by Schäfer (2008). In the following subsections we highlight the key factors suggested by existing research, which can be divided into four groups following Schäfer's work: *the music*, *the listener*, *the context*, and *the use of music*. For additional information, we refer interested reader to a number of literature reviews on the topic (McDermott, 2012; Uitdenbogerd & van Schyndel, 2002).

2.2.2 Facets of music preference

Factors related with the music.

Specific characteristics of music, such as loudness, tempo, pitch, timbre, harmonicity, melody, and complexity level are evidenced to be of fundamental importance for listeners in a number of statistical studies (Finnäs, 1989; McDermott, 2012; North & Hargreaves, 2008; Teo, 2003). Higher preference generally tends to be related with fast tempos and distinct rhythm, coherent melodies, absence of pronounced dissonances, and a moderate degree of complexity.[1]

[1] According to the studies working with Western listeners. These preferences may vary between different populations according to social and culturally-conditioned habits.

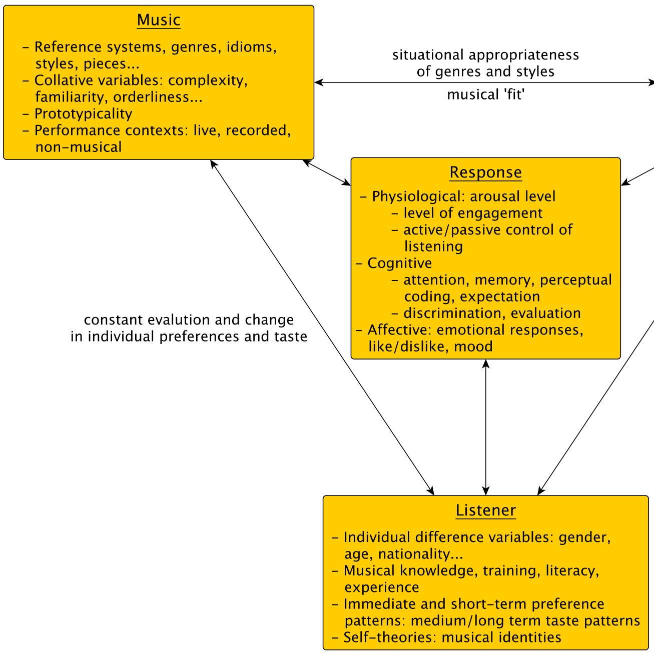

Music
- Reference systems, genres, idioms, styles, pieces...
- Collative variables: complexity, familiarity, orderliness...
- Prototypicality
- Performance contexts: live, recorded, non-musical
situational appropriateness of genres and styles
musical 'fit'
Situations and contexts
- Social and cultural contexts
- Everyday situations: work, leisure, consumer, education, health, media, entertainment
- Presence/absence of others
- Other ongoing activities
Response
- Physiological: arousal level
- level of engagement
- active/passive control of listening
- Cognitive
- attention, memory, perceptual coding, expectation
- discrimination, evaluation
- Affective: emotional responses, like/dislike, mood
constant evalution and change in individual preferences and taste
Individual use of music resource in different situations: goals in specific environments
Listener
- Individual difference variables: gender, age, nationality...
- Musical knowledge, training, literacy, experience
- Immediate and short-term preference patterns: medium/long term taste patterns
- Self-theories: musical identities

Figure 2.2: The reciprocal feedback model of musical response (Hargreaves et al., 2005).

Fast and lively tempo was found to be preferred by subjects across a wide range of ages when listening to a range of musical styles including classical, jazz, pop and folk (Teo, 2003). Preference of tempo has been found to interact with other variables such as the ability to discriminate tempo, affective association of tempo, music styles, sub-division of beats and performance medium. In addition, music with well-defined rhythm and clear regular meter, with an unchanging pulse easy to detect, was found to be preferred to those with unmarked and irregular rhythm. Music with moderate rhythm complexity was preferred to the one perceived as too simple or too complex.

Regarding pitch, the studies revealed that preference of pitch correlates significantly with the ability to discriminate it. Intensity of pitch was found to be an important correlating factor. Timbre characterizes particular instrument sounds, which matter greatly to music listeners. Till now, there are no scientific conclusion about why people prefer particular instrument sounds, but it is likely that such a preference varies across genre and culture, and that individual differences are substantial (McDermott, 2012). Studies revealed preference for instrumental over vocal timbre, especially in the case of preference for the classical music or non-western traditional music, with an exception of pop music, for which, oppositely, listeners indicated higher preference for vocal timbre.

In respect to musical harmony, a clear preference of consonant chords over dissonant ones is found, at least, for western listeners (McDermott, 2012) independently of the instrument used. An underlying perceptual explanation is that harmonic frequencies are supposed to be preferred over inharmonic ones due to their resemblance to single tones and a lower amount of beating phenomena, preponderant for the dissonant chords. Subjects on average prefer harmonic over inharmonic spectra and stimuli without beating over those with beats. The observed higher preference for consonance can be explained by long-term exposure to music of that kind, and is culture-dependent. Analyzing melodic redundancy (frequency at which the notes in a melody are repeated), its low and intermediate levels together with moderate amount of different pitches (pentatonic or diatonic) present in the stimuli are correlated with higher preference, while high level of redundancy and high number of pitches were correlated to lower preference of music piece (Teo, 2003).

Apart from purely auditory/perceptual factors, music can invoke referential meaning (McDermott, 2012). For example, melody, harmony, rhythm and mode are usually associated with certain ideas and emotional content by listener (Koelsch et al., 2004), and these associations can further affect appreciation of aesthetic value of music (Finnäs, 1989). It is believed that emotion content of music is one of the main reasons why people listen to it, and that typically listeners identify the emotion that a piece of music was intended to convey. Moreover, there is a tendency in preference of music that induce emotions over those that do not. Both emotional content of the music and emotional reaction of the listener affect music preferences. Listeners report using music for mood regulation with the goal of altering current emotional state or, oppositely, enhancing it.

Complexity is another important factor of music preference. The idea behind is that musical stimuli, that are too simple or too complex for a listener, might not be aesthetically pleasing, while a moderate amount of complexity can provide greater appreciation[2] (McDermott, 2012; North & Hargreaves, 1995). More generally, the dependence of preference of musical stimuli from its complexity follows the inverted U-shape (Berlyne, 1974). Research studies measure complexity in terms of the number of chords, degree of syncopation, temporal correlation of melodic sequences, and human ratings. Naturally, the effect of complexity on preferences interacts with the musical training/expertise of a listener. Studies suggest that the people with a higher degree of musical ability tend to prefer more complex music. A level of familiarity with music also influences appreciation: preference of somewhat complicated music can be increased by repeated listening. Other evidence indicates that expertise reduces the influence of complexity on preference decisions in return for more importance of other aesthetic factors.

Rentfrow et al. (2011) decompose music preferences into 5 latent dimensions

[2] More generally, it is widely discussed in experimental aesthetics that the aesthetic response is related to complexity.

based on listeners' affective reactions to excerpts of music from a wide variety of genres: (1) Mellow (smooth and relaxing styles); (2) Unpretentious (sincere and rootsy music such as is often found in country and singer–songwriter genres); (3) Sophisticated (classical, operatic, world, and jazz); (4) Intense (loud, forceful, and energetic music); and (5) Contemporary (rhythmic and percussive music, such as is found in rap, funk, and acid jazz). Both acoustical (dense, distorted, electric, fast, instrumental, loud, and percussive) and psychologically oriented attributes (aggressive, complex, inspiring, intelligent, relaxing, romantic, and sad) contribute to these dimensions.

Finally, the already mentioned familiarity, or prior exposure to music, has a large influence on preference (Finnäs, 1989; McDermott, 2012; North & Hargreaves, 2008). Typically, a listener is inclined to like music heard before, and to dislike unknown one. In particular, familiarity can explain cultural differences in music preferences in the sense that a listener usually prefers the music from the culture he/she comes from. Still, particular tracks within the familiar culture and genre can be disliked by a listener upon first listen as well, but further appreciated with repeated listens. In general, across genres, familiar music pieces are liked more than unfamiliar ones, but repeated listening also increases liking of similar unfamiliar music.

Factors related to the listener

The listener's age seems to have high impact on music preferences, as the importance of music in life supposedly increases until adolescence and then decreases slowly over life-span. Holbrook & Schindler (1989) provides analysis of this correlation and shows that people tend to prefer music that they were exposed at their critical life period, culminating at the age around 23.5. Other studies suggest similar critical period between 20 and 25 years old. This effect might be explained by certain experiences in the individual's development, such as coping with problems, social activities, identifying with artists, that were formative for music preference. Furthermore, age is found to have a negative effect on music consumption (Chamorro-Premuzic et al., 2010).

Numerous studies have shown the existence of correlation between listener's personality and music preferences, although attempts of their systematization reveal some inconsistencies (Dunn et al., 2011). In general, standard personality trait assessments were found to correlate with self-reports on preferred music (Chamorro-Premuzic et al., 2010; Glasgow et al., 1985; Kemp, 1996; Pearson & Dollinger, 2004; Rentfrow & Gosling, 2003; Zweigenhaft, 2008), explicitly given by listeners, and with implicitly gathered statistics of listening behavior over time (Dunn et al., 2011). Therefore, we can think about music preferences as a reflection of the listener's personality, at least, in certain aspects.

A fundamental study on this topic, including large-scale experiments, was done by Rentfrow & Gosling (2003). Their model relates four music dimensions to personality traits. Music preferences of 3500 participants were measured by means of the proposed Short Test of Musical Preferences. In this test participants were asked to rate their preference toward 14 genres. The obtained ratings were mapped to a revealed 4-dimensional preference space that included Reflective and Complex (e.g., classical), Intense and Rebellious (e.g., rock), Upbeat and Conventional (e.g., pop), and Energetic and Rhythmic (e.g., rap) dimensions. Thereafter, personality dimensions (such as Openness), self-views (such as political orientation, lifestyle), and cognitive abilities (such as verbal IQ) were matched to these music dimensions. Specifically, personality dimensions were inferred from a number of psychological tests, including the so-called Big Five Inventory (Gosling et al., 2003). Experimental results showed that the Reflective and Complex dimension was positively related to Openness to New Experiences, self-perceived intelligence, verbal ability, and political liberalism and negatively related to social dominance orientation and athleticism. In turn, the Intense and Rebellious dimension was positively related to Openness to New Experiences, athleticism, self-perceived intelligence, and verbal ability. The Energetic and Rhythmic dimension was positively related to Extraversion, Agreeableness, blirtatiousness, liberalism, self-perceived attractiveness, and athleticism and negatively related to social dominance orientation and conservatism.

Furthermore, lifestyle information can be associated with music preferences. North & Hargreaves (2007a,b,c) conducted an extensive study assessing the correlation between musical preferences and lifestyle aspects on 2532 participants. To this end, participants provided information about their interpersonal relationships, living arrangements, moral and political beliefs, criminal behavior, media preferences, leisure interests, music usage, travel, personal finances, education, employment, health, drinking, and smoking. It is concluded that music preferences of participants provided a meaningful way of distinguishing different lifestyle choices. The authors observed a very broad dichotomy of music, media, and literature preferences and leisure interests, broadly dividing stimuli into intellectually promising (which are referred as "high-art") and intellectually undemanding (referred as "low-art"). The fans of "high-art" (such as classical music, opera) and "low-art" musical styles demonstrated a preference for other "high-art" and "low-art" media and literature preferences, and leisure interests, respectively. Concerning social conditions, "high-art" music was found to be associated with upper-middle/upper class, while "low-art" music was associated with lower-middle/lower class. Furthermore, liberal-conservative dichotomy was found. Unfortunately, we believe such a dichotomy of preferences and social extraction to be oversimplified, while more complex approaches for the analysis of correlation between lifestyle and music preferences are necessary. For additional information, we refer interested reader to Perkins (2008) and Schäfer (2008), who provided extensive reviews of existing research studies on the correlation between music preferences, identity, and lifestyle.

Finally, it is important to stress the role of musical training, which we already mentioned in the context of music complexity. Considering how individuals listen to music, Kemp (1996) highlights "objective-analytic" and "affective" listening strategies. The former includes objective or technical reactions and is followed by musically experienced listeners. The latter corresponds to more emotional, and generally more musically naïve, appreciation. These different strategies correlate with preference, which can be demonstrated by the fact that music experts and novices respond to different aspects of musical pieces, not just the same aspects at different levels of complexity.

Factors related to the listener's context

It is sometimes argued that music listening is a social activity, and, therefore, music preference can be seen as a social phenomenon (Lonsdale & North, 2011; North & Hargreaves, 2008). Existing research support the relation of social ties to music preference and an intended use of music for social expression of an individual (Finnäs, 1989; MacDonald et al., 2002; North & Hargreaves, 1999; Rentfrow & Gosling, 2006). This research suggest that individuals might see music preferences as important indicators of personality traits. Moreover, individuals use music preferences as their badges of identity helping to communicate their personality to others, and to associate themselves better with the desired social groups, especially in the case of adolescents (North & Hargreaves, 1999; Rentfrow & Gosling, 2003). In particular, in order to get an impression on the personality of another person, adolescents tend to talk about their music preferences more than on other topics (Rentfrow & Gosling, 2006).

Interestingly, personal choices of music are strongly influenced by opinions of other people (McDermott, 2012). For example, in the context of online music distribution systems, it was shown that users' ratings on tracks are highly dependent on the judgment of other users, even though the users are not familiar with each other. The conformity in the development of music preference can be explained by compliance. The reasoning behind is that people intend to belong to certain social groups, which share similar values, opinions, or activities. To be a member of such a group (e.g., a group of friends), individual can adapt his/her music preferences to the ones expressed by its members. Another possible explanation is related to "informational influence" (or "prestige effect") highlighted by North & Hargreaves (1999). This effect implies that people tend to form their own preference to unfamiliar music based upon judgments of others or contextual information about music (such as a description of the composer).

Besides the context of social groups, at a more broad scope, music preferences of the individual can be dependent from his/her cultural environment (MacDonald et al., 2002; McDermott, 2012; Schäfer, 2008). The cultural background and ethnicity of the listener may influence the perception of aesthetic quality of genres, styles, or particular music pieces. Finally, apart from long-term context, concrete listening situation, including factors such as ongoing activities, presence or absence of people, or location, has a great impact on music appreciation (Schäfer, 2008).

Factors related to the use of music

Humans also practice different uses of music to serve their needs, such as the ones related to cognitive, emotional, socio-cultural, and physiological functions (Schäfer, 2008; Schäfer & Sedlmeier, 2009). Music functions referring to social communication and self-reflection are found to be substantial. Music can be used by the individual to improve or modify social ties, e.g., get in contact with other people. Moreover, there is a strong evidence that music can be used for the individual's own mood altering or enhancement (Schäfer, 2008; Ter Bogt et al., 2010), or self-socialization, especially among adolescents, when the individual searches for reflection and possible alleviation of his/her life problems in music (Schwartz & Fouts, 2003). Rentfrow & Gosling (2003) speculates that understanding the functions of music, the individual benefit of listening to music, may be the key for understanding listening behavior of the individual. However, current experimental research is still away from definite conclusions on the role of the use of music, suggesting that preferences are much more complex (Schäfer & Sedlmeier, 2009), and that the short-term context of the listener should be given a high attention.

Interestingly, a recent study by Chamorro-Premuzic et al. (2010) concluded that in order to predict music consumption it is more relevant to focus on the reasons why individuals use music rather than individual difference factors in personality or demographics. The authors devised the Uses of Music Inventory in order to assess three distinct motives for using music: emotional use (mood inducing in listener), cognitive use (enjoyment from analysis of music in an intellectual or rational manner), and background use (enjoyment of music while being involved in other activity, such as working, studying or socializing). Results showed significant positive effects of all music uses factors onto music consumption. A similar study by Ter Bogt et al. (2010) suggested a Typology of Music Listeners based on level of involvement with music and four types of uses of music: mood enhancement, coping with problems, defining personal identity, and marking social identity. The emotional use of music was found to be the most popular among listeners with any degree of musical involvement.

2.2.3 Computational approaches

As we have seen, a large number of possible factors is suggested by research on the nature of music preferences. The conclusions, however, are not unambiguous, due to differences in sampled population, considered variables, and methodology of their measurement and consequent analysis. Further systematization efforts and large-scale experiments will be necessary from the researchers in the fields of music perception and music psychology in order to advance in understanding music preferences.

From the engineering point of view, natural questions may arise of how important all these factors are and how they can be better specified, measured, and integrated into music information systems. The design of such systems is conditioned by a compromise between what should be done from the theoretic point of view and what can be technically achieved. Reviewing the above-mentioned factors, some of them appear to be more technically difficult to exploit than others although they might be of more fundamental relevance.

The majority of factors related to the music can be addressed by computational musicology and music information retrieval disciplines. There exist algorithms for measuring objective acoustic properties of music, including the ones we highlighted as suggested by research on music perception (i.e., loudness, tempo, pitch, timbre, harmony, melody, and complexity). These algorithms allow to approximate different acoustical and musical properties, and they might provide a music preference model under the assumption that a listener is primally driven by acoustic properties of music. Still, they are yet far away from a fully automatic description of all melodic, rhythmic or harmonic characteristics, comparable to a description which an experienced listener, or a musicologist, can provide.

Factors related to semantic referential meaning evoked in a listener by music are much more complex and harder to assess. While it is important to address subjective referential meaning, the current state of the art solely addresses a much more simpler problem of automatic assessment of generic semantic concepts that can be associated with music, i.e., concepts of genres, styles, instrumentation, emotional content, and cultural context, which have a common sense within a certain large population of listeners. This itself constitutes a challenge as the existing approaches are yet far from being perfect. Researchers strive to employ both audio content analysis and web-mining for cultural metadata associated with music to assess such a common referential meaning (Bertin-Mahieux et al., 2010; Sordo, 2012).

Focusing on emotion, there is evidence of commonality in emotional response to music on the basic level (arousal/valence), and the existing approaches achieved success in predicting common generic mood categories directly from audio (Laurier, 2011; Yang & Chen, 2012). However, recognition of more detailed emotional categories is still a complicated task, specifically, due to their intrinsic subjectivity. Furthermore, there are types of referential meaning, which may be related to the individual's personal experiences, including past events, situations, and communication with other persons. In this case, applying computational approaches to measure such factors can be very problematic, if not unfeasible.

Factors related to the listener can be obtained via the human-computer interaction. Listener's musical abilities and listening experience, which in particular interact with factors of music complexity and familiarity, can be technically measured by explicit surveying or implicit monitoring of user behavior. Demographic factors are generally suggested by research on recommender systems, and can be used for music recommenders (Celma, 2010). For example, a common technique in marketing is to apply data mining approaches over large datasets of demographic data and cluster customers according to their patterns. In particular, listener's age can be integrated into music recommender systems following this principle. Personality traits are more difficult to integrate, as this would require passing psychological tests by the users of a system, which arouses privacy issues, meanwhile demographic data can be gathered, at least partially, in a more unobtrusive way by monitoring the users' online identities in information services and social networks. The social context of the listener can be retrieved in a similar way, and peer groups can be identified from the graph of relations in social networks. However, we may expect a higher noisiness of the data when mined from online identities rather than provided explicitly by the listener.

Addressing factors related to current listening situation and the uses of music seems even more complicated. While it is possible to monitor unobtrusively current location, time, and weather conditions (Baltrunas & Amatriain, 2009; Baur & Butz, 2009; Herrera et al., 2010; Kim et al., 2008a; Lee & Lee, 2008; Stober & Nürnberger, 2009), it is much more difficult to understand current activity and emotional state of the listener, and his/her uses of music, which will require explicit feedback of the user.

To conclude, it is currently problematic to measure a large part of the above-mentioned factors, specifically, user-related. Measuring these factors may require large user effort and is technically very complicated. In order to minimize user effort, unobtrusive implicit monitoring strategies are required, however they are expected to produce noisy data. Therefore, it is reasonable to focus on the factors related to the music itself as the ones that can be measured more easily. Developing computational approaches to automatic preference inference and music recommendation by means of music audio content and associated referential meanings is very challenging, especially if one is interested in using generic semantic concepts only. This problem has not yet received enough attention among researchers, and addressing it can contribute to both fundamental understanding of music preferences and to a technical improvement of recommender systems.

2.3 User modeling and music recommendation

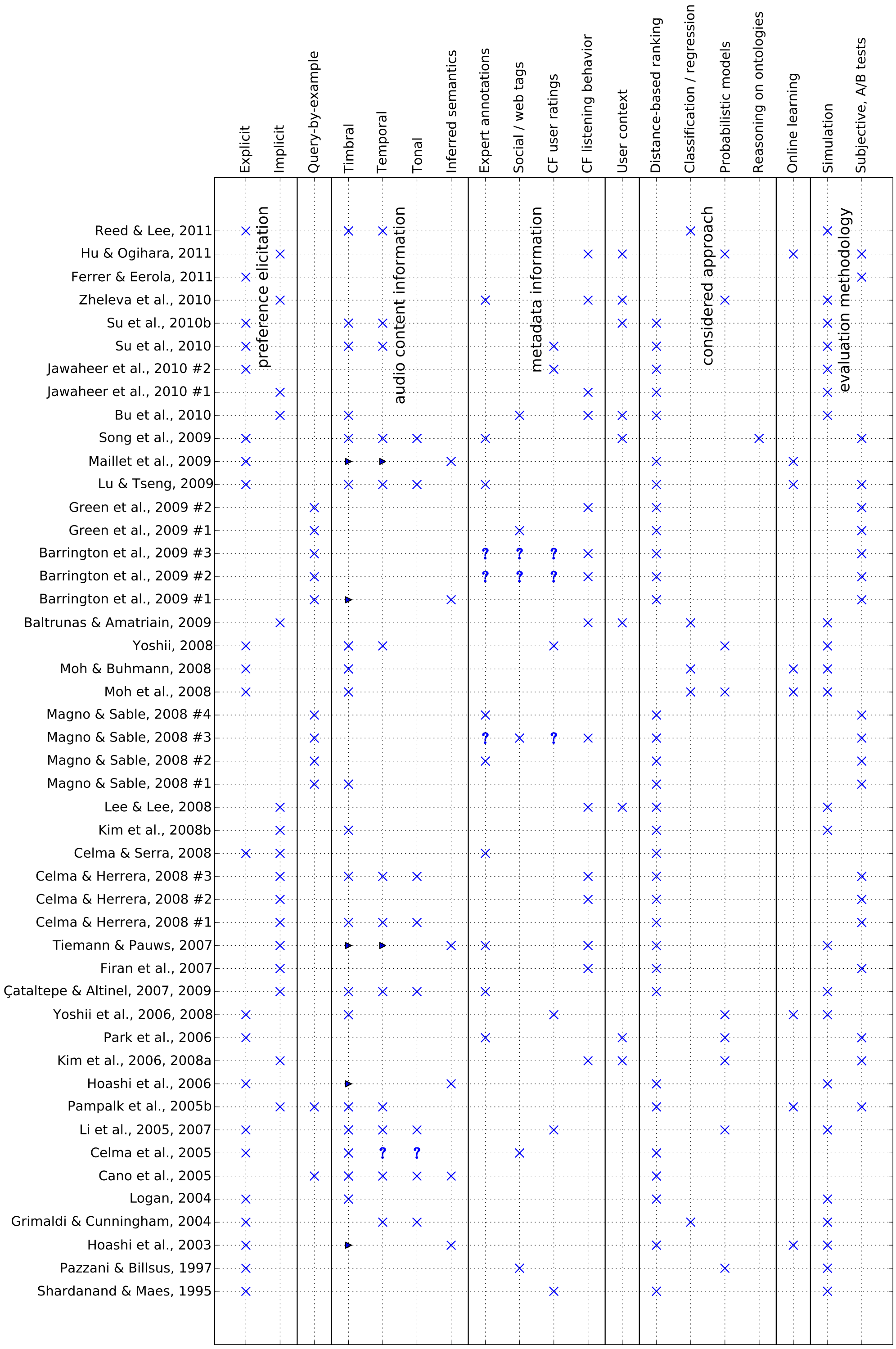
2.3.1 Approaches to music recommendation

There exist a considerable amount of works which address, directly or indirectly, the problem of recommending music content, a relatively new trend of research within MIR community. First and foremost, the existing studies related to this problem can be divided in two categories: the ones which address *music recommendation* and the ones which address *music similarity.* The former studies require understanding of the listener's preferences and have to operate with the measurements of user satisfaction with the provided recommendations as an evaluation criteria. The goal of the latter ones is not necessarily to provide quality recommendations, but to assess the aspects that can connect or relate two musical excerpts, which may have different applications apart from recommendation. As one may expect, the concept of music similarity is often employed as a basic tool to produce recommendations (Celma, 2008) following the idea that one might like the tracks similar in some aspects to the tracks he/she liked before. Development of more accurate similarity measures can presumably benefit the quality of music recommendations. However, the terms "similarity" and "recommendation" cannot be substituted, and a good performance of similarity measures does not necessarily equate to good recommendations (McNee et al., 2006). For this reason, we believe that the advances in measuring music similarity should be supported with proper evaluations in the context of recommendation, which is not always the case. In this section, we proceed by presenting research studies focused and evaluated in the context of music recommendation. A brief summary of related studies is presented in Figure 2.3. We also highlight approaches to music similarity working with metadata and review audio content-based similarity measures.

Preference elicitation strategies

The majority of studies on music recommendation present approaches which are grounded on certain former knowledge about the listener's preferences and predict preference for music items within a target music collection (e.g., music tracks or artists). In the extreme case, the only knowledge about music preferences can be a single music item explicitly given by the listener (the henceforth called as the *query-by-example use-case*) (Barrington et al., 2009; Cano et al., 2005; Green et al., 2009; Magno & Sable, 2008; Pampalk et al., 2005b). We can divide the existing studies with respect to *explicit* and *implicit* sources of knowledge about the listener's preferences (Hanani et al., 2001), that is, explicit and implicit preference elicitation strategies. Explicit information can be obtained by directly querying the user, meanwhile implicit information can be obtained from the usage of the system by the user, measuring the interaction with different items. Conceptually the difference in both types of feedback is that a numerical value of explicit feedback indicates preference, whereas the numerical value of implicit feedback indicates confidence in that the observed user interaction with an item can be associated with a higher preference of that item (Hu et al., 2008). Figure 2.4 summarizes a number of possible strategies by information sources. These strategies vary in amount of user effort and the granularity of captured information.

Explicit sources typically include:



	preference elicitation			audio content information				metadata information					considered approach					evaluation methodology	
	Explicit	Implicit	Query-by-example	Timbral	Temporal	Tonal	Inferred semantics	Expert annotations	Social / web tags	CF user ratings	CF listening behavior	User context	Distance-based ranking	Classification / regression	Probabilistic models	Reasoning on ontologies	Online learning	Simulation	Subjective, A/B tests
Reed & Lee, 2011	×			×	×									×				×	
Hu & Ogihara, 2011		×									×	×			×		×		×
Ferrer & Eerola, 2011	×																		×
Zheleva et al., 2010		×						×			×	×			×			×	
Su et al., 2010b	×			×	×							×	×					×	
Su et al., 2010	×			×	×					×			×					×	
Jawaheer et al., 2010 #2	×									×			×					×	
Jawaheer et al., 2010 #1		×									×		×					×	
Bu et al., 2010		×		×					×		×	×	×					×	
Song et al., 2009	×			×	×	×		×				×				×			×
Maillet et al., 2009	×			▸	▸		×						×				×		
Lu & Tseng, 2009	×			×	×	×		×					×				×		×
Green et al., 2009 #2			×								×		×						×
Green et al., 2009 #1			×						×				×						×
Barrington et al., 2009 #3			×					?	?	?	×		×						×
Barrington et al., 2009 #2			×					?	?	?	×		×						×
Barrington et al., 2009 #1			×	▸			×						×						×
Baltrunas & Amatriain, 2009		×									×	×		×				×	
Yoshii, 2008	×			×	×					×					×			×	
Moh & Buhmann, 2008	×			×										×			×	×	
Moh et al., 2008	×			×										×	×		×	×	
Magno & Sable, 2008 #4			×					×					×						×
Magno & Sable, 2008 #3			×					?	×	?	×		×						×
Magno & Sable, 2008 #2			×					×					×						×
Magno & Sable, 2008 #1			×	×									×						×
Lee & Lee, 2008		×									×	×	×					×	
Kim et al., 2008b		×		×									×					×	
Celma & Serra, 2008	×	×						×					×						
Celma & Herrera, 2008 #3		×		×	×	×					×		×						×
Celma & Herrera, 2008 #2		×									×		×						×
Celma & Herrera, 2008 #1		×		×	×	×							×						×
Tiemann & Pauws, 2007		×		▸	▸		×	×			×		×					×	
Firan et al., 2007		×									×		×						×
Çataltepe & Altinel, 2007, 2009		×		×	×	×		×					×					×	
Yoshii et al., 2006, 2008	×			×						×					×		×	×	
Park et al., 2006	×							×				×			×				×
Kim et al., 2006, 2008a		×									×	×			×				×
Hoashi et al., 2006	×			▸			×						×					×	
Pampalk et al., 2005b		×	×	×	×								×				×		×
Li et al., 2005, 2007	×			×	×	×				×					×			×	
Celma et al., 2005	×			×	?	?			×				×						
Cano et al., 2005			×	×	×	×	×						×						
Logan, 2004	×			×									×					×	
Grimaldi & Cunningham, 2004	×				×	×								×				×	
Hoashi et al., 2003	×			▸			×						×				×	×	
Pazzani & Billsus, 1997	×								×						×			×	
Shardanand & Maes, 1995	×									×			×					×	

Figure 2.3: Studies on music recommendation. Triangles mark the approaches which use low-level features indirectly as a source for high-level inference. Questions mark the publications with missing details (e.g., employing black-box approaches).

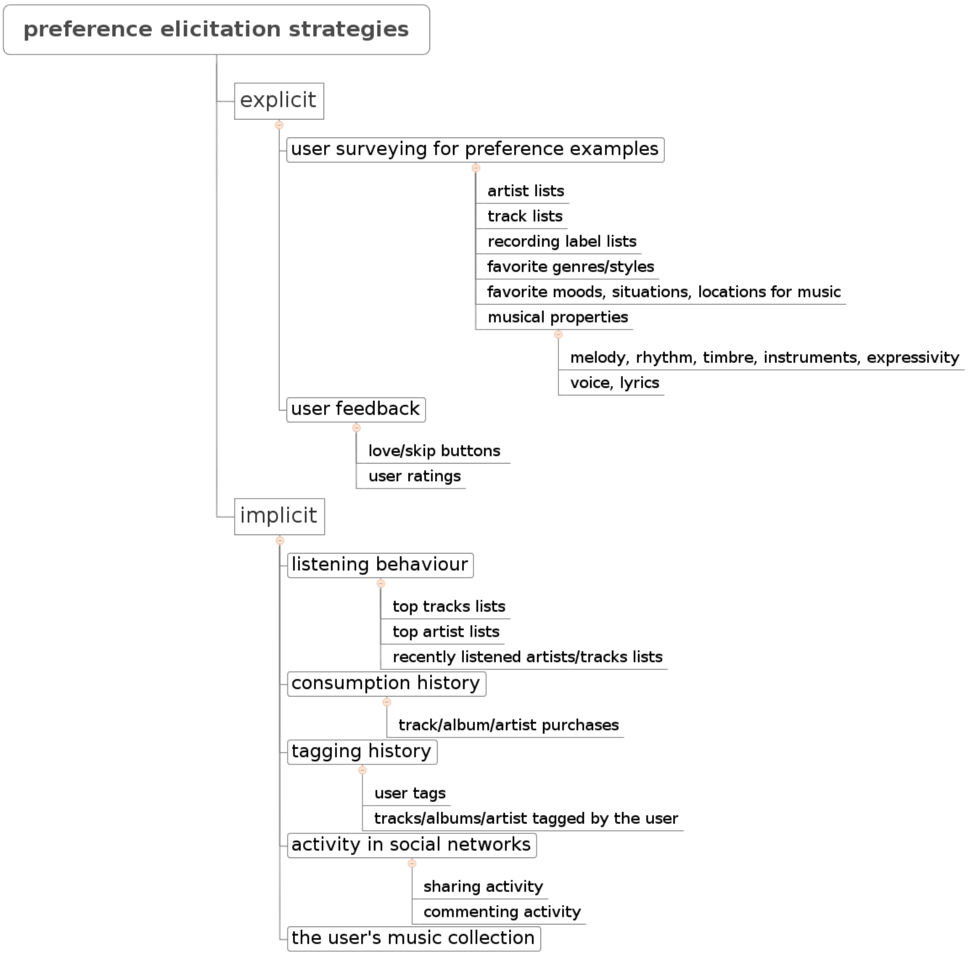

preference elicitation strategies
explicit
user surveying for preference examples
artist lists
track lists
recording label lists
favorite genres/styles
favorite moods, situations, locations for music
musical properties
melody, rhythm, timbre, instruments, expressivity
voice, lyrics
user feedback
love/skip buttons
user ratings
implicit
listening behaviour
top tracks lists
top artist lists
recently listened artists/tracks lists
consumption history
track/album/artist purchases
tagging history
user tags
tracks/albums/artist tagged by the user
activity in social networks
sharing activity
commenting activity
the user's music collection

Figure 2.4: Diagram of music preference elicitation strategies.

- artist ratings (Pazzani & Billsus, 1997; Reed & Lee, 2011; Shardanand & Maes, 1995),
- album ratings (Su et al., 2010a,b),
- track ratings (Grimaldi & Cunningham, 2004; Hoashi et al., 2003, 2006; Li et al., 2005, 2007; Lu & Tseng, 2009; Park et al., 2006; Yoshii, 2008; Yoshii et al., 2006, 2008),
- examples of liked/disliked tracks (Jawaheer et al., 2010; Logan, 2004; Moh & Buhmann, 2008; Moh et al., 2008; Song et al., 2009),
- examples of liked/disliked artists and genres (Ferrer & Eerola, 2011),
- user-specified keywords of interest (Celma et al., 2005; Celma & Serra, 2008; Maillet et al., 2009).

In turn, consumption statistics and listening behavior (e.g., track/artist playcounts with associated timestamps and skipping behavior gathered via a music player plugin) are usually used as an implicit source of information about music preferences (Baltrunas & Amatriain, 2009; Barrington et al., 2009; Bu et al., 2010; Celma & Herrera, 2008; Celma & Serra, 2008; Firan et al., 2007; Hu & Ogihara, 2011; Jawaheer et al., 2010; Kim et al., 2008a, 2006, 2008b; Lee & Lee, 2008; Pampalk et al., 2005b; Tiemann & Pauws, 2007; Zheleva et al., 2010; Çataltepe & Altinel, 2007, 2009).

From a common sense, implicit information is inherently noisy (Hu et al., 2008). For example, listening statistics based on playcounts for artists or tracks might not represent real preferences since they ignore the difference between track durations or users' activities when listening to the music (Jawaheer et al., 2010). Furthermore, low artist/track playcounts do not necessarily mean actual dislike of music, i.e., such information does not bring negative feedback. In contrast, explicit feedback is expected to be more reliable. However there was found that rating behavior can be inconsistent when rates are repeatedly asked for after some time lapses (Amatriain et al., 2009; Celma, 2010), at least for a group of users. Furthermore, ratings can be biased by the precision of a rating scale and by decisions on the design of the recommender interface (Cosley et al., 2003). D'Elia & Piccolo (2005) revealed two group of raters: thoughtful and instinctive, with the latter being possibly biased by the current context and therefore less consistent in their rating approach. The main problem of explicit strategies, however, is in the scarcity of data as users are not necessarily eager to provide ratings.

We might expect both types of feedback to be correlated to some extent. A study by Parra & Amatriain (2011) found such a correlation between user ratings and listening statistics. However, prediction of preference ratings from implicit data by linear regression resulted in less than 14% of explained variance. This user-based study employed 114 participants, who had accounts on *Last.fm*, comparing their listening behavior with 10122 album ratings gathered in an explicit survey. Another study be Jawaheer et al. (2010) compared explicit feedback (artist love/ban tags on *Last.fm*) to listening behavior and found no difference in the performance recommendation approaches, however the results were considered as inconclusive due to the size of the dataset. The dataset included 527 *Last.fm* users with associated 2167 artists and 8242 love tags. Baltrunas & Amatriain (2009) and Celma (2010) propose recoding listening behavior into preference ratings for music recommendation.

The choice of the strategy is often reasoned by the type of user data available to academic researchers, who often prefer working with existing datasets rather than employing real subjects for evaluation due to economic and time cost, not to mention methodological issues. As these data are typically limited being difficult to gather, researchers often have no opportunity to compare benefits of different strategies. To the best of our knowledge, the question of how a proper elicitation strategy can increase user satisfaction with music recommenders remains unexplored being yet out of focus in academic studies.

Information sources

The choice of available information about music is crucial in the design of any music retrieval system. We can divide research studies into three groups: approaches working with information extracted from *metadata* associated with music, approaches employing *audio content* and, finally, their hybrid combinations. Possible metadata sources include:

- manual expert annotations, e.g.,
  - editorial metadata, such as artist-track-album relations (Bu et al., 2010; Song et al., 2009) or artist relations (Celma & Serra, 2008),
  - genre (Magno & Sable, 2008; Park et al., 2006; Song et al., 2009; Tiemann & Pauws, 2007; Zheleva et al., 2010; Çataltepe & Altinel, 2007, 2009),
  - tempo, mood, and instrumentation (Lu & Tseng, 2009; Magno & Sable, 2008; Park et al., 2006; Song et al., 2009; Tiemann & Pauws, 2007);
- annotations automatically mined from the Internet, e.g.,
  - social tags (Magno & Sable, 2008),
  - keywords extracted from web-pages (Green et al., 2009; Pazzani & Billsus, 1997), and RSS feeds (Celma et al., 2005);
- collaborative filtering data generated by users, e.g.,
  - artist/track rating datasets (Jawaheer et al., 2010; Li et al., 2005, 2007; Shardanand & Maes, 1995; Su et al., 2010a; Yoshii, 2008; Yoshii et al., 2006, 2008),
  - listening behavior information, such as artist and track playcounts (Baltrunas & Amatriain, 2009; Barrington et al., 2009; Bu et al., 2010; Celma & Herrera, 2008; Ferrer & Eerola, 2011; Firan et al., 2007; Green et al., 2009; Jawaheer et al., 2010; Kim et al., 2008a, 2006; Lee & Lee, 2008; Magno & Sable, 2008; Tiemann & Pauws, 2007; Zheleva et al., 2010).

Among different types of metadata, collaborative filtering data is probably the most established as it can be successfully applied to virtually any domain of recommendation (e.g., video, image, text, or goods apart from music items). Similarly to recommender systems in other fields (Sarwar et al., 2001), collaborative filtering approaches can be applied for music recommendation. Shardanand & Maes (1995) proposes such an approach in their study, which may be considered as the earliest, to the best of our knowledge, study on automatic music recommendation. In addition, other types of metadata (expert annotations, social tags) are also very popular among researchers and industry. Although they might not be as efficient as collaborative filtering data according to some studies (Green et al., 2009), they can be used to extend or replace it. In general, using metadata can be an effective way to build music recommender systems when working with popular music items. However, there might be not enough of such data when the target music collection contains long-tail items which lack user ratings or listening behavior information, annotations by experts, or social tags, due to their unpopularity. Furthermore, new items introduced to the target collection will suffer from the cold-start problem, i.e., they will lack metadata until it is finally gathered from experts or built by users. Approaches working with information extracted from audio content challenge to solve this problem.

We highlight four types of content information which can be used by recommendation approaches:

- *timbral* information (Barrington et al., 2009; Cano et al., 2005; Celma & Herrera, 2008; Celma et al., 2005; Hoashi et al., 2003, 2006; Kim et al., 2008b; Li et al., 2005, 2007; Logan, 2004; Lu & Tseng, 2009; Magno & Sable, 2008; Maillet et al., 2009; Moh & Buhmann, 2008; Moh et al., 2008; Pampalk et al., 2005b; Reed & Lee, 2011; Su et al., 2010a,b; Tiemann & Pauws, 2007; Yoshii, 2008; Yoshii et al., 2006, 2008; Çataltepe & Altinel, 2007, 2009);

- *temporal* information, characterizing temporal evolution of loudness and timbral characteristics, dynamics, rhythmic properties, and musical structure (Cano et al., 2005; Celma & Herrera, 2008; Grimaldi & Cunningham, 2004; Li et al., 2005, 2007; Lu & Tseng, 2009; Maillet et al., 2009; Pampalk et al., 2005b; Reed & Lee, 2011; Song et al., 2009; Su et al., 2010a,b; Tiemann & Pauws, 2007; Yoshii, 2008; Çataltepe & Altinel, 2007, 2009);

- *tonal* information (Cano et al., 2005; Celma & Herrera, 2008; Grimaldi & Cunningham, 2004; Li et al., 2005, 2007; Lu & Tseng, 2009; Song et al., 2009; Çataltepe & Altinel, 2007, 2009);

- *inferred semantic* information, e.g., automatic genre classification (Cano et al., 2005), more extensive autotagging by genre, mood, instrumentation, and other categories using audio content (Barrington et al., 2009; Maillet et al., 2009), and unsupervised inference of clusters of music similar in the audio feature space (Hoashi et al., 2003, 2006; Tiemann & Pauws, 2007).

We will refer to the first three types as being low-level when compared to semantic information. The latter can be inferred from low-level information relying on annotated ground truth music collections or by means of unsupervised clustering. Low-level timbral, temporal, and tonal information can provide a solid ground for recommendation algorithms, addressing different aspects of music, but is rarely used altogether in academic studies till recently (see Figure 2.3), following the advances of music analysis tools. Instead, a large number of existing studies are patently incomplete as they are focused solely on timbre (most frequently, MFCCs, representing spectral envelope) ignoring other acoustical and musical aspects. In turn, high-level semantic information is very rarely employed although there is some evidence of its advantage in the existing research in music similarity (Barrington et al., 2007b; West & Lamere, 2007), and its usage is supported by the importance of referential meaning on music appreciation (Section 2.2.2).

For further improvement of the quality of recommendations, a number of studies are focused on hybrid approaches, merging both audio content and metadata, and report on advantage of such approaches over solely content-based or metadata-based ones (Su et al., 2010a; Tiemann & Pauws, 2007; Yoshii, 2008; Yoshii et al., 2006).

In addition to music information, a number of studies stress the importance of the *listener's context* (Section 2.2.2) and propose context-aware music recommenders (Baltrunas & Amatriain, 2009; Baur & Butz, 2009; Bu et al., 2010; Herrera et al., 2010; Hu & Ogihara, 2011; Kim et al., 2008a, 2006; Lee & Lee, 2008; Park et al., 2006; Song et al., 2009; Stober & Nürnberger, 2009; Su et al., 2010b; Zheleva et al., 2010). They employ information about current time (hour, morning/evening, day of week, weekend/working day) and weather (atmospheric conditions, temperature, humidity, month and season), location, physiological and emotional state of the listener (age, gender, pulse, health conditions, mood), her/his current activity (situation, event, weekend/working day) and social interaction (friendship relations, membership to online groups).

Computational approaches

Different computational approaches can be followed to address the problem of prediction of the listener's interest in music items and recommendation (see Section 1.2) given the information about his/her music preferences. In regard to the algorithms proposed in the literature, we can highlight:

- distance-based ranking, e.g.,
    - distance from preferred tracks, or a query-by-example, to the tracks in a target music collection (Cano et al., 2005; Celma & Herrera, 2008; Celma et al., 2005; Celma & Serra, 2008; Logan, 2004; Magno & Sable, 2008),
    - user-to-user collaborative filtering distances (Bu et al., 2010; Lu & Tseng, 2009);

- discriminative models, e.g.,
  - classification into liked/disliked music, based on support vector machines (SVMs) (Moh & Buhmann, 2008; Moh et al., 2008) or k-nearest neighbor algorithm (kNN) (Grimaldi & Cunningham, 2004),
  - ordinal regression for user rating prediction (Reed & Lee, 2011);
- probabilistic generative models, e.g.,
  - Gaussian mixture models (GMMs) (Hu & Ogihara, 2011; Li et al., 2005, 2007; Moh et al., 2008),
  - Bayesian networks (Park et al., 2006; Pazzani & Billsus, 1997; Yoshii et al., 2006, 2008),
  - hidden Markov models (HMM) (Kim et al., 2008a),
  - latent Dirichlet allocation (LDA) (Zheleva et al., 2010);
- automated reasoning on ontologies (Song et al., 2009).

As we can see, a large amount of approaches are based on similarity estimation between items in a target collection and a representation of the listener's music taste, which can be a set of preferred items or clusters of items. In particular, timbral distances can be implemented with a basic idea of comparing spectral shapes of the tracks, which can be represented as probability distributions of the frame-wise Mel-frequency cepstrum coefficients (MFCCs):

- a simple Euclidean distance between vectors of means and variances of MFCCs in each distribution (Tzanetakis & Cook, 2002),
- the Earth Mover's Distance comparing GMMs of MFCCs, trained with k-means clustering (Logan & Salomon, 2001),
- distance based on Monte-Carlo sampling comparing GMMs, trained with the Expectation-Maximization algorithm initialized with k-means clustering (Aucouturier et al., 2005).

Magno & Sable (2008) compared these approaches together with a simple metadata baseline, searching tracks of the same genre label, in a query-by-example scenario of music recommendation. The approach based on Monte-Carlo sampling was found to perform best among the three timbral distances. Recommendations driven by these approaches hardly surpassed mean rating of 3 (on a 1-to-5 Likert-type scale) representing only average user satisfaction ($\approx$ 3.03 for the best approach, 13 participants). Interestingly, recommendations based on commercial metadata-based black-box recommenders (*Pandora* and *Last.fm*) achieved only marginally better results (up to $\approx$ 3.22), that is, no statistically significant difference was found comparing their performance with the best of the content-based approaches. In turn, Logan (2004) considered distances to a user profile and, specifically, considered average, median, and minimum distance from tracks in a target music collection to the preferred tracks. Alternatively, she proposed to compute distance to a summarized MFCC distribution of all preferred tracks (a simplified and noisier representation of user preferences).

There are studies that implement more acoustic features, other than the standard MFCCs, expanding towards temporal and tonal dimensions of music, and complementing it with metadata. Pampalk et al. (2005b) expanded timbral similarity between GMMs of MFCCs (Aucouturier et al., 2005) with temporal information, which included fluctuation patterns and derived "focus" (distinctiveness of the fluctuations at specific frequencies) and "gravity" (the overall perceived tempo) descriptors (Pampalk et al., 2005a). The proposed approach generates playlists starting from a query-by-example and incorporating skipping behavior feedback during the playback by the listener. Two sets of liked and disliked tracks are gradually formed, and a distance to them is used as a criterion for adding new tracks (recommendations) to the playlist.

Celma & Herrera (2008) proposed such an approach based on an Euclidean distance, which utilizes such an expanded set of features, describing dynamics, tempo, meter, rhythmic patterns, tonal strength, key and mode information (Cano et al., 2005). Artist-level recommendations are generated starting from a set of favorite artists, inferred according to the listening behavior (artist playcounts from *Last.fm* for 288 users). This approach is compared to an item-based collaborative filtering distance using listening statistics from Last.fm, and a hybrid approach combining both measures is proposed. A large-scale evaluation on real participants was conducted in this study, and it suggested that collaborative filtering approach scores higher than hybrid and content-based ones in liking, however producing more familiar recommendations (28.3% vs 21.7% and 19%). Importantly, this study corroborates that content-based approaches can be effectively incorporated in order to increase novelty of recommendations without a devastating decrease in their quality. In addition, the effect of familiarity with recommendations is shown to correlate significantly with appreciation, with a more familiar music being rated higher in liking (which corroborates familiarity factor of preference presented in Section 2.2.2). Again average liking ratings were only satisfactory on average (≈ 3.39 and ≈ 2.87 for collaborative filtering and content-based approaches, respectively, on a 1-to-5 Likert-type liking scale).

Bu et al. (2010) proposed to compute a hybrid distance from the hypergraph, which combines timbral similarities between tracks (the above-mentioned Earth Mover's Distance between GMMs of MFCCs), user similarities according to collaborative filtering of listening behavior from *Last.fm*, and similarities on the graph of *Last.fm* users, groups, tags, tracks, albums, and artists (all possible interactions crawled from the *Last.fm* web-pages). Therefore, this approach employs social interaction between users together with editorial metadata (relations between artists, albums and tracks). The proposed approach was compared with user-based collaborative filtering, a content-based timbral approach, and their hybrid combination, on a listening behavior dataset. Again, the performance of a timbral approach fell behind the ones working with metadata, while incorporation of social information and editorial metadata showed the best results.

Lu & Tseng (2009) proposed a recommendation approach based on hybrid combination of three rankings:

- a weighted Manhattan distance over audio features including pitch, tempo, rhythmic speed (number of pitches per minute), meter density (number of distinct meters), key, key density (number of distinct keys), and chord density (number of distinct chords);
- a ranking of music according to user-based collaborative filtering over a dataset of user surveys;
- emotion-based ranking in accordance with manual emotion annotations by an expert.

Interestingly, this combination is personalized for each particular user, according to the initial survey, in which users need to specify preference assessments (likes/dislikes) and the underlying reasons (such as preference by tonality, rhythm, etc.) for a sample of tracks, and posteriorly by re-weighting distance components according to user feedback. The scope of this system is considerably limited: its audio content-based component is based on score analysis instead of real audio while the emotion-based component requires manual expert annotations.

Instead of computing similarity between music items and a user profile, one can train discriminative models which would allow to classify items into liked and disliked categories, or predict values of user ratings. For example, Grimaldi & Cunningham (2004) proposed such a classification using the tracks rated by a user as “good” and “bad” examples. The authors employ kNN and feature sub-space ensemble classifiers working on a set of temporal (12 tempo features derived from beat-histograms) and tonal (32 features describing harmony) features. The employed classifiers and features were originally suited for the task of genre classification, and authors found that the proposed approach fails in the case when user preferences are not driven by certain genres, leading to a worse classification performance. Moh et al. (2008) propose to classify music into liked and disliked using several modifications of support vector machines (SVMs). The peculiarity of their study is in including online learning in accordance to the user’s feedback. Boosting by an additive expert ensemble of a number of least squares support vector machines is proposed as one of the abovementioned SVM modifications. In addition, two generalizations for online learning, incremental SVMs and Online Passive-Aggressive algorithm, are considered (Moh & Buhmann, 2008).

Additionally, ordinal regression can be applied for user rating prediction. Reed & Lee (2011) propose such a regression to predict ratings, assigned by the user, from audio content of the tracks, more specifically, from temporal evolution of the MFCCs within each track. To this end, they propose acoustic segment modeling, which consists in three stages. The initial stage segments each training track by a maximum-likelihood procedure. Next, a universal set of acoustic units, called acoustic segment models (ASMs), are found and modeled with a hidden Markov model. Finally, latent semantic analysis (LSA) converts each music track into a vector of weighted ASM counts. Minimum classification error (MCE) algorithm is used to train the regression model.

A number of studies make use of more complex probabilistic generative models in order to predict user preferences (i.e., ratings) for music items. For example, preference can be seen as a distribution over a feature space comprised of audio features. Moh et al. (2008) trains a full covariance Gaussian model in such a space of temporal and tonal features (specified above). Alternatively, more complex models can be built. Li et al. (2005, 2007) propose a probabilistic model, in which music tracks are classified into groups by means of audio content and collaborative data (user ratings), and the predictions are made for users considering Gaussian distribution of user ratings. Authors utilize timbral features (MFCCs, spectral centroid, spectral rolloff, spectral flux, sum of scale factor), temporal (relative amplitude of the first and second peak in the beat histogram, amplitude ratio of the second peak to the first peak, first peak and second peak BPMs, beat strength), and tonal (the amplitudes and periods of maximum peaks in the pitch histogram, pitch intervals between the two most prominent peaks, the overall sums of the histograms) features.

Pazzani & Billsus (1997) present an approach, which uses a naive Bayesian classifier in order to predict user preference for artists, based on related semantic tags extracted from web pages. Yoshii et al. (2006, 2008) propose a hybrid probabilistic model incorporating user ratings (collaborative filtering) and timbral audio information. Each music track is represented as a vector of weights of timbres (a "bag-of-timbres"), i.e., as a GMM of MFCCs, where each Gaussian correspond to a single timbre. The Gaussian components are chosen universally across tracks, being predefined on a certain music collection. Ratings and "bags-of-timbres" are associated with latent variables, conceptually corresponding to genres, and music preferences of a particular listener can be represented in terms of proportions of the genres. A three-way aspect model (a Bayesian network) is proposed for this mapping, with an idea that a user stochastically chooses a genre according to her/his preference, and then the genre stochastically "generates" pieces and timbres.

Zheleva et al. (2010) presents a probabilistic graphical model based on latent Dirichlet allocation. The approach employs listening behavior data, segmenting it into sessions for each user by means of playback timestamps, and captures associated latent “listening moods” common across users. Therefore, it is possible to detect groups of tracks and groups of listeners from/with similar listening sessions. The model assumes that each user is represented as a distribution over different moods, and for each session, there is a latent mood which guides the choice of tracks. Park et al. (2006) and Kim et al. (2008a) propose probabilistic models focused on predicting music preference with respect to the listener's context and metadata (manual semantic annotations or listening behavior data). Their approaches are using fuzzy Bayesian networks and hidden Markov models, taking into account such factors as weather, time, location, age, gender, and cardiac pulse.

In addition, online learning, or real time adaptivity to the listener's relevance feedback, is another topic covered in several studies (Hoashi et al., 2003; Hu & Ogihara, 2011; Lu & Tseng, 2009; Maillet et al., 2009; Moh & Buhmann, 2008; Moh et al., 2008; Pampalk et al., 2005b; Yoshii et al., 2006, 2008).

Metadata-based music similarity

There are many more approaches to music similarity which can serve for music recommendation but that have not been yet evaluated properly in this context to the best of our knowledge. Here we highlight a number of such approaches working with metadata. Schedl et al. (2011a) provides a comprehensive overview of music similarity measures working with data mined from the web. They include distances based on:

- co-occurrence of music items (tracks or artists) shared by users of peer-to-peer networks (Koenigstein et al., 2010);
- co-occurrence patterns of music items in expert-built or user-built music playlists;
- page counts and Web co-occurrences (Schedl & Knees, 2009);
- similarity between vector space models (i.e., vectors of normalized keyword occurrences) for music artists extracted from
  - song lyrics;
  - web-pages associated with artists;

Authors create a large-scale music search system, which operates on an index of term profiles formed by search and retrieval of web-pages related to music artists (Schedl et al., 2011b). Alternatively, vector space models can be built using information from collaborative tagging services specifically focused on music, such as *Last.fm* (Levy & Sandler, 2008, 2009), where large amount of users tag music items. Distance measures can also be built by collaborative filtering of user ratings (Slaney et al., 2008) or listening behavior (Celma, 2010). However, all types of collaborative data, considered in existing studies, are costly as they require a large user base.

Content-based music similarity

Focusing on audio content-based similarity, there exist a wide variety of approaches for providing a distance measurement between music tracks. These approaches comprise both the selection of audio descriptors and the choice of an appropriate distance function. A brief overview of the existing approaches in terms of employed information is presented in Table 2.1.

Measuring similarity by comparing track spectra is probably one of the most popular and early approaches. There exist specific timbral representations, the most prominent one being modeling the tracks as probability distributions of vectors of MFCCs, calculated on a frame basis and characterizing spectral shapes of the tracks (as we have already discussed in Section 2.3.1). Tzanetakis & Cook (2002) compares vectors of means and variances of MFCCs in each distribution by a simple Euclidean distance. Logan & Salomon (2001) represent MFCC clouds as cluster models by means of k-means clustering and compare them with the Earth Mover's Distance. Mandel & Ellis (2005) compare means and covariances of MFCCs applying the Mahalanobis distance. Furthermore, GMMs can be used to represent the probability distributions, and then these models can be compared by the symmetrized Kullback-Leibler divergence. However, in practice, approximations are required for the case of several Gaussian components in a mixture. To this end, Aucouturier and collaborators (2002, 2005) create GMMs with an Expectation-Maximization algorithm initialized with k-means clustering, and then compare the models by means of Monte Carlo sampling. In contrast, Mandel & Ellis (2005) and Flexer et al. (2008) simplify the models to single Gaussian representations, for which a closed form of the Kullback-Leibler divergence exists. Pampalk (2006) gives a global overview of these approaches. As well, Jensen et al. (2009) provide an evaluation of different GMM configurations. Besides MFCCs, more descriptors can be used for timbral distance measurement. For example, Li & Ogihara (2006) apply a Euclidean metric on a set of descriptors, including MFCCs, spectral centroid, rolloff, flux, Daubechies wavelet coefficient histograms (DWCH), and zero-crossing rate (ZCR).

Temporal (or rhythmic) representation of music is another important aspect. A number of works propose specific temporal distances in combination with timbral ones. For example, Pampalk (2006); Pampalk et al. (2005a) exploit fluctuation patterns (FP), which describe spectral fluctuations (amplitude modulations of loudness in individual critical bands) over time, together with several derivative descriptors, modeling overall tempo ("gravity") and fluctuation information at specific frequencies ("focus"). They define a hybrid distance as a linear combination of a Euclidean distance on fluctuation patterns together with a timbral distance, based on GMMs of MFCCs. Pohle & Schnitzer (2007) follow this idea, but propose a cosine similarity distance for fluctuation patterns together with a specific distance measure related to cosine similarity for GMMs of MFCCs. Furthermore, they propose an alternative temporal de-

Table 2.1: The overview of state-of-the-art approaches to music similarity.

Study	Timbral	Temporal	Tonal	Inferred patterns	Inferred semantics	Metric learning	Features
Logan & Salomon (2001)	yes						MFCCs
Tzanetakis & Cook (2002)	yes						MFCCs
Aucouturier & Pachet (2002); Aucouturier et al. (2005)	yes						MFCCs
Berenzweig et al. (2003)	yes						MFCCs; artist anchors, genres
Cano et al. (2005)	yes	yes	yes				timbre; dynamics, tempo, meter, rhythmic patterns; tonal strength, key and mode
Mandel & Ellis (2005)	yes						MFCCs
Li & Ogihara (2006)	yes						MFCCs, DWCH, spectral centroid, rolloff, flux, ZCR
Pampalk (2006)	yes	yes					MFCCs; FP, gravity, focus
Barrington et al. (2007a, 2009)	yes				yes		MFCCs; genres, moods, rhythm, instrumentation, vocals
Pohle & Schnitzer (2007)	yes	yes					MFCCs, spectral contrast, harmonicness, and attackness; modified FP
West & Lamere (2007)	yes					yes	MFSI; genres
Bertin-Mahieux et al. (2008)	yes	yes				yes	MFCCs, spectrogram, AC; social tags
Flexer et al. (2008)	yes						MFCCs
Hoffman et al. (2008)	yes						MFCCs; bags-of-timbres
Marolt (2008)			yes				chroma
Slaney et al. (2008)	yes	yes				yes	loudness, tempo, beat regularity, tatums, time signature, time signature stability
Song & Zhang (2008)	yes	yes					MFCCs, spectrum histogram; FP
Jensen et al. (2009)	yes						MFCCs
Levy & Sandler (2009)	yes	yes		yes			MFCCs; AC; bags-of-timbres/-rhythms
Maillet et al. (2009)	yes	yes				yes	MFCCs, AC, danceability, loudness
McFee & Lanckriet (2009)	yes		yes		yes		MFCCs; chroma; genres, moods, rhythm, instrumentation, vocals
Serrà et al. (2009)			yes				chroma
Seyerlehner et al. (2010)	yes	yes					spectral shape, contrast, band correlation, flux; modified FP
Charbuillet et al. (2011a)	yes			yes			MFCCs, spectral flatness; bags-of-timbres
Garcıa-Dıez et al. (2011)			yes				chords progressions
McFee et al. (2012a)	yes			yes		yes	MFCCs; bags-of-timbres

scriptor set, including a modification of fluctuation patterns (onset patterns and onset coefficients), and additional timbral descriptors (spectral contrast coefficients, harmonicness, and attackness) along with MFCCs for single Gaussian modeling (Pohle & Schnitzer, 2009; Pohle et al., 2009). Song & Zhang (2008) present a hybrid distance measure, combining a timbral Earth Mover's Distance on MFCC cluster models, a timbral Euclidean distance on spectrum histograms, and a temporal Euclidean distance on fluctuation patterns. Seyerlehner et al. (2010) propose a distance working on timbral and temporal features computed on a block level (comprised of several frames). They represent tracks using patterns of the spectral shape, spectral contrast, and correlation between bands, onset detection information (spectrum magnitude increments in individual bands for consequent blocks) and modified fluctuation pattern features. L_1 metric is then applied separately for different features, and the resulting distances are combined with a distance space normalization.

There also exist some attempts to exploit tonal representation of tracks. Ellis & Poliner (2007), Marolt (2008), and Serrà et al. (2009) present specific melodic and tonality distance measurements, not addressed to the task of music similarity, but to version (cover) identification. In principle, their approaches are based on matching sequences of pitch class profiles, or chroma feature vectors, representing the pitch class distributions (including the melody) for different tracks. Garcıa-Dıez et al. (2011) proposes a similarity measure matching chords progressions by means of comparing their graph representations.

Other approaches operate on a higher level by inferring a vocabulary of patterns of music in an unsupervised manner and representing the tracks via this vocabulary. For example, Charbuillet et al. (2011a) propose timbral modeling by using the GMM-supervector approach, which allows to represent complex statistical models by a Euclidean vector. The main idea it to build a generic GMM with a very large number of components, called Universal Background Model (UBM), by using a large dataset of representative music tracks. Therefore, it is possible to model the overall data distribution, and represent specific tracks in terms of the components of the UBM by their weight. Distance between vectors of GMM weights is then used for similarity measurement based on MFCC and spectral flatness. Similarly, Hoffman et al. (2008) develop a method for discovering the latent structure in MFCC feature data using the Hierarchical Dirichlet Process. This model also represents each track as a mixture of globally defined multivariate Gaussians and the similarity between tracks is computed by the symmetrized Kullback-Leibler divergence. Levy & Sandler (2009) represent tracks as the bags of audio "muswords" (i.e., as vector space models) describing timbral characteristics of the signal. Starting from the MFCCs, musical events are detected within each track and are consequently mapped onto a global vocabulary of muswords built on a music collection. The mapping to a particular musword is done by means of a trained self-organizing map. Similarity between tracks can be computed by using cosine distance between the vectors of weights of muswords or between the vectors of latent aspect probabilities estimated from the probabilistic latent semantic analysis. Authors similarly extract rhythmic muswords using temporal autocorrelation features (AC), which however did not improve the performance.

It is very promising to combine these types of representations. Cano et al. (2005) demonstrate a straightforward approach using a Euclidean metric after a principal component analysis (PCA) transformation of an expanded set of empirically selected features, describing timbre, dynamics, tempo, meter, rhythmic patterns, tonal strength, key and mode information. As well, one might want to adapt the measure to be a true metric, and furthermore incorporate semantic information into the metric learning process. Slaney et al. (2008) propose learning a Mahalanobis metric on loudness and temporal features. Features describing loudness dynamics over the time segments of the tracks, overall tempo, beat regularity, tatum durations and their number per beat, rhythmic time signature and its stability are used. In addition to the unsupervised metric learning using whitening transformation, the authors consider algorithms to integrate semantic relations between tracks (same artist, same album, co-occurrences in music blogs) into the low-level metric space. These algorithms include linear discriminant analysis (LDA), relevant component analysis (RCA) (Shental et al., 2002), neighborhood components analysis, and large-margin nearest neighbor classification (Weinberger & Saul, 2009). Maillet et al. (2009) utilize information about playlist co-occurrences and incorporate it into the space of autocorrelation features, MFCCs, track danceability and loudness, via neural networks. McFee et al. (2012a) recodes MFCCs to a vocabulary of generic codewords, representing the tracks as the histograms of words which can be than compared. Furthermore, authors apply metric learning to optimize the distance by additional collaborative information (listening behavior from *Last.fm*).

Though common approaches for content-based music similarity may include a variety of perceptually relevant descriptors related to different musical aspects, such descriptors are, in general, relatively low-level and not directly associated with a semantic explanation (Celma et al., 2006). In contrast, research on computing high-level semantic features from low-level audio descriptors exists. In particular, in the context of MIR classification problems, genre classification (Sturm, 2012; Tzanetakis & Cook, 2002), mood detection (Huq et al., 2010; Laurier et al., 2009a,b), and artist identification (Mandel & Ellis, 2005) have gathered much research attention.

We hypothesize that the combination of classification problem outputs can be a relevant step to overcome the so-called semantic gap (see Section 1.2) between human judgments and low-level machine learning inferences, specifically in the case of content-based music similarity. A number of works support this hypothesis. Berenzweig et al. (2003) propose to infer high-level semantic dimensions, such as genres and “canonical” artists, from low-level timbral descriptors, such as MFCCs, by means of neural networks. The inference is done on a frame basis, and the resulting clouds in high-level feature space are compared by centroids with a Euclidean distance. Barrington et al. (2007a, 2009, 2007b) train GMMs of MFCCs for a number of semantic concepts, such as genres, moods, instrumentation, vocals, and rhythm. Thereafter, high-level descriptors can be obtained by computing the probabilities of each concept on a frame basis. The resulting semantic clouds of tracks can be represented by GMMs as well, and compared with a Kullback-Leibler divergence. McFee & Lanckriet (2009) propose a hybrid low-dimensional feature transformation embedding musical artists into Euclidean space subject to a partial order, based on a set of manually annotated artist similarity triplets, over pairwise low-level and semantic distances. As such, the authors consider low-level timbral distance, based on MFCCs, tonal distance, based on chroma descriptors, and the above-mentioned semantic distance (Barrington et al., 2007b). The evaluation includes the embeddings, which merge timbral and tonal distances, and, alternatively, timbral and semantic distances. West & Lamere (2007) apply classifiers to infer semantic features of the tracks. In their experiment, mel-frequency spectral irregularities (MFSI) are used as an input for a genre classifier. The output class probabilities form a new high-level feature space, and are compared with a Euclidean distance. The authors propose to use classification and regression trees or LDA for classification. Bertin-Mahieux et al. (2008) propose a content-based method for predicting social tags collected from the Web. Authors implement 360 tag-specific classifiers using boosting, that is a combination of the simplest one-feature decision trees, by the ensemble learning algorithm FilterBoost. Authors make use of MFCCs, spectrogram coefficients, and temporal autocorrelation features. Cosine similarity is then used to compare vectors of autotags. Auto-tagging brings rich semantic capabilities to the semantic measure. However, the proposed approach blindly includes all popular tags found on *Last.fm* regardless of the actual ability to predict them. This adds significant noise to the similarity measurements.

In spite of having a variety of potential content-based approaches to music similarity, there exist certain open issues yet. The distances, operating solely on low-level audio descriptors, lack semantic explanation of similarity on a level at which human judgments operate. The majority of approaches, both low-level and high-level, focus mostly on timbral descriptors, whereas other types of low-level descriptors, such as temporal and tonal, are potentially useful as well. Furthermore, comparative evaluations are necessary, especially those carried out comprehensively and uniformly on large music collections. In existing research, there is a lack of such comparative evaluations, taking into consideration different approaches. Objective evaluation criteria of music similarity are generally reduced to co-occurrences of genre (Charbuillet et al., 2011b; Hoffman et al., 2008; Jensen et al., 2007; Levy & Sandler, 2009; Logan & Salomon, 2001), album (Logan & Salomon, 2001), and artist labels (Levy & Sandler, 2009; Logan & Salomon, 2001), performance metrics in the kNN classification tasks (Berenzweig et al., 2003; Pampalk et al., 2005a; Schedl et al., 2011a; Schnitzer et al., 2011; Slaney et al., 2008), or correlation with the ground-truth similarity by collaborative filtering (Bertin-Mahieux et al., 2008). These criteria are tested on relatively small ground truth collections. In turn, subjective evaluations with human raters are not common (Barrington et al., 2009; Berenzweig et al., 2003). In our study we will focus on filling these open issues and will employ comprehensive music collections, objective criteria for similarity, and human listeners for subjective evaluations. As the majority of existing approaches still perform poorly (at the moment of starting our research in 2008, systems' performance was very unsatisfactory) we hypothesize that better performance may be achieved by combining conceptually different distance measurements, which will help to jointly exploit different aspects of music similarity.

2.3.2 Evaluating automatic recommendation techniques

As we have seen, there is a considerable amount of research works addressing the problem of music recommendation. Authors of these works are primarily focused on designing suitable algorithms and features. A great part of research employs offline evaluations in which recommendation approaches are compared by means of objective metrics of performance without any user interaction (see Figure 2.3). Datasets of user ratings or listening behavior are typically used as they are generally easier to obtain and utilize than to conduct very costly human evaluations. Therefore, many researchers consider music recommendation in the context of optimization problem for rating prediction. The choice of the evaluation strategy is dependent on the dataset at hand, and preference elicitation strategies are not in the direct focus of researchers, being simply forced to match this data. The users and associated track/artist ratings can be obtained from online music services (Grimaldi & Cunningham, 2004; Reed & Lee, 2011; Su et al., 2010a; Yoshii, 2008; Yoshii et al., 2006, 2008). Alternatively, listening behavior data, including track/artist play-counts, and associated timestamps, (Baltrunas & Amatriain, 2009; Tiemann & Pauws, 2007; Zheleva et al., 2010) can be used. However, this data is very difficult to obtain for academic researches and the commercial services are not always ready to provide these data for research purposes. Researchers also create their own datasets to work with, surveying participants in order to rate music items (Hoashi et al., 2003, 2006; Li et al., 2005, 2007; Shardanand & Maes, 1995; Su et al., 2010b). Classification accuracy (Grimaldi & Cunningham, 2004; Reed & Lee, 2011; Su et al., 2010a), mean absolute error (Li et al., 2007; Reed & Lee, 2011), or ranking-based measures, such as discounted cumulative gain (Reed & Lee, 2011) or mean average precision (McFee et al., 2012b), are among the typical performance metrics applied in order to minimize prediction error.

Objective evaluations allow us to collect quantitative insights on the performance of prediction algorithms. However, they do not provide clues on the perceived quality of recommendations and their actual usefulness for the listener (Shani & Gunawardana, 2009), and there is some research evidence that high recommender accuracy does not always correlate with user satisfaction (McNee et al., 2006). Furthermore, as recommender systems are targeted towards music discovery, it is fundamental to assess the listener's familiarity with the recommended items apart from their relevance. The studies employing offline objective evaluations a significantly limited because they address the perceived relevance of recommendations only vaguely meanwhile measuring serendipity, that is the amount of novel and relevant recommendations, is almost impossible using the objective metrics available to researchers. This motivates the necessity of subjective evaluations, which would allow to measure user satisfaction in A/B listening tests. Such a measurement can be done using Likert-type scales, which are usually suggested as a very efficient way of collecting self-report data in usability evaluation practice (Tullis & Albert, 2008), but which give way to additional problems such as computing statistics and inferences from discrete or non-metrical scales.

Current research counts only a few user-centric studies, measuring liking or satisfaction (Hu & Ogihara, 2011; Kim et al., 2008a, 2006; Lu & Tseng, 2009; Magno & Sable, 2008; Pampalk et al., 2005b; Park et al., 2006; Song et al., 2009), and only a few include familiarity in consideration (Celma & Herrera, 2008; Firan et al., 2007). Primarily this is due to high expense of such studies: collecting a large set of subjects and using them to evaluate a large enough set of algorithms via associated recommended items is a very costly procedure in terms of required user effort. Therefore, evaluations are typically restricted to a small set of subjects and a relatively small set of tested approaches. Celma & Herrera (2008) provide the largest user-based study till the present date, being conducted on 288 participants. Each subject provided liking and familiarity ratings for ≈ 19 tracks recommended by three approaches in a blind evaluation. A large total number of evaluated tracks served as a solid basis for statistical testing by within-subjects ANOVA. Furthermore, the authors analyzed novelty of provided recommendations, and compared liking for the considered approaches taking only novel or familiar recommendations.

The study by Celma & Herrera (2008) may serve as an example of a proper subjective evaluation methodology, taken on a larger scale. However, the majority of existing research works are significantly limited: designing evaluation, one may typically encounter a problem of recruiting a large number of participants and has to agree on a trade-off between this number and the number of evaluated tracks per subject. A brief summarization of user-centered studies is presented in Table 2.2.

In general, the main problem academic practitioners are faced with in the field of music recommendation is the lack of data, which is usually proprietary. Meanwhile designers of commercial music recommender systems can implement A/B tests on large amount of their users, this becomes a critical problem for academic researchers. Therefore, researchers often opt for small-scale evaluations and create their own datasets, e.g., mining ratings from Internet music services. Yet, there exist few datasets suited for objective offline

Table 2.2: User-centered evaluations of music recommendation approaches conducted in academic research. Question marks stand for missing details.

Study	Subjects	Feedback type	Approaches	Effort (tracks/subject)
Pampalk et al. (2005b)	?	skipping behavior	4	?
Park et al. (2006)	10	satisfaction rating	2	4
Kim et al. (2008a, 2006)	50	satisfaction rating	2	≈ 12
Firan et al. (2007)	18	familiarity and liking ratings	7	70
Celma & Herrera (2008)	288	familiarity and liking ratings	3	≈ 19
Magno & Sable (2008)	13	satisfaction rating	8	24
Lu & Tseng (2009)	27	like/dislike rating	1	50
Song et al. (2009)	30	satisfaction rating	2	?
Hu & Ogihara (2011)	11	skipping behavior	2	< 140

evaluations (Celma, 2008; Dror et al., 2011; McFee et al., 2012b) created in the collaborating with the industry.[3][4][5] Remarkable examples suited for such a methodology, and taken on the large-scale level, are the recent *KDD-Cup'2011* challenge based on the *Yahoo! Music* dataset (Dror et al., 2011) and the *Million Song Dataset Challenge* (McFee et al., 2012b). Both initiatives finally allowed access to large-scale datasets of music data and users, but, nevertheless, they are limited, being decoupled from the actual audio content.[6]

Furthermore, it is problematic to systematize the existing approaches, as the considered baselines, employed music collections, performance metrics, and even rating scales, differ significantly from study to study. The majority of reported subjective evaluations lack statistical data about their participants (e.g., surveying for demographic data, interest in music, qualitative and quantitative clues on music preferences) and/or the information about employed music collections (e.g., number of tracks and genre distribution). In addition, some research works lack a proper statistical testing of hypothesis. Furthermore, we believe that it might be advisable to extend the amount of ratings per track in the poll, as there is research evidence of the noisiness of single user ratings (Amatriain et al., 2009). Using several ratings addressing different aspects of preference instead of a single liking or satisfaction rating would allow to assess consistency of preference decisions and reduce such a noise.

[3] http://ocelma.net/MusicRecommendationDataset/index.html

[4] http://webscope.sandbox.yahoo.com/

[5] http://labrosa.ee.columbia.edu/millionsong/

[6] The Million Song Dataset provides metadata and a limited set of precomputed audio features for the tracks associated with listening behavior from *Last.fm*, giving researchers the opportunity to experiment with content-based and metadata-based approaches. This dataset did not exist at the moment of conducting our research. The KDD-Cup'2011 dataset solely provides user ratings associated with anonymized music items, making impossible any approach different that collaborative filtering.

2.3.3 Conclusions

The main problem of existing research on music recommendation is the lack of user-centered studies, i.e., subjective evaluations. Those studies that employed such evaluations revealed only average-quality listener satisfaction with both metadata and content-based approaches (Bu et al., 2010; Celma & Herrera, 2008; Magno & Sable, 2008). Therefore, there is a large room for improvement of existing approaches. In addition, very few research works addressed music discovery and analyzed the familiarity of listeners with the provided recommendations. They have found that audio content can serve recommender systems in order to increase the novelty of recommendations, but the existing approaches are not accurate enough. This results in a lower user satisfaction than that achieved with the state-of-the-art metadata-based approaches.

We conclude with the necessity of improvement of content-based approaches and their hybrids with metadata. To accomplish that, we hypothesize that timbral, temporal, and tonal audio features are to be used altogether, which would lead to a richer music representation. Furthermore, it seems promising to introduce semantic audio descriptors, bringing referential meaning in the system, in order to reduce the semantic gap between human concepts and low-level audio features. To the best of our knowledge, no research with subjective evaluations has been done before in order to corroborate this hypothesis. We will focus on filling in this open issue, and, to this end, we will employ evaluations in the context of the problems of music similarity and music recommendation. On the other side, current metadata-based approaches require large amounts of collaborative filtering information or tags, which makes them expensive. Exploring cheaper alternatives is another research challenge that we want to pursue.

Finally, the existing studies do not include qualitative validation of user models themselves in terms of their understandable interpretation, but only examine the prediction power of the models or the quality of the provided recommendations. Furthermore, no research has been done in order to link computational user models to the state-of-the-art understanding of music preferences in the field of music cognition and psychology, a challenge that we will undertake too.

CHAPTER 3

Preference elicitation strategies

3.1 Introduction

Initial knowledge about music preferences of the listener will be required in order to form a user profile and generate recommendations. Part of this information can be gathered by polling users at the time they start using the system. Further user behavior within the system, and user feedback on provided recommendations can help to expand the profile. In this chapter we present the preference elicitation strategy followed in our study. We define user profile in terms of particular tracks preferred by the listener, and propose its content-based representation including semantic descriptors automatically inferred from audio.

3.2 Track-level vs artist-level recommendation

Considerable amount of approaches work on artist level, using artist preference information and returning artist recommendations. In contrast, other recommenders provide track-level recommendations using preference information about particular tracks. Moreover, there can be hybrid approaches, taking both types of artist and track information about preferences to produce track and artist recommendations.

Conceptually, artists-level can be seen as more generic than track-level, as the music by an artist can be described as a set of tracks by this artist. However, many artists do not play similar music over the years or even in a single album. Assuming that “artist” is a valid unit to provide recommendations (if you like artist “A” than you will like artist “B”) is a strong assumption that can be debated for many artists, especially those not belonging to pop genre. In contrast, track-level goes to the very basic micro-unit of musical taste. A listener might like only specific songs of artist “A” and “B”. In such a case, when a user prefers listening to particular tracks rather than artists, we expect track-level recommendations to have higher user satisfaction. In addition, we run into difficulties, when applying audio content-based approaches to artist level (e.g., it is no clear how to summarize information from different tracks). Conceptually, if the problem of recommendation were solved on a track-level, it would be possible to recommend artists based on tracks, i.e., effectively exploit hybrid artist/track approaches.

In our studies we will consider both track-level and artist-level approaches. However, there is a problem of how to compare track-level recommendations with the artist-level recommendations. For consistency, we will always consider track recommendations as particular tracks are easier to evaluate in a listening test than artists. In the latter case it is not clear which track, or group of tracks, to present to a listener for evaluation. Therefore, for the considered artist-based approaches, we will return randomly-selected tracks by recommended artists.

3.3 Explicit preference elicitation based on preference examples

We describe our explicit preference elicitation strategy, which we follow to build the ground for the music recommendation approaches further considered in Chapter 5, and in the posterior analysis of music preferences in Chapter 6.

3.3.1 Proposed approach

We propose a preference elicitation strategy which consists in gathering music preference examples explicitly from the listener in the form of particular tracks. As stated in Section 3.2, track examples represent musical preferences more precisely than artist lists, allowing us to address the problem of preference elicitation by using audio content. Our assumption is that, although the proposed strategy requires more user effort, the listeners will be finally rewarded by recommendations of a better quality.

To this end, we ask the listener to gather the *minimal* set of music tracks which is *sufficient* to grasp and convey their musical preferences (the henceforth called "*preference set*"). Ideally, the selection of representative music should not be biased by any user expectations about a final system or interface design issues. Therefore, for evaluation purposes, we do not inform the listener about any further usage of the gathered data, such as giving music recommendations or preference visualization. Furthermore, we do not specify the number of required tracks, leaving this decision to the user.

Generally, example gathering could be performed by either asking the user to provide the selected tracks in audio format (e.g., mp3) or by means of editorial metadata sufficient to reliably identify and retrieve each track (i.e., artist, piece title, edition, etc.). For the content-based recommendation, the music pieces are informative even without any additional metadata (e.g., artist names and track titles). In contrast, metadata-based approaches require editorial information sufficient to reliably identify each track (for track-level approaches) or, at least, artist of each track (for artist-level approaches). In our study we will consider both content-based and metadata-based approaches. Therefore, for our evaluation purposes only, users are obliged to provide audio files and optionally provide metadata. We then, by means of audio fingerprinting[1], retrieve and clean metadata for all provided tracks including the ones solely submitted in audio format. After following these procedure, we will be able to compare content-based approaches working on track level to metadata-based recommendation approaches working on track and artist level.

For user analysis purposes, we also ask the listeners for additional information, including personal data (gender, age, interest in music, musical background), a description of the strategy followed to select the music pieces, and the way they would describe their musical preferences. The exact text of the questionnaire is presented in Appendix B.

3.3.2 User data analysis

In order to evaluate the proposed strategy, we worked with a group of 39 participants (26 male and 13 female) selected from the authors' colleagues and acquaintances without disclosing any detail of the targeted research. They were aged between 19 and 46 years old (average $\mu = 31.35$ and standard deviation $\sigma = 6.4$) and showed a very high interest in music (rating around $\mu = 9.34$, with $\sigma = 0.96$, where 0 means no interest in music and 10 means passionate about music). 34 of the 39 participants play at least one musical instrument, including violin, piano, guitar, accordion, synthesizers, ukulele, and drums, or sing. Taking into account this information, we consider that the population represented by our participants corresponds to that of music enthusiasts, but not necessarily mainstream music consumers. Therefore, we may not expect that the conclusions from our further experiments can be generalized to a general public. Nevertheless, these conclusions can be applied for the population of music enthusiasts, which represents 21% of general public of the 16–45 age group according to some estimations (Celma, 2008), and which comprises a large percentage of the users of music recommender systems.

The number of tracks selected by the participants to convey their musical preferences was very varied, ranging from 8 to 178 music pieces ($\mu = 56.08$, $\sigma = 41.59$) with the median being 50 tracks. The time spent for this task also differed a lot, ranging from 12 minutes to 60 hours ($\mu = 6.7$ hours, $\sigma = 14.61$) with the median being 2 hours.

It is interesting to analyze the provided verbal descriptions about the strategy followed to select the music tracks. Some of the participants were selecting

[1] We use MusicBrainz service: http://musicbrainz.org/doc/MusicBrainz_Picard.

one track per artist, while some others did not apply this restriction. They also covered various uses of music such as listening, playing, singing or dancing. Other participants mentioned musical genre, mood, expressivity, and musical qualities as driving criteria for selecting the tracks. Furthermore, some participants implemented an iterative procedure by gathering a very large amount of music pieces from their music collections and performing a further refinement to obtain the final selection. Finally, all participants provided a set of labels to define their musical preferences. We asked them to provide labels related to the following facets: musical genre, mood, instrumentation, rhythm, melody/harmony, and musical expression. We also included a free category for additional labels on top of the proposed musical facets.

The number of labels provided by the participants ranged from 3 to 94 labels ($\mu = 15.68$, $\sigma = 16.87$). The distribution of the number of labels that participants provided for each facet (normalized by the total number of labels provided by each participant) is presented in Figure 3.1. We observe that most of them where related to genre, mood, and instrumentation, some of them to rhythm and few to melody, harmony, or musical expression. Other suggested labels were related to lyrics, year, and duration of the piece. The participants' preferences covered a wide range of musical styles (e.g., classical, country, jazz, rock, pop, electronic, folk), historical periods, and musical properties (e.g., acoustic vs. synthetic, calm vs. danceable, tonal vs. atonal).

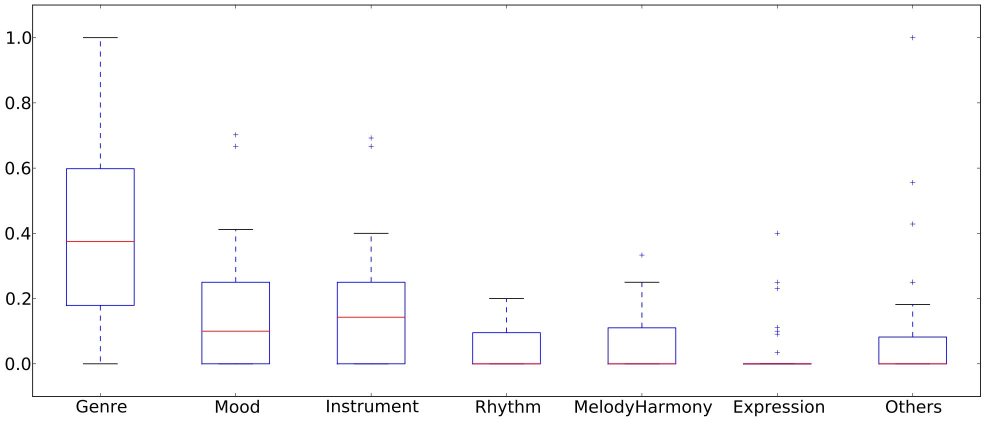


Figure 3.1: Box plot of the proportions of provided labels per musical facet, normalized by the total number of labels per participant. Categories from left to right correspond to genre, moods, instruments, rhythm, melody and harmony, musical expression, and other labels respectively. Blue crosses stand for extreme outliers.

Finally, the music provided by the participants was very diverse. Figure 3.2 presents an overall tag cloud of music preferences of our population (mostly genre-based). The tag cloud was generated using artist tags found on *Last.fm* tagging service for all tracks provided by the participants with a normalization by the number of tracks provided by each participant.

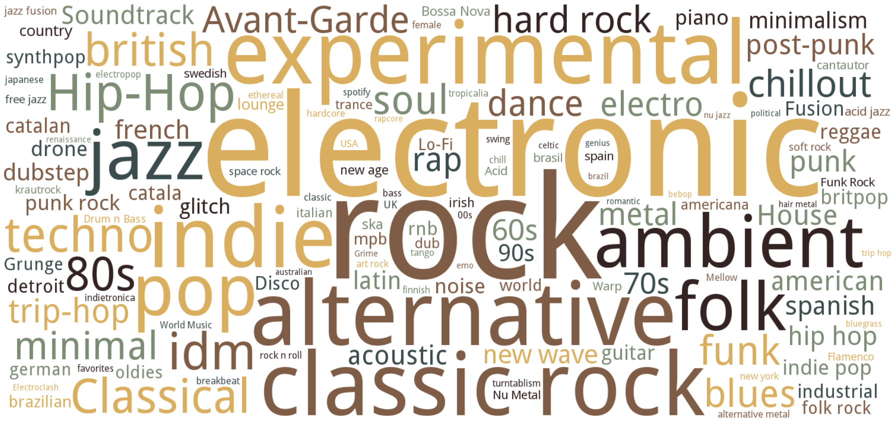


Figure 3.2: Tag cloud representing overall music preferences of our participants, based on artist tags found on *Last.fm*.

3.3.3 Discussion

Following the proposed strategy, we can assure the quality of information about preferences provided explicitly by the listener. We required two conditions on the preference set to be fulfilled by the user: compactness and sufficiency. These requirements guarantee that we obtain a maximum possible (complete) image of the listener's preferences condensed in a compact form. Therefore, we will be able to estimate the upper bound for performance of track-level music recommendation approaches, and we can expect worse performance when working on implicit data.

Our strategy solves the problem of user cold start from the beginning of user interaction with a system, but naturally requires a considerable user effort as can be seen from the obtained mean for the time spent by the participants on building their preference sets. Informal post experimental inquiry revealed that much of the effort was actually spent on finding audio files rather than deciding which music to select. Indeed, for a considerable amount of users in a real world industrial scenario, providing metadata might be easier than finding and uploading audio. In this case, the audio, including full tracks or previews, can be obtained from the associated digital libraries. Moreover, the explicit selection process followed by users can be facilitated by an intelligent system. For example, this can be achieved by analyzing implicit information (listening behavior or files in a personal music collection) and generating suggestions for preference examples to be validated and/or refined by the listeners.

Interestingly, we noticed that some listeners have difficulties assessing artists as favorites, as they like only particular tracks. Oppositely, for some listeners it is easier to provide artist names rather than concrete tracks. This further motivates our intention to consider both track-level and artist-level recommendation approaches.

A possible disadvantage of the proposed strategy is that negative preference examples are left out of consideration. We believe this information to be useful, but there is a cold-start problem of gathering such a data: not all users are able to provide tracks they dislike. In contrast, negative examples can be gradually gathered from further user feedback on provided recommendations.

3.4 Audio feature extraction

Here we describe the procedure followed to obtain a low-level timbral, temporal, and tonal and high-level semantic representation of each music track from the user's preference set.

3.4.1 Low-level audio features

For each music track, we calculate a low-level feature representation using an in-house audio analysis tool Essentia.[2] In total, this tool provides over 60 commonly used low-level audio feature classes, characterizing global properties of the given tracks, related to timbral, temporal, and tonal information. The majority of these features are extracted on a frame-by-frame basis with a 46 ms frame size, and 23 ms hop size, and then summarized by their means and variances across these frames. In the case of multidimensional descriptors, covariances between components are also considered (e.g., with MFCCs). We provide a brief overview of the feature classes we use in Table 3.1. Since it is not the objective of this thesis to review existing methods for feature extraction, the interested reader is referred to the literature cited in this table for further details.

3.4.2 High-level semantic descriptors

We use the described low-level features to infer semantic descriptors. To this end, we perform a regression by suitably trained classifiers producing different semantic dimensions such as genre, musical culture, moods, instrumentation, rhythm, and tempo. Support vector machines (SVMs) have been shown to be an effective tool for various classification tasks in MIR (Gómez & Herrera, 2008; Laurier et al., 2009a,b; Mandel & Ellis, 2005; Xu et al., 2003). We opt for multi-class SVMs with a one-versus-one voting strategy (Bishop, 2006), and use the libSVM implementation.[3] In addition to simple classification, this implementation extends the capabilities of SVMs making available class probability estimation (Chang & Lin, 2011), which is based on the improved algorithm by Platt (2000). The classifiers are trained on 20 ground truth music collections (including full tracks and excerpts) presented in Table 3.2, corresponding

[2] http://mtg.upf.edu/technologies/essentia

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

to 20 classification tasks. For some descriptors we used existing collections in the MIR field (Cano et al., 2006; Gómez & Herrera, 2008; Homburg et al., 2005; Laurier et al., 2009a; Tzanetakis & Cook, 2002), while for other descriptors we created manually labeled in-house collections.

For each given track, each classifier returns the probabilistic estimates of classes on which it was trained. The classifiers operate on optimized low-level feature representations of tracks. More concretely, each classifier is trained on a reduced set of features, which is individually selected based on correlation-based feature selection (CFS) (Hall, 2000) over all available $[0, 1]$-normalized features (Section 3.4.1) according to the underlying music collection. For example, a classifier using the G1 collection is trained on an optimized descriptor space, according to the collection's classes and the CFS process, and returns genre probabilities for the labels "alternative", "blues", "electronic", "folk/country", etc. Moreover, the parameters of each SVM are found by a grid search with 5-fold cross-validation. As a rule of thumb, supported by similar results in Laurier et al. (2009b), we generally use the C-SVC method and a radial basis function kernel with default parameters.

Classification results form a high-level semantic descriptor space, which contains the probability estimates for each class of each classifier. The accuracy of classifiers varies between 60.3% and 98.2% with the median accuracy being 88.2%. Classifiers trained on G1 (alternative, blues, electronic, folk/country, funk/soul/rnb, jazz, pop, rap/hiphop, and rock classes) and RBL (chachacha, jive, quickstep, rumba, samba, tango, viennese waltz, and waltz classes) show the worst performance, close to 60%,[4] while classifiers for CUL (western, non-western), MAG (aggressive, non-aggressive), MRE (relaxed, non-relaxed), MAC (acoustic, non-acoustic), OVI (voice, instrumental), and OTB (bright timbre, dark timbre) show the best performance, greater than 93%. Table 3.3 provides complete information on the accuracies of the employed classifiers. Some of the classifiers are based on imbalanced datasets. Their further optimization can be achieved by compensating the imbalance, e.g., by a linear combination of SVMs trained on balanced amount of examples (Lin et al., 2011).

With the described procedure we obtain 62 semantic descriptors, shown in Table 3.2, for each track in the user's preference set including categories of genre, musical culture, moods, instrumentation, rhythm, and tempo.

Proposing the aforementioned inference of semantic description, we address the problem of semantic gap between low-level audio features and human-level music description (see Section 1.2). We will be able to consider music similarity measures and music preference in term of such high-level description. The proposed feature extraction brings a track-wise semantic knowledge about music preferences, and can be used to form our user profile.

[4] Still, note the amount of classes in G1 and RBL classifiers is 9 and 8, respectively.

Table 3.1: Overview of musical features. Source references: (1) Peeters (2004), (2) Laurier et al. (2009b), (3) Logan (2000), (4) Pampalk (2006), (5) Brossier (2007), (6) Gómez et al. (2009), (7) Gouyon (2005), (8) Gómez (2006), (9) Sethares (2005)

Feature group	Feature class
Timbral	Bark bands (1,2) MFCCs (3,1,4,2) Spectral centroid, spread, kurtosis, rolloff, decrease, skewness (1,7,2) High-frequency content (7,5) Spectral complexity (2) Spectral crest, flatness, flux (1,7) Spectral energy, energy bands, strong peak, tristimulus (7) Inharmonicity, odd to even harmonic energy ratio (1)
Rhythmic	BPM (1st and 2nd peaks' BPM, weight and spread), onset rate (5,7,2) Beats loudness, beats loudness bass (6)
Tonal	Transposed and untransposed harmonic pitch class profiles, key, key strength (8,2) Tuning frequency (8) Dissonance (9,2) Chord change rate (2) Chords histogram, equal tempered deviations, non-tempered/tempered energy ratio, tuning diatonic strength, key, scale, and key strength, chords key, scale, and strength (6)
Miscellaneous	Average loudness (1) Pitch (5), pitch centroid (6) Zero-crossing rate (1,4) Silence rate Spectral RMS variance

Table 3.2: Ground truth music collections employed for semantic regression. Source references: (1) Homburg et al. (2005), (2) in-house, (3) Tzanetakis & Cook (2002), (4) Gómez & Herrera (2008), (5) Laurier et al. (2009a) + in-house, (6) Hu & Downie (2007), (7) Cano et al. (2006).

Name	Category	Classes (semantic descriptors)	Size (tracks)	Source
G1	Genre & Culture	Alternative, blues, electronic, folk/country, funk/soul/rnb, jazz, pop, rap/hiphop, rock	1820 track excerpts, 46 - 490 per genre	(1)
G2		Classical, dance, hip-hop, jazz, pop, rhythm'n'blues, rock, speech	400 tracks, 50 per genre	(2)
G3		Blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock	993 track excerpts, 100 per genre	(3)
GEL		Ambient, drum'n'bass, house, techno, trance	250 track excerpts, 50 per genre	(2)
CUL		Western, non-western	1640 track excerpts, 1132/508 per class	(4)
MHA	Moods & Instruments	Happy, non-happy	302 full tracks + excerpts, 139/163 per class	(5)
MSA		Sad, non-sad	230 full tracks + excerpts, 96/134 per class	(5)
MAG		Aggressive, non-aggressive	280 full tracks + excerpts, 133/147 per class	(5)
MRE		Relaxed, non-relaxed	446 full tracks + excerpts, 145/301 per class	(5)
MAC		Acoustic, non-acoustic	321 full tracks + excerpts, 193/128 per class	(5)
MEL		Electronic, non-electronic	332 full tracks + excerpts, 164/168 per class	(5)
MCL		5 mood clusters	269 track excerpts, 32 - 74 per cluster	(6)
RPS	Rhythm & Tempo	Perceptual speed: slow, medium, fast	3000 full tracks, 1000 per class	(2)
RBL		Chachacha, jive, quickstep, rumba, samba, tango, viennese waltz, waltz	683 track excerpts, 60 - 110 per class	(7)
ODA	Other	Danceable, non-danceable	306 full tracks, 124/182 per class	(2)
OPA		Party, non-party	349 full tracks + excerpts, 198/151 per class	(2)
OVI		Voice, instrumental	1000 track excerpts, 500 per class	(2)
OTN		Tonal, atonal	345 track excerpts, 200/145 per class	(2)
OTB		Timbre: bright, dark	3000 track excerpts, 1000 per class	(2)
OGD		Voice gender: male, female	3311 full tracks, 1508/1803 per class	(2)

Table 3.3: Accuracies of the classifiers employed for semantic regression. N.C. stands for "not computed" due to technical difficulties.

Classifier (dataset)	Classes	Accuracy
G1	Alternative, blues, electronic, folk/country, funk/soul/rnb, jazz, pop, rap/hiphop, rock	60.29%
G2	Classical, dance, hip-hop, jazz, pop, rhythm'n'blues, rock, speech	88.22%
G3	Blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock	77.74%
GEL	Ambient, drum'n'bass, house, techno, trance	89.33%
CUL	Western, non-western	93.47%
MHA	Happy, non-happy	84.90%
MSA	Sad, non-sad	86.96%
MAG	Aggressive, non-aggressive	98.21%
MRE	Relaxed, non-relaxed	91.78%
MAC	Acoustic, non-acoustic	93.42%
MEL	Electronic, non-electronic	86.38%
MCL	5 mood clusters	62.83%
RPS	Perceptual speed: slow, medium, fast	77.64%
RBL	Chachacha, jive, quickstep, rumba, samba, tango, viennese waltz, waltz	60.03%
ODA	Danceable, non-danceable	N.C.
OPA	Party, non-party	89.21%
OVI	Voice, instrumental	96.0%
OTN	Tonal, atonal	N.C.
OTB	Timbre: bright, dark	93.93%
OGD	Voice gender: male, female	N.C.

3.4.3 Content-based user profile

After applying audio analysis to each track from the user's preference set, we are able to represent her/his music preferences as a set of feature vectors characterizing each particular track. In the case of using solely semantic descriptors, we can define a semantic user model as a set U:

$$U = \Big\{ \Big(P(C_{1,1}|T_i), ..., P(C_{1,N_1}|T_i), ..., P(C_{17,1}|T_i) ..., P(C_{17,N_{17}}|T_i) \Big) \Big\}, \quad (3.1)$$

where $P(C_{k,l}|T_i)$ stands for the probability of track T_i from a preference set belonging of l-th class $C_{k,l}$ of the k-th classifier having N_k classes.

CHAPTER 4

Content-based music similarity measures

4.1 Introduction

Music similarity measures are commonly applied in the context of the problem of music recommendation. As we have discussed in Section 1.2, approaches based on content-based music similarity overcome the problem of popularity bias, typical for metadata-based systems. In this chapter we focus on the ways to measure such a content-based similarity between tracks. We are specifically interested in non-personalized measures. Adaptive measures would have required user feedback, while in our study we opted for designing similarity measures common for all users and applicable from scratch. We will start by designing such similarity measures, to be later applied together with our preference elicitation strategy for music recommendation. We propose two simple approaches (i.e., non-hybrid approaches as opposed to complex measures consisting of different similarity measures themselves) working on audio content and evaluate them objectively against a number of baselines. Furthermore, we present subjective evaluations using comparative A/B listening tests, which are often absent in the majority of the existing studies. We explore the possibility of creating a hybrid approach, based on the considered simple approaches as potential components. In the considered approaches we rely on the audio features described in Section 3.4.

4.2 Baseline simple approaches

In our study, we consider a number of conceptually different simple approaches to music similarity. Among them we indicate several baselines, which will be used in objective and subjective evaluations, and moreover will be regarded as potential components of the hybrid approach.

4.2.1 Euclidean distance based on principal component analysis (L2-PCA)

As a starting point, we follow the ideas proposed by Cano et al. (2005), and apply an unweighted Euclidean metric on a manually selected subset of the low-level features outlined above.[1] This subset includes bark bands, pitch, spectral centroid, spread, kurtosis, rolloff, decrease, skewness, high-frequency content, spectral complexity, spectral crest, flatness, flux, spectral energy, energy bands, strong peak, tristimulus, inharmonicity, odd to even harmonic energy ratio, beats loudness, beats loudness bass, untransposed harmonic pitch class profiles, key strength, average loudness, and zero-crossing rate.

Preliminary steps include descriptor normalization in the interval $[0, 1]$ and principal component analysis (PCA) (Witten & Frank, 2005) to reduce the dimension of the descriptor space to 25 variables. The choice of the number of target variables is conditioned by a trade-off between target descriptiveness and the curse of high-dimensionality (Aggarwal, 2005; Beyer et al., 1999; Korn et al., 2001), typical for L_p metrics, and is supported by research work on dimension reduction for music similarity (Wack et al., 2006) and autotagging (Sordo, 2012). Nevertheless, through our PCA dimensionality reduction, an average of 78% of the information variance was preserved on our music collections, reducing the number of 201 native descriptors by a factor of 8.

4.2.2 Euclidean distance based on relevant component analysis (L2-RCA-1 and L2-RCA-2)

Along with the previous measure, we consider more possibilities of descriptor selection. In particular, we perform relevant component analysis (RCA) (Shental et al., 2002). Similar to PCA, RCA gives a rescaling linear transformation of a descriptor space, but is based on preliminary training on a number of groups of similar tracks. Having such training data, the transformation reduces irrelevant variability in the data while amplifying relevant variability. As in the L_2-PCA approach, the output dimensionality is chosen to be 25. We consider both the descriptor subset used in L_2-PCA and the full descriptor set of Table 3.1 (L_2-RCA-1 and L_2-RCA-2, respectively).

4.2.3 Kullback-Leibler divergence based on GMM of MFCCs (1G-MFCC)

Alternatively, we consider timbre modeling with GMM as another baseline approach (Aucouturier et al., 2005). We implement the simplification of this timbre model using single Gaussian with full covariance matrix (Flexer et al., 2008; Mandel & Ellis, 2005; Pohle et al., 2006). Comparative research of timbre

[1] Specific details not included in the cited reference were consulted with P. Cano in personal communication.

distance measures using GMMs indicates that such a simplification can be used without significantly decreasing performance while being computationally less complex (Jensen et al., 2009; Pampalk, 2006). As a distance measure between single Gaussian models for tracks X and Y we use a closed form symmetric approximation of the Kullback-Leibler divergence,

$$\begin{aligned} d(X, Y) = & \\ & Tr(\Sigma_X^{-1}\Sigma_Y) + Tr(\Sigma_Y^{-1}\Sigma_X) + \\ & Tr((\Sigma_X^{-1} + \Sigma_Y^{-1})(\mu_X - \mu_Y)(\mu_X - \mu_Y)^T) - \\ & 2N_{MFCC}, \end{aligned} \tag{4.1}$$

where μ_X and μ_Y are MFCC means, Σ_X and Σ_Y are MFCC covariance matrices, and N_{MFCC} is the dimensionality of the MFCCs. This dimensionality can vary from 10 to 20 (Jensen et al., 2009; Laurier et al., 2009b; Pampalk et al., 2003). To preserve robustness against different audio encodings, the first 13 MFCC coefficients are taken (Sigurdsson et al., 2006).

4.3 Proposed simple approaches

Concerning simple approaches to music similarity, here we propose two novel distance measures that are conceptually different than what has been reviewed. We regard both approaches as potential components of the hybrid approach.

4.3.1 Tempo-based distance (TEMPO)

The first approach we propose is related to the exploitation of tempo-related musical aspects with a simple distance measure. This measure is based on two descriptors, beats per minute (BPM) and onset rate (OR), the latter representing the number of onsets per second. These descriptors are fundamental for the temporal description of music. Among different implementations, we opted for BPM and OR estimation algorithms presented by Brossier (2007).

For two tracks X and Y with BPMs X_{BPM} and Y_{BPM}, and ORs X_{OR} and Y_{OR}, respectively, we determine a distance measure by a linear combination of two separate distance functions,

$$d(X, Y) = w_{\mathrm{BPM}} d_{\mathrm{BPM}}(X, Y) + w_{\mathrm{OR}} d_{\mathrm{OR}}(X, Y), \tag{4.2}$$

defined for BPM as

$$d_{\mathrm{BPM}}(X, Y) = \min_{i \in \mathbb{N}} \left(\alpha_{\mathrm{BPM}}^{i-1} \left| \frac{max(X_{\mathrm{BPM}}, Y_{\mathrm{BPM}})}{min(X_{\mathrm{BPM}}, Y_{\mathrm{BPM}})} - i \right| \right), \tag{4.3}$$

and for OR as

$$d_{\mathrm{OR}}(X, Y) = \min_{i \in \mathbb{N}} \left(\alpha_{\mathrm{OR}}^{i-1} \left| \frac{max(X_{\mathrm{OR}}, Y_{\mathrm{OR}})}{min(X_{\mathrm{OR}}, Y_{\mathrm{OR}})} - i \right| \right), \tag{4.4}$$

where $X_{\mathrm{BPM}}, Y_{\mathrm{BPM}}, X_{\mathrm{OR}}, Y_{\mathrm{OR}} > 0$, $\alpha_{\mathrm{BPM}}, \alpha_{\mathrm{OR}} \geq 1$. The parameters w_{BPM} and w_{OR} of Eq. 4.2 define the weights for each distance component. Eq. 4.3 is based on the assumption that tracks with the same BPMs or multiples of the BPM, e.g. $X_{\mathrm{BPM}} = iY_{\mathrm{BPM}}$, are more similar than tracks with non-multiple BPMs. For example, the tracks X and Y with $X_{\mathrm{BPM}} = 140$ and $Y_{\mathrm{BPM}} = 70$ should have a closer distance than the tracks X and Z with $Z_{\mathrm{BPM}} = 100$. Our assumption is motivated by research on the perceptual effects of double or half tempo (McKinney & Moelants, 2006). Eq. 4.4 is based on the similar assumption for ORs. The strength of this assumption depends on the parameter α_{BPM} (α_{OR}). Moreover, such a distance can be helpful in relation to the common problem of tempo duplication (or halving) in automated tempo estimation (Gouyon et al., 2006; Smith, 2010). In the case of $\alpha_{\mathrm{BPM}} = 1$, all multiple BPMs are treated equally, while in the case of $\alpha_{\mathrm{BPM}} > 1$, preference inversely decreases with i. In practice we use $i = 1, 2, 4, 6$.

Equations 4.2, 4.3, and 4.4 formulate the proposed distance in the general case. In a parameter-tuning phase we performed a grid search with one of the ground truth music collections (RBL) under the objective evaluation criterion described in Section 4.4.1. Using this collection, which is focused on rhythmic aspects and contains tracks with various rhythmic patterns, we found $w_{\mathrm{BPM}} = w_{\mathrm{OR}} = 0.5$ and $\alpha_{\mathrm{BPM}} = \alpha_{\mathrm{OR}} = 30$ to be the most plausible parameter configuration. Such values reveal the fact that in reality both components are equally meaningful and that mainly a one-to-one relation of BPMs (ORs) is relevant for the music collection and descriptors we used to evaluate such rhythmic similarity. In the case our BPM (OR) estimator had increased duplicity errors (e.g. a BPM of 80 was estimated as 160), we should expect lower α values as the most plausible.

4.3.2 Classifier-based distance (CLAS)

The second approach we propose derives a distance measure from diverse classification tasks. In contrast to the aforementioned methods, which directly operate on a low-level descriptor space, we first infer high-level semantic descriptors using suitably trained classifiers, as described in Section 3.4.2, and then define a distance measure operating on this newly formed high-level semantic space. We operate on a reduced subset of semantic descriptors as in our experiments we initially had no access to a part of our ground truth collections (Table 3.2). The subset includes the descriptors inferred using the G1, G2, G3, CUL, MHA, MSA, MAG, MRE, MAC, MEL, RPS, RBL, OPA, and OVI collections.

We define a distance operating on a formed high-level semantic space (i.e., the one of the descriptor probabilities). To this end, we consider different measures frequently used in collaborative filtering systems. Among the standard ones, we select the *cosine distance* (CLAS-Cos), *Pearson correlation distance* (CLAS-Pears) (Celma, 2008; Gibbons & Chakraborti, 2003; Sarwar et al., 2001), and *Spearman's rho correlation distance* (CLAS-Spear) (Gibbons & Chakraborti, 2003; Herlocker et al., 2004).

Cosine similarity is a measure of similarity between two vectors by measuring the cosine of the angle between them. Equation 4.5 shows the definition of the cosine distance between two tracks:

$$d(X,Y) = 1 - cos(\vec{X},\vec{Y}) = 1 - \frac{\vec{X}\cdot\vec{Y}}{||X|| * ||Y||} = 1 - \frac{\sum\limits_i P_{i,X}P_{i,Y}}{\sqrt{\sum\limits_i P_{i,X}^2}\sqrt{\sum\limits_i P_{i,Y}^2}}, \quad (4.5)$$

where $\{P_{i,X}\}$ and $\{P_{i,Y}\}$ form vectors $\vec{X}$ and $\vec{Y}$ of computed semantic descriptors for tracks X and Y, respectively.

Pearson correlation measures the extent to which there is a linear relationship between two variables. Equation 4.6 defines the Pearson correlation distance between tracks:

$$d(X,Y) = 1 - \frac{Cov(X,Y)}{\sigma_X\sigma_Y} = 1 - \frac{\sum\limits_i (P_{i,X} - \overline{P_X})(P_{i,Y} - \overline{P_Y})}{\sqrt{\sum\limits_i (P_{i,X} - \overline{P_X})^2}\sqrt{\sum\limits_i (P_{i,Y} - \overline{P_Y})^2}}, \quad (4.6)$$

where $\overline{P_X}$ and $\overline{P_Y}$ are the averages of descriptor values for track X and Y. In contrast, Spearman's rho measure the extent to which two different rankings agree independently of the actual values of the variables. It is computed in the same manner as the Pearson correlation, except that the $P_{i,X}$'s and $P_{i,Y}$'s are transformed into ranks, and the correlations are computed on the ranks.

Moreover, we consider a number of more sophisticated measures. In particular, the *adjusted cosine distance* (CLAS-Cos-A, Equation 4.7) (Celma, 2008; Sarwar et al., 2001) is computed by taking into account the average probability for each class, i.e. compensating distinction between classifiers with different number of classes:

$$d(X,Y) = 1 - \frac{\sum\limits_i (P_{i,X} - P_i)(P_{i,Y} - \overline{P_i})}{\sqrt{\sum\limits_i (P_{i,X} - \overline{P_i})^2}\sqrt{\sum\limits_i (P_{i,Y} - \overline{P_i})^2}}, \quad (4.7)$$

where $\bar{P_i}$ stands for average probability value for the descriptor i estimated by the associated classifier.

Alternatively, we can prioritize the importance of the semantic descriptors by assigning them weights and applying *weighted cosine distance* (CLAS-Cos-W) (Cripps et al., 2006) or *weighted Pearson correlation distance* (CLAS-Pears-W) (Abdullah, 1990). Equation 4.8 presents the weighted cosine distance:

$$d(X,Y) = 1 - \frac{\sum\limits_i w_i P_{i,X}P_{i,Y}}{\sqrt{\sum\limits_i w_i P_{i,X}^2}\sqrt{\sum\limits_i w_i P_{i,Y}^2}}, \quad (4.8)$$

where w_i are descriptor weights. In turn, weighted Pearson correlation distance can be defined with Equation 4.9:

$$d(X,Y) = 1 - \frac{\sum_i w_i(P_{i,X} - \overline{P_{X,w}})(P_{i,Y} - \overline{P_{Y,w}})}{\sqrt{\sum_i w_i(P_{i,X} - \overline{P_{X,w}})^2}\sqrt{\sum_i w_i(P_{i,Y} - \overline{P_{Y,w}})^2}}, \tag{4.9}$$

with weighted averages of descriptor values $\overline{P_{X,w}} = \frac{\sum_i w_i P_{i,X}}{\sum_i w_i}$ and $\overline{P_{Y,w}} = \frac{\sum_i w_i P_{i,Y}}{\sum_i w_i}$ for tracks X and Y.

We study both manual weighting (W_M) and automatic weighting based on the classification accuracy (W_A) computed for each descriptor. For W_M, we split the collections into 3 groups of musical dimensions, namely genre and musical culture, moods and instruments, and rhythm and tempo. We empirically assign weights 0.50, 0.30, and 0.20 to these groups respectively. Our choice is supported by research on the effect of genre in terms of music perception (Cupchik et al., 1982; Novello et al., 2006; Rentfrow & Gosling, 2003) and the fact that genre is the most common aspect of similarity used to evaluate distance measures in the MIR community (Section 2.3.1). For W_A, we evaluate the accuracy of each classifier (Table 3.3), and assign proportional weights which sum to 1.

With this setup, the problem of content-based music similarity can be seen as a collaborative filtering problem of item-to-item similarity (Sarwar et al., 2001) (see Section 2.3.1). Such a problem can generally be solved by calculating a correlation distance between rows of a track/user rating matrix with the underlying idea that similar items should have similar ratings by certain users. Transferring this idea to our context, we can state that similar tracks should have similar probabilities of certain classifier labels (i.e., semantic descriptors). To this extent, we compute track similarity on a track/user rating matrix with class labels (semantic descriptors) playing the role of users, and probabilities playing the role of user ratings, so that each N-class classifier corresponds to N users.

4.4 Experiment 1: Evaluation of simple approaches

We evaluated all considered approaches with a uniform methodological basis, including an objective evaluation on comprehensive ground truths and a subjective evaluation based on ratings given by real listeners. As an initial benchmark for the comparison of the considered approaches we used a random distance (RAND), i.e., we selected a random number from the standard uniform distribution as the distance between two tracks.

Table 4.1: Additional ground truth music collections employed for objective evaluation of the simple approaches.

Name	Category	Classes	Size	Source
G4	Genre & Culture	Alternative, blues, classical, country, electronica, folk, funk, heavy metal, hip-hop, jazz, pop, religious, rock, soul	140 full tracks, 10 per genre	Rentfrow & Gosling (2003)
ART	Artist	200 different artist names	2000 track excerpts, 10 per artist	In-house
ALB	Album	200 different album titles	2000 track excerpts, 10 per album	In-house

4.4.1 Objective evaluation methodology

In our evaluations we covered different musical dimensions such as genre, mood, artist, album, culture, rhythm, or presence or absence of voice. A number of ground truth music collections (including both full tracks and excerpts) were employed for that purpose. To this end, we used the same subset of our ground truth collections as we employed for our CLAS measure (i.e., G1, G2, G3, CUL, MHA, MSA, MAG, MRE, MAC, MEL, RPS, RBL, OPA, and OVI collections). Moreover, we included two new collections, ART and ALB, representing tracks by particular artists or albums, and an additional genre collection G4 (see Table 4.1). As we have noticed in Section 2.3.1, existing research studies on music similarity usually rely on genre, artist and album ground truths in their objective evaluations and typically take only a few datasets in consideration. In contrast, we are able to evaluate our approaches on the extended set of 17 ground truths, covering aspects rarely considered in existing evaluations, such as rhythm or moods.

For each collection, we considered tracks from the same class to be similar and tracks from different classes to be dissimilar, and assessed the relevance of the tracks' rankings returned by each approach. To this end, we used the mean average precision (MAP) measure (Manning et al., 2008). The MAP is a standard information retrieval measure used in the evaluation of many query-by-example tasks. For each approach and music collection, MAP was computed from the corresponding full distance matrix. The average precision (AP) (Manning et al., 2008) was computed for each matrix row (for each track query) and the mean was calculated across queries (columns). That is, for a collection of the size of N tracks, the average precision was computed for each track, defined as

$$\text{AveP} = \frac{\sum_{k=1}^{N-1}(P(k) \times \text{rel}(k))}{\text{number of relevant tracks}}, \tag{4.10}$$

where k is the rank in the sequence of retrieved tracks, $N - 1$ is the total number of retrieved tracks (all tracks in the collection except for the query), $P(k)$ is the precision at cut-off k in the list, $\text{rel}(k)$ is an indicator function equaling 1 if the track at rank k is relevant to the query (i.e., from the same class), zero otherwise. Mean average precision was then computed as

$$\text{MAP} = \frac{\sum_{q=1}^{N} \text{AveP(q)}}{N}. \tag{4.11}$$

As three of the considered approaches (L_2-RCA-1, L_2-RCA-2, and CLAS) require training, cross-validation is necessary. Moreover, the CLAS approach needs to work on 14 training datasets simultaneously. For consistency, we applied the same procedure to each of the considered distances, whether they required training or not: the results for RAND, L_2-PCA, L_2-RCA-1, L_2-RCA-2, 1G-MFCC, TEMPO, and CLAS-based distances were averaged over 5 iterations of 3-fold cross-validation. On each iteration, all 17 ground truth collections were split into training and testing sets. For each testing set, the CLAS-based distances were provided with 14 out of 17 training sets. The G4, ART, and ALB collections were not included as training sets due to the insufficient size of their class samples. In contrast, for each testing set, L_2-RCA-1, and L_2-RCA-2 were provided with a single complementary training set belonging to the same collection.

4.4.2 Objective evaluation results

The average MAP results are presented in Fig. 4.1 and Table 4.2. Additionally, the approaches with statistically non-significant difference in MAP performance according to the independent two-sample t-tests are presented in Table 4.3. These t-tests were conducted to separately compare the performances for each music collection. In the cases that are not reported in Table 4.3, we found statistically significant differences in MAP performance ($p < 0.05$).

We first see that all considered distances outperform the random baseline (RAND) for most of the music collections. When comparing baseline approaches (L_2-PCA, L_2-RCA-1, L_2-RCA-2, 1G-MFCC), we find 1G-MFCC to perform best on average. Still, L_2-PCA performs similarly for some collections (MHA, MSA, MRE, and MEL) or slightly better for other collections (MAC and RPS). With respect to tempo-related collections, TEMPO performs similarly (RPS) or significantly better (RBL) than baseline approaches. Indeed, it is the best performing distance for the RBL collection. Surprisingly, TEMPO yielded accuracies which are comparable to some of the baseline approaches for music collections not strictly related to rhythm or tempo

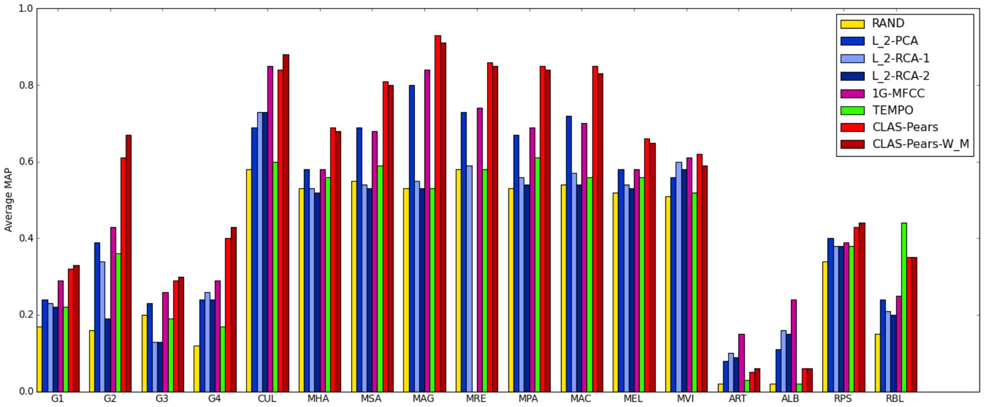


Figure 4.1: Objective evaluation results (MAP) of the simple approaches for the different music collections considered.

such as G2, MHA, and MEL. In contrast, no statistically significant difference was found in comparison with the random baseline for the G3, MAG, MRE, and ALB collections. Finally, we saw that classifier-based distances achieved the best accuracies for the majority of the collections. Since all CLAS-based distances (CLAS-Cos, CLAS-Pears, CLAS-Spear, CLAS-Cos-W, CLAS-Pears-W, CLAS-Cos-A) showed comparable accuracies, we only report two examples (CLAS-Pears, CLAS-Pears-W$_M$). In particular, CLAS-based distances achieved large accuracy improvements with the G2, G4, OPA, MSA, and MAC collections. In contrast, no improvement was achieved with the ART, ALB, and RBL collections. The distance 1G-MFCC performed best for the ART and ALB collections. We hypothesize that the success of 1G-MFCC for the ART and ALB collections might be due to the well known “album effect” (Mandel & Ellis, 2005). This effect implies that, due to production process, tracks from the same album share much more timbral characteristics than tracks from different albums of the same artist and, moreover, of different artists.

Table 4.2: Objective evaluation results (MAP) of the simple approaches (Section 4.4.2) and the hybrid approach (Section 4.7.1) for the different music collections considered. N.C. stands for “not computed" due to technical difficulties. For each collection, the MAPs of the approaches, which perform best without statistically significant difference between them, are marked in bold.

Method	G1	G2	G3	G4	CUL	MHA	MSA	MAG	MRE	OPA	MAC	MEL	OVI	ART	ALB	RPS	RBL
RAND	0.17	0.16	0.12	0.20	0.58	0.53	0.55	0.53	0.58	0.53	0.54	0.52	0.51	0.02	0.02	0.34	0.15
L_2-PCA	0.24	0.39	0.24	0.23	0.69	0.58	0.69	0.80	0.73	0.67	0.72	0.58	0.56	0.08	0.11	0.40	0.24
L_2-RCA-1	0.23	0.34	0.26	0.13	0.73	0.53	0.54	0.55	0.59	0.56	0.57	0.54	0.60	0.10	0.16	0.38	0.21
L_2-RCA-2	0.22	0.19	0.24	0.13	0.73	0.52	0.53	0.53	N.C.	0.54	0.54	0.53	0.58	0.09	0.15	0.38	0.20
1G-MFCC	0.29	0.43	0.29	0.26	0.85	0.58	0.68	0.84	0.74	0.69	0.70	0.58	0.61	**0.15**	**0.24**	0.39	0.25
TEMPO	0.22	0.36	0.17	0.19	0.60	0.56	0.59	0.53	0.58	0.61	0.56	0.56	0.52	0.03	0.02	0.38	**0.44**
CLAS-Pears	0.32	0.61	0.40	0.29	0.84	**0.69**	**0.81**	**0.93**	**0.86**	**0.85**	**0.85**	**0.66**	0.62	0.05	0.06	0.43	0.35
CLAS-Pears-W$_M$	**0.33**	**0.67**	**0.43**	**0.30**	**0.88**	**0.68**	0.80	0.91	**0.85**	**0.84**	**0.83**	**0.65**	0.59	0.06	0.06	**0.44**	0.35
HYBRID	0.29	0.62	0.37	**0.33**	**0.89**	0.65	0.77	0.89	0.81	0.76	0.79	**0.65**	**0.64**	0.06	0.07	0.43	0.37

Table 4.3: The approaches with statistically non-significant difference in MAP performance according to the independent two-sample t-tests. The L_2-RCA-2 approach was excluded from the analysis due to technical difficulties.

Collection	Compared approaches	P-value
G3	RAND, TEMPO	0.40
MHA	RAND, L_2-RCA-1	1.00
	L_2-PCA, 1G-MFCC	1.00
	CLAS-Pears, CLAS-Pears-W$_M$	0.37
MSA	L_2-PCA, 1G-MFCC	0.37
	CLAS-Pears, CLAS-Pears-W$_M$	0.50
MAG	RAND, TEMPO	1.00
MRE	RAND, TEMPO	0.33
	L_2-PCA, 1G-MFCC	0.09
	CLAS-Pears, CLAS-Pears-W$_M$	0.37
OPA	CLAS-Pears, CLAS-Pears-W$_M$	0.50
MAC	CLAS-Pears, CLAS-Pears-W$_M$	0.08
MEL	L_2-PCA, 1G-MFCC	1.00
	CLAS-Pears, CLAS-Pears-W$_M$	0.37
ALB	RAND, TEMPO	0.33
	CLAS-Pears, CLAS-Pears-W$_M$	0.33
RPS	L_2-RCA-1, TEMPO	1.00

4.4.3 Subjective evaluation methodology

In the light of the results of the objective evaluation (Sec. 4.4.2), we selected 4 conceptually different approaches (L_2-PCA, 1G-MFCC, TEMPO, and CLAS-Pears-W$_M$) together with the random baseline (RAND) for the listeners' subjective evaluation. We designed a web-based survey where registered listeners performed a number of iterations blindly voting for the considered distance measures, assessing the quality of how each distance reflects perceived music similarity. In particular, we evaluated the resulting sets of most similar tracks produced by the selected approaches, hereafter referred as "playlists". Such a scenario is an effective way to assess the quality of music similarity measures (Barrington et al., 2009; Slaney & White, 2007). It increases discrimination between approaches in comparison with a pairwise track-to-track evaluation. Moreover, it reflects the common applied context of music similarity measurement, which consists of playlist generation.

During each iteration, the listener was presented with 5 different playlists (one for each measure) generated from the same seed track. A screenshot of the web-based evaluation is presented in Figure 4.2. Each playlist consisted of the 5 nearest-to-the-seed tracks. The entire process used an in-house collection of 300K music excerpts (30 sec.) by 60K artists (5 tracks/artist) covering a wide range of musical dimensions (different genres, styles, arrangements, geographic locations, and epochs). No playlist contained more than one track from the same artist. Independently for each playlist, we asked the listeners to provide two ratings (participants were informed about the meaning of the ratings):

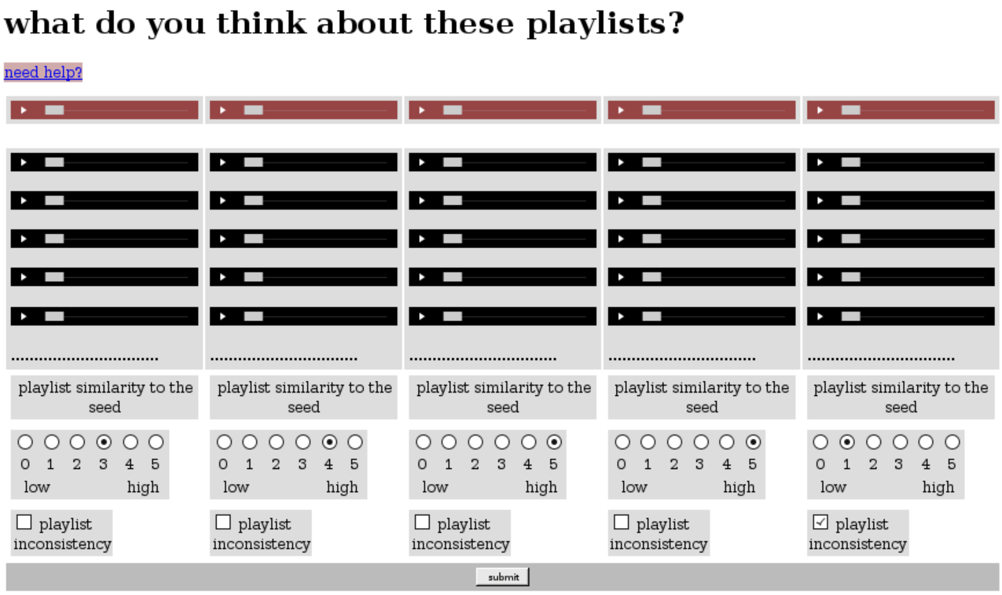
- *playlist similarity* rating indicating the appropriateness of the tracks in the playlist in respect to the seed track;

- *playlist inconsistency* boolean answer indicating that the playlist contains inconsistent results (e.g., speech mixed with music, really different tempos, completely opposite feelings or emotions, distant musical genres, etc.)

For playlist similarity ratings we used a 6-point Likert-type scale (0 corresponding to the lowest similarity, 5 to the highest) to evaluate the appropriateness of the playlist with respect to the seed. Likert-type scales (Saris & Gallhofer, 2007) are bipolar scales used as tools-of-the-trade in many disciplines to capture subjective information, such as opinions, agreements, or disagreements with respect to a given issue or question. The two opposing positions occupy the extreme ends of the scale (in our case, low-high similarity of the playlist to the seed), and several ratings are allocated for intermediate positions. We explicitly avoided a “neutral” point in order to increase the discrimination between positive and negative opinions.

We wanted to gather ratings of our participants on a number shared playlists based on the preselected seeds covering a variety of types of music, avoiding the variability in ratings that could be attributed to the difference between participants’ playlists. In addition, we wanted to include the playlists based on random seeds individual for each participant. Therefore, we divided the test into two phases: in the first, 12 seeds and corresponding playlists were shared between all listeners; in the second one the seeds for each listener (up to a maximum of 21) were randomly selected. Listeners were never informed of this distinction. Additionally, we asked each listener about his/her musical background, which included musicianship and listening expertise information (each measured in 3 levels). Altogether we collected playlist similarity ratings, playlist inconsistency indicators, and background information from 12 listeners.[2]

[2] Due to confidential reasons, the survey was conducted on a limited closed set of participants, and was unavailable to general public.



what do you think about these playlists?

need help?

playlist similarity to the seed	playlist similarity to the seed	playlist similarity to the seed	playlist similarity to the seed	playlist similarity to the seed
○ ○ ○ ● ○ ○	○ ○ ○ ○ ● ○	○ ○ ○ ○ ○ ●	○ ○ ○ ○ ○ ●	○ ● ○ ○ ○ ○
0 1 2 3 4 5	0 1 2 3 4 5	0 1 2 3 4 5	0 1 2 3 4 5	0 1 2 3 4 5
low high	low high	low high	low high	low high
□ playlist inconsistency	□ playlist inconsistency	□ playlist inconsistency	□ playlist inconsistency	☑ playlist inconsistency

Figure 4.2: A screenshot of the subjective evaluation web-based survey.

4.4.4 Subjective evaluation results

In experimental situations such as our subjective evaluation, analysis of variance (ANOVA) is the usual methodology employed to assess the effects of one variable (like the similarity computation approach) on another one (such as the similarity rating obtained from listeners) with possible interaction effects (like the effect of testing phase). ANOVA provides a statistical test of whether or not the means of several groups (in our case, the ratings obtained using a specific similarity computation approach) are equal. In our case it should be preferred to t-test because it provides a compact overview and decreases the possibility of false-rejecting a null hypothesis (i.e., that the similarity computation approaches yield no difference on similarity ratings) as it includes the possibility to assess variable interaction effects that a series of t-tests will not address. In addition to the effect of the different similarity computation methods of similarity ratings, in our evaluation we wanted to know the possible effect of the musicianship and listening experience of the participants. Furthermore, we also wanted to know the effect produced by the two consecutive testing phases used: the one presenting the same tracks to all the listeners and the other using different tracks for each of them (in fact we wanted to be sure that there would be no effect on the results because of this two-phase setup). Therefore a mixed-design ANOVA with two between-subjects factors (musicianship and listening expertise) and two within-subjects factors (similarity computation approach and testing phase) was used.

In order to proceed with the multivariate within-subject ANOVA tests, a so-called "sphericity" assumption (stating that all the variances of the differences between the levels of the within-subjects factors are equal) was required (Huberty & Olejnik, 2006). According to the Mauchly's sphericity test, similarity ratings achieved such requirement whereas the inconsistency ratings did not and a Greenhouse-Geiser correction was needed to test the effects where inconsistency was involved.

The results from the analysis revealed that the effect of the similarity computation method on the ratings was statistically significant (Wilks Lambda = 0.005, $F(4, 2) = 93.943$, $p < 0.05$). Pairwise comparisons (a Fisher's least-significant difference test with Bonferroni correction, which conservatively adjusts the observed significance level based on the fact that multiple comparisons are made) separated the methods into 3 different groups: RANDOM and L_2-PCA (which yielded the lowest similarity ratings) versus TEMPO versus 1G-MFCC and CLAS-Pears-W$_M$ (which yielded the highest similarity ratings). The same pattern was obtained for the effects on the inconsistency ratings. The effect of the testing phase, also found to be significant, reveals that ratings yielded slightly lower values in the second phase. This could be due to the "tuning" of the similarity ratings experienced by each subject as the experiment proceeded. Fortunately, the impact of phase was uniform and did not depend on or interact with any other factor. Hence, the similarity ratings are only made "finer" or more "selective" as the experiment progresses, but irrespective of the similarity computation approach. On the other hand, the potential effects of musicianship and listening expertise revealed no impact on the similarity ratings.

Overall, we conclude that the L_2-PCA and TEMPO distances, along with a random baseline, revealed poor performance, tending to provide disruptive examples of playlist inconsistency. Contrastingly, CLAS-Pears-W$_M$ and 1G-MFCC revealed acceptable performance with slightly positive user satisfaction. Average playlist similarity ratings and proportion of inconsistent playlist for each considered approach are presented in Figure 4.3. In particular, the observed mean similarity ratings for both 1G-MFCC and CLAS-Pears-W$_M$ were equal to 3.0; the average playlist inconsistency ratings were 3.0 for 1G-MFCC and 0.3 for CLAS-Pears-W$_M$.

4.5 Semantic explanation of music similarity

Here we discuss the proposed CLAS distance and its semantic application. An interesting aspect of this proposed approach is the ability to provide a user of the final system with a concrete motivation for the retrieved tracks starting from a purely audio content-based analysis. To the best of the authors' knowledge, this aspect was rare among other music content-processing systems (Maillet et al., 2009) at the moment of accomplishing our research. However, there is evidence that retrieval or recommendation results perceived

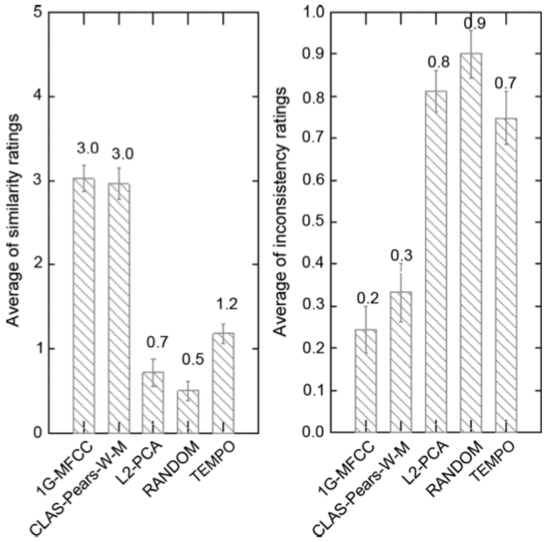
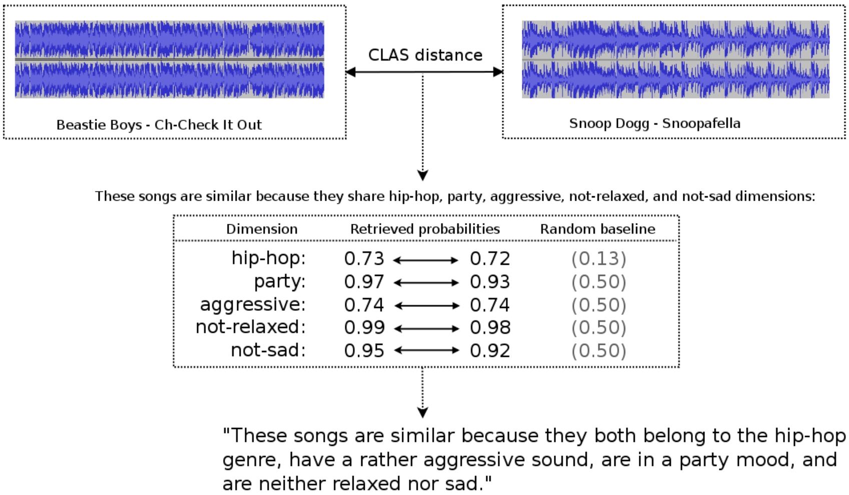


Figure 4.3: Average playlist similarity rating and proportion of inconsistent playlists for the subjective evaluation of the simple approaches. Error bars indicate one standard error of the mean.



These songs are similar because they share hip-hop, party, aggressive, not-relaxed, and not-sad dimensions:

Dimension	Retrieved probabilities	Random baseline
hip-hop:	0.73 ⟷ 0.72	(0.13)
party:	0.97 ⟷ 0.93	(0.50)
aggressive:	0.74 ⟷ 0.74	(0.50)
not-relaxed:	0.99 ⟷ 0.98	(0.50)
not-sad:	0.95 ⟷ 0.92	(0.50)

"These songs are similar because they both belong to the hip-hop genre, have a rather aggressive sound, are in a party mood, and are neither relaxed nor sad."

Figure 4.4: A real example of a semantic explanation of the similarity between two tracks retrieved from our music collection for the classifier-based distance.

as transparent (getting an explanation of why a particular retrieval or recommendation was made) are preferred by users, increasing there confidence in a system (Aman & Liikkanen, 2010; Celma & Herrera, 2008; Cramer et al., 2008; Lee, 2011; Sinha & Swearingen, 2002; Swearingen & Sinha, 2001; Tintarev & Masthoff, 2007).

Remarkably, the proposed classifier-based distance gives the possibility of providing high-level semantic descriptions for the similarity between a pair of tracks along with the distance value itself. In a final system, such annotations can be presented in terms of probability values of the considered descriptors that can be understood by a user. Alternatively, automatic text generation can be employed to present the tracks' qualities in a textual way. For a brief justification of similarity, a subset of dimensions with the highest impact on overall similarity can be selected. A simple use-case example is shown in Figure 4.4. For a pair of tracks and the CLAS-Pears-W$_M$ distance measure, a subset of 15 dimensions was determined iteratively by greedy distance minimization. In each step the best candidate for elimination was selected from different dimensions, and its weight was zeroed. Thereafter, the residual dimension probabilities that exceeded corresponding random baselines[3] can be presented to a user. Notice however that, as random baselines differ for different descriptors depending on the number of output classes of the corresponding classifier, the significance of dimension probabilities cannot be treated equally. For example, the 0.40 probability of a descriptor inferred by an 8-class classifier is considerably more significant than the 0.125 random baseline. Though not presented, the descriptors with probabilities below random baselines also have an impact on the distance measurement (probabilities close to zero in a multiclass classifier are informative because they imply that a music track does not belong to the corresponding classes). Still, such negative statements (in the sense of a low probability of a regressed dimension) are probably less suitable than positive ones for justification of music similarity to a user.

4.6 Proposed hybrid approach (HYBRID)

According to the results and observations derived from our first experiment, we advanced the possibility that a hybrid approach, combining conceptually different methods covering timbral, rhythmic, and semantic aspects of music similarity, can lead to the improvement of the similarity measurement. We select these 4 conceptually different approaches relying on the results of the objective evaluation of potential components (Section 4.4.2) and propose a hybrid distance measure. We define the distance as a weighted linear combination of L2-PCA, 1G-MFCC, TEMPO, and CLAS-Pears-W$_M$ distances.

For each selected component, we apply score normalization, following ideas presented by Fernández et al. (2006) and Arevalillo-Herráez et al. (2008). More concretely, each original distance variable d_i is equalized to a new variable $\overline{d_i} = E_i(d_i)$, uniformly distributed in $[0, 1]$. The equalizing function E_i is given by the cumulative distribution function of d_i, which can be obtained from a distance matrix on a given representative music collection. As such, we use an aggregate collection of 16K full tracks and music excerpts, composed

[3] Under the assumptions of the normal distribution of each classifier's labels for a music collection.

from the ground truth collections previously used for objective evaluation of simple approaches (Tables 3.2 and 4.1). The final hybrid distance is obtained by a weighted linear combination of component distances. As we are mostly interested in improving subjective quality of the measure, the weights are based on the results of previous subjective evaluation (Section 4.4.4) so that the measures have higher weights if they were rated higher. For each component, we have assigned an average playlist similarity rating given by listeners for this method as its weight: 0.7 for L_2-PCA, 3.0 for 1G-MFCC, 1.2 for TEMPO, and 3.0 for CLAS-Pears-W$_M$ distances.

4.7 Experiment 2: Evaluation of hybrid approach

4.7.1 Objective evaluation methodology

As a first step, we have proceeded with the same methodology as for simple approaches (Section 4.4.2), evaluating the MAP of the HYBRID measure on 17 music collections. In addition, we decided to follow a different evaluation methodology. This methodology comes from the fact that the ground truth music collections available to our evaluation, both in-house and public, can have different biases (due to different collection creators, music availability, audio formats, covered musical dimensions, how the collection was formed, etc.). Therefore, in order to minimize these effects, we carried out a large-scale cross-collection evaluation of the hybrid approach against its component approaches, namely L_2-PCA, 1G-MFCC, TEMPO, and CLAS-Pears-W$_M$, together with the random baseline (RAND) on two new large music collections. Cross-collection comparison implies that the queries and their answers belong to different music collections (out-of-sample results), thus making evaluation results more robust to possible biases.

Solely the genre musical dimension was covered in this experiment. Two large in-house ground truth music collections were employed for that purpose: (i) a collection of 299K music excerpts (30 sec.) (G-C1), and (ii) a collection of 73K full tracks (G-C2). Both collections had a genre label associated with every track. In total, 218 genres and subgenres were covered. The size of these music collections is considerably large, which makes evaluation conditions closer to a real world scenario. As queries, we randomly selected tracks from the 10 most common genres from both collections G-C1 and G-C2. The distribution of the selected genres among the collections is presented in Table 4.4. More concretely, for each genre, 790 tracks from collection G-C1 were randomly selected as queries. The number of queries per genre corresponds to a minimum number of genre occurrences among the selected genres.

Each query was sent to the collection G-C2, forming a full row in a distance matrix. As with the objective evaluation of simple approaches (Section 4.4.1), MAP was used as an evaluation measure, but was calculated with a cutoff (similarly to pooling techniques in text retrieval (Croft et al., 2010; Radlinski

Table 4.4: Number of occurrences of 10 most frequent genres, common for collections G-C1 and G-C2.

Genre	G-C1	G-C2
Reggae	2991	790
New Age	4294	1034
Blues	6229	2397
Country	8388	1699
Folk	10367	1774
Pop	15796	4523
Electronic	16050	4038
Jazz	22227	5440
Classical	43761	4802
Rock	49369	11486

& Craswell, 2010; Turpin & Scholer, 2006)) equal to the 10 closest matches due to the large dimensionality of the resulting distance matrix. The evaluation results were averaged over 5 iterations. In the same manner, a reverse experiment was carried out, using tracks from the G-C2 collection as queries, and applied to the collection G-C1. As the evaluation was completely out-of-sample, the full ground truth collections were used to train the CLAS approach.

4.7.2 Objective evaluation results

The evaluation results on the 17 music collections for the HYBRID distance are presented in Table 4.2. No statistically significant difference was found between HYBRID and CLAS-Pears-W_M for the G4, CUL, MSA, MEL, ALB, ART, RBL collections ($p > 0.05$ in the independent two-sample t-tests). HYBRID achieved an improvement in MAP for the OVI collection, but underperformed for the rest nine collections. Still, significantly better performance was achieved in comparison with the baselines and the TEMPO measure for the majority of the collections. This suggests that adding a semantic component to the measure was a relevant step in order to improve performance, and its higher weight might be appropriate for certain music collections. Using semantic similarity solely was the most efficient.

Furthermore, the results of cross-collection evaluation are presented in Table 4.5. We analyzed the obtained MAPs with a series of independent two-sample t-tests. All the approaches were found to perform with statistically significant difference ($p < 0.001$). We see that all considered distances outperform the random baseline (RAND). We found 1G-MFCC and CLAS-Pears-W$_M$ to have comparable performance, being the best among the simple approaches. As well, the TEMPO distance was found to perform similarly or slightly better than L_2-PCA. Overall, the results for simple approaches conform with our

Table 4.5: Objective cross-collection evaluation results (MAP with cutoff at 10) averaged over 5 iterations.

Distance	G-C1 → G-C2	G-C2 → G-C1
RANDOM	0.07	0.08
L_2-PCA	0.09	0.11
1G-MFCC	0.23	0.22
TEMPO	0.11	0.12
CLAS-Pears-W$_M$	0.21	0.23
HYBRID	**0.25**	**0.28**

previous objective evaluation. Meanwhile, our proposed HYBRID distance achieved the best accuracy in the cross-collection evaluation in both directions.

4.7.3 Subjective evaluation methodology

We repeated the listening experiment, conducted for simple approaches (Section 4.4.3) to evaluate the hybrid approach against its component approaches. The same music collection of 300K music excerpts (30 sec.) by 60K artists (5 tracks/artist) was used for that purpose. Each listener was presented with a series of 24 iterations, which, according to the separation of the experiment into two phases, included 12 iterations with seeds and corresponding playlists shared between all listeners, and 12 iterations with randomly selected seeds, different for each listener. In total, we collected playlist similarity ratings, playlist inconsistency indicators, and background information about musicianship and listening expertise from 21 listeners.

4.7.4 Subjective evaluation results

An ANOVA with two between-subjects factors (musicianship and listening expertise) and two within-subjects factors (similarity computation approach and testing phase) was used to test their effects on the similarity ratings and on the inconsistency ratings given by the listeners (Figure 4.5). In the case of similarity ratings, the Mauchly's sphericity test revealed that the required sphericity assumption was not achieved, and therefore a Greenhouse-Geiser correction was applied to test the effects. The only clearly significant factor explaining the observed variance in the similarity ratings was the similarity computation approach (Wilks lambda = 0.43, $F(4, 11) = 9.158$, $p < 0.005$). The specific pattern of significant differences between the tested computation approaches makes the HYBRID metric to clearly stand out from the rest, while L_2-PCA and TEMPO score low (but without statistical differences between them), and CLAS-Pears-W$_M$ and 1G-MFCC (again without statistically significant differences between them) score between the two extremes. We did not find any

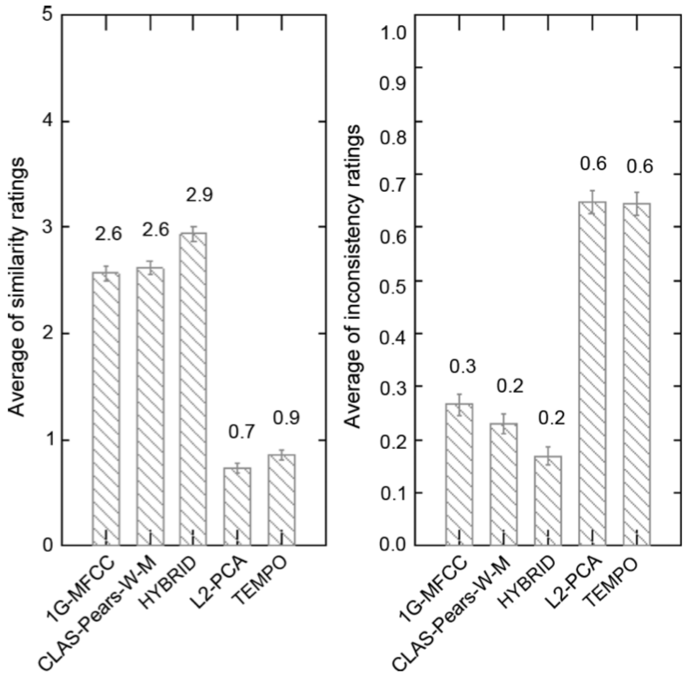


Figure 4.5: Average playlist similarity rating and proportion of inconsistent playlists for the subjective evaluation of the hybrid approach. Error bars indicate one standard error of the mean.

significant effect of musicianship and listening expertise on the similarity ratings.

In the case of inconsistency ratings, the Mauchly's sphericity test confirmed the assumption of sphericity. The same pattern and meaning, as for the similarity ratings, was found for the inconsistency ratings: they were dependent on the similarity computation approach, and most of them were generated by the L_2-PCA and TEMPO methods, whereas the HYBRID method provided significantly lower inconsistency ratings. No other factor or interaction between factors was found to be statistically significant, but a marginal interaction effect of similarity computation approach and testing phase was found. This effect means that some similarity computation methods (but not all) elicited lower ratings as the evaluation progressed. The same pattern was obtained for the inconsistency ratings. In conclusion, we found a similarity computation method (HYBRID) that was clearly preferred over the rest and no effect other than the computation method was responsible for that preference.

Again, only slightly above average user satisfaction with the similarity measures was achieved. Considering the means, the observed mean similarity ratings were 2.9 for HYBRID and 2.6 for both 1G-MFCC and CLAS-Pears-W$_M$. The observed mean playlist inconsistency was 0.2 for CLAS-Pears-W$_M$ and

HYBRID, and 0.3 for 1G-MFCC approach. The observed rating values for 1G-MFCC and CLAS-Pears-W$_M$ are close to the ones obtained in the previous experiment (Section 4.4.4) tending to being equal.

4.8 Evaluations: Audio Music Similarity at MIREX

In addition to the evaluations presented before, which are the standard or most acknowledged methods, algorithms can be evaluated in the annual research community-based initiative MIREX, coupled to the International Society for Music Information Retrieval conference (ISMIR). MIREX provides a framework for the formal evaluation of MIR systems and algorithms (Downie et al., 2010; Downie, 2008). Among other tasks, MIREX allows for the comparison of different algorithms for artist identification, genre classification, or music transcription. In particular, MIREX allows for a subjective human assessment of the accuracy of different approaches to music similarity by community members, this being one of the central task within the framework (Audio Music Similarity and Retrieval task). For that purpose, participants can submit their algorithms as binary executables and the MIREX organizers determine and publish the algorithms' accuracies in predicting human-based similarity ratings and runtimes. The underlying music collections are never published or disclosed to the participants, neither before or after the contest. Therefore, participants cannot tune their algorithms to the music collections used in the evaluation process. The history of Audio Music Similarity and Retrieval task counts six annual evaluations (in 2006, 2007, and 2009-2012) till the date of writing this thesis. We submitted the proposed HYBRID and CLAS-Pears-W$_M$ approaches during the 2009 and 2010 evaluation campaigns.

4.8.1 Methodology

In the MIREX'2009 edition, the evaluation of each submitted algorithm was performed on a music collection of 7000 tracks (30 sec. excerpts), which were chosen from IMIRSEL's[4] collections (Downie et al., 2010) and pertained to 10 different genres (700 tracks from each genre). The genres included Blues, Jazz, Country/Western, Baroque, Classical, Romantic, Electronica, Hip-Hop, Rock, and HardRock/Metal.

For each participant's approach, a 7000×7000 distance matrix was calculated. A query set of 100 tracks was randomly selected from the music collection, representing each of the 10 genres (10 tracks per genre). For each query and participant approach, the 5 nearest-to-the-query tracks out of the 7000 were chosen as candidates (after filtering out the query itself and all tracks of the same artist). All candidates were evaluated by human graders using the

[4] http://www.music-ir.org/evaluation/

Evalutron 6000 grading system (Gruzd et al., 2007). For each query, a single grader was assigned to evaluate the derived candidates from all approaches. Thereby, the uniformity of scoring within each query was ensured. For each query/candidate pair, a grader provided (i) a categorical broad score in the set {0, 1, 2} (corresponding to "not similar", "somewhat similar", and "very similar" categories), and (ii) a fine score in the range from 0 (failure) to 10 (perfection). The listening experiment was conducted by 50 graders (selected from the authors of submitted systems and their colleagues), and each one of them evaluated two queries. As this evaluation was completely out-of-sample, our submitted systems were trained on the full ground truth collections required for the CLAS distance to infer semantic descriptors. The MIREX'2010 edition followed the same evaluation methodology on the same music collection and with the same amount of graders expect for a minor change of the fine score scale to a scale from 0 (failure) to 100 (perfection).

In addition to the subjective evaluation, a number of objective measurements were conducted including neighborhood clustering by genre, number of triangle inequality violations, hubs (tracks similar to many tracks) and orphans (tracks that are not similar to any other tracks at N results) statistics.

4.8.2 MIREX 2009 results

We submitted both CLAS-Pears-W$_M$ and HYBRID distances. The overall evaluation results are reproduced in Table 4.6.[5] Our measures are noted as BSWH1 for CLAS-Pears-W$_M$, and BSWH2 for HYBRID. The results of the Friedman test against the summary data of fine scores are presented in Figure 4.6a.

First, and most importantly, we found the HYBRID measure to be one of the best performing distances in the MIREX 2009 audio music similarity task. HYBRID was very close to PS1 (Pohle & Schnitzer, 2007), but worse than the leading PS2 distance (Pohle & Schnitzer, 2009). However, no statistically significant difference between PS2, PS1 and our HYBRID measure was found in the Friedman test. Second, the CLAS-Pears-W$_M$ measure revealed satisfactory average performance comparing to other distances with no statistically significant difference to the majority of the participant approaches. In contrast to our subjective evaluation in Experiment 2 (Section 4.9.4), no statistically significant difference between our HYBRID and CLAS-Pears-W$_M$ approaches was found. Nevertheless, CLAS-Pears-W$_M$ outperformed a large group of poor performing distances with a statistically significant difference. Despite the fact that we do not observe examples of stable excellent performance in any of the participants' algorithms, above-average user satisfaction

[5] Detailed results can be found on the official results webpage for MIREX'2009: http://www.music-ir.org/mirex/2009/index.php/Audio_Music_Similarity_and_Retrieval_Results

Table 4.6: MIREX 2009 overall summary results sorted by average fine score. The proposed approaches CLAS-Pears-W_M and HYBRID are highlighted in gray.

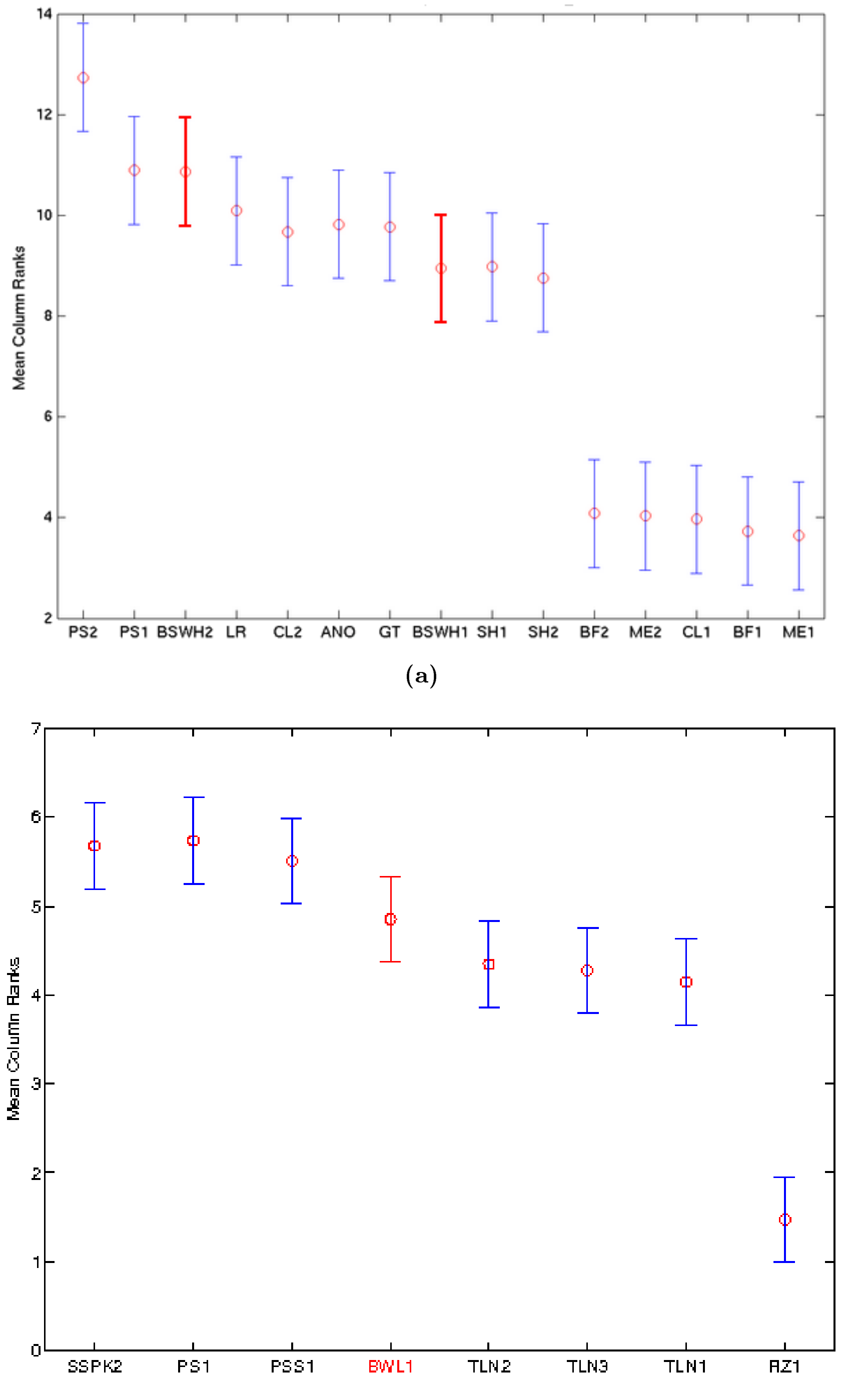
Acronym	Authors (measure)	Average fine score	Average broad score
PS2	Tim Pohle, Dominik Schnitzer (2009)	6.458	1.448
PS1	Tim Pohle, Dominik Schnitzer (2007)	5.751	1.262
BSWH2	Dmitry Bogdanov, Joan Serrà, Nicolas Wack, and Perfecto Herrera (HYBRID)	5.734	1.232
LR	Thomas Lidy, Andreas Rauber	5.470	1.148
CL2	Chuan Cao, Ming Li	5.392	1.164
ANO	Anonymous	5.391	1.126
GT	George Tzanetakis (Marsyas)	5.343	1.126
BSWH1	Dmitry Bogdanov, Joan Serrà, Nicolas Wack, and Perfecto Herrera (CLAS-Pears-W_M)	5.137	1.094
SH1	Stephan Hübler	5.042	1.012
SH2	Stephan Hübler	4.932	1.040
BF2	Benjamin Fields (mfcc10)	2.587	0.410
ME2	François Maillet, Douglas Eck (sda)	2.585	0.418
CL1	Chuan Cao, Ming Li	2.525	0.476
BF1	Benjamin Fields (chr12)	2.401	0.416
ME1	François Maillet, Douglas Eck (mlp)	2.331	0.356

Table 4.7: Metric quality of the best performing approaches submitted to MIREX'2009: mean genre match with a query within top 5, 10, 20, and 50 retrieved results, the percent of orphans, maximum hub size, and a percent of triangle inequality holdings.

Distance	CLAS	HYBRID	PS2
Mean genre match accuracy (with an artist filter):			
at top-5 results	0.445	0.510	0.499
at top-10 results	0.432	0.495	0.496
at top-20 results	0.416	0.476	0.498
at top-50 results	0.394	0.448	0.496
% of files never similar at 5 results	0.038	0.097	0.795
% of files never similar at 10 results	0.008	0.041	0.718
% of files never similar at 20 results	0.002	0.017	0.582
% of files never similar at 50 results	< 0.001	0.004	0.251
Maximum number of times a track was similar:			
at 5 results	20	43	314
at 10 results	34	93	327
at 20 results	61	170	355
at 50 results	155	310	437
Triangular inequality holding rate	90.69%	95.91%	99.96%

(with the maximum being ≈ 6.5 out of 10) was achieved by the majority of the approaches, including our HYBRID and CLAS-Pears-W_M distances.

Unlike our approach, PS1 proposes a combination of a MFCC-based audio similarity measure (Mandel & Ellis, 2005) with a Fluctuation-Pattern based similarity measure (Pampalk, 2006). In turn, PS2 expands this measure and



(a)

(b)

Figure 4.6: MIREX'2009 (a) and MIREX'2010 (b) Friedman's test on the fine scores. Figure obtained from the official results web page.

employs additional timbral information (spectral contrast features, harmonicness and percussiveness). Both measures account for symmetry, which may be promising for further performance improvement. Similarly, LR (Lidy & Rauber, 2009) employs rhythm histograms together with timbral information. We may conclude that it is advisable to incorporate richer rhythmic representation for better similarity measurement. Interestingly, our CLAS approach was found to perform comparably to the GT (Tzanetakis, 2009) based on an Euclidean distance over spectral centroid, rolloff, flux and MFCCs together with their temporal evolution on short audio segments of the tracks (average fine scores of 5.137 and 5.343, respectively). This result corroborates our findings of comparable above-average subjective quality of CLAS and timbral 1GMFCC approaches (Section 4.4.4: average similarity ratings 2.6 vs 2.6 out of 5, respectively; Section 4.7.4: average similarity ratings of 3.0 vs 3.0).

In addition, MIREX results provide insights on the quality of the proposed similarity measures (see Table 4.7). We can see that our approaches violated triangular inequalities more than the leading PS2 measure. Nevertheless, they provided a lower amount of ”orphans”, i.e., the tracks retrieval of which is significantly complicated as they never occur in the *top-N* lists. We can also assess the hubness effect, that is undesired frequent appearance of some tracks in the nearest neighbor lists of many other tracks, using the k-occurrence measure (Flexer et al., 2012). This measure represents the number of times the track occurs in the first k nearest neighbors of all the other tracks in the data base. We evidenced lower maximum k-occurrence at top 5, 10, 20, and 50 lists. This might suggest that the CLAS approach is effective against hubs.

4.8.3 MIREX 2010 results

In contrast to MIREX’2009, we solely submitted an updated version of HYBRID distance (HYBRID-2010). More exactly, we updated the classifier-based distance component adding 4 new collections, covering genre, moods, and voice gender, and timbre color: the GEL, MCL, OGD, and OTB (see Table 3.2). The overall evaluation results are presented in Table 4.8.[6] Some of the MIREX’2009 approaches were resubmitted by other participants: PS2 measure from previous evaluation is now noted as PS1; GT measure is resubmitted in three versions, TLN1, TLN2, and TLN3. In addition, some of the approaches were updated: PSS1 stands for an updated version of PS2; our updated HYBRID-2010 measure is noted as BWL1. The results of the Friedman test against the summary data of fine scores are shown in Figure 4.6b.

As expected, all submitted systems outperformed a random baseline (RZ1). We found our HYBRID-2010 distance to perform adequately well comparing to other submitted systems. Though the SSPK2, PS1, PSS1 systems outperform

[6] Detailed results can be found on the official results webpage for MIREX’2010: http://www.music-ir.org/mirex/2010/index.php/Audio_Music_Similarity_and_Retrieval_Results

Table 4.8: MIREX 2010 overall summary results sorted by average fine score. The proposed HYBRID-2010 approach is highlighted in gray.

Acronym	Authors (measure)	Average fine score	Average broad score
SSPK2	Klaus Seyerlehner, Markus Schedl, Tim Pohle, and Peter Knees	56.642	1.248
PS1	Tim Pohle, Dominik Schnitzer (2009)	55.080	1.228
PSS1	Tim Pohle, Klaus Seyerlehner, and Dominik Schnitzer (2010)	54.984	1.212
BWL1	Dmitry Bogdanov, Joan Serrà, Nicolas Wack, and Perfecto Herrera (HYBRID-2010)	49.704	1.078
TLN3	George Tzanetakis, Mathieu Lagrange, and Steven Ness (Marsyas)	46.604	0.968
TLN2	George Tzanetakis, Mathieu Lagrange, and Steven Ness (Marsyas)	46.544	0.970
TLN1	George Tzanetakis, Steven Ness, and Mathieu Lagrange (Marsyas)	45.842	0.940
RZ1	Rainer Zufall (random baseline)	16.668	0.240

our approach according to the average fine and broad scores, no statistically significant difference between these approaches and our approach was found in the Friedman test. The newly introduced SSPK2 approach (Seyerlehner et al., 2010) utilizes patterns of the spectral shape, spectral contrast, and correlation between bands, onset detection information (spectrum magnitude increments in individual bands for consequent blocks) and modified fluctuation pattern features. Again, as in the case of MIREX'2009 evaluation, none of the submitted systems performed with the user satisfaction greater than above-average (with the maximum average fine score being ≈ 56.6 out of 100).

4.8.4 Discussion

Comparing the results of MIREX'2009 and MIREX'2010 evaluations, we can see that lower average scores were obtained for the HYBRID-2010 approach than for HYBRID. Nevertheless, the same applied to other participants who have also presented their approaches to both '2009 and '2010 evaluations (e.g., PS1 and PSS1 in MIREX'2010 vs PS2 in MIREX'2009, TLN1, TLN2, and TLN3 vs GT) with no changes or minor changes. In fact, no approach submitted to the latter evaluation has achieved the best performance of the '2009. This can be explained due to the change in the ratings scales, or due to insufficient size of participants and evaluated queries in both evaluations.

In general, relying on the obtained subjective evaluation results, we may conclude that HYBRID approach is better or comparable to CLAS-Pears-W$_M$ and that both approaches are comparable to the state-of-the-art music similarity measures. Only the winning approach proposed by Pohle & Schnitzer (2009) and, expectedly, by Seyerlehner et al. (2010) overcome our classifier-based distance with statistical significance. Furthermore, our approaches are probably less hub-prone, as they returned lower k-occurrence metric. Further analysis of hubness of our measures following other metrics in the existing methodology (Flexer et al., 2012) will be of interest, but has been left out of this thesis's scope.

On the other side, the results that no statistically significant difference was found between many of approaches in the '2009 and '2010 evaluation might be taken with caution. Further insights on the statistical problems of conducted evaluations can be found in the study by Urbano et al. (2011), where discriminative power and stability of the results in the context of Audio Music Similarity and Retrieval task are discussed.

4.9 Conclusions

In this chapter we have presented, studied, and comprehensively evaluated, both objectively and subjectively, a number of new and existing content-based music similarity measures. We studied a number of simple approaches, each of which apply a uniform distance measure for overall similarity. We considered 5 baseline distances, including a random one, and explored the potential of two new conceptually different distances not strictly operating on the often exclusively used musical timbre aspects. More concretely, we presented a simple tempo-based distance which can be especially useful for expressing music similarity in collections where rhythm aspects are predominant. Using only two low-level temporal descriptors, BPM and OR, this distance is computationally inexpensive, yet effective for such collections. To this respect, our subjective evaluation experiments revealed a slight preference by listeners of tempo-based distance over a generic euclidean distance. In addition, we investigated the possibility of benefiting from the results of classification problems and transferring this gained semantic knowledge to the context of music similarity. To this end, we presented a classifier-based distance (CLAS) which makes use of high-level semantic descriptors inferred from the low-level ones. This distance covers diverse groups of musical dimensions such as genre and musical culture, moods and instruments, and rhythm and tempo. The classifier-based distance outperformed all the considered simple approaches in most of the ground truth music collections used for objective evaluation. Contrastingly, this performance improvement was not seen in the subjective evaluation when compared with the best performing baseline distance considered. In general, the classifier-based distance represents a semantically rich approach to music similarity comparable to the state-of-the-art. In spite of being based solely on audio content information, this approach can overcome the so-called "semantic gap" in content-based music similarity and provide a semantic explanation to justify the retrieval results to a user.

We explored the possibility of creating a hybrid approach, based on the studied simple approaches as potential components. We presented a new distance measure (HYBRID), which combines a low-level Euclidean distance based on principal component analysis (PCA), a timbral distance based on single Gaussian MFCC modeling, our tempo-based distance, and a high-level semantic classifier-based distance. The HYBRID distance outperformed the CLAS in the conducted large-scale cross-collection evaluation and the objective evaluation within MIREX'2009, both of which were based on a genre match between the queries and the results as the criteria. In contrast, the evaluation on the 17 other ground truth collections revealed better performance just for a single collection, comparable performance on 7 collections, and worse performance on the rest of collections. Such different results can be explained by the fact that we evaluate conceptually different similarity facets, and by the possible biases in ground truth music collections. Nevertheless, the proposed hybrid approach revealed the best performance for listeners in a subjective evaluation. Moreover, we participated in a subjective evaluation against a number of state-of-the-art distance measures, within the bounds of the MIREX'2009 and MIREX'2010 audio music similarity and retrieval tasks. The results revealed high performance of our hybrid measure, with no statistically significant difference from the best performing methods submitted. In general, the hybrid distance represents a combinative approach, benefiting from timbral, rhythmic, and high-level semantic aspects of music similarity.

The proposed CLAS and HYBRID approaches have several limitations, and we outline a number of improvements to be considered in further research:

- Improvement of the classifier-based distance by addition of more semantic musical descriptors (expanding ground truths and classifiers). Descriptors characterizing specific sub-genres, instruments, and epochs of music, will be of interest. We expect this information to be very important to be included in our CLAS approach. Given that separate dimensions can be straightforwardly combined with this distance, such additional improvements are feasible and potentially beneficial.

- Improvement of the quality of our ground truth collections and classifiers. We are aware that some of our current classifiers might not be yet enough effective due to their considerably low accuracies. In particular, some of the datasets we use are unbalanced while it is known that better classification strategies can be applied to deal with such datasets. For instance, Lin et al. (2011) propose to use easyEnsemble approach to compensate imbalance with a linear combination of SVMs trained on balanced amount of examples.

- Inclusion of metadata information for music similarity. We did not consider metadata in our study, while it is known to improve music similarity measurements, as our main intention was to improve content-based approaches. Nevertheless, the proposed CLAS approach can be straightforwardly fused with collaborative filtering approaches by adding extra dimensions in the form of user ratings or social tags.

- Further improvement of the classifier-based distance by alternative combinations of the classifiers' output probabilities.

- Improvement of the TEMPO distance component of the HYBRID approach. In our study, we have employed only basic rhythmic information (BPMs and onset rates) to construct this distance. Nevertheless, we have evidenced an importance of a richer rhythmic description (Pohle & Schnitzer, 2009) in MIREX evaluations.

- Weighting optimization for the components of HYBRID approach (in the present study it is done empirically).

- Metric optimization to reduce hubness and triangle inequality violations. In particular, mutual proximity optimization (Schnitzer et al., 2011) is shown to improve the quality of music similarity measures.

CHAPTER 5

Music recommendation based on preference examples

5.1 Introduction

In the previous chapter we considered non-personalized approaches to music similarity. We have seen that the state-of-the-art approaches, including the proposed semantic and hybrid content-based approaches are moderately effective (i.e., providing up to above-average user satisfaction) in measuring similarity between music tracks starting from raw audio. However, solving a problem of music similarity is only a part of a more complex problem of music recommendation as we have noted in Section 2.3.1. In other words, existing research studies focused on solving a problem of music similarity might not be necessarily transferable to a problem of music recommendation. While such studies can represent the query-by-example use-case of a recommender system, their evaluations are focused on measuring similarity instead of the relevance and user satisfaction by the provided recommendations. Therefore, if one wants to build a recommender system rather than an impartial search engine, evaluation of similarity approaches in the context of music recommendation, i.e., incorporating knowledge about music preferences and assessing the relevance of the recommended items, is of crucial importance. An interesting question we might ask to illustrate the conceptual difference between music similarity and distance-based music recommendation is whether user satisfaction with recommendations will correlate with the quality of underlying similarity measurements. We will address this question by comparing some of the similarity measures considered in previous chapter in the context of music recommendation.

In this chapter, we focus on recommendation approaches starting from the proposed preference elicitation strategy. Specifically, we are interested in the use-case of music recommendation and discovery when generating a list of recommendations is based on a set of preferred tracks provided by the user, as opposed to querying by a single example, and study methods based on the user profiles obtained with the proposed preference elicitation strategy. We employ both audio content and metadata information sources and mostly consider distance-based approaches. Distance-based approaches operate on track representations in a feature space. They employ similarity measurement between tracks in a target music collection and track in the preference set as a criteria for recommendation. In addition, we apply probabilistic decisions, which rely on a probability distribution of user preferences in a feature space and are able to estimate likelihood of a track being liked by the listener.

5.2 Formalization of a track recommendation problem

Let us provide a formal definition of the problem of music track recommendation, in the context of which we will consider all recommendation approaches in the present study. Given a preference set of tracks of a listener: $U = \{U_1, ..., U_M\}$ by artists $A_U = \{A_{U_1}, ..., A_{U_M}\}$ and a set of tracks in a target music collection $C = \{C_1, ..., C_L\}$ by artists $A_C = \{A_{C_1}, ..., A_{C_L}\}$, such as $U_k \neq U_l, C_k \neq C_l$ if $k \neq l$, recommend N tracks $R = \{R_1, ..., R_N\}$ by artists $A_R = \{A_{R_1}, ..., A_{R_N}\}$ such as $R_i \in C, R_i \notin U, A_R \notin A_U$ and $R_k \neq R_l, A_{R_k} \neq A_{R_l}$ when $k \neq l$. The set R should contain tracks relevant to the user. In particular, we require an artist filter to be applied for each approach for evaluation reasons. In what follows we will call the tracks in R *recommendation outcomes*, and the tracks in a preference set U *recommendation sources*.

Distance-based approaches can vary in their underlying principles, e.g., employing a distance to the user centroid (as a rough approximation of user preferences) or a distance to the user's preference set (maintaining information about all the tracks in the preference set). In turn, probabilistic approaches may utilize a probability distribution function in order to model user preference, and rank tracks in music collection with respect to this function (a Gaussian mixture model in this thesis). We highlight the difference between such strategies in Figure 5.1.

5.3 Formalization of distance-based recommendation approaches

Here we provide a formal recommendation algorithm which we follow in the majority of considered distance-based approaches. It relies on the tracks in the user's preference set and searches for N recommendations within a target music collection. For each track X in the user's preference set (a recommendation source), we apply a distance measure to retrieve the closest track C_X (a recommendation outcome candidate) from the music collection and form

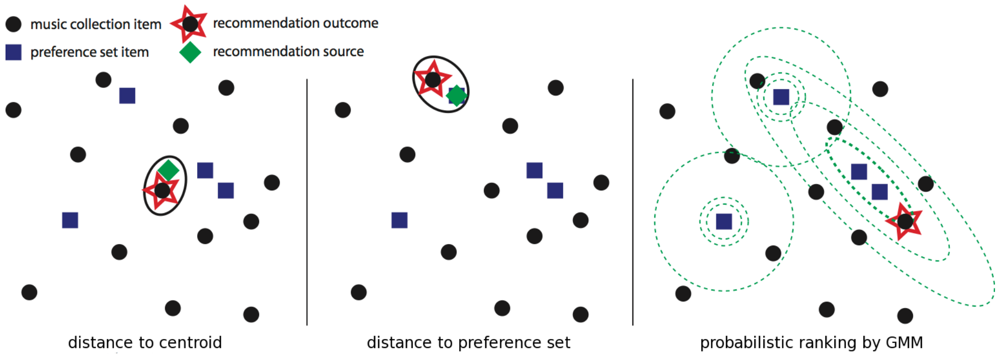


Figure 5.1: Graphical representation distance-based approaches and approaches working with probability distribution function of preference on the example of GMM (here a feature space is reduced to two dimensions and the case of one recommended track for illustration purpose). Solid lines outline recommendation outcomes (items marked by stars) and the respective recommendation sources. The dashed lines indicate regions of equal probability of the respective components of the GMM in the case of the probabilistic approach.

a triplet $(X, C_X, distance(X, C_X))$. We sort the triplets by the obtained distances, delete the duplicates of the recommendation sources (i.e., each track from the preference set produces only one recommendation outcome), and apply an artist filter. We return, as recommendations, the recommendation outcome candidates from the top N triplets. If it is impossible to produce N recommendations due to the small size of the preference set (less than N tracks) or because of the applied artist filter, we increase the number of possible recommendation outcome candidates per recommendation source (N_{O}).

Pseudo-code of the distance-based recommendation procedure.

```
set IGNORE_ARTISTS to artists in preference set
remove tracks by IGNORE_ARTISTS from music collection
set N_O to 1
set N to 15

while true:
 c set POSSIBLE_RECS to an empty list
  for track X in preference set:
    set X_NNS to N_O closest to X tracks in music collection
    for track C_X in X_NNS:
      append triplet(X,C_X,distance(X,C_X)) to POSSIBLE_RECS
  sort POSSIBLE_RECS by increasing distance

  set RECS to an empty list
  for triple(SOURCE,OUTCOME,DISTANCE) in POSSIBLE_RECS:
    if OUTCOME occurs in RECS:
      next iteration
    if SOURCE occurs in RECS >= N_O times:
      next iteration
```

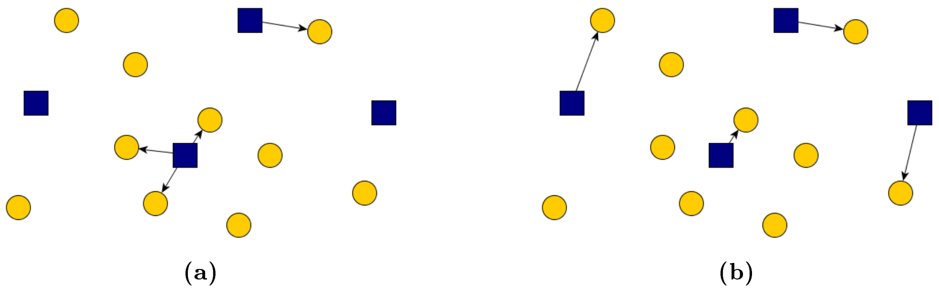


Figure 5.2: Graphical representation of distance-based approaches for $N = 4$, and $N_{\mathrm{O}} = 4$ (a) vs $N_{\mathrm{O}} = 1$ (b). The squares represent tracks from the preference set, while the circles represent tracks in target music collection. The arrows correspond to the minimum distances to the preference set resulting in recommendations in accordance with N_{O} limit.

```
    append triple (SOURCE,OUTCOME,DISTANCE) to RECS
    if length of RECS list is N:
      return outcomes from RECS as recommendations
  set N_O to N_O + 1
```

Therefore, we consider the parameter N_{O}, which limits a number of possible outcomes per preference example. The larger N_{O} is, the closer the recommendation outcomes are to the recommendation sources, which leads to more accurate (in terms of similarity) but less diverse (in terms of employed recommendation sources) recommendations. When N_{O}=N, any track from the preference set can hypothetically be a recommendation source for all N_{R} recommendations. Contrastingly, N_{O}=1 produces the most diverse results, reducing the bias produced by track density of the target music collection in the feature space. Figure 5.2 demonstrates this difference.

5.4 Proposed approaches

In this section we propose the content-based, metadata-based and hybrid approaches based on preference examples.

5.4.1 Semantic/hybrid content-based approaches

For our content-based approaches, we apply audio feature extraction as described in Section 3.4. These features cover multiple facets of music, such as timbre, rhythm, tonality, and semantic categories. We propose a number of approaches to generate music recommendations, operating on the computed low-level features and a subset of the retrieved semantic descriptors. Our idea is to apply the proposed and validated semantic classifier-based similarity

(CLAS, see Section 4.3.2). Therefore, we use the subset of descriptors, corresponding to this semantic distance (i.e., the descriptors inferred using the G1, G2, G3, CUL, MHA, MSA, MAG, MRE, MAC, MEL, RPS, RBL, OPA, and OVI collections described in Table 3.2). The distance is defined as a weighted Pearson correlation distance between vectors. Moreover, we consider the proposed hybrid distance (HYBRID, see Section 4.6) being a linear combination of CLAS with an Euclidean distance based on principal component analysis, a Kullback-Leibler divergence based on single Gaussian MFCC modeling, and a tempo-based distance. In addition, we consider a probabilistic model working on the same subset of semantic descriptors.

The proposed approaches are:

1. *Semantic distance to user centroid (SEM-MEAN).* As the simplest approach, we can summarize the user model across individual tracks to a single point in the semantic descriptor space, which can be seen as a considerably rough representation of user preferences. As such, we compute the mean point, i.e., the centroid (Salton et al., 1975), for the user's preference set. Therefore, we are able to rank the tracks according to the semantic distance to the mean point and return N nearest tracks as recommendations.

2. *Semantic distance to preference set (SEM-N and SEM-1).* Alternatively, we consider all individual tracks instead of simplifying the user model to a single point. Thus, we take into account all possible areas of preferences, explicitly specified by the user, while searching for the most similar tracks. We define a track-to-set semantic distance as a minimum semantic distance from a track to any of the tracks in the preference set. N nearest tracks are returned as recommendations according to this distance. To this end, we apply the formal procedure described in Section 5.3. As we have seen, this procedure can be tuned by specifying N_{O} to produce different levels of diversity in recommendations. We decided to evaluate both $N_{\mathrm{O}}=N$ and $N_{\mathrm{O}}=1$ scenarios marked as SEM-N and SEM-1 approaches, respectively.

3. *Semantic/low-level distance to preference set (HYBRID-1).* This approach is a counterpart of SEM-1 with the only difference in the underlying hybrid distance measure (HYBRID, see Section 4.6).

4. *Semantic Gaussian mixture model (SEM-GMM).* Finally, we propose to represent the user model as a probability density of preferences in the semantic space. We employ a Gaussian mixture model (GMM) (Bishop, 2006), which estimates a probability density as a weighted sum of a given number of simple Gaussian densities (components). The GMM is initialized by k-mean clustering, and is trained with an expectation-maximization algorithm (Bishop, 2006). We select the number of com-

ponents in the range between 1 and 20, using a Bayesian information criterion (Bishop, 2006). Once we have trained the model, we compute the probability density for each of the tracks. We rank the tracks according to the obtained density values[1] and return the N most probable tracks as recommendations.

5.4.2 Refinement by genre metadata

We consider the inclusion of metadata in purpose to refine recommendations yielded by SEM-1 and HYBRID-1. Our intention is to include the minimum amount of metadata, preferably being low-cost to gather and maintain, but however sufficiently descriptive for effective filtering. As such we decided to focus on genre/style information. Genre is often used by the listeners as a unit of expression of music preferences (Rentfrow & Gosling, 2003), and may be considered as an important preference factor related to the referential meaning of music (Section 2.2.2). A number of studies select genre to describe listeners' music preference (Dunn et al., 2011; Hoashi et al., 2003) and as an evaluation criterion for music similarity (Section 2.3.1). Current content-based classification approaches are still considerably weak in recognizing some genres (see Table 3.3 in Section 3.4.2). Furthermore, most of the existing models only deal with broad genres (such as jazz, classic, folk), and with small amount of categories (Sturm, 2012), meanwhile the models with the amount of categories larger than 10 perform poorly (for example, see Schindler et al. (2012)). To the best of our knowledge, we are not aware of research studies on classification of particular sub-genres (styles) of music, leaving aside the attempts for auto-tagging (Sordo, 2012). In an informal preliminary study, we considered such a task of sub-genre classification. In our evaluation we have observed a low accuracy ($\approx$ 20%, with a random baseline being 2%) of an SVM classifier trained to differentiate 50 musical styles on the same features as used in Section 3.4.2. We hypothesize that simple genre/style metadata tags can be a reasonable source of such information, and that it would differ from the information captured by the state-of-the-art content-based distances. While micro-level detailed genre/style information still cannot be inferred reliably by means of audio, this information can be obtained for the music collections by manual expert annotations, from social tagging services, or can be already available in the ID3[2] tags for audio files or in other metadata description formats generated in the music production stage.

Therefore, we propose a simple filtering to refine the SEM-N and HYBRID-1 approaches (marked as *SEM-GENRE-1* and *HYBRID-GENRE-1*, respectively). We apply the same sorting procedure (Section 5.3), but we solely consider the tracks of the same genre labels as possible recommendation outcomes. More-

[1] Under the assumption of a uniform distribution of the tracks in the universe within the semantic space.

[2] http://www.id3.org/

over, as discussed, we suppose that increasing the specificity of genre tags to certain degree (e.g., from “rock” to “prog rock”) would increase the quality of filtering.

To this end, we annotate the target music collection and the user’s preference set with genre tags. As a proof-of-concept, we opt for obtaining artist tags with the *Last.fm* API[3] to simulate manual single-genre annotations of each track. *Last.fm* provides tag information for both artists and tracks. We opt for artist tags due to the fact that track tags tend to be more sparse, generally more difficult to obtain, and can be insufficient for music retrieval in the long tail. To this end, we assign the same tags to the tracks as were assigned to the artists. We analyze a set of possible tags suitable for the target music collection. For each track, we select the *Last.fm* artist tags with the maximum weight (100.0) and add them to the pool of possible tags for genre annotation (“top-tags”). We then filter the pool deleting the tags with less than 100 occurrences (this threshold was selected in accordance with the top-tag histogram and the collection size) and blacklisting the tags which do not correspond to genres (“60s”, “80s”, “under 2000 listeners”, “japanese”, “spanish”, etc.) We then revise the music collection to annotate each track with a single top-tag. For each track, we consider the candidates among its artist tags, selecting the tags with the maximum possible weight, which are also present in the top-tag pool. If there are several candidates (e.g. both “rock” and “prog rock” have weight 100.0 and are present in the top-tag pool), we select the top-tag, which is the least frequent in the pool. Thereafter, we annotate the tracks from the user’s preference set in the same manner using the created pool. The idea behind this procedure is to select the most salient tags (top-tags) for the music collection, skip possible tag outliers, and annotate each track with the most specific of these top-tags keeping the maximum possible confidence level.

Pseudo-code of the genre-annotation procedure.

```
set TOP_TAGS to an empty list
for track X in music collection:
  retrieve a list X_TAGS of artist tags and their weights on Last.fm
  for tag T and weight W in X_TAGS:
    if W == 100:
      append T to TOP_TAGS

 compute histogram TOP_TAGS_HIST of tags in TOP_TAGS
 remove tags with less than 100 occurences in TOP_TAGS_HIST from TOP_TAGS
 remove tag dublicates in TOP_TAGS
 remove blacklisted tags from TOP_TAGS

for track X in music collection:
  retrieve a list X_TAGS of artist tags and their weights on Last.fm
  remove tags not present in TOP_TAGS from X_TAGS
  set X_TAGS_MAX_W to the maximum weight among tags in X_TAGS
```

[3] http://www.last.fm/api

```
remove tags with weight < X_TAGS_MAX_W from X_TAGS
sort the list X_TAGS by the ascending number of occurences in TOP_TAGS_HIST
annotate track X by the first tag in the list X_TAGS
```

5.4.3 Artist similarity based on editorial metadata (DISCOGS-1)

In addition to our previous approaches, which employed audio content information, we also consider a purely metadata-based approach. We aim for a lightweight method suitable for large-scale music collections, in particular containing the long-tail of artists and tracks, while working with publicly available data. We propose a novel artist-level recommendation approach which is based exclusively on editorial metadata. To this end, we propose to use a public database of music releases, *Discogs.com*,[4] which contains extensive user-built information on artists, labels, and their recordings. We construct a user profile using editorial metadata about the artists from the user's preference set instead of computing audio features for each track. More concretely, for each artist we retrieve a descriptive tag cloud, containing information about particular genres, styles, record labels, years of release activity, and countries of release fabrication. We then employ latent semantic analysis (LSA) (Deerwester et al., 1990; Levy & Sandler, 2008; Sordo et al., 2008) to compactly represent each artist as a vector, and match the user's preference set to a music collection to produce recommendations.

The approach we proposed works exclusively on editorial metadata found in the *Discogs.com* database. The dump of this database is released under the Public Domain license,[5] which makes is useful for different music applications, and in particular for research purposes of the MIR community. While there exist similar music services, such as public *MusicBrainz*[6] database, or proprietary *Last.fm* or *AllMusic*,[7] we opt for *Discogs* as it contains the largest catalog of music releases and artists, while being known for accurate curated metadata, which includes comprehensive annotations of particular releases.

The database contains the extensive information about up to 3,932K releases, 2,848K artists, and 468K labels.[8] In particular, for each artist this information includes a list of aliases, members (in the case an artist is a group), and group memberships (in the case an artist is a single person). Moreover it contains a list of releases featured by the artist, including albums, singles and EPs, and a list of appearances on the releases headed by other artists or compilations. A release corresponds to a particular edition of an album, single, EP, etc., and the releases related to the same album, single, or EP, can be

[4] http://discogs.com

[5] http://www.discogs.com/data/

[6] http://musicbrainz.org

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

[8] As on May 21, 2013.

grouped together into a “master release”. Each release contains genre, style, country and year information. Genres are broad categories (such as classical, electronic, funk/soul, jazz, rock, etc.) while styles are more specific categories (such as neo-romantic, tech house, afrobeat, free jazz, viking metal, etc.) In total the database counts up to 15 genre categories and 329 styles.

For each artist in the database we create a tag-cloud using genre, style, label, country, and year information related to this artist. To this end, we retrieve three lists of releases (*MAIN*, *TRACK*, *EXTRA*), where the artist occurs, respectively, as (1) main artist, heading the release, (2) track artist, for example being on a compilation or with a guest appearance on a release, (3) extra artist, being mentioned in the credits of a release (usually related to the activity such as remixing, performing, writing, arranging, producing, etc.)

For each found release related to the artist, we retrieve genre, style, label, country, and year tags. For each of the three lists, we merge releases accordingly to their master release, keeping the genres, styles, and countries, which are present in at least one of the releases (i.e., applying a set union). Concerning the release years, we attempt to approximate the authentic epoch, when the music was firstly recorded, produced, and consumed. As a master release can contain reissues along with original releases, we keep the earliest (the original) year and, moreover, propagate it with descending weights as following:

$$W_{y \pm i} = W_y * 0.75^i, i \in \{1, 2, 3, 4, 5\} \tag{5.1}$$

where W_y is the original year y, and 0.75 is a decay coefficient. For example, if the original year “y” is 1995, the resulting year-tag weights will be $W_{1995} = 1.0, W_{1994} = W_{1996} = 0.75, W_{1993} = W_{1997} \approx 0.56, W_{1992} = W_{1998} \approx 0.42, W_{1991} = W_{1999} \approx 0.32, W_{1990} = W_{2000} \approx 0.24$.

Thereafter, we summarize *MAIN*, *TRACK*, and *EXTRA* lists of the artist to a single tag-cloud. We assume a greater importance of tag annotations for the main artist role in comparison to track artists or extra artists; e.g., tags found on an artist’s album are more important than the ones found on a compilation. We empirically assign the weights to these three groups of artist roles: 1.0 for main artists and 0.5 for both track and extra artists. As well, we assign further weights to tags according to their category: 1.0 for genres, styles, and labels, and 0.5 for years and countries, rescaling the artist tag-cloud. In particular, we decided to give equal importance to label information as to genres and styles. The rational behind grounds on the hypothesis that record label information gives a very valuable clue to a type of music, especially in the long-tail for the case of niche labels.

Finally, we propagate artist tags using the artist relations found in the database, such as aliases and membership relations. We suppose related artists to share similar musical properties and, therefore, assure that artists with low amount of releases will obtain reasonable amount of tags. To this end, for each artist we add a set of weighted tag-clouds of all related artists to the associated tag-cloud. We select a propagation weight of 0.5 and apply only 1-step propagation; i.e. tags will be propagated only between artists sharing a direct relation. Figure 5.3 presents an example of the proposed annotation procedure.

Following the described procedure we are able to construct tag-clouds for each artist in the *Discogs* database which together form a sparse tag matrix. An example of generated tag-cloud is presented in Figure 5.4. More examples can be found in Appendix B. To simplify the obtained matrix, for each artist we apply additional filtering by means of erasing the tags with weight less than 1% of the artist's tag with the maximum weight. We then apply latent semantic analysis (Deerwester et al., 1990) to reduce the dimensionality of the obtained tag matrix to 300 latent dimensions. Originally being applied for natural language processing, this technique allows for finding a set of concepts (latent variables) by analysis of relationships between a set of documents and the terms they contain. LSA relies on singular value decomposition of a document/term matrix and, in principle, assumes that words that are close in meaning will occur close together in text. Levy & Sandler (2008) and Sordo et al. (2008) have demonstrated the application of LSA to music domain, in particular, for tag-based music similarity. Similarly, we apply LSA on our artist tag matrix. Afterwards, Pearson correlation distance (Celma, 2008; Gibbons & Chakraborti, 2003) can be computed on the resulting topic space as a measure of similarity between artists. Once we have matched the annotated artists to the tracks in our music collection and the user's preference set, we retrieve recommendations applying the tag-based distance by the formal procedure (Section 5.3).

5.4.4 Possible approaches out of consideration

In the pre-analysis we have also considered several other probabilistic models, which included kernel density estimation and Bayesian networks (Bishop, 2006), and discriminative approaches, such as one-class SVM (Chang & Lin, 2011), representation of music preferences as a convex hull in the feature space, and density-based clustering (Ertoz et al., 2003) starting from the preference set. These approaches were discarded from subjective evaluation due to no evident advantage in an informal evaluation by the author and close collaborators.

5.5 Baseline approaches

We consider a number of baseline approaches to music recommendation working on audio content and metadata. Specifically, we take two content-based approaches working on the low-level timbral description (MFCCs), which are the standard descriptors for lots of MIR tasks including music recommendation (Section 2.3.1). Considering metadata-based approaches, one of the baselines

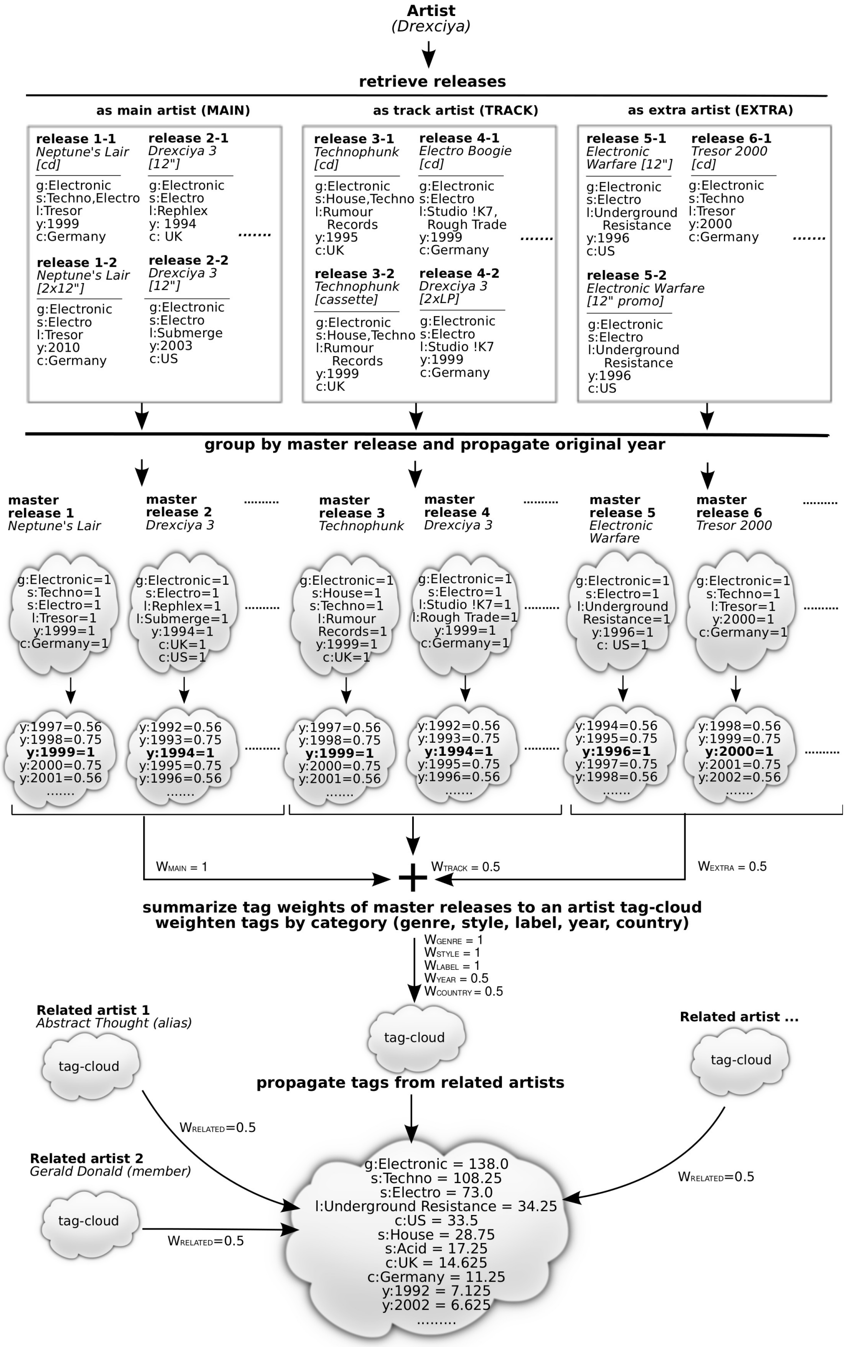

Artist
(Drexciya)
retrieve releases
as main artist (MAIN)
as track artist (TRACK)
as extra artist (EXTRA)
release 1-1 Neptune's Lair [cd]
g:Electronic s:Techno,Electro l:Tresor y:1999 c:Germany
release 2-1 Drexciya 3 [12"]
g:Electronic s:Electro l:Rephlex y: 1994 c: UK
release 1-2 Neptune's Lair [2x12"]
g:Electronic s:Electro l:Tresor y:2010 c:Germany
release 2-2 Drexciya 3 [12"]
g:Electronic s:Electro l:Submerge y:2003 c:US
release 3-1 Technophunk [cd]
g:Electronic s:House,Techno l:Rumour Records y:1995 c:UK
release 4-1 Electro Boogie [cd]
g:Electronic s:Electro l:Studio !K7, Rough Trade y:1999 c:Germany
release 3-2 Technophunk [cassette]
g:Electronic s:House,Techno l:Rumour Records y:1999 c:UK
release 4-2 Drexciya 3 [2xLP]
g:Electronic s:Electro l:Studio !K7 y:1999 c:Germany
release 5-1 Electronic Warfare [12"]
g:Electronic s:Electro l:Underground Resistance y:1996 c:US
release 6-1 Tresor 2000 [cd]
g:Electronic s:Techno l:Tresor y:2000 c:Germany
release 5-2 Electronic Warfare [12" promo]
g:Electronic s:Electro l:Underground Resistance y:1996 c:US
group by master release and propagate original year
master release 1 Neptune's Lair
master release 2 Drexciya 3
master release 3 Technophunk
master release 4 Drexciya 3
master release 5 Electronic Warfare
master release 6 Tresor 2000
g:Electronic=1 s:Techno=1 s:Electro=1 l:Tresor=1 y:1999=1 c:Germany=1
g:Electronic=1 s:Electro=1 l:Rephlex=1 l:Submerge=1 y:1994=1 c:UK=1 c:US=1
g:Electronic=1 s:House=1 s:Techno=1 l:Rumour Records=1 y:1999=1 c:UK=1
g:Electronic=1 s:Electro=1 l:Studio !K7=1 l:Rough Trade=1 y:1999=1 c:Germany=1
g:Electronic=1 s:Electro=1 l:Underground Resistance=1 y:1996=1 c: US=1
g:Electronic=1 s:Techno=1 l:Tresor=1 y:2000=1 c:Germany=1
y:1997=0.56 y:1998=0.75 y:1999=1 y:2000=0.75 y:2001=0.56
y:1992=0.56 y:1993=0.75 y:1994=1 y:1995=0.75 y:1996=0.56
y:1997=0.56 y:1998=0.75 y:1999=1 y:2000=0.75 y:2001=0.56
y:1992=0.56 y:1993=0.75 y:1994=1 y:1995=0.75 y:1996=0.56
y:1994=0.56 y:1995=0.75 y:1996=1 y:1997=0.75 y:1998=0.56
y:1998=0.56 y:1999=0.75 y:2000=1 y:2001=0.75 y:2002=0.56
WMAIN = 1
WTRACK = 0.5
WEXTRA = 0.5
summarize tag weights of master releases to an artist tag-cloud
weighten tags by category (genre, style, label, year, country)
WGENRE = 1
WSTYLE = 1
WLABEL = 1
WYEAR = 0.5
WCOUNTRY = 0.5
Related artist 1
Abstract Thought (alias)
tag-cloud
tag-cloud
Related artist ...
tag-cloud
propagate tags from related artists
WRELATED=0.5
Related artist 2
Gerald Donald (member)
tag-cloud
WRELATED=0.5
WRELATED=0.5
g:Electronic = 138.0
s:Techno = 108.25
s:Electro = 73.0
l:Underground Resistance = 34.25
c:US = 33.5
s:House = 28.75
s:Acid = 17.25
c:UK = 14.625
c:Germany = 11.25
y:1992 = 7.125
y:2002 = 6.625

Figure 5.3: An example of the proposed artist annotation based on editorial metadata from *Discogs*. Three lists of releases (MAIN, TRACK, EXTRA) are retrieved according to an artist's role. Particular releases are summarized into master releases, merging all found genre, style, label, and country tags, and selecting and propagating original year. Thereafter, tags are weighted to form a tag-cloud of an artist, and suimmed with the propagated tags of all related artists. Letters "g", "s", "l", "y", "c" stand for genre, style, label, year and country tags respectively.

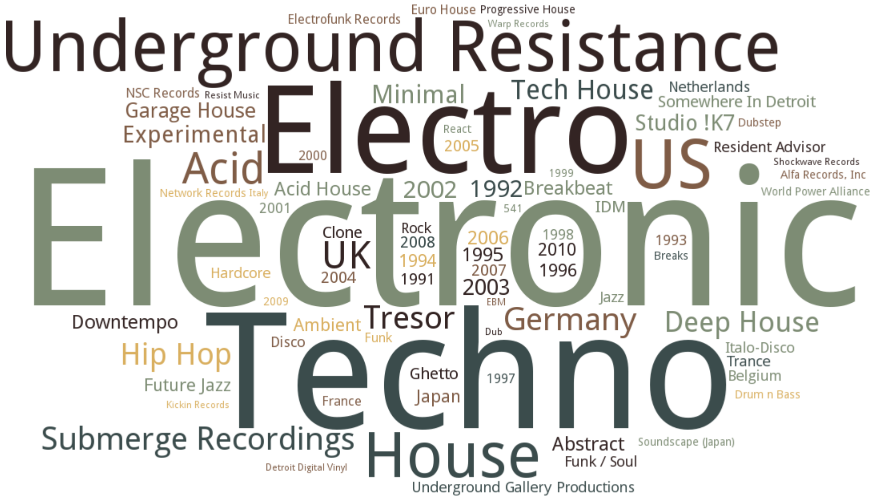


Figure 5.4: An example of tag-cloud generated for the artist Drexciya following the described annotation process.

is constructed exclusively using information about the listener's genre preferences. The other two are based on the information about preferred tracks and artists (taken from the editorial metadata provided by the user for the preference set). They partially employ collaborative filtering information, querying commercial state-of-the-art music recommenders for similar music tracks.

5.5.1 Low-level timbral approaches

We consider two audio content-based baseline approaches. These approaches apply the same ideas as the proposed semantic approaches, but operate on low-level timbral features, frequently used in the related literature.

1. *Timbral distance to preference set (MFCC-N).* This approach is a counterpart to the proposed *SEM-N* approach using a common low-level timbral distance (Aucouturier et al., 2005; Pampalk, 2006) instead of the semantic one. The tracks are modeled by probability distributions of MFCCs using single Gaussian with full covariance matrix. For such representations a distance measure can be defined using a closed form approximation of the Kullback-Leibler divergence. This baseline resembles the state-of-the-art timbral user model, proposed by Logan (2004), which uses the Earth-Mover's Distance between MFCC distributions as a distance.

2. *Timbral Gaussian mixture model (MFCC-GMM).* Alternatively, we consider a counterpart to the proposed *SEM-GMM* probabilistic approach: we use a population of mean MFCC vectors (one vector per track from the user's preference set) to train a timbral GMM.

5.5.2 Random genre-based recommendation (GENRE-1)

This simple and low-cost approach provides quasi-random recommendations relying on genre categories of the user's preference set. We assume that all tracks in the target music collection are manually tagged at least with a genre label by an expert (as it was in the case of the collections we employed for our evaluations). We randomly preselect N tracks from the preference set and obtain their genre labels. For each of the N preselected tracks, we return a random track of the same genre label.

Ideally, tracks from the preference set should contain manual genre annotations by an expert as well. Moreover, the annotations should be consistent with the ones in the music collection to be able to match the tracks by genre. Nevertheless, the tracks from the preference set, since they were submitted by the user, do not necessarily contain a genre tag, and the quality of such tags and their consistency with the genres in the music collection cannot be assured. Therefore, we retrieve this information from the Web. We use track pages or artist pages from the social music tagging system *Last.fm* as the source of genre information. We run queries using metadata of the preselected tracks, and select the most popular genre tag, which is presented among genre tags of the given music collection.

Note that oppositely to our SEM-GENRE-1 and HYBRID-GENRE-1 approaches, we now assumed that the target collection is tagged by an expert (as it is in the case of our in-house collection) and, therefore, we did not retrieve any genre tags for the target collection from *Last.fm*. When retrieving them for the preference set, we do not aim for greater specificity and take the most popular tags.

5.5.3 Artist similarity based on social tags (LASTFM-TAGS-1)

We consider a purely metadata-based similarity measure working on the artist level. This approach is based on social tags provided by the *Last.fm* API, retrieved for the artists from the user's preference set and the target music collection. Using the API, we obtain a weight-normalized tag list for each artist. The weight ranges in the $[0, 100.0]$ interval, and we select a minimum weight threshold of 10.0 to filter out possibly inaccurate tags. The resulting tags are then assigned to each track in the preference set and the music collection. We then apply latent semantic analysis (Deerwester et al., 1990; Levy & Sandler, 2008; Sordo et al., 2008) to reduce dimensionality to 300 latent dimensions, similarly to the proposed DISCOGS-1. Pearson correlation distance can be applied on the resulting topic space. We retrieve recommendations following the proposed formal distance-based procedure.

5.5.4 Collaborative filtering black-box approaches

As we highlighted in Section 2.3.2, it is often problematic to employ collaborative filtering approaches in research studies, as such data is generally proprietary. The existing listening behavior dataset provided by Celma (2008), and the recent The Million Song Dataset[9] (McFee et al., 2012b), are not suited to our needs due to poor intersection with the preference sets of our participants and our target collection. In contrast, we have decided to use black box recommendations provided by *iTunes Genius*, which provided almost ideal intersection. We also used *Last.fm* as another baseline for our evaluation.

1. *Black-box track similarity by iTunes Genius (GENIUS-BB-1).* We consider commercial black-box recommendations obtained from the *iTunes Genius*[10] playlist generation algorithm. Given a music collection and a query, this algorithm is capable to generate a playlist by means of the underlying music similarity measure, which works on metadata and partially employs collaborative filtering of large amounts of user data (music sales, listening history, and track ratings) (Barrington et al., 2009). From the preference set we randomly select N tracks annotated by artist, album, and track title information, sufficient to be recognized by Genius. When some tracks appear to be not identified by Genius (in rare occasions), they are ignored from the consideration by this approach, and other recognizable tracks are selected instead. For each of the selected tracks (a recommendation source), we generate a playlist, apply the artist filter, and select the top track in the playlist as the recommendation outcome. We increase the amount of possible outcomes per source when it is impossible to produce N recommendations.

2. *Black-box track similarity from Last.fm (LASTFM-BB-1). Last.fm* is an established music recommender with an extensive number of users, and a large playable music collection, providing means for both monitoring listening statistics and social tagging (Jones & Pu, 2007). In particular, it provides track-to-track[11] and artist-to-artist[12] similarity computed by an undisclosed algorithm, which is partially based on collaborative filtering, but does not use any audio content.[13] It is important to notice that the underlying music collection of *Last.fm* used in this baseline approach differs (being significantly larger and broader) from the collection used by the other approaches in our evaluation. As *Last.fm* do not provide concrete distance values but only ranked lists of artists or tracks, we are not able to apply our formal distance-based procedure. Instead, we

[9] This dataset was not yet available at the moment of our experiment.

[10] http://www.apple.com/itunes/features/

[11] For example, |http://last.fm/music/Grandmaster+Flash/_/The+Message/+similar

[12] For example, http://last.fm/music/Baby+Ford/+similar

[13] At least, at the moment of conducting the present research.

randomly preselect N tracks from the preference set and independently query *Last.fm* for each of them to receive a recommendation. For each track we select the most similar recommended track among the ones with an available preview.[14] If no track-based similarity information is available (e.g., when the query track is an unpopular long-tail track with a low number of listeners), we query for similar artists. In this case we choose the most similar artist and select its most popular track that has an available preview.

5.6 User-based evaluation methodology

5.6.1 Procedure

Here we describe the methodology for the subjective evaluation of our proposed approaches to music recommendation. Ideally, subjective evaluations should employ considerably large amount of listeners and recommended tracks. Nevertheless, they are very costly because they require a huge user effort, and therefore large-scale evaluations are hardly feasible in academic research (as discussed in Section 2.3.2). In this situation, we decided to work with small groups of participants (we were only able to employ a small amount of participants due to the demanding task asked to the subjects), splitting our evaluation into a series of experiments each of which was conducted following the same procedure. For consistency, we focus on the task of retrieving $N = 20$ or $N = 15$ music tracks from a target music collection as recommendations for each participant. Selecting a larger number of recommended tracks (N) was problematic as this number should be balanced with an amount of approaches under consideration. As the source for recommendations, we employed two large in-house music collections, covering a wide range of genres, styles, arrangements, geographic locations, and musical epochs. These collections consist of $100,000$ and $68,000$ music excerpts, respectively.

In each experiment we evaluate a subset of the proposed approaches and baselines. For each subject, we compute the user profiles from the provided preference set as required by each approach under evaluation. Each of the considered approaches generates a playlist containing N music tracks. Following a usual procedure for evaluation of music similarity measures and music recommendations, we apply an artist filter (Pampalk, 2006) to assure that no playlist contained more than one track from the same artist nor tracks by the artists from the preference set for a particular user. These playlists are merged into a single list, in which tracks are randomly ordered and anonymized, including filenames and metadata. The tracks offered as recommendations are equally likely to come from each single recommendation approach. This allows us to

[14] These previews are downloadable music excerpts (30 sec.), which are later used in our subjective evaluation for the case of the LASTFM-BB-1 approach.

avoid any response bias due to presentation order, recommendation approach, or contextual recognition of tracks (by artist names, etc.) by the participants. In addition, the participants are not aware of the amount of recommendation approaches, their names and their rationales. We adopt such a blind evaluation process in order to avoid any possible biases, including the factor of familiarity with the music which is evidenced to affect music appreciation (Section 2.2). We do not restrict the duration of the evaluation process. Therefore, each subject can spend any amount of time to assess subjective appraisal for the recommended music.

5.6.2 Subjective ratings

To gather feedback on recommendations, we provide a questionnaire for the subjects to express their subjective impressions related to the recommended music (see Table 5.1). For each recommended track the participants are asked to provide four ratings:

- *Familiarity* ranged from 0 to 4; with 0 meaning absolute lack of familiarity, 1 feeling familiar with the music, 2 knowing the artist, 3 knowing the title, and 4 being able to identify the artist and the title.

- *Liking* measured the enjoyment of the presented music with 0 and 1 covering negative liking, 2 representing a neutral position, and 3 and 4 representing increasing liking for the musical excerpt.

- *Listening intentions* measured the readiness of the participant to listen to the same track again in the future. This measure is more direct and behavioral than the *liking*, as an intention is closer to action than just the abstraction of liking. Again the scale contained 2 positive and 2 negative steps plus a neutral one.

- *"Give-me-more"* with 1 indicating request for more music like the presented track, and 0 indicating reject of such music.

The users are also asked to provide the track title and artist name for those tracks rated high in the familiarity scale.

We propose using three ratings related to the listener satisfaction with provided recommendations in contrast to the existing studies which utilize only one rating, e.g., "satisfaction" (Section 2.3.2). Our choice can be motivated by the fact that preference is multifaceted (Rentfrow et al., 2011). We consider not only the fact of liking, but also behavioral aspects, which we believe to be important. Three proposed ratings are not necessarily correlated: our evaluations revealed that an discrepancy in subjective ratings occurred in approximately 19% of the cases. Furthermore, as we are focused on music discovery, we also measure familiarity of the listener with provided recommendations. This is rarely done in the existing literature as it is hard to assess objectively

Table 5.1: Meaning of *familiarity*, *liking*, *listening intentions*, and *"give-me-more"* ratings as given to the participants.

Rating	Value	Meaning
Familiarity	4	I know the song and the artist
	3	I know the song but not the artist
	2	I know the artist but not the song
	1	It sounds familiar to me even I ignore the title and artist (maybe I heard it in TV, in a soundtrack, long time ago...)
	0	No idea
Liking	4	I like it a lot!
	3	I like it
	2	I would not say I like it, but it is listenable
	1	I do not like it
	0	It is annoying, I cannot listen to it!
Listening intentions	4	I am going to play it again several times in the future
	3	I probably will play it again in the future
	2	It doesn't annoy me listening to it, although I am not sure about playing it again in the future
	1	I am not going to play it again in the future
	0	I will skip it in any occasion I find in a playlist
Give-me-more	1	I would like to be recommended more songs like this one
	0	I would not like to be recommended more songs like this one

without real participants. In contrast, we employ this subjective rating in our evaluations.

5.6.3 Recommendation outcome categories

After gathering questionnaires filled by participants, we propose to recode the provided subjective ratings to four outcome categories: *hits*, *trusts*, *fails*, and *unclear* recommendations. First, we manually correct familiarity ratings when the artist/title provided by a user is incorrect compared to the actual ones. In such situations, a familiarity rating of 3, or, more frequently, 4 or 2, is lowered to 1 (in the case of incorrect artist and track title) or 2 (in the case of correct artist, but incorrect track title). We expected a low number of corrections to be done (above 5%), which is supported by our further experiments.

The four gathered subjective ratings can be used to characterize different aspects of the considered recommendation approaches. We expect a good recommender system to provide high liking, listening intentions, and "give-me-more" ratings. Moreover, if we focus on music discovery, low familiarity ratings are desired, which will guarantee the novelty of relevant (liked) recommendations. We recode the participants' ratings for each evaluated track into three categories which refer to the type of the recommendation: *hits*, *fails*, and *trusts*. We define a recommended track to be a hit when it received low familiarity ratings (< 2) and high liking (> 2), listening intentions (> 2), and "give-me-more" ($= 1$) ratings simultaneously. Similarly, trusts are the tracks with high liking, listening intentions, "give-me-more", but as well high familiarity (> 1). Trusts, provided their overall amount is low, can be useful for a user to feel that the recommender is understanding his/her preferences (Barrington et al., 2009; Cramer et al., 2008) (i.e., a user could be satisfied by getting a trust track from time to time, but annoyed if every other track is a trust). Fails are the tracks which received low liking (< 3), listening intentions (< 3) and "give-me-more" ($= 0$) ratings. In any other case (e.g., a track received high liking, but low listening intentions and "give-me-more") the outcome category is considered to be "unclear".

We expect a good recommender to get a large amount of hits, and considerable, though not excessive, amount of trusts, in the case of music discovery. In the case of playlist generation, more trusts are acceptable. In general, the desired amount of trusts is dependent on the final application. Considering the unclear category, we may not expect such tracks to be as relevant as hits and trust categories because such recommendations consisted of the tracks with inconsistent ratings. Still, such tracks can be useful for certain scenarios (e.g., playlist generation), but are probably not well suited for others (e.g., digital music vending). In the extreme case, we can assume both fails and unclear categories to be unwanted outcomes in contrast to trusts and hits, which are wanted outcomes. Recoding subjective ratings into proposed categorical variables brings us qualitative advantages in the analysis of results: it is easier to assess the amount of novel and trusted recommendations, and the character of such recommendations as we can do further analysis inside each category.

5.6.4 Discussion

5.7 Experiment 3: Advantage of semantic content description

5.7.1 Rationale and procedure

In our first experiment on music recommendation, we hypothesize the benefits of the proposed semantic description of audio content over the exclusive use of low-level timbral description. We evaluate here three proposed approaches, SEM-MEAN, SEM-N, and SEM-GMM, against the baselines working on timbral audio features (MFCCs): MFCC-N and MFCC-GMM. In addition, we consider two metadata-based baselines: the simplest genre-based recommender (GENRE-1) and a commercial black-box recommender working on collaborative filtering information and social tags (LASTFM-BB-1[15]). It is important to notice that the underlying music collection of *Last.fm* recommender used in this baseline approach differs (being significantly larger and broader) from the collection used by the other approaches in our evaluation. Therefore we will consider the results obtained for LASTFM-BB-1 only as tentative.

We performed subjective listening tests on 12 subjects in order to evaluate the considered approaches with N being set to 20. Our population consisted of 8 males and 4 females with the average age of 34 years ($\mu = 33.83, \sigma = 5.2$) and a high interest in music ($\mu = 9.58, \sigma = 0.67$) being a subset of the population described in Section 3.3.2. The mean size of the provided preference sets was $\mu = 73.58$, $\sigma = 45.66$, and the median was 57 tracks. As the source for recommendations, we employed a large in-house music collection, covering a wide range of genres, styles, arrangements, geographic locations, and musical epochs. This collection consists of $100,000$ music excerpts (30 sec.) by $47,000$ artists with approximately 2 tracks per artist.

5.7.2 Results

Following the described methodology (see Section 5.6.3), manual corrections of familiarity rating represented only 3% of the total familiarity judgments. Recoding subjective ratings to outcome categories, 18.3% of all the recommendations were considered as "unclear". Most of the unclear recommendations (41.9%) consisted of low liking and intention ratings (< 3 in both cases) followed by a positive "give-me-more" request; other frequent cases of unclear recommendation consisted of a positive liking (> 2) that was not followed by positive intentions and positive "give-me-more" (15.5%) or positive liking not followed by positive intentions though positive "give-me-more"(20.0%). We excluded the unclear recommendations from further analysis.

We report the percent of each outcome category per recommendation approach in Table 5.2 and Figure 5.5. An inspection of it reveals that the approach which yields the largest amount of hits (41.2%) and trusts (25.4%) is LASTFM-BB-1. The trusts found with other approaches were scarce, all below 4%. The approaches based on the proposed semantic user model (*SEM-N*, *SEM-MEAN* and *SEM-GMM*) yielded more than 30% of hits, and the remaining ones did not surpass 25%. In our experiments we need to ensure statistically the existence of association between the recommendation approach and the obtained recommendation outcome. To this end, we used Pearson chi-square test, which demonstrated that the existence of an association between

[15] All experiments were conducted on May 2010.

Table 5.2: The percent of fail, trust, hit, and unclear categories per recommendation approach. Note that the results for the *LASTFM-BB-1* approach were obtained on a different underlying music collection.

Approach	fail	hit	trust	unclear	hit+trust
SEM-MEAN	49.2	31.2	2.5	17.1	33.7
SEM-N	42.5	34.6	3.3	19.6	37.9
SEM-GMM	48.8	30.0	2.5	18.7	32.5
MFCC-N	64.1	15.0	2.1	18.8	17.1
MFCC-GMM	69.6	11.7	1.2	17.5	12.9
LASTFM-BB-1	16.7	41.2	25.4	16.7	66.6
GENRE-1	53.8	25.0	1.2	20.0	26.2

Table 5.3: Mean ratings per recommendation approach. Note that the results for the *LASTFM-BB-1* approach were obtained on a different underlying music collection.

Approach	liking	intentions	give-me-more	familiarity
SEM-MEAN	2.18	2.01	0.46	0.23
SEM-N	2.34	2.14	0.53	0.34
SEM-GMM	2.30	2.13	0.45	0.25
MFCC-N	1.78	1.65	0.30	0.31
MFCC-GMM	1.59	1.45	0.24	0.17
LASTFM-BB-1	2.99	2.91	0.77	1.31
GENRE-1	1.98	1.84	0.43	0.15

recommendation approach and the type of outcome of the recommendation was statistically significant ($\chi^2(18) = 351.7$, $p < 0.001$). Additionally, we performed three separate between-subjects ANOVA tests in order to assess the effects of the recommendation approaches on the liking, intentions, and "give-me-more" subjective ratings, coupled with a Tukey's test for pairwise comparisons. The effect was confirmed in all of them ($F(6, 1365) = 55.385$, $p < 0.001$ for the liking rating, $F(6, 1365) = 48.89$, $p < 0.001$ for the intentions rating, and $F(6, 1365) = 43.501$, $p < 0.001$ for the "give-me-more" rating). Pairwise comparisons using Tukey's test revealed the same pattern of differences between the recommendation approaches, irrespective of the 3 tested indexes. This pattern highlights the *LASTFM-BB-1* approach as the one getting the highest overall ratings. It also groups together the timbral *MFCC-GMM* and *MFCC-N* approaches (those getting the lowest ratings), and the remaining approaches (*SEM-N*, *SEM-MEAN*, *SEM-GMM*, and *GENRE-1*) are grouped in-between. The normality assumption required for between-subject ANOVA (normal distribution of the ratings for each approach) was considered as trusted by visual inspection of the respective Q-Q plots. The mean values of the obtained liking, listening intentions, and "give-me-more" ratings per each approach are presented in Table 5.3 and Figure 5.6. Additionally, Figure 5.7 presents histograms of these subjective ratings.

Finally, focusing on the discovery use-case of music recommender, a measure of the quality of the hits was computed by multiplying the difference of liking and familiarity by listening intentions for each recommended track. This quality score ranks recommendations considering that the best ones correspond to the tracks which are highly-liked though completely unfamiliar, and intended to be listened again (i.e., a very highly liked track, which was totally unfamiliar, and intended to be listened again would yield the highest quality score; contrastingly, a very highly liked track which was highly familiar and intended to be listened again would yield a lower quality score, as the recommender would be unnecessary in this case). Selecting only the hits, an ANOVA on the effect of the recommendation approach on this quality measure revealed no significant differences between any of the approaches. Therefore, considering the quality of hits, there is no recommendation approach granting better or worst recommendations than any other. The same pattern was revealed by solely using the liking as a measure of the quality of the hits.

5.7.3 Discussion

The evaluation results revealed the users' preference for the proposed semantic approaches over the low-level timbral baselines. This fact supports our hypothesis on the advantage of using a semantic description for music recommendation. Moreover, it complements the outcomes of our research on semantic music similarity measures presented in Chapter 4: we have previously observed the advantage of semantic similarity over MFCC-based similarity in objective evaluations, although no statistically significant differences were found in subjective listening tests. In the present experiment, we may conclude that the high-level semantic description outperforms the low-level timbral description in the task of music recommendation and that it is well-suited for music preference elicitation.

Interestingly, considering the amount of hits+trusts, SEM-N performed the best compared to SEM-MEAN and SEM-GMM. In the case of SEM-MEAN, we expected such results, as a centroid can be considered a rough approximation of music preferences in comparison to the complete preference set. In the case of SEM-GMM, such results corroborate previous research studies which reveal that distance-based approaches are to be preferred to the probabilistic models in the case of noisy audio features, at least in the task of automatic content-based tag-propagation (Sordo, 2012). We conclude that it is reasonable to maintain and exploit information about particular tracks in the listener's profile and apply distance-based approaches to music recommendation.

Comparing with the baselines working on metadata, we found that the proposed approaches perform better than the simple genre-based recommender

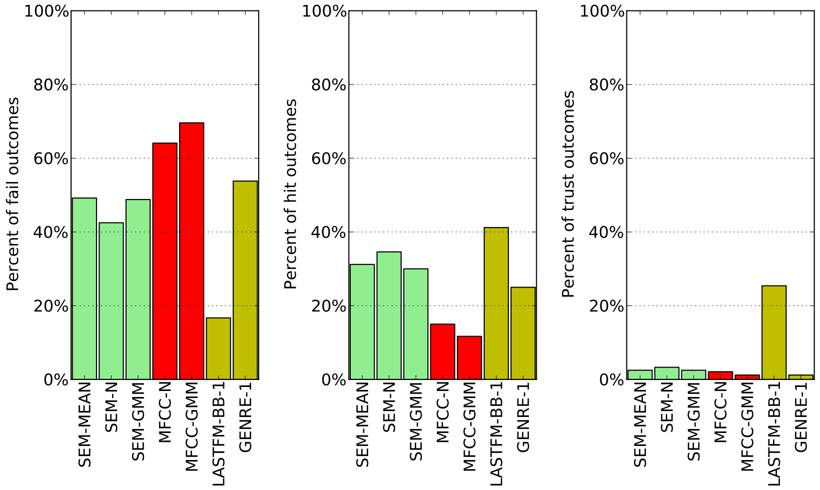


Figure 5.5: The percent of fail, trust, hits, and clear categories per recommendation approach. The results for the *LASTFM-BB-1* approach were obtained on a different underlying music collection. The proposed content-based approaches, the content-based baselines, and the metadata-based baselines are differentiated by color (green, red, and pistachio).

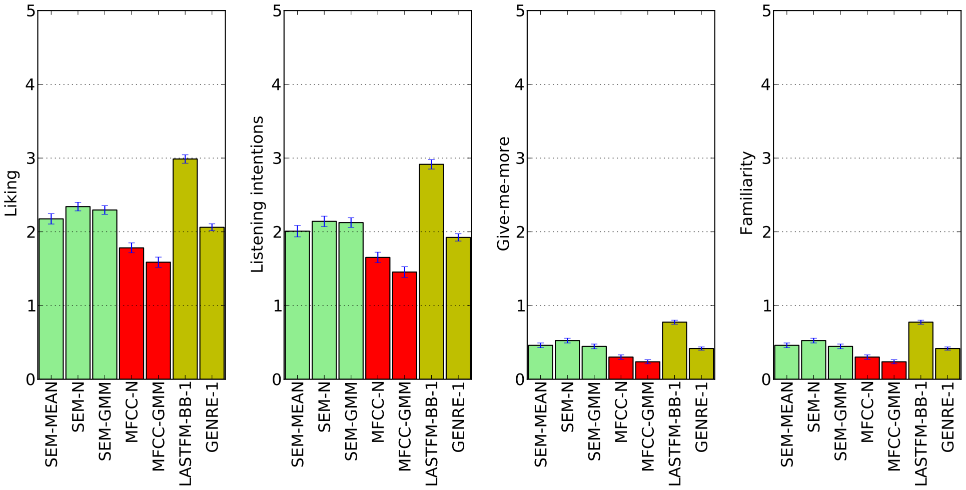
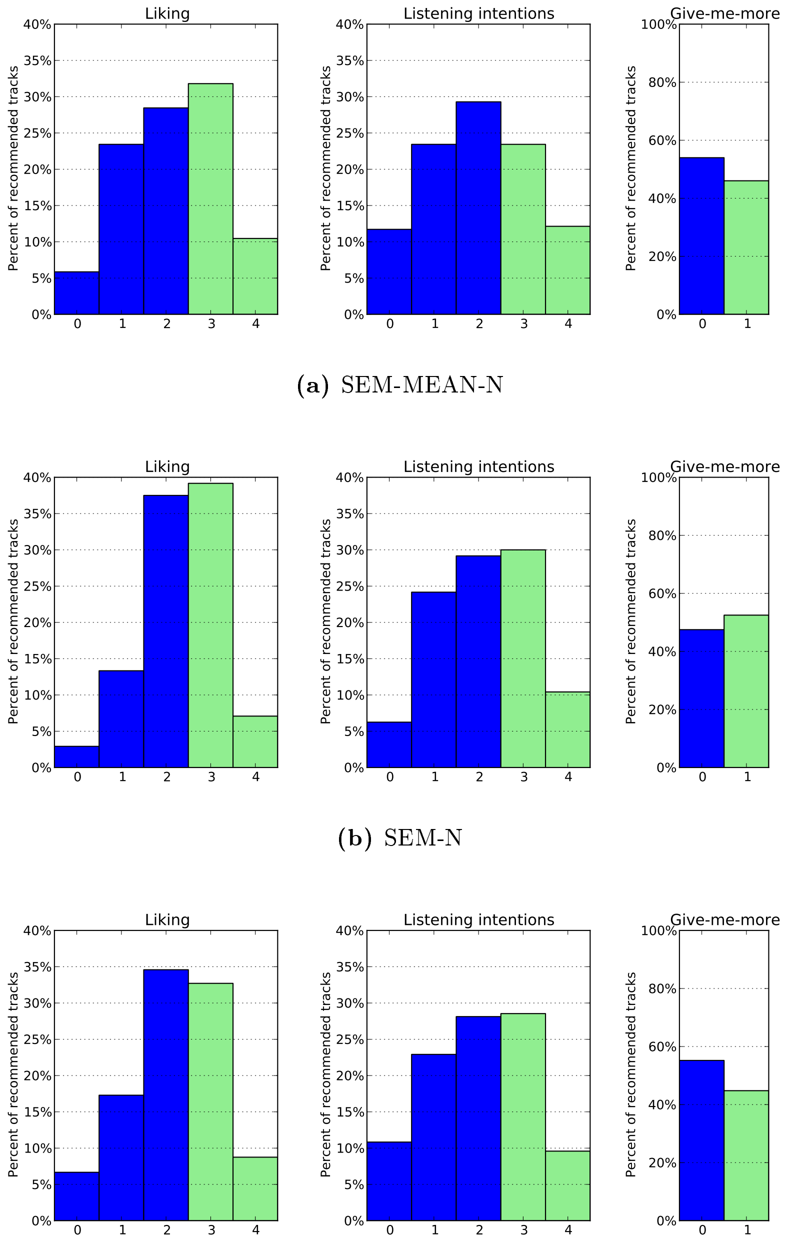


Figure 5.6: Mean ratings per recommendation approach. The results for the *LASTFM-BB-1* approach were obtained on a different underlying music collection. The “give-me-more” rating varies in the $[0, 1]$ interval. Error bars indicate one standard error of the mean. The approaches are differentiated by color similarly to the previous figure.

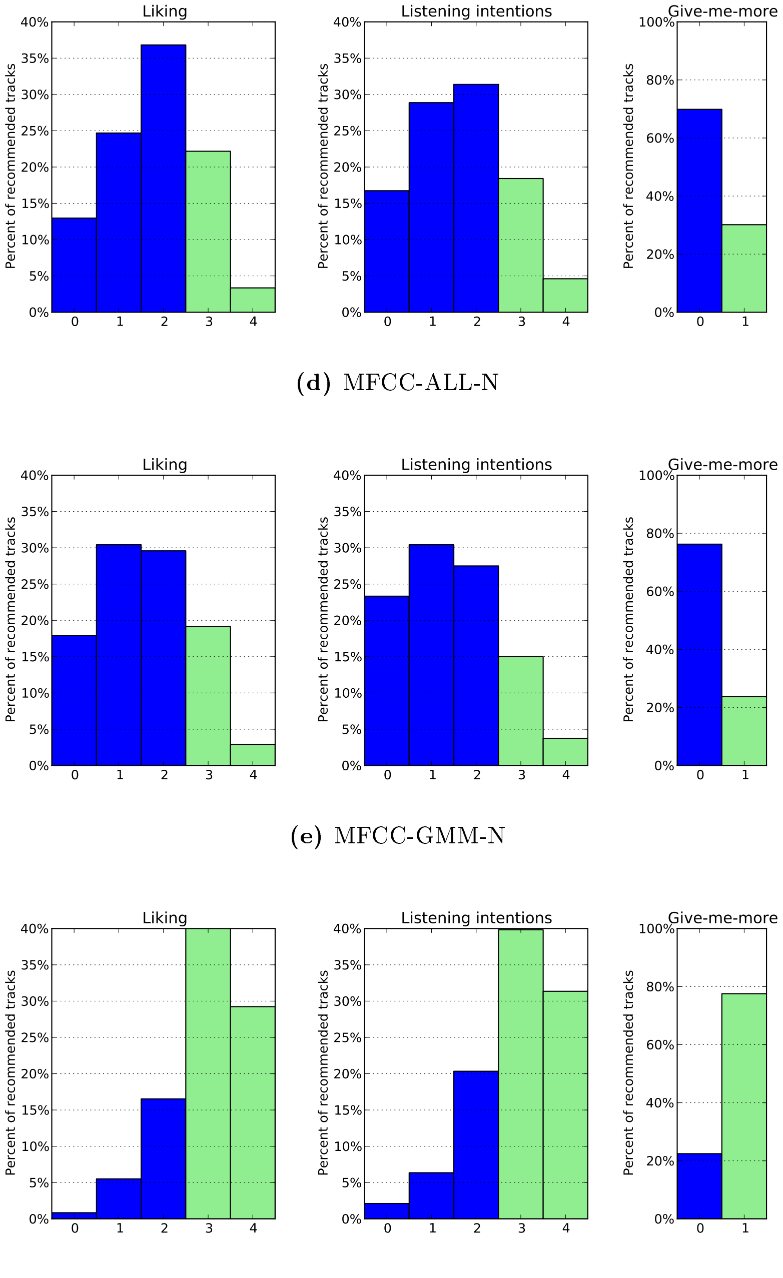


(a) SEM-MEAN-N

(b) SEM-N

(c) SEM-GMM

Figure 5.7: Histograms of liking, listening intentions, and "give-me-more" ratings gathered for the (a) SEM-MEAN-N, (b) SEM-N, (c) SEM-GMM, (d) MFCC-ALL-N, (e) MFCC-GMM-N, (f) LASTFM-BB-1, and (g) GENRE-1 approaches. Green bars stand for high (i.e., desired) ratings while blue bars stand for unsatisfactory ratings. Note that the results for the *LASTFM-BB-1* approach were obtained on a different underlying music collection. (Continued on next page.)

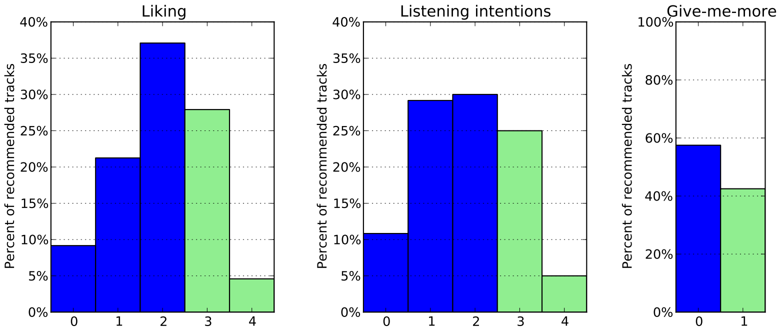


(d) MFCC-ALL-N

(e) MFCC-GMM-N

(f) LASTFM-BB-1

Figure 5.7: Continued (caption shown on previous page.)



(g) GENRE-1

Figure 5.7: Continued (caption shown on previous page.)

(although no statistically significant differences were found in terms of liking, listening intentions, and "give-me-more" ratings). Interestingly, this naive genre-based recommender still outperformed the timbre-based baselines. This could be partially explained by the fact that genre was one of the driving criteria for selecting the users' preference sets according to their own reports (see Section 3.3.2), and suggests us that manually annotated genre and subgenre labels entail more information and diversity than timbral information automatically extracted from MFCCs.

On the other hand, the proposed approaches were found to be inferior to the considered commercial recommender (LASTFM-BB-1) in terms of the number of successful novel recommendations (hits) and trusted recommendations (trusts). Still, LASTFM-BB-1 yielded only 7 absolute percentage points more hits than one of our proposed semantic methods (*SEM-1*). Considering trusted recommendations, the LASTFM-BB-1 baseline provided about 22% more recommendations already known by the participants. Interestingly, one track out of four recommended by the LASTFM-BB-1 baseline was already familiar to the participants, which might be an excessive amount considering the music discovery use-case. In particular, the larger amount of both hits and trusts provided by the LASTFM-BB-1 baseline can be partly explained by the fact that the recommendations were generated using the *Last.fm* music collection. Due to the extensive size of this collection and the large amount of available collaborative filtering data, we can hypothesize that the obtained performance of this approach is an upper bound in both hits and trusts and expect a lower performance on our smaller in-house collection.

5.8 Experiment 4: Improving content-based approaches by genre metadata

5.8.1 Evaluation

In this experiment we consider how a minimum amount of metadata information can improve purely content-based recommendations, and propose a filtering approach relying on single, but sufficiently descriptive, genre tags to refine recommendations. As we have found in Experiment 3, a genre-based recommender was able to surpass baseline timbral recommenders. Moreover, we did not find statistically significant differences between the proposed semantic content-based approaches and genre-based recommender in terms of subjective ratings. Therefore, we may hypothesize that genre metadata can be very valuable for content-based music recommender as an additional source. To this end, we propose an improvement of the HYBRID-1 approach by such a filtering (HYBRID-GENRE-1). In addition, in this experiment we evaluate if hybrid content-based music similarity would lead to better recommendations by comparing the proposed SEM-1 and HYBRID-1 approaches.

We evaluate these approaches against three metadata-based baselines: simple recommendations by genre (GENRE-1), black-box track-level collaborative filtering recommendations provided by GENIUS-BB-1,[16] and tag-based artist-level approach LASTFM-TAGS-1.[17]

We performed subjective listening tests on the 19 participants with N_{R} being set to 15. Again, the population was as subset of the population presented in Section 3.3.2. It consisted of 14 males and 5 females with the average age of 33 years ($\mu = 33.0, \sigma = 4.67$) and a high interest in music ($\mu = 9.24, \sigma = 1.01$). The mean size of the provided preference sets was $\mu = 67.26$, $\sigma = 42.53$, and the median was 61 tracks. This time we employed our second in-house target music collection[18], covering a wide range of genres, styles, arrangements, geographic locations, and musical epochs. This collection contains $68,000$ music excerpts (30 sec.) by $16,000$ artists with a maximum of 5 tracks per artist.

5.8.2 Results

Again, our manual corrections of the familiarity rating represented less than 3% of the total familiarity judgments. Recoding subjective ratings into recommendation outcome categories resulted in 20.4% of "unclear" recommendations. We excluded these recommendations from further analysis.

Table 5.4 and Figure 5.5 report the percent of each outcome category per recommendation approach. As we can see, the proposed HYBRID-GENRE-1 approach yielded the largest amount of hits (32.0%), followed by LASTFM-

[16] All experiments were conducted using iTunes 10.1.1.4 on March, 2011.

[17] All tags were obtained on March, 2011.

[18] We could not continue using the first collection due to technical difficulties.

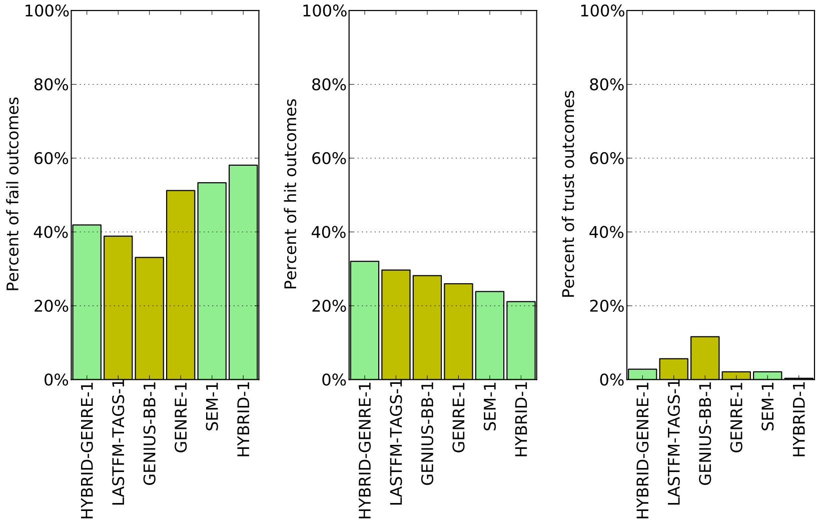


Figure 5.8: The percent of fail, trust, hits, and clear categories per recommendation approach. Metadata-based baselines and the proposed approaches are differentiated by color (green and pistachio, respectively).

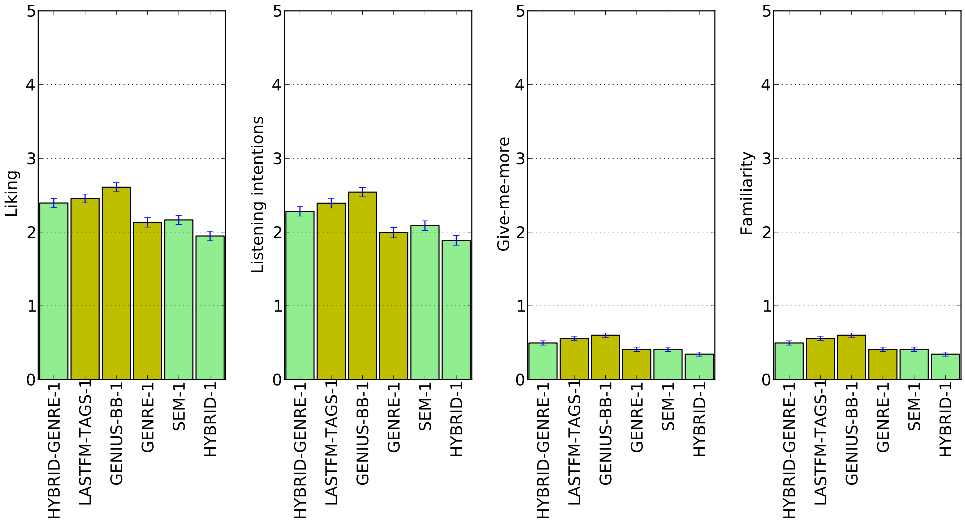


Figure 5.9: Mean ratings per recommendation approach. The "give-me-more" rating varies in the $[0, 1]$ interval. Error bars indicate one standard error of the mean. Metadata-based baselines and the proposed approaches are differentiated by color (green and pistachio, respectively).

Table 5.4: The percent of fail, trust, hit, and unclear categories per recommendation approach.

Approach	fail	hit	trust	unclear	hit+trust
HYBRID-GENRE-1	41.9	32.0	5.3	20.8	37.3
LASTFM-TAGS-1	38.9	29.7	10.6	20.8	40.3
GENIUS-BB-1	33.1	28.2	18.3	20.4	46.5
GENRE-1	51.2	26.0	2.8	20.0	28.8
SEM-1	53.3	23.9	2.8	20.0	26.7
HYBRID-1	58.1	21.1	0.4	20.4	21.5

Table 5.5: Mean ratings per recommendation approach.

Approach	liking	intentions	give-me-more	familiarity
HYBRID-GENRE-1	2.39	2.28	0.50	0.39
LASTFM-TAGS-1	2.46	2.39	0.56	0.63
GENIUS-BB-1	2.61	2.54	0.60	1.09
GENRE-1	2.13	1.99	0.41	0.33
SEM-1	2.16	2.09	0.41	0.27
HYBRID-1	1.95	1.89	0.35	0.23

TAGS-1 (29.7%) and GENIUS-BB-1 (28.2%), and was the only (partially) content-based approach that provided considerably large amount of successful recommendations. We can evidence that inclusion of genre metadata improved the amount of hits by 11% for the HYBRID-1, making its refined version comparable to the metadata-based baselines. On the other side, the GENIUS-BB-1 and LASTFM-TAGS-1 approaches provided the largest amount of trusts (18.3% and 10.6% respectively), while the rest of approaches yielded only scarce trusts (5.3% for HYBRID-GENRE-1, the rest below 3%). Finally, we can see that all recommendation approaches provided more than 33% of fails, which means that at least each third recommendation was possibly annoying for the user. The Pearson chi-square test confirmed the association between the recommendation approach and the outcome category ($\chi^2(15) = 131.5$, $p < 0.001$).

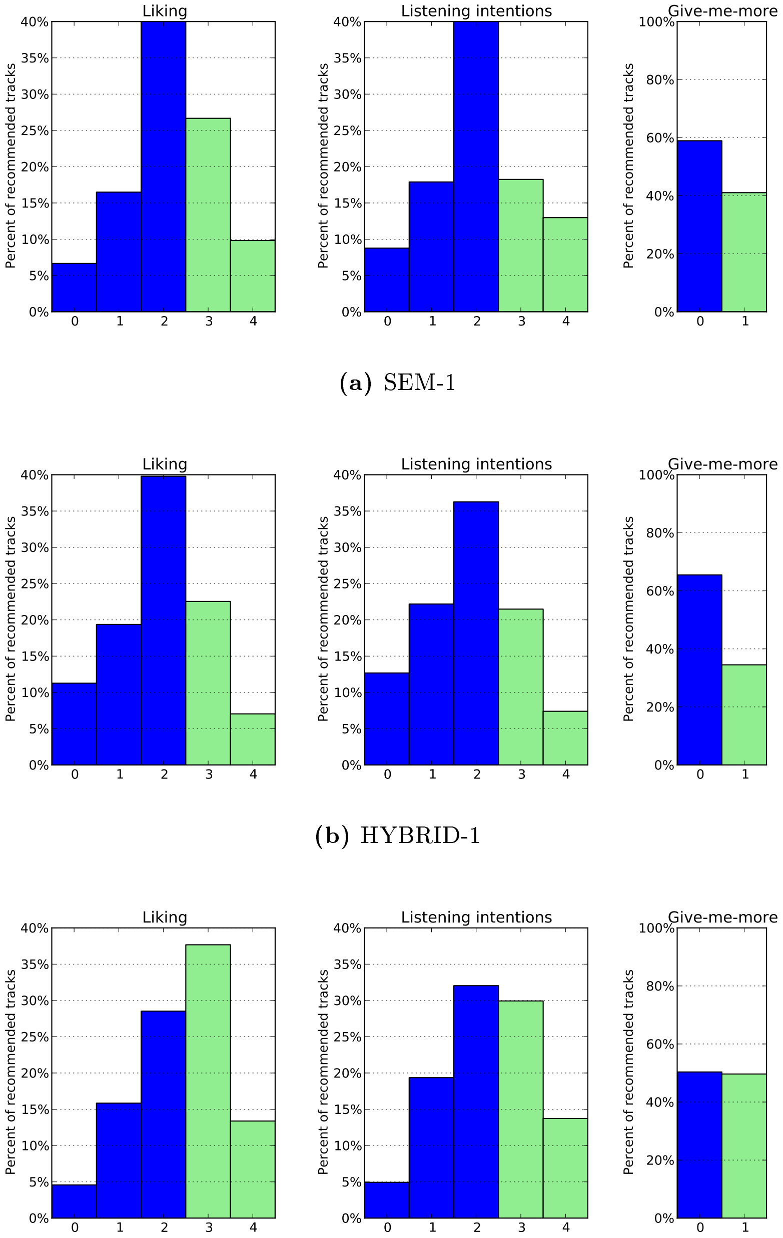
Three separate between-subjects ANOVA tests confirmed the effects of the recommendation approaches on the liking, intentions, and “give-me-more” subjective ratings: $F(5, 1705) = 15.237$, $p < 0.001$ for the liking rating, $F(5, 1705) = 14.578$, $p < 0.001$ for the intentions rating, and $F(5, 1705) = 11.420$, $p < 0.001$ for the “give-me-more” rating (normality assumptions required for ANOVA could be trusted according to the conducted Shapiro-Wilk’s test). Pairwise comparisons using Tukey’s test revealed the same pattern of differences between the approaches, irrespective of the 3 tested indexes. It highlights the following groups with no statistically significant difference inside each group: 1) GENIUS-BB-1, LASTFM-TAGS-1, and HYBRID-GENRE-1 having the highest ratings, 2) SEM-1 and HYBRID-GENRE-1, and 3) SEM-1, GENRE-1, and HYBRID-1 having the lowest. Note that the HYBRID-GENRE-1 and SEM-1 are both belonging to two different groups. The mean liking and listening intentions ratings are presented in Table 5.5 and Figure 5.9. In addition, Figure 5.10 presents the histograms for the liking, listening intentions, and "give-me-more" ratings.

5.8.3 Discussion

We have evidenced that simple filtering by genre significantly improves recommendations on the example of HYBRID-GENRE-1. Furthermore, such a refined approach surpasses the considered metadata-based recommenders LASTFM-TAGS-1 and GENIUS-BB-1 in terms of successful novel recommendations (hits) and provides satisfying recommendations, comparable to these baselines with no statistically significant difference. Considering the content-based approaches without genre filtering, LASTFM-TAGS-1 and GENIUS-BB-1 work significantly better than SEM-1 and HYBRID-1. In particular, artist-level recommendations based on social tags still produce recommendations more accurate than content-based methods working on track-level.

Interestingly, we did not find any improvements over the proposed semantic content-based recommender using instead a complex low-level/semantic distance. On the contrary, HYBRID-1 approach performed worse than semantic SEM-1 in terms of amount of hits+trusts (although the difference in the obtained subjective ratings was not statistically significant). This suggests that such a complex distance, previously found to overcome the semantic distance in the task of music similarity, is not well suited for the music recommendation use-case. A possible explanation is the fact that listeners might prefer semantically similar rather than purely acoustically similar music content as recommendations. That is, while low-level audio similarity (which is the prevalent component of our HYBRID-1 approach) provide acoustically similar music, the listener's preferences might require a higher level of abstraction above acoustical properties to judge the relevance and suitability of the recommended items. Further optimization of the hybrid approach, increasing the importance of its semantic component.

Similarly to Experiment 3, we have observed the fact of no statistically significant difference between purely content-based approaches (SEM-1 and HYBRID-1) and a simple genre-based baseline. This fact corroborates the importance of genre metadata again. Indeed, the quality of recommendations improves significantly after applying the proposed filtering by genre, overcoming the gap between content-based approaches and commercial metadata-based recommenders. Adding audio content-information to the simple genre infor-

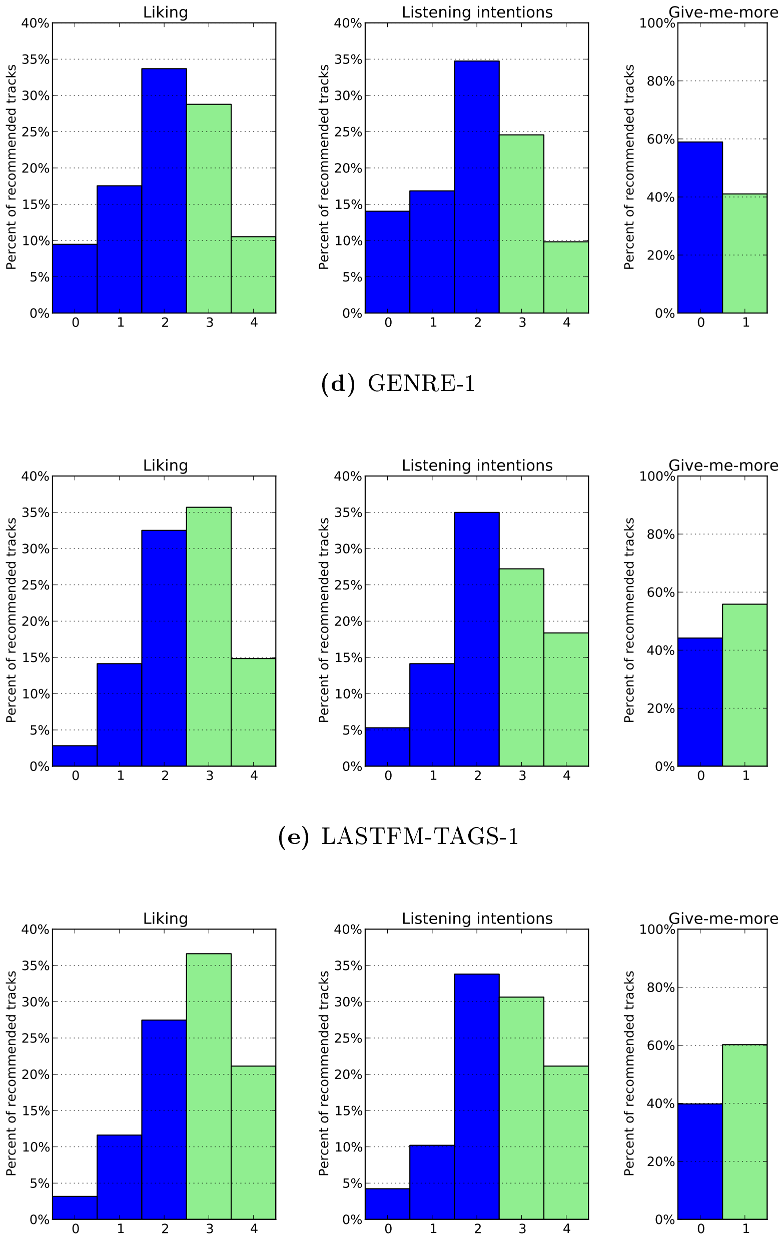


(a) SEM-1

(b) HYBRID-1

(c) HYBRID-GENRE-1

Figure 5.10: Histograms of liking, listening intentions, and “give-me-more” ratings gathered for the (a) SEM-1, (b) HYBRID-1, (c) HYBRID-GENRE-1, (d) GENRE-1, (e) LASTFM-TAGS-1, and (f) GENIUS-BB-1 approaches. Green bars stand for high (i.e., desired) ratings while blue bars stand for unsatisfactory ratings. (Continued on next page.)



(d) GENRE-1

(e) LASTFM-TAGS-1

(f) GENIUS-BB-1

Figure 5.10: Continued (caption shown on previous page.)

mation (HYBRID-GENRE-1) boosts the performance significantly compared to the genre-based baseline.

Finally, we would like to discuss the effect of the N_{O} parameter responsible for the compromise between accuracy of the employed similarity measure and diversity of recommendations (see Section 5.3). As expected, comparing results of Experiments 3 and 4, we can see a performance-worsening when selecting N_{O}=1 rather than N_{O}=N. In particular, we can see similar performance of our HYBRID-GENRE-1 and SEM-N (although the results are not directly comparable due to different target music collections). That is, selecting the N_{O} equal to 1, we evaluate a lower-bound of the considered approaches' performances. In general, relying on the obtained results, we may conclude that the proposed approach, operating on complex content-based distance, refined by simple genre metadata is well suited for the use-case of music discovery.

5.9 Experiment 5: Employing editorial metadata

5.9.1 Evaluation

In this experiment we consider how editorial metadata can be used to describe artists in terms of genres, styles, recording labels, release years, and geographical locations, to provide music recommendation, and evaluate our proposed DISCOGS-1[19] approach. We consider it as a cheaper alternative to GENIUS-BB-1[20] and LASTFM-TAGS-1 approaches, which we use as the baselines. In addition, following the conducted Experiment 4, we evaluate the proposed content-based semantic approach refined by genre metadata (SEM-GENRE-1). We performed subjective listening tests on the 27 participants using the same target music collection as in Experiment 4, being additionally filtered by previously recommended tracks. The population was a subsect of the population presented in Section 3.3.2 with the mean size of the provided preference sets being $\mu = 51.41$, $\sigma = 38.38$, and the median being 50 tracks. It consisted of 17 males and 10 females with the average age of 31 years ($\mu = 31.04, \sigma = 5.76$) and a high interest in music ($\mu = 9.43, \sigma = 0.91$).

5.9.2 Results

This time, manually corrections of the familiarity represented 4.5% of the total familiarity judgments (73 corrections out of 1620 tracks). "Unclear" outcome category amounted to 17.3% of all recommendations. Again, we excluded these recommendations from further analysis.

We report the percent of hit fail, trust, and unclear outcomes per recommendation approach in Table 5.6 and Figure 5.11. According to the results of the Pearson chi-square test, an association between the approaches and the

[19] In our experiments, we used a *Discogs* monthly dump dated by January, 2011.

[20] All experiments were conducted using iTunes 10.3.1 on December, 2011.

Table 5.6: The percent of fail, trust, hit, and unclear categories per recommendation approach.

Approach	fail	hit	trust	unclear	hit+trust
LASTFM-TAGS-1	32.8	38.8	7.4	21.0	46.2
DISCOGS-1	34.4	31.9	16.4	17.3	48.3
GENIUS-BB-1	36.2	35.7	13.2	14.9	48.9
SEM-GENRE-1	41.6	37.9	4.4	16.1	42.3

Table 5.7: Mean ratings per recommendation approach.

Approach	liking	intentions	give-me-more	familiarity
LASTFM-TAGS-1	2.52	2.45	0.63	0.49
DISCOGS-1	2.63	2.57	0.63	0.83
GENIUS-BB-1	2.60	2.50	0.59	0.80
SEM-GENRE-1	2.45	2.33	0.52	0.37

outcome categories ($\chi^2(9) = 46.879$, $p < 0.001$) can be accepted. In general, the proposed DISCOGS-1 approach performed well comparing to the baselines. The DISCOGS-1 provided a considerably low (34.4%) amount of fails, being in between of the metadata-based baselines LASTFM-TAGS-1 (with the lowest amount of fails, 32.8%) and GENIUS-BB-1. In contrast, the SEM-GENRE-1 approach, which is partially content-based, provided the largest (over 41%) amount of fails. Considering hits, the LASTFM-TAGS (38.8%) and SEM-GENRE-1 (37.9%) are the leaders followed by GENIUS-BB-1, and lastly, the DISCOGS-1. That is, our proposed approach provided the least amount of novel relevant recommendations (31.9%). Nevertheless this fact is compensated by the largest amount of trusts, gathered by the DISCOGS-11 (16.4%) followed by the GENIUS-BB-1 (13.2%), LASTFM-TAGS-1 (7.4%), and the SEM-GENRE-1 (4.4%). The amount of unclear recommendations ranged as well. Considering the extreme case, when fails and unclear categories are both unwanted outcomes, the metadata-based GENIUS-BB-1 and DISCOGS-1 result as approaches with the least amount of unwanted recommendations (51.1% and 51.7%, respectively), followed by the M-TAGS, and lastly by the partially content-based SEM-GENRE-1 approach (57.7%). In contrast, considering trusts and hits as wanted outcomes, the GENIUS-BB-1 and DISCOGS-1 provide their largest amount (48.9% and 48.3%, respectively), followed by the LASTFM-TAGS-1 and SEM-GENRE-1.

Considering subjective ratings, we conducted three separate between-subjects ANOVAs (normality assumptions required for ANOVA could be trusted according to the conducted Shapiro-Wilk's test). Tested approaches were shown to have an impact on these ratings ($F(3, 1612) = 3.004$, $p < 0.03$ for the

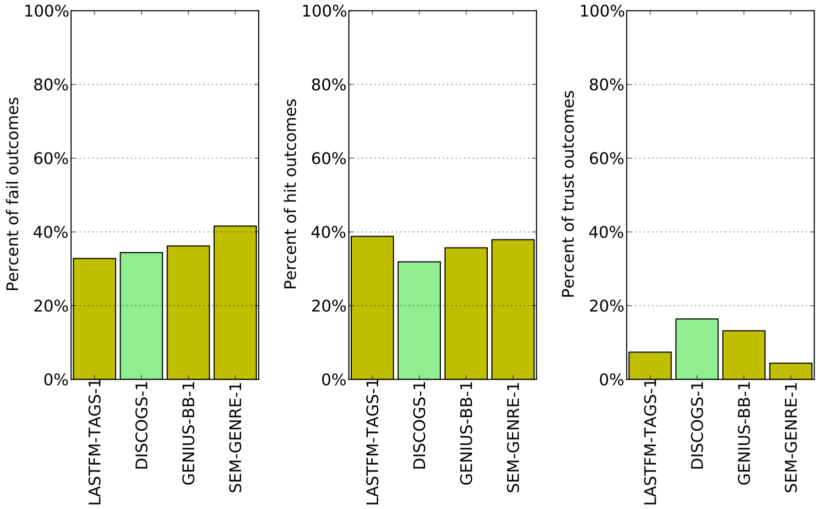


Figure 5.11: The percent of fail, trust, hits, and clear categories per recommendation approach. The proposed approach based on editorial metadata is differentiated by color.

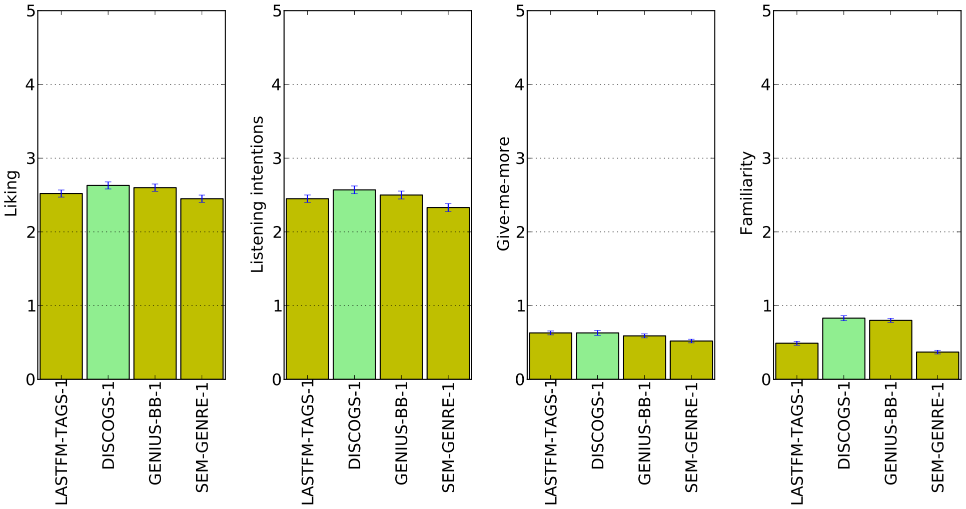
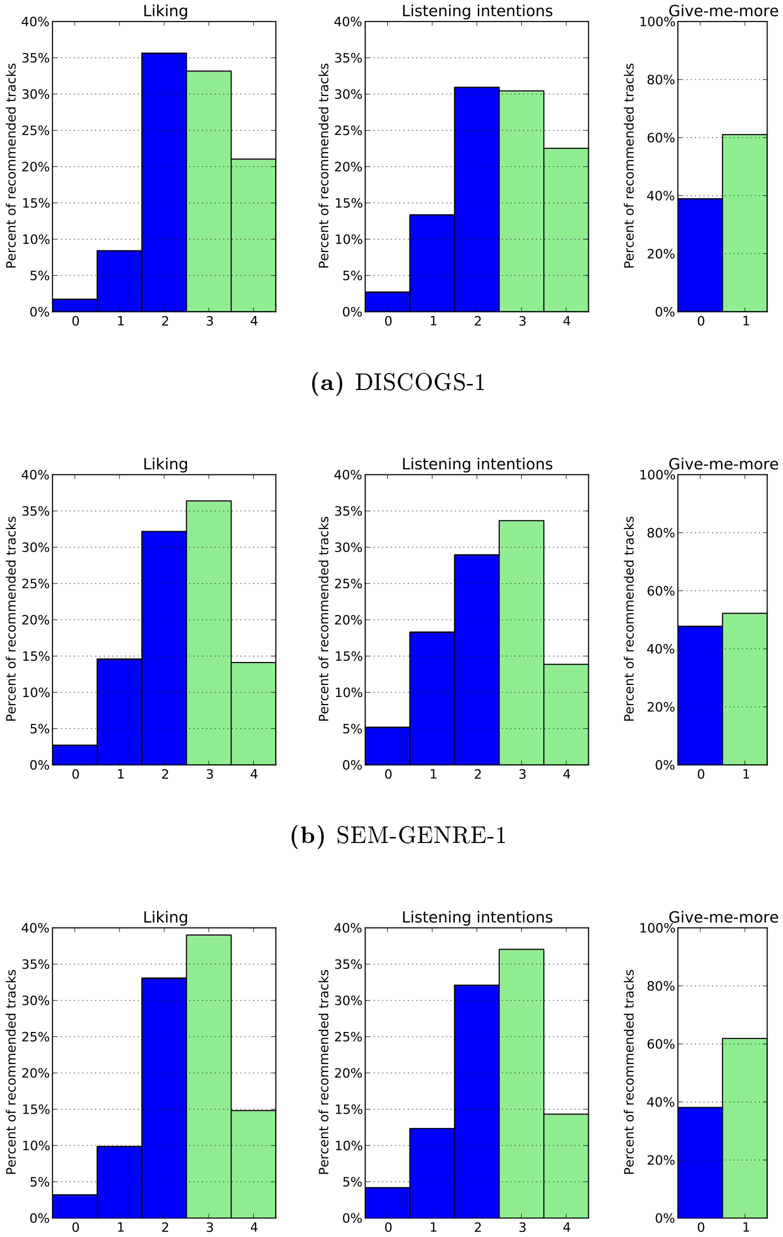


Figure 5.12: Mean ratings per recommendation approach. The "give-me-more" rating varies in the $[0, 1]$ interval. Error bars indicate one standard error of the mean. The proposed approach based on editorial metadata is differentiated by color.

liking rating, $F(3, 1612) = 3.660$, $p < 0.02$ for the intentions rating, and $F(3, 1612) = 3.363$, $p < 0.02$ for the "give-me-more" rating). Pairwise comparisons using Tukey's test revealed differences only between DISCOGS-1 vs SEM-GENRE-1 for the case of all three ratings, and, in addition, a difference between LASTFM-TAGS-1 vs SEM-GENRE-1 in the case of the "give-me-more" rating. In Figure 5.13 we present the histograms for the liking, listening

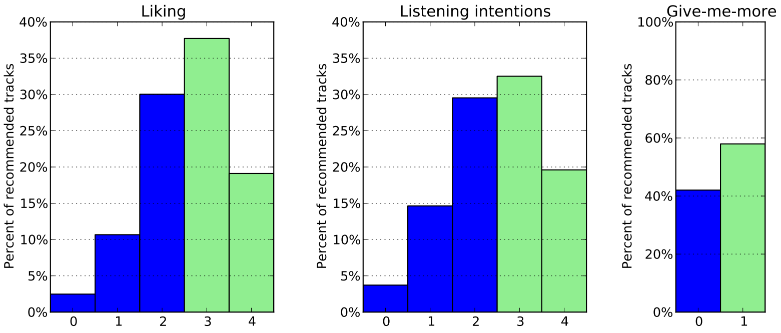


(a) DISCOGS-1

(b) SEM-GENRE-1

(c) LASTFM-TAGS-1

Figure 5.13: Histograms of liking, listening intentions, and "give-me-more" ratings gathered for the (a) DISCOGS-1, (b) SEM-GENRE-1, (c) LASTFM-TAGS-1, (d) GENIUS-BB-1 approaches. Green bars stand for high (i.e., desired) ratings while blue bars stand for unsatisfactory ratings. (Continued on next page.)



(d) GENIUS-BB-1

Figure 5.13: Continued (caption shown on previous page.)

intentions, and "give-me-more" ratings. Mean values of these ratings are provided in Table 5.7 and Figure 5.12. Inspecting the means, we see that all considered approaches performed with a user satisfaction slightly above average. Almost half of the provided recommendations were favorably evaluated, i.e., received high liking and listening intentions ratings (> 2) and a positive "give-me-more" request. An inspection of histograms shows that the proposed DISCOGS-1 approach receives the highest amount of maximum ratings for liking and listening intentions ($\simeq 21\%$ and $\simeq 22.5\%$, respectively). In contrast, the amount of received negative ratings is lower. Still, returning to the ANOVA results, the only clear difference in performance, as measured by our 3 indexes, happens between DISCOGS-1 and SEM-GENRE-1. In other words, the proposed DISCOGS-1 approach is able to achieve similar liking, listening intentions and willingness to get recommended music as existing (and commercial) state-of-the-art systems.

5.9.3 Discussion

Subjective evaluation demonstrated that the proposed DISCOGS-1 approach, operating solely on editorial metadata, performs comparably to the state-of-the-art metadata systems, which require social tags (LASTFM-TAGS-1) or/and collaborative filtering datasets (GENIUS-BB-1). In particular, our approach provided large amount of trusted and novel relevant recommendations, which suggests that the proposed approach is well suited for music discovery and playlist generation.

Interestingly, the evaluated content-based approach filtered by simple genre metadata (SEM-GENRE-1) revealed performance comparable to the metadata-based approaches as well. In terms of statistically significant differences in the subjective ratings, SEM-GENRE-1 is surpassed only by the proposed DISCOGS-1 approach.

Finally, we observe the SEM-GENRE-1 approach to be better than HYBRID-GENRE-1 similarly to the non-filtered SEM-1 and HYBRID-1 (hypothetically, as the results are not directly comparable due to the applied collection filtering) in terms of hits+trusts and subjective ratings.

5.10 Summary of results and general discussion

In order to summarize the results, we have grouped data from Experiments 3, 4, and 5 together, and conducted between-subject ANOVA with one additional factor being the number of experiment. The results confirmed statistical significance of the effect of approach on the liking rating ($F(13, 4993) = 29.260$, $p < 0.001$). Furthermore, we can reject the effect of experiment ($F(2, 4993) = 1.758$, $p = 0.172$) and the effect of interaction between approach and experiment variables ($F(1, 4993) = 1.758$, $p = 0.504$). Similar results were evidenced in the case of listening intentions rating ($F(13, 4993) = 29.984$, $p < 0.001$ for the effect of approach; $F(2, 4993) = 1.274$, $p = 0.280$ for the experiment number; $F(1, 4993) = 0.754$, $p = 0.385$ for the interaction between approach and experiment, respectively) and give-me-more rating ($F(13, 4993) = 17.339$, $p < 0.001$ for the approach; $F(2, 4993) = 0.622$, $p = 0.537$ for the experiment; $F(1, 4993) = 2.191$, $p = 0.139$ for the interaction between both). Therefore, we can be assured that there is no possible effect of using different music collections in the experiments, and that the results of these experiments are directly comparable.

Post-hoc Tukey's test showed similar pattern of differences to the ones we have evidenced before in Sections 5.7.1, 5.8.1 and 5.9.1. Tables 5.8 and 5.9 present subsets of approaches with no statistically significant difference found between approaches in each group according to each of the three tested ratings. The observed mean rating values are reported for each approach. In particular, content-based methods filtered by genre metadata (SEM-GENRE-1, HYBRID-GENRE-1) and metadata-based LASTFM-TAGS-1, GENIUS-BB-1, and DISCOGS-1, are grouped together in the best performing group apart from LASTFM-BB-1. The latter one achieved the maximum performance, but however can be considered as an exception due to the significantly larger underlying music collection.

In what follows, we summarize our conclusions on the conducted Experiments 3, 4, and 5. Firstly, we have proposed semantic content-based approaches for music recommendation, and observed statistically significant improvement over baseline timbral approaches. In addition, we have revealed the fact that simple genre/style tags can be a reasonable source of information to provide recommendations superior to the common low-level timbral music similarity based on MFCCs.

Secondly, we have proposed filtering by simple genre metadata. We found that it greatly improves the proposed semantic/hybrid content-based approaches

Table 5.8: Subsets of approaches with no statistically significant difference found between approaches in each group in respect to the liking (a) and listening intentions (b) ratings. The observed mean rating values are reported for each approach. The results for the *LASTFM-BB-1* approach were obtained on a music collection different from that of the other approaches.

(a)

Approach	#cases	Subset									
		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
MFCC-GMM	240	1.59									
MFCC-N	239	1.78	1.78								
HYBRID-1	284		1.95	1.95							
GENRE-1	525			2.06	2.06						
SEM-1	285			2.16	2.16	2.16					
SEM-MEAN	239			2.18	2.18	2.18	2.18				
SEM-GMM	240				2.30	2.30	2.30	2.30			
SEM-N	240					2.34	2.34	2.34	2.34		
HYBRID-GENRE-1	284					2.39	2.39	2.39	2.39	2.39	
SEM-GENRE-1	404						2.45	2.45	2.45	2.45	
LASTFM-TAGS	688							2.50	2.50	2.50	
GENIUS-BB-1	687								2.61	2.61	
DISCOGS-1	404									2.63	
LASTFM-BB-1	235										2.99
Significance		.482	.751	.219	.182	.213	.056	.438	.070	.160	1.000

(b)

Approach	#cases	Subset							
		#1	#2	#3	#4	#5	#6	#7	#8
MFCC-GMM	240	1.45							
MFCC-N	239	1.65	1.65						
HYBRID-1	284		1.89	1.89					
GENRE-1	525		1.92	1.92					
SEM-MEAN	239			2.01	2.01				
SEM-1	285			2.09	2.09	2.09			
SEM-GMM	240			2.13	2.13	2.13			
SEM-N	240			2.14	2.14	2.14	2.14		
HYBRID-GENRE-1	284				2.28	2.28	2.28	2.28	
SEM-GENRE-1	404					2.33	2.33	2.33	
LASTFM-TAGS	688						2.43	2.43	
GENIUS-BB-1	687							2.52	
DISCOGS-1	404							2.57	
LASTFM-BB-1	235								2.91
Significance		.572	.105	.173	.098	.258	.068	.066	1.000

Table 5.9: Subsets of approaches with no statistically significant difference found between approaches in each group in respect to the give-me-more ratings. The observed mean rating values are reported for each approach. The results for the *LASTFM-BB-1* approach were obtained on a music collection different from that of the other approaches.

Approach	#cases	Subset						
		#1	#2	#3	#4	#5	#6	#7
MFCC-GMM	240	.24						
MFCC-N	239	.30	.30					
HYBRID-1	284	.35	.35	.35				
SEM-1	285		.41	.41	.41			
GENRE-1	525		.42	.42	.42			
SEM-GMM	240			.45	.45			
SEM-MEAN	239			.46	.46	.46		
HYBRID-GENRE-1	284				.50	.50	.50	
SEM-GENRE-1	404				.52	.52	.52	
SEM-N	240				.53	.53	.53	
GENIUS-BB-1	687					.60	.60	
LASTFM-TAGS	688						.60	
DISCOGS-1	404						.63	
LASTFM-BB-1	235							.77
Significance		.332	.217	.226	.234	.063	.054	1.000

up to the level of statistically non-significant differences with the state-of-the-art metadata-based recommenders. The obtained number of hits+trusts outcomes is fewer than for such baselines. Nevertheless, the amount of hits is comparable, or even greater than of these baselines, which suggests that our content-based approaches filtered by genre metadata are well-suited for the use-case of music discovery. One possible problem of such filtering is that it can theoretically lower the serendipity of the provided recommendations. We are aware of this problem, but our counterargument is the supposition that there is enough music, unexpected and potentially attractive for the listener, by the preferred styles of music to be recommended. Moreover, basing on tag ontology or simple relations between tags (Sordo, 2012), it is possible to expand a set of style tags for filtering with related tags in the case one wants broader results.

Thirdly, we have proposed an approach working exclusively on editorial metadata taken from publicly available music database, *Discogs.com.* Relying on user-built information about music releases present in this database, we demonstrated how this information can be applied to create descriptive tag-based artist profiles, containing information about particular genres, styles, record labels, years of release activity, and countries. Furthermore, to overcome the problem of tag sparsity, such artist profiles can be compactly represented as vectors in a latent semantic space of reduced dimension. Applying a distance measure between the resulting artist vectors for the tracks in the preference set of a user and the tracks within a music collection, we are able to generate recommendations. We observed, that the performance of this approach is also comparable to the state-of-the-art metadata-based approaches and is well-suited especially for the case of playlist generation. The proposed approach has a number of advantages over common metadata-based approaches. Firstly, it is able to provide a compact profile for each artist found in *Discogs* database. Matching these profiles to music collections, large-scale recommendation systems can be built. Secondly, the proposed approach is based only on open public data, meanwhile the majority of successful recommender systems operate on commercially undisclosed metadata. As a consequence, our approach is easy to re-create and reproduce.

As expected, in all our experiments we evidenced a high number of trusted recommendations for the metadata-based approaches, and fewer in the case of content-based recommendations which is in line with results reported elsewhere (Celma & Herrera, 2008). In general, we may conclude that it is possible to achieve recommendations with a (slightly) above-average user satisfaction, which would not require large datasets of social tags or collaborative filtering data, but would work on the basis of audio analysis and simpler editorial metadata. Notably, the best performing metadata approaches we considered are suffering from the same "above-average" ceiling in all of the experiments. This fact highlights a lot of room for improvement of music recommender systems. Although we have considered and evaluated the proposed approaches in the context of "passive discovery", relying on preference sets provided by listeners, we expect our conclusions to be applicable for the query-by-example use-case.

CHAPTER 6

Content-based analysis of music preferences

6.1 Introduction

In this chapter, we provide a further study on how audio content information can be exploited to provide interesting insights on the nature of music preferences from both acoustical and semantic perspectives. We conduct a regression analysis in order to capture the relation between the listener's preference of particular tracks and respective low-level audio features and inferred high-level semantic descriptors (the henceforth called "predictors"). Our main goal is to reveal "important" (or "key") predictors defining the music preferences of each of our participants, i.e., those that are best candidates for explaining the outcome (the subjective preference ratings) and helping us to build a model. By finding important predictors for each participant we then will be able to compare preference patterns of different participants and find similarities or differences among them. To this end, we construct personal user models for a set of participants according to their preference ratings and preference set. The results of this analysis provide insights on how important are particular audio features and semantic descriptors to describe, and distinguish, the overall preferences of a particular listener. We also reveal general patterns suitable for the whole sample of our participants.

6.2 Experiment 7: Predictive audio features of individual preferences

6.2.1 User data

We gathered preference sets (see Section 3.3) and preference feedback (see Section 5.6.2) for 31 out of our 39 participants we previously worked with for Experiments 3, 4, and 5 (Sections 5.7.1, 5.8.1, 5.9.1). The rest of participants were excluded, as we were missing their subjective preference feedback, therefore, having only preference sets. For each participant, we assigned maximum liking, listening intentions, and give-me-more ratings (4, 4, and 1, respectively) to all tracks from her/his preference set as if they were rated by the participant. Furthermore, we used actual received ratings for the tracks previously given to the participant for subjective evaluation. To reduce possible noise, we considered the tracks with inconsistent preference ratings as unclear, similarly to Sections 5.7.1, 5.8.1 and 5.9.1, and excluded them from further consideration. We then matched audio features and semantic descriptors of each track (i.e., the predictors) with the associated liking rating (i.e, the dependent variable).[1]

6.2.2 Building user models

Linear multivariate regression models can be applied to study relationships between variables, some of them to be predicted, some of them being potential predictors, and the respective importance of the latter. However, this common approach is vulnerable to high collinearity of data, i.e., a large number of highly correlated predictors. If two or more of the predictors are correlated to the dependent variable, then the estimates of coefficients in a regression model tend to be unstable or counterintuitive. Indeed, among the 386 predictors available in our study, lots are highly (and even absolutely) correlated. Moreover, we are faced to the "*large p, small n problem*", which refers to the situation when the number of predictors p is larger than the number of observations n, i.e., the ill-conditioned data, which is the case for our subjects.

Therefore, we need to select a regression approach with regularization (i.e., penalization of models based on the number of their parameters) which is not prone to multicollinearity, preserving similar importance levels for highly correlated values, and which is able to deal with the problem of ill-conditioned data. Moreover, we want to obtain a sparse solution that contains a small amount of predictors with non-zero weights for better model stability and less over-fitting. Regularization introduces a second factor which weights the penalty against more complex models with an increasing variance in the data errors. This gives an increasing penalty as model complexity increases (Alpaydin, 2004). In particular, penalization can be in a form of the L_1 or L_2 Euclidean norm of the vector of model coefficients. We considered searching for models with L_1 (LASSO) and L_2 regularization (ridge regression), and their hybrid combination (elastic net):

- *Ridge regression* (Hoerl, 1962) is applied to deal with unstable parameter estimates caused by multicollinearity. Ridge regression generally yields better predictions than ordinary least squares solution, through a better

[1] Similar analysis can be done using listening intentions and give-me-more ratings, but they have been omitted for simplicity.

compromise between bias and variance. Its main drawback is that all predictors are kept in the model.

- *LASSO* (Tibshirani, 1996) is more appropriate to achieve a sparse solution due to L_1 penalization but it will not necessarily yield good results in presence of high collinearity. It has been observed that if predictors are highly correlated, the prediction performance of the LASSO is dominated by ridge regression. Moreover, LASSO tends to select only one predictor among a group of predictors with high pairwise correlations, and ignore the others. The second problem with L_1 penalty is that the LASSO solution is not uniquely determined when the number of predictors is greater than the number of observations, which is not the case of ridge regression.

- *Elastic net* (Zou & Hastie, 2005) is suggested to use to obtain sparse results in the presence of multicollinearity and/or high dimensional data, when the LASSO often fails. Elastic net overcomes these limitations by employing both L_1 and L_2 penalties. It encourages a grouping effect, where strongly correlated predictors tend to be in (out) the model together. The elastic net is particularly useful when the number of predictors is much larger than the number of observations.

We opt for elastic net regularization, which conceptually suits our needs. In the simplest case, for each participant we can directly create a regression model with this regularization. Two parameters need to be estimated: $\rho \in [0, 1]$, defining a balance between LASSO (L_1) and ridge (L_2) regression (only LASSO when $\rho = 1$), and λ defining the amount of penalty for both L_1 and L_2. Equation 6.1 provides exact formula of the objective function to be minimized by a coordinate descent in the regression process (Friedman et al., 2010):

$$\frac{1}{2n}\|y - Xw\|_2^2 + \lambda\rho\|w\|_1 + \lambda\frac{(1-\rho)}{2}\|w\|_2^2, \tag{6.1}$$

where X is a $p \times n$ matrix of predictor observations, p is number of predictors, n is number of observations, y is a dependent variable, and w are regression coefficients for the predictors.

Parameter estimation

The best ρ and λ can be found in a grid search with a stratified 10-fold cross-validation. Thereafter, we can additionally validate the built model if we preliminary split the participant's data into the training set (90%), to be used by grid parameter search, and the holdout test set (10%). However, due to the small number of observations per participant, this modeling scenario can lead to unstable results as the elastic net parameters and the selected predictors can be highly dependent on the particular train/test split. Bagging approach

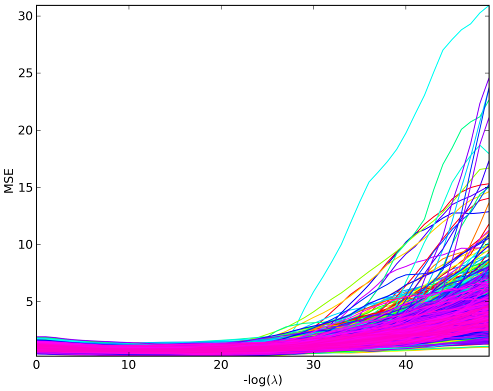


Figure 6.1: Computed MSE for 2000 models (10-fold cross-validation of 200 bagged models) of a particular user and $\rho = 0.9$ as a function of amount of penalization λ. Larger λ values (on the left) lead to heavier shrinkage of predictor coefficients and, in particular, less amount of non-zero predictors.

to modeling can be used to cope with this problem (Breiman, 1996), which is a common practice in different research fields (Bosnić & Kononenko, 2008; Bühlmann, 2002). To this end, we repeatedly run regression on 100 random 90% train/10% test splits of the participant's data (uniform sampling with replacement). The 100 models are fitted using their 100 train samples and can be combined by averaging the outputs. As we are mostly interested in finding important predictors, we introduce a predictor selection step instead of averaging. To this end, we compute a stability score of each predictor as a number of times this predictor was selected (i.e., received a non-zero coefficient) by the 100 bagged regression models. A predictor is considered as stable if it occurred in $\geq$ 95% of bagged models. We intentionally exaggerate requirements for stability to assure that the selection is not biased by the dataset splits.

For each bagged model, the grid search of ρ and λ was conducted on the corresponding training set[2] considering $n_\rho \times n_\lambda \times n_{\text{CV}}$ possible combinations of parameter values and training folder number. We empirically selected $n_\rho = 7, n_\lambda = 50, n_{\text{CV}} = 10$ for the grid resolution. The evaluated values for ρ included an interval $[0.1, 0.9]$ with a linear step 0.1, while λ was evaluated on an automatically selected log-scale. The criterion for best parameters is mean value of mean square error (MSE) over the test folds.

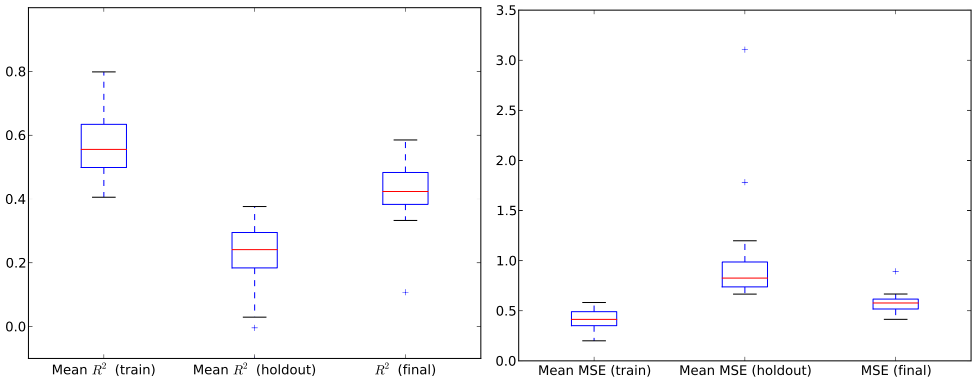
In general, we have found that the best fitted bagged models for all participants were obtained for $\rho = 0.9$. This evidences that the LASSO component of penalization was very important for effective regression, i.e., the original set of predictors was highly redundant as we expected. We, nevertheless, did not consider the case of $\rho = 1$ as we wanted to keep the ridge penalization

[2] We used scikits-learn scientific library for Python, `http://scikit-learn.org`

component (L2) to preserve grouping of correlated variables to some extent. Indeed, inspecting the obtained predictor coefficients for the user models, we found the groupings to be roughly preserved and we, therefore, did not miss possible important but highly correlated predictors in our analysis. For example, both "instrumental" and "voice" descriptors, associated with the OVI classifier, are present in the model for user #11 with similar absolute values for the coefficients but opposite signs as these two predictors are perfectly inversely correlated (-0.0345 and 0.0345, respectively). Figure 6.1 presents the dependency between the MSE and λ on the example of all models computed in 10-fold cross-validation for 20 bagged models of a particular user. As λ increases (toward the left), MSE increases as well. The predictor coefficients are reduced too much and they do not adequately fit the responses. As λ decreases, the models are more complex (have more non-zero coefficients). The increasing MSE suggests that the models are over-fitted. In addition, we can see the scatter of the MSE curves, which might indicate a lack of stability of the trained models due to small number of available observations.

Model validation

To assess quality of each particular bagged model, we computed the coefficient of determination R^2 and MSE on both the training data (measuring goodness-of-fit) and the holdout data (measuring the generalization ability of the model). In particular, the coefficient of determination is useful because it gives the proportion of the variance of the dependent variable that can be explained by a predictor or a set of them. It is a measure that allows us to determine how certain one can be in making predictions from a certain model. Having selected



(a)

(b)

Figure 6.2: Box plots of the obtained R^2 (a) and MSE (b) on the training data, used for grid search, and holdout testing data averaged across bagged models for each participant, and the R^2 and MSE for final user models on the entire participant data.

only the stable predictors, we are able to construct the final ridge regression model for the participant and assess its goodness-of-fit. Unfortunately, due to the limited amount of the available preference data, we decided to use all of it for the process of selection of stable predictors and had no possibility to assess the goodness-of-fit for the final models on additional holdout data. Figures 6.2a and 6.2b present box plots of the computed R^2 and MSE measures. Their inspection reveals a considerably high goodness-of-fit for each model (with the median R^2 of ≈ 0.555) and but relatively low measures on the holdout data (the median R^2 of ≈ 0.2409). The goodness-of-fit for the final model is expectedly lower than for the bagged models, due to predictor selection, but it is on the reasonable level with the median R^2 being ≈ 0.4228. Low R^2 values obtained for the holdout data indicate that, unfortunately, our models might still lack generality, which can be expected in the case of ill-conditioned data as ours.

Quality of user models

Figure 6.3 presents bagged testing, bagged holdout, and final model's coefficients of determination and MSEs for particular users. In addition, the number of track observations available per user and the number of selected predictors for the final model are presented. No stable predictors were found, and, therefore, final models were not constructed for 8 "problematic" participants (#23-31). This can be mostly explained by the lack of preference data (less than 170 tracks), not enough to identify stable descriptors, or by the fact that actual factors of music preferences of these participants cannot be addressed by the predictors at our reach. An inspection of the figure reveals satisfactory goodness-of-fit for a large part of user models. Among the "difficult" participants, i.e., those with low R^2 (< 0.4) for final models, we may highlight participants #3, 4, 7, 10, 12, 14, 15, 18, and 21. The most problematic is the user #4, whose R^2 scores is less than 0.2. MSE of final models reveals the same pattern of problematic users. Considering the difference between train/holdout estimations, which characterizes generalization quality of the constructed models, we highlight the models for participants #19, 18, 11, 12, and 5 as the most unstable (with the difference of R^2 being greater than 0.4) in order of decreasing severity. We therefore include participants #11, 19, and 5 to the list of problematic participants. In total, final models built for 38.7% of our participants are probably unsatisfactory and, moreover, we were not able to build stable models for %25.8 of participants.

Considering model simplicity (i.e., the number of stable predictors in the final model which corresponds to the number of predictors selected consistently across the bagged models among our large set of 386 predictors), we observe that some participants have a very low number of selected predictors. We were able to select only less than 5 predictors for participants #4, 5, 11, 15, and 19, which amounts to 16.1% of participants. Figure 6.4a presents a box plot of the number of predictors selected for each user model. The most complex models (that is, those with the greatest number of predictors, > 20) were obtained for 12.9% of participants with the maximum being 30 stable predictors for participant #1.

In addition, we analyzed the effect of training data size on model complexity and goodness-of-fit. Figure 6.4b demonstrates a positive correlation between the number of samples and the number of selected predictors. A linear trend (the observed Pearson correlation $\rho \approx 0.88$) suggests that the models' complexity seems to increase with the number of samples, i.e. the available preference examples. This might imply that a model's specificity increases with more additional information available, thereby, capturing more nuances of preferences. Figure 6.5a and 6.5b shows a negative correlation (Pearson's $\rho \approx -0.33$) between the MSE of final models and the number of samples and similar positive correlation for R^2 of final models. This suggests that model accuracy may increase with larger amount of training data, which is also somehow expected. According to the line of best fit, more than 150 tracks might

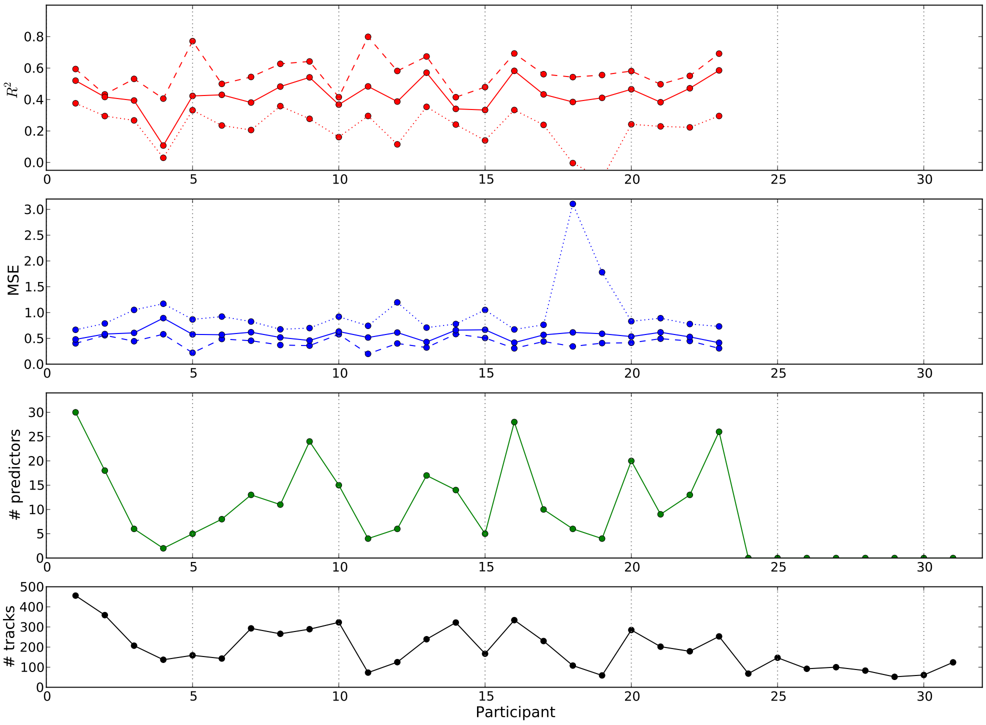
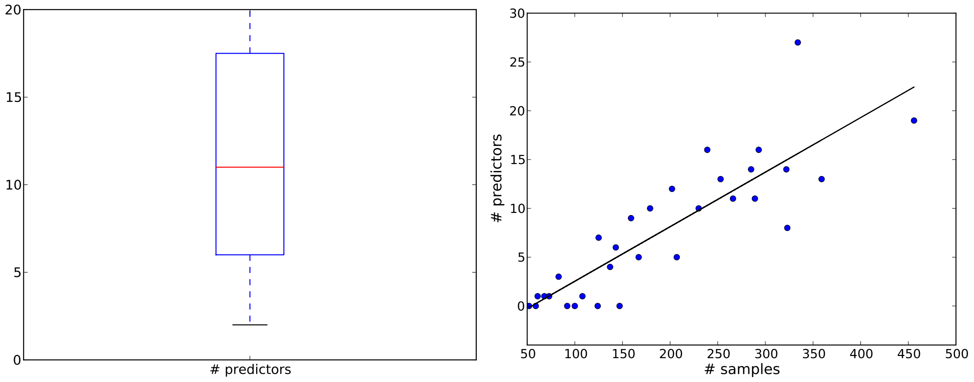
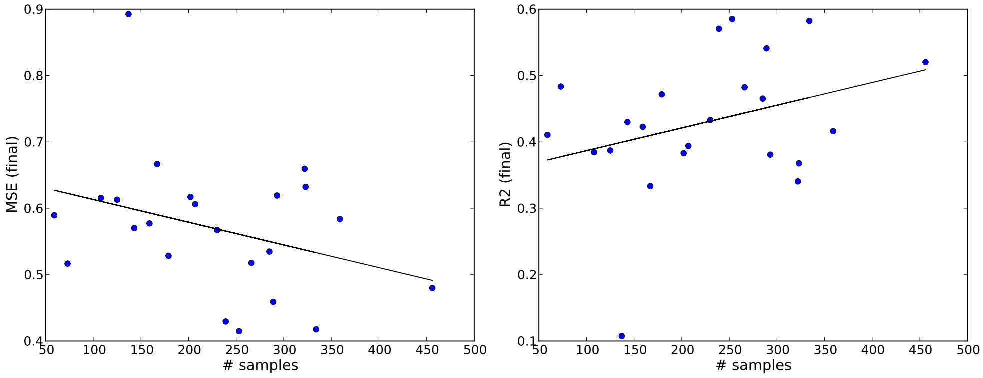


Figure 6.3: Coefficient of determination and mean square error for bagged and final models per participant, supplemented with the number of available track observations and the number of selected predictors. Dashed lines stand for the average R^2 and MSE across bagged models on training dataset (goodness-of-fit), while dotted lines stand for the average R^2 and MSE estimation on holdout testing dataset. Solid lines stand for the goodness-of-fit of final models. No final models were constructed for participants #26-31 due to absence of stable predictors.



(a) (b)

Figure 6.4: Box plot of the number of predictors included in the final user models of our participants (a). Scatter plot and the line of best fit representing the correlation between the number of samples (i.e., training data size) and the number of selected predictors (i.e., model complexity) (b).



(a) (b)

Figure 6.5: Scatter plot and the line of best fit representing the correlation between the number of samples (i.e., training data size) and the obtained MSE (c) and R^2 (d) of final models (i.e., model goodness-of-fit).

be required to build a minimally satisfactory model ($R^2 > 0.4$) and more than 400 tracks are desirable for a satisfactory model ($R^2 > 0.5$).

6.2.3 Important predictors

For each participant, we rank predictors found in the respective final user model by their *univariate* R^2 score (henceforth called as *"importance score"*), which we consider as a measure of importance. As we noted before, using directly the predictor's coefficients from final models as an importance indicator can be misleading in the case of multicollinearity. Instead, we can use coefficients of determination obtained for the univariate linear models. If the predictor was selected for the final model, we create a linear model with this single predictor and assess its goodness-of-fit (R^2) on the participant's dataset. We assign zero importance scores to all predictors not included into the final model of the participant.

In addition, we compute the following two scores, which provide us clues on importance of particular predictors for predicting music preferences in general (i.e., for our entire population):

- *"Generality"* score: the number of times a predictor was included into final user models. In addition, we assessed the number of times when the predictor effect was positive or negative in respect to the sign of the corresponding coefficient in the final model.

- *Mean importance* score: an average of importance scores of the predictor across all participants.

We identified several particular predictors, presented in Table 6.1, with the generality score greater from the rest. More informative conclusions can be driven based on the detailed per-user lists of important predictors in Table 6.3. In total, 143 out of 386 predictors served to build our user models.

Firstly, we have evidenced that tempo related features (first and second peaks' BPM, spread, and weight characterizing BPM probability distribution) were frequent in the user models (61.3% of our participants). A common pattern can be traced for models including these predictors: a positive effect for the second peak BPM and both peaks' spread, a negative effect of both peaks' weight. Higher spread values with lower weight might suggest more dynamic structure of the tracks with a varying rhythm speed, i.e. with a higher rhythm complexity, and a dislike for music with a prominent steady pulse. Together with other descriptors (beats loudness, rhythm type and perceptual speed), rhythmic features occurred in 21 out of 24 built user models (67.7% of participants). We conclude that once the factors of the listener's music preference are related to acoustic properties of music, rhythm might be of the primary importance among them.

Secondly, tonality was found to be another important aspect of preference. Predictors related to HPCP (untransposed and transposed) appeared in 51.6% of models, with particular predictors being very user-specific. Similarly, chords histogram also appeared in models for 45.2% of listeners, not necessarily intersected with HPCP predictors. Tuning equal tempered deviation was present in models for 32.3% participants with a consistently negative effect. This predictor indicates whether the track's scale may be considered as equal-tempered or not by computing the deviation of HPCP local maxima with respect to equal-tempered bins. Our findings suggest a dislike of music with non-western

Table 6.1: Predictors with the highest generality score. Associated ground truths (see Section 3.4.2), and corresponding musical facets, are mentioned in parenthesis in the case of semantic descriptors. Positive effect implies that higher values of a predictor are associated with a higher preference. Oppositely, lower predictor values are associated with a higher preference in the case of negative effect.

Predictor	Musical dimension	Number of participants	Positive effect	Negative effect
First peak spread	Rhythmic	12	11	1
Tuning equal tempered deviation	Tonal	10	0	10
Second peak BPM	Rhythmic	10	10	0
Chords scale=minor	Tonal	8	8	0
Chords scale=major	Tonal	8	0	8
Second peak weight	Rhythmic	8	0	8
Second peak spread	Rhythmic	8	8	0
First peak weight	Rhythmic	8	0	8
Chords histogram #22	Tonal	5	2	3
Chords strength mean	Tonal	4	0	4
Chords histogram #2	Tonal	4	0	4
Chords histogram #18	Tonal	4	4	0
Beats loudness mean	Rhythmic	4	0	4
Spectral complexity variance	Timbral	4	4	0
Folk/country (G1, genre)	Semantic	4	0	4
Alternative (G1, genre)	Semantic	4	0	4
Male (OGD, instrumentation)	Semantic	4	1	3
Female (OGD, instrumentation)	Semantic	4	3	1
Quickstep (RBL, rhythm)	Semantic	4	1	3
Chords key=F#	Tonal	3	2	1
Chords key=A	Tonal	3	1	2
Chords histogram #17	Tonal	3	3	0
First peak BPM	Rhythmic	3	0	3
Beats loudness bass variance	Rhythmic	3	3	0
Spectral flatness dB variance	Timbral	3	3	0
Voice (OVI, instrumentation)	Semantic	3	2	1
Instrumental (OVI, instrumentation)	Semantic	3	1	2
Disco (G3, genre)	Semantic	3	2	1
Country (G3, genre)	Semantic	3	1	2
Trance (GEL, genre)	Semantic	3	0	3
House (GEL, genre)	Semantic	3	3	0
Blues (G1, genre)	Semantic	3	3	0
Western (CUL, musical culture)	Semantic	3	0	3
Non-western (CUL, musical culture)	Semantic	3	3	0
Chachacha (RBL, rhythm)	Semantic	3	3	0

scaling for a group of our participants. In addition, scale and key related predictors both independently appeared for 29% and 38.7% of participants. In total, tonality related predictors (including HPCP, key, scale, and key strength, chords key, scale, and strength, tuning diatonic strength) have been found in 22 out of 24 built user models, which accounts for 71% of participants.

In turn, semantic descriptors were also frequent within the models (19 models, 61.3% of participants). In particular, association with folk/country and

Table 6.2: Top 35 commonly important predictors according to the computed mean importance score. Associated ground truths (see Section 3.4.2) and corresponding musical facets are mentioned in parenthesis in the case of semantic descriptors.

Predictor	Musical dimension	Mean importance score
First peak spread	Rhythmic	0.0485
Second peak BPM	Rhythmic	0.0362
First peak weight	Rhythmic	0.0331
Tuning equal tempered deviation	Tonal	0.0277
Beats loudness mean	Rhythmic	0.0193
Second peak spread	Rhythmic	0.0175
Spectral flatness dB variance	Timbral	0.0157
Country (G3, genre)	Semantic	0.015
Trance (GEL, genre)	Semantic	0.0144
Alternative (G1, genre)	Semantic	0.0133
Chachacha (RBL, rhythm)	Semantic	0.0131
Male (OGD, instrumentation)	Semantic	0.0129
Female (OGD, instrumentation)	Semantic	0.0129
Quickstep (RBL, rhythm)	Semantic	0.0122
Spectral RMS variance	Timbral	0.012
Folk/country (G1, genre)	Semantic	0.0118
Chords scale=minor	Tonal	0.0113
Chords scale=major	Tonal	0.0113
House (GEL, genre)	Semantic	0.0112
Jive (RBL, rhythm)	Semantic	0.0103
Non-western (CUL, musical culture)	Semantic	0.00964
Western (CUL, musical culture)	Semantic	0.00964
Second peak weight	Rhythmic	0.00945
Beats loudness bass variance	Rhythmic	0.00903
Blues (G1, genre)	Semantic	0.00899
Voice (OVI, instrumentation)	Semantic	0.00855
Instrumental (OVI, instrumentation)	Semantic	0.00855
Acoustic (MAC, mood)	Semantic	0.00838
Non-acoustic (MAC, instrumentation)	Semantic	0.00838
Bark bands variance #23	Timbral	0.00835
Spectral flux mean	Timbral	0.00813
THPCP #3	Tonal	0.00759
Bark bands mean #8	Timbral	0.00731
First peak BPM	Rhythmic	0.00728
Spectral flux variance	Timbral	0.0072

alternative rock, voice gender, and quickstep rhythm were among the frequent ones (12% of participants each). If we do not consider particular descriptors (classes) but associated ground truths, the genre-related G1 and G3, and GEL, and the the rhythm-related RBL were among the most frequent (occurred in 11, 9, 6, and 7 models, respectively), followed by OGD (voice gender, 4 models), CUL (western/non-western culture, 3 models), and OVI (voice/instrumental, 3 models). To sum up, genre was found to be an important factor for 48.4% of our participants, followed by rhythm and instrumentation (25.8% each), and mood (12.9%).

Finally, predictors related to timbre were found to contribute to 19 models (61.3% of participants). Specifically, bark bands occurred in 12 models (38.7%), but MFCCs solely in 5 models (16.1% of participants). Partially, this might be explained by that timbral information is already integrated into high-level descriptors related to genres and instrumentation. Building models solely on low-level predictors (see Section 6.2.4) did not provide significant increase in the number of models utilizing timbral predictors.

Our "importance score" reveals a similar list of top ranked predictors, including first peak spread, second peak BPM, first and second peak spread and weight, second peak BPM, and tuning equal tempered deviation, among the others, providing alternative rankings. We report top rankings by the mean importance score in Table 6.2.

Table 6.3: Signed importance scores of predictors selected for each one of the final user models of participants. The sign of the score (positive/negative) corresponds to the sign of the respective predictor coefficient, representing positive or negative effect of the predictor on the liking rating. Associated ground truths (see Section 3.4.2) are mentioned in parenthesis in the case of semantic descriptors.

Participant	Predictor	Musical dimension	Signed importance score
#1	Tuning equal tempered deviation	Tonal	-0.0578
	Second peak spread	Rhythmic	0.0496
	First peak weight	Rhythmic	-0.0477
	Onset rate	Rhythmic	-0.0392
	Jazz (G2, genre)	Semantic	0.0288
	Electronic (G1, genre)	Semantic	-0.028
	Blues (G1, genre)	Semantic	0.0218
	Chords strength mean	Tonal	-0.0218
	HPCP variance #2	Tonal	-0.0204
	Bark bands mean #5	Timbral	-0.0184
	Chords key=F#	Tonal	0.0179
	Reggae (G3, genre)	Semantic	0.014
	Chords histogram #19	Tonal	0.00894
	Second peak weight	Rhythmic	-0.00743
	Spectral complexity variance	Timbral	0.00325
#2	Tuning equal tempered deviation	Tonal	-0.158
	Alternative (G1, genre)	Semantic	-0.108
	Spectral flux variance	Timbral	0.0772
	Bark bands variance #26	Timbral	-0.049
	Second peak BPM	Rhythmic	0.0357
	Chords histogram #5	Tonal	0.0289
#3	Second peak BPM	Rhythmic	0.123
	Chachacha (RBL, rhythm)	Semantic	0.106
	Alternative (G1, genre)	Semantic	-0.069
	Second peak weight	Rhythmic	-0.0663
	Key strength	Tonal	0.0395
	Chords scale=major	Tonal	-0.0378
	Chords scale=minor	Tonal	0.0378

Table 6.3 – *Continued from previous page*

Participant	Predictor	Musical dimension	Signed importance score
	Mood cluster #5 (MCL, mood)	Semantic	-0.034
	Rumba (RBL, rhythm)	Semantic	-0.0246
	First peak spread	Rhythmic	0.0171
	HPCP variance #24	Tonal	0.0149
	Chords histogram #3	Tonal	0.0136
	Chords histogram #2	Tonal	-0.0134
	HPCP variance #33	Tonal	0.0114
	Silence rate 60dB variance	Miscellaneous	0.00801
	HPCP variance #8	Tonal	-0.00783
	Chords histogram #17	Tonal	0.00655
#4	Beats loudness bass variance	Rhythmic	0.203
	Metal (G3, genre)	Semantic	-0.147
	Bark bands variance #4	Timbral	-0.119
	Chords key=G#	Tonal	-0.0794
#5	Alternative (G1, genre)	Semantic	-0.111
	Second peak BPM	Rhythmic	0.0655
	MFCC mean #3	Timbral	0.0597
	Tuning equal tempered deviation	Tonal	-0.0523
	Chords histogram #10	Tonal	-0.0338
#6	Spectral RMS variance	Timbral	0.151
	Quickstep (RBL, rhythm)	Semantic	0.0798
	First peak BPM	Rhythmic	-0.0656
	MFCC mean #11	Timbral	0.0604
	HPCP variance #25	Tonal	0.0567
	Chords histogram #18	Tonal	0.054
	Chords histogram #22	Tonal	-0.0341
	Chords scale=major	Tonal	-0.0242
	Chords scale=minor	Tonal	0.0242
#7	Bright timbre (OTB, instrumentation)	Semantic	-0.107
	Dark timbre (OTB, instrumentation)	Semantic	0.107
#8	First peak weight	Rhythmic	-0.167
	Fast (RPS, rhythm)	Semantic	-0.146
	Second peak BPM	Rhythmic	0.111
	First peak spread	Rhythmic	0.0984
	Pop (G3, genre)	Semantic	-0.092
	Female (OGD, instrumentation)	Semantic	0.0895
	Male (OGD, instrumentation)	Semantic	-0.0895
	Quickstep (RBL, rhythm)	Semantic	-0.0834
	Chords changes rate	Tonal	-0.0732
	HPCP variance #29	Tonal	-0.0394
	Non-western (CUL, musical culture)	Semantic	0.0321
	Western (CUL, musical culture)	Semantic	-0.0321
	Key scale=major	Tonal	-0.0217
	Key scale=minor	Tonal	0.0217
	Folk/country (G1, genre)	Semantic	-0.0214
	Chords histogram #0	Tonal	-0.0207
	Spectral complexity variance	Timbral	0.0149
	Spectral strong peak variance	Timbral	0.011
	Chords key=F	Tonal	0.0108
	Bark bands variance #3	Timbral	0.00776

Table 6.3 – *Continued from previous page*

Participant	Predictor	Musical dimension	Signed importance score
	Chords histogram #22	Tonal	-0.00719
	Electronic (MEL, instrumentation)	Semantic	0.00678
	Non-electronic (MEL, instrumentation)	Semantic	-0.00678
	Chords histogram #18	Tonal	0.00579
	Beats loudness bass variance	Rhythmic	0.00511
	Second peak weight	Rhythmic	-0.00191
	Key key=F#	Tonal	-0.00103
	Bark bands variance #8	Timbral	-9.28e-05
	Rap/hip-hop (G1, genre)	Semantic	5.78e-05
	Funk/soul/rnb (G1, genre)	Semantic	2.01e-05
#9	Country (G3, genre)	Semantic	-0.164
	Non-western (CUL, musical culture)	Semantic	0.138
	Western (CUL, musical culture)	Semantic	-0.138
	Bark bands variance #10	Timbral	-0.101
	Folk/country (G1, genre)	Semantic	-0.0843
	Rock (G3, genre)	Semantic	-0.0738
	First peak weight	Rhythmic	-0.0514
	HPCP variance #34	Tonal	0.0426
#10	Bark bands variance #23	Timbral	-0.192
	Spectral flux variance	Timbral	0.0885
	Chords histogram #16	Tonal	-0.0817
	Chords scale=major	Tonal	-0.0537
	Chords scale=minor	Tonal	0.0537
	HPCP variance #7	Tonal	-0.0234
#11	Acoustic (MAC, mood)	Semantic	0.193
	Non-acoustic (MAC, instrumentation)	Semantic	-0.193
	Spectral flux mean	Timbral	-0.137
	Spectral flatness dB variance	Timbral	0.115
	Second peak BPM	Rhythmic	0.0888
	THPCP #1	Tonal	-0.0798
	Chords key=F#	Tonal	0.0456
	First peak spread	Rhythmic	0.0427
	HPCP variance #12	Tonal	0.0382
	Instrumental (OVI, instrumentation)	Semantic	-0.0345
	Voice (OVI, instrumentation)	Semantic	0.0345
	Bark bands variance #19	Timbral	0.00625
	Tristimulus variance #1	Timbral	-0.00235
#12	Country (G3, genre)	Semantic	0.114
	Mood cluster #1 (MCL, mood)	Semantic	-0.0919
	Key key=C#	Tonal	-0.0764
	Dance (G2, genre)	Semantic	-0.0511
	Spectral spread variance	Timbral	-0.0459
	THPCP #5	Tonal	-0.0347
	Chords histogram #22	Tonal	0.0266
	First peak spread	Rhythmic	0.0219
	Jazz (G3, genre)	Semantic	0.02
	Male (OGD, instrumentation)	Semantic	0.0153
	Female (OGD, instrumentation)	Semantic	-0.0153
	Chords histogram #21	Tonal	-0.0142

Table 6.3 – *Continued from previous page*

Participant	Predictor	Musical dimension	Signed importance score
	MFCC mean #5	Timbral	0.00213
#13	Bark bands mean #24	Timbral	-0.119
	Trance (GEL, genre)	Semantic	-0.106
	First peak weight	Rhythmic	-0.0964
	Second peak BPM	Rhythmic	0.0918
	First peak spread	Rhythmic	-0.0914
	Beats loudness mean	Rhythmic	-0.0769
	House (GEL, genre)	Semantic	0.0653
	Key key=G#	Tonal	-0.0527
	Bark bands mean #0	Timbral	-0.0461
	Tuning equal tempered deviation	Tonal	-0.0454
	Bark bands kurtosis variance	Timbral	0.0333
	Rumba (RBL, rhythm)	Semantic	0.0258
	Bark bands variance #13	Timbral	0.0192
	Chords histogram #18	Tonal	0.0186
	Alternative (G1, genre)	Semantic	-0.0179
	HPCP variance #4	Tonal	-0.0131
	Chords key=G	Tonal	0.0116
	Bark bands variance #16	Timbral	-0.0114
	Chords strength mean	Tonal	-0.0102
	Chords scale=major	Tonal	-0.01
	Chords scale=minor	Tonal	0.01
	Chords key=A	Tonal	-0.00697
	Second peak weight	Rhythmic	-0.00348
	MFCC mean #11	Timbral	0.00228
	First peak BPM	Rhythmic	-0.00221
	Chords histogram #2	Tonal	-0.00132
	HPCP variance #2	Tonal	0.000823
	Disco (G3, genre)	Semantic	9.61e-05
#14	Trance (GEL, genre)	Semantic	-0.18
	Blues (G1, genre)	Semantic	0.161
	First peak spread	Rhythmic	0.119
	Second peak BPM	Rhythmic	0.0876
	Second peak spread	Rhythmic	0.071
	Electronic (G1, genre)	Semantic	-0.0621
	Tuning equal tempered deviation	Tonal	-0.0468
	Instrumental (OVI, instrumentation)	Semantic	-0.0323
	Voice (OVI, instrumentation)	Semantic	0.0323
	Chords key=A	Tonal	0.0233
#15	Folk/country (G1, genre)	Semantic	-0.166
	Pop (G1, genre)	Semantic	-0.146
	Rock (G1, genre)	Semantic	-0.14
	Tuning diatonic strength	Tonal	-0.0989
	Quickstep (RBL, rhythm)	Semantic	-0.0871
	Chachacha (RBL, rhythm)	Semantic	0.0856
	Jive (RBL, rhythm)	Semantic	-0.0761
	Chords strength mean	Tonal	-0.0656
	Spectral skewness variance	Timbral	0.0582
	Key key=D	Tonal	-0.0534
	Non-western (CUL, musical culture)	Semantic	0.0518

Table 6.3 – *Continued from previous page*

Participant	Predictor	Musical dimension	Signed importance score
	Western (CUL, musical culture)	Semantic	-0.0518
	Trance (GEL, genre)	Semantic	-0.0438
	Chords scale=major	Tonal	-0.0423
	Chords scale=minor	Tonal	0.0423
	THPCP #6	Tonal	-0.0159
	Chords key=E	Tonal	0.00701
	Disco (G3, genre)	Semantic	0.0013
#16	House (GEL, genre)	Semantic	0.157
	Spectral flatness dB variance	Timbral	0.108
	Tristimulus mean #1	Timbral	-0.0692
	Spectral skewness variance	Timbral	0.066
	Second peak BPM	Rhythmic	0.0586
	Spectral flux mean	Timbral	-0.0504
	Disco (G3, genre)	Semantic	-0.0402
	Jazz (G1, genre)	Semantic	0.0384
	Tuning equal tempered deviation	Tonal	-0.0242
	Chords histogram #15	Tonal	0.0217
	Male (OGD, instrumentation)	Semantic	-0.0205
	Female (OGD, instrumentation)	Semantic	0.0205
	Chords histogram #1	Tonal	0.02
	Funk/soul/rnb (G1, genre)	Semantic	0.0171
	Chords scale=major	Tonal	-0.0166
	Chords scale=minor	Tonal	0.0166
	MFCC mean #5	Timbral	0.0138
	Jazz (G3, genre)	Semantic	-0.0129
	Second peak weight	Rhythmic	-0.00672
	Chords histogram #22	Tonal	0.00645
	Second peak spread	Rhythmic	0.0054
	Bark bands mean #14	Timbral	0.00444
	Chords key=G	Tonal	-0.00272
	Spectral complexity variance	Timbral	0.00185
#17	Beats loudness mean	Rhythmic	-0.209
	First peak spread	Rhythmic	0.204
	Second peak weight	Rhythmic	-0.125
	HPCP mean #32	Tonal	-0.072
	Country (G3, genre)	Semantic	-0.0665
#18	Tuning equal tempered deviation	Tonal	-0.131
	Voice (OVI, instrumentation)	Semantic	-0.13
	Instrumental (OVI, instrumentation)	Semantic	0.13
	First peak weight	Rhythmic	-0.114
	Beats loudness mean	Rhythmic	-0.0959
	Second peak BPM	Rhythmic	0.0817
	First peak spread	Rhythmic	0.0801
	THPCP #26	Tonal	-0.0748
	HPCP variance #25	Tonal	-0.0534
	Second peak spread	Rhythmic	0.0501
	HPCP variance #7	Tonal	-0.0297
	Chords key=D	Tonal	-0.025
	Chords strength variance	Tonal	0.0237
	Blues (G1, genre)	Semantic	0.0237

Table 6.3 – *Continued from previous page*

Participant	Predictor	Musical dimension	Signed importance score
	Rhythm'n'blues (G2, genre)	Semantic	-0.0202
	Chords histogram #17	Tonal	0.018
	Chords key=F#	Tonal	-0.014
	Key key=F	Tonal	-0.0109
	Chords histogram #23	Tonal	0.0105
	Key key=A	Tonal	0.00789
	Chords histogram #2	Tonal	-0.00603
	Tristimulus mean #0	Timbral	-0.00235
	Chords key=A	Tonal	-0.00152
	Second peak weight	Rhythmic	-0.000693
	Beats loudness bass variance	Rhythmic	0.000104
	Folk/country (G1, genre)	Semantic	-1.43e-05
#19	Chachacha (RBL, rhythm)	Semantic	0.109
	First peak spread	Rhythmic	0.0624
	Key scale=major	Tonal	-0.0605
	Key scale=minor	Tonal	0.0605
	Onset rate	Rhythmic	-0.0489
	Chords scale=major	Tonal	-0.0463
	Chords scale=minor	Tonal	0.0463
	THPCP #27	Tonal	-0.0448
	First peak weight	Rhythmic	-0.0359
	Tuning equal tempered deviation	Tonal	-0.0343
	House (GEL, genre)	Semantic	0.0342
	THPCP #9	Tonal	0.0314
	Quickstep (RBL, rhythm)	Semantic	-0.031
	Second peak spread	Rhythmic	0.0309
#20	Spectral flatness dB variance	Timbral	0.139
	Second peak spread	Rhythmic	0.0995
	Chords strength variance	Tonal	0.0844
	Tuning equal tempered deviation	Tonal	-0.0782
	Chords histogram #23	Tonal	0.0669
	THPCP #24	Tonal	0.0484
#21	Male (OGD, instrumentation)	Semantic	-0.173
	Female (OGD, instrumentation)	Semantic	0.173
	Jive (RBL, rhythm)	Semantic	-0.16
	Spectral RMS variance	Timbral	0.125
	Techno (GEL, genre)	Semantic	0.0791
	Beats loudness mean	Rhythmic	-0.0629
	Rumba (RBL, rhythm)	Semantic	0.058
	Chords strength mean	Tonal	-0.0309
	Chords scale=major	Tonal	-0.0287
	Chords scale=minor	Tonal	0.0287
	Chords histogram #2	Tonal	-0.0276
	First peak weight	Rhythmic	-0.0261
	Chords histogram #22	Tonal	-0.0235
	Second peak spread	Rhythmic	0.0224
	First peak spread	Rhythmic	0.022
	Mood cluster #4 (MCL, mood)	Semantic	-0.0189
	Bark bands mean #26	Timbral	0.0145
	Tuning equal tempered deviation	Tonal	-0.00976

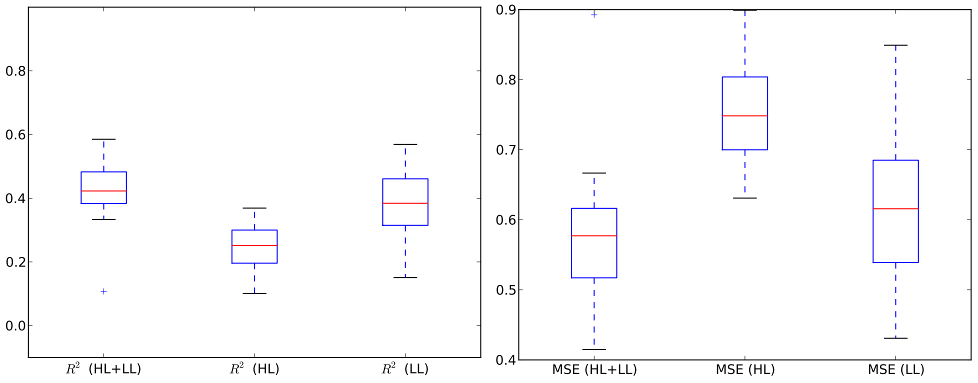
Table 6.3 – *Continued from previous page*

Participant	Predictor	Musical dimension	Signed importance score
	Chords histogram #19	Tonal	0.00425
	Spectral complexity variance	Timbral	0.00219
#22	First peak weight	Rhythmic	-0.223
	First peak spread	Rhythmic	0.174
	Second peak BPM	Rhythmic	0.0896
	Second peak spread	Rhythmic	0.0741
	Bark bands mean #4	Timbral	-0.0715
	THPCP #10	Tonal	0.0626
	Bark bands variance #11	Timbral	0.0539
	Chords histogram #9	Tonal	0.0502
	Chords histogram #18	Tonal	0.0456
	Chords histogram #17	Tonal	0.00635
	Second peak weight	Rhythmic	-0.0058
#23	First peak spread	Rhythmic	0.182
	THPCP #3	Tonal	-0.175
	Bark bands mean #8	Timbral	-0.168
	First peak BPM	Rhythmic	-0.0996

6.2.4 Low-level and semantic models

Finally, we would like to provide further insights on advantages of high-level vs low-level audio content description. We already evidenced the advantage of high-level semantic description over common approaches working with low-level features in the task of measuring non-personalized music similarity (Chapter 4). In contrast, in the present experiment we create a personalized model for each listener. In order to analyze if high-level descriptors provide similar benefits, we have repeated the regression procedure described above with two different predictor sets containing solely high-level semantic descriptors or low-level timbral, temporal and tonal features. Figures 6.6a and 6.6b show goodness-of-fit of final ridge models obtained on all predictors, semantic descriptors, and low-level features. In addition box-plots of the number of selected predictors are presented in Figures 6.7 and 6.7.

In contrast to our previous observations, models with semantic predictors preformed worse than models with low-level or both types of predictors according to the obtained median values of R^2 and MSE. We conducted three pairwise T-tests in order to assess statistical significance of the observed differences in R^2 values. No statistically significant differences were found between models working on all predictors and solely on low-level predictors ($t(44) = 1.60; p \approx 0.116$). In contrast, statistical differences were found between high-level and low-level models ($t(45) = 4.76; p < 0.001$) and between models working on all predictors and high-level models ($t(46) = 6.84, p < 0.001$). The median number of selected predictors also differed in respect to the predictor set. The non-parametric Wilcoxon signed-rank test revealed the same pattern of differences ($p = 0.082$, $p < 0.001$, and $p < 0.001$, respectively). The median R^2



(a)

(b)

Figure 6.6: Box plots of the obtained R^2 (a) and MSE (b) for final user models trained on all predictors (HL+LL), semantic descriptors (HL), and low-level features (LL).

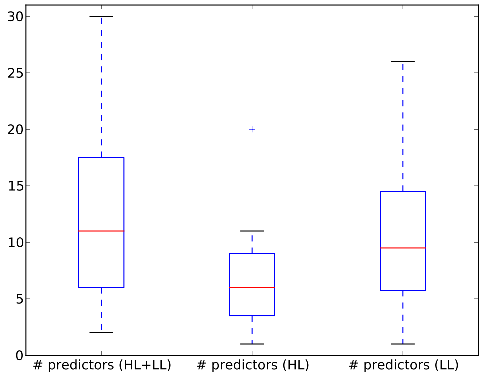


Figure 6.7: Box plot of the number of predictors selected for final user models when trained on all predictors (HL+LL), semantic descriptors (HL), and low-level features (LL).

for the final models using all predictors was ≈ 0.423, while the median R^2 for low-level and high-level models was 0.384 and 0.252, respectively. On average, simple models (6-11 predictors) were selected using all three sets of predictors. More predictors were selected for the largest set (all predictors), while high-level models contained the least amount of predictors.

The obtained results suggest better applicability of low-level audio features in the task of preference modeling. However, such features are often difficult to explain. While the computed semantic user models had a low coefficient of regression (less than 0.14 decrease in the R^2 median value), the addition of semantic descriptors did not decrease the performance of user models at all, but extended them with additional semantic facets for 61.2% of our participants. Therefore, we may expect semantic descriptors to be a considerably effective and robust way to extend, but not substitute, low-level acoustic preference models. A similar conclusion was reached in Section 4.9, concerning the advantage of hybrid low-level/semantic music similarity measure. Even more, there might be listeners for whom preference models based solely on low-level features are not feasible and an inclusion of additional semantic categories is essential.

6.2.5 Conclusions

In this chapter we addressed the task of identifying important predictors of music preferences specific to our participants. To this end, we analyzed both low-level audio features and semantic descriptors available from our audio analysis tool for creating user models. We proposed a new approach based on a modeling using linear regression with elastic net regularization (to handle problem of ill-conditioned data) and bagging (to improve stability).

We were able to create user models for 81% of our participants, with an acceptable, though not excellent, goodness-of-fit (median $R^2 \approx 0.4376$ for final models, that is, the models explained approximately 43.8% of variance). For comparison, the best models proposed for the recent Netflix and Yahoo Music competitions (large-scale collaborative filtering on movie and music recommendation) explained 42.6% and 59.3% of variance, respectively (Dror et al., 2011). These models worked on large datasets of metadata (user ratings), and we can expect worse performance for current content-based approaches. We believe the fact of ill-conditioned data (low number of examples, large number of predictors) to be the main reason for lower performance. To address this problem, we deliberately simplified the models by using elastic net regression and by introducing a stability measure. The trade-off between goodness-of-fit and stability naturally provoked lower R^2 and MSE measures. We expect higher goodness-of-fit when more user data is available ($>$350 tracks with associated subjective ratings per user), which was surely problematic in the context of our experiment. We should highlight that the main goal of the present analysis is not to create models with high prediction accuracy, suitable for music recommendation, but to reveal important predictors of music preferences.

The models we constructed ranged in their simplicity, and the simpler models were associated with a poorer performance. Again, this problem can be associated with the lack of user data impeding the stability of selected predictors. Low goodness-of-fit coupled with model simplicity might also signify that not all the real factors of preference are addressed by the predictors at our reach.

In general, we evidenced that there are low-level and high-level audio features which are fundamental for explaining music preferences. Low-level features related to rhythm, tonality, and timbre, and semantic descriptors corresponding to instrumentation and rhythm were found to be important predictors corroborating results from studies on music perception (Section 2.2.2). Our models included semantic descriptors which can be associated with importance of generic referential meaning. In particular, we found genre to be of high importance, but have not found mood to be as important which contradicts the evidence from the psychological studies. However, we still observed the importance of tonal information (major/minor key) which is related to the emotional content of music.

We presented a computational study of music preferences grounded on audio features and semantic descriptors as the predictors of preference. This study is unique in the sense that, to the best of our knowledge, there are no other similar research works trying to infer preference models using a large variety of audio features. The main limitation of this study lies in the amount of available data. Further analysis on a larger (and therefore, well-conditioned) user dataset will be of interest. As well, we hypothesize that complex non-linear patterns of preferences may occur, but they are out of scope of this thesis.

CHAPTER 7

Visualization of music preferences

7.1 Introduction

In previous chapters we have researched on how audio content information, in particular, including inferred semantic categories, can be used for music recommendation and user preference modeling. In this chapter we consider possible applications of the proposed semantic user profile suited to enrich interaction with music recommender systems. Specifically, we focus on the problem of how music preferences can be visualized in a convenient way. To this end, we study how the proposed semantic user profile can be mapped to a visual domain. To the best of our knowledge, this task of translating music-oriented user models into visual counterparts has not been explored previously. This study has been done in collaboration with other authors, to whom we are very grateful for their valuable contributions. We propose a novel approach to depict a user's preferences in form of a *Musical Avatar*, a humanoid cartoon-like character. Although such a task is not directly related to music recommendation, it might be a useful enhancement for recommender systems. In particular, automatic user visualization can provide means to increase user engagement in the system, justify recommendations (e.g., by visualizing playlists), and facilitate social interaction among users.

We operate on the semantic description of the listener's preference set, inferred from low-level timbral, temporal and tonal audio features, and consider three descriptor integration methods to represent user preferences in a compact form suitable for mapping it to a visual domain. We evaluate this visualization approach on 12 subjects and discuss the obtained results. More precisely, we show that the generated visualizations are able to reflect the subjects' core preferences and are considered by the users as a closely resembling, though not perfect, representation of their musical preferences. We would like to acknowledge our colleagues from the Music Technology Group as this research has benefited from several collaborators.

7.2 Semantic representation of music preferences

We follow the proposed preference elicitation strategy (Section 3.3) and semantic descriptor extraction for each track in the listener's preference set (Section 3.4). The retrieved semantic descriptors provide a rich representation of user preferences, which in particular can give valuable cues for visualization. Instead of using their full potential, in this proof-of-concept application we map a reduced subset of descriptors for simplicity reasons. To this end, we select this subset considering the classifiers' accuracy against ground truth values provided by a subset of 5 participants. When selecting the subset, we also intend to preserve the representativeness of the semantic space. We asked these participants to manually annotate their own music collections with the same semantic descriptors as those inferred by the classifiers. We then compared these manual annotations with the classifiers' outputs by Pearson correlation and selected the best performing descriptors. The observed correlation values for all semantic descriptors varied between -0.05 and 0.70 with the median being 0.40. The subset of 17 descriptors was selected with the majority of correlations (for 14 descriptors) being greater than 0.40. The resulting descriptors, which are used by the proposed visualization approach, are presented in Table 7.1.[1]

7.3 Descriptor summarization

Having refined the semantic descriptors for the computed user profile, we are faced with a problem of their summarization. Our user profile consists of a set of vectors of semantic descriptors, but we consider a single image, representing general preferences of the listener, as an output of our visualization system. Therefore, we opt to use the centroid of the user's preference set as a rough approximation of the overall music preferences.

[1] Note that in this study was done before the analysis presented in the previous Chapter 6, and we did not take an advantage of the results of the analysis presented there.

Table 7.1: Selected descriptors, and the corresponding music collections used for regression, per category of semantic descriptors (i.e., genre, moods & instruments, and others) used for visualization.

Genre	Moods & Instruments	Others
Electronic (G1)	Happy (MHA)	Party (OPA)
Dance (G2)	Sad (MSA)	Vocal (OVI)
Rock (G2)	Aggressive (MAG)	Tonal (OTN)
Classical (G3)	Relaxed (MRE)	Bright (OTB)
Jazz (G3)	Electronic (MEL)	Danceable (ODA)
Metal (G3)	Acoustic (MAC)	

We consider different summarization methods to obtain a compact representation which can be mapped to the visual domain. With these summarization strategies we explore the degree of descriptor resolution necessary for optimal visual representation. These strategies can be based on continuous or discrete values, and therefore lead to visual elements of continuous or discrete nature (e.g., size). The idea behind this exploration is related to the possibility that users might prefer simpler objects (discrete visual elements such as presence or absence of a guitar) or more complex ones (continuous elements such as guitars of different sizes) depicting subtle variations of preferences.

We summarize the user model across individual tracks to a single multidimensional point in a semantic descriptor space as in the case of the *SEM-MEAN* approach we considered for music recommendation (Section 5.4.1). We first standardize each descriptor to remove global scaling and spread; i.e., for each track from the user's preference set we subtract the global mean and divide by the global standard deviation. We estimate the reference means ($\mu_{R,i}$) and standard deviations ($\sigma_{R,i}$) for each descriptor from the representative in-house music collection of $100,000$ music excerpts used for the subjective evaluation of music recommendation approaches in Experiment 3 (Section 5.7.1). Moreover, we range-normalize the aforementioned standardized descriptor values according to the following equation:

$$N_i = \frac{d_i - min}{max - min}, \tag{7.1}$$

where d_i is the standardized value of descriptor i, and since d_i has zero mean and unit variance, we set the respective min and max values to -3 and 3, since according to Chebyshev's inequality at least 89 % of the data lies within 3 standard deviations from its mean value (Grimmett & Stirzaker, 2001). We clip all resulting values smaller than 0 or greater than 1. The obtained scale can be seen as a measure of preference for a given category, and is used by the visualization process (see Section 7.4). We then summarize the descriptor values across tracks by computing the mean for every normalized descriptor ($\mu_{N,i}$).

At this point, we consider three different methods to quantize the obtained mean values. These quantization methods convey different degrees of data variability, and are defined as follows:

- *Binary* forces the descriptors to be either 1 or 0, representing only two levels of preference (i.e., 100% or 0%). We quantize all $\mu_{N,i}$ values below 0.5 to zero and all values above (or equal) 0.5 to one.

- *Ternary* introduces a third value representing a neutral degree of preference (i.e., 50%). We perform the quantization directly from the original descriptor values, that is, we calculate the mean values for every descriptor (μ_i) and quantize them according to the following criteria:

$$Ternary_i = \begin{cases} 1 & \text{if } \mu_i > (\mu_{R,i} + th_i), \\ 0.5 & \text{if } (\mu_{R,i} - th_i) \leq \mu_i \leq (\mu_{R,i} + th_i), \\ 0 & \text{if } \mu_i < (\mu_{R,i} - th_i), \end{cases} \tag{7.2}$$

where $th_i = \sigma_{R,i}/3$.

- *Continuous* preserves all possible degrees of preference. We maintain the computed $\mu_{N,i}$ values without further changes.

At the end of this process we obtain three simplified representations of the user model, each of them consisting of the same 17 semantic descriptors.

7.4 Visualization approach

In order to generate the *Musical Avatar*, we convert the summarized semantic descriptors to a set of visual features. According to MacDonald et al. (2002), individual, cultural and sub-cultural musical identities emerge through social groups concerning different types of moods, behaviors, values or attitudes. Further evidence of relation between identities and music preferences, and an important role of social ties and expression through music, was presented in Section 2.2.2. We apply the cultural approach of representing urban tribes (Maffesoli, 1996), since in these tribes, or subcultures, music plays a relevant role in both personal and cultural identities. Moreover, they are often identified by specific symbolisms which can be recognized visually.

Therefore, we decided to map the semantic descriptors into a basic collection of cultural symbols. As a proof-of-concept, we opt for an iconic cartoon style of visualization. This choice is supported by a number of reasons; firstly, this style is a less time-consuming technique compared to other approaches more focused on realistic features (Ahmed et al., 2005; Petajan, 2005; Sauer & Yang, 2009). Secondly, it is a graphical medium which, by eliminating superfluous features, amplifies the remaining characteristics of a personality (McCloud, 2009). Thirdly, there are examples of existing popular avatar collections of this kind such as Meegos[2] or Yahoo Avatars.[3]

In our approach the relevant role is played by the graphical symbols, which are filled with arbitrary colors related to them. Although colors have been successfully associated with musical genres (Holm et al., 2009) or moods (Voong & Beale, 2007), the disadvantage of using only colors is the difficulty to establish a global mapping due to reported cultural differences about their meaning.

In our design, we consider the information provided by the selected descriptors and the design requirements of modularity and autonomy. Starting from

[2] http://meegos.com

[3] http://avatars.yahoo.com

a neutral character,[4] we divide the body into different parts (e.g., head, eyes, mouth). For each of the parts we define a set of groups of graphic symbols (graphic groups) to be mapped with certain descriptors. Each of these graphic groups always refers to the same set of descriptors. For example, the graphic group corresponding to the mouth is always defined by the descriptors from the categories “Moods and Instruments” and “Others” but never from “Genre” category. The relation between graphic groups and categories of the semantic descriptors is presented in Table 7.2. For this mapping, we consider the feasibility of representing the descriptors (e.g., the suit graphic group is more likely to represent a musical genre compared to the other descriptor categories). We also bear in mind a proportional distribution between the three main descriptor categories vs. each of these graphic groups in order to notice them all. However, in accordance with the cartoon style some of these graphic groups refer to all three main descriptor categories because they can highlight better the most prominent characteristics of the user’s profile, and also they can represent a wide range of descriptors (e.g., the head and complement graphic groups). In addition to the listed graphic groups, we introduce a label to identify the gender of the avatar, each providing a unique set of graphic symbols.

Besides the body elements, we also add a set of possible backgrounds to the graphic collection in order to support some descriptors of the “Others” category such as “party”, “tonal”, or “danceable”. In addition, the “bright” descriptor is mapped to a grey background color that ranges from RGB(100,100,100) to RGB(200,200,200). We note that our decisions on the design, and in particular on the descriptor mapping presented in Table 7.2 are arbitrary, being a matter of choice, of visual and graphic sense, and common sense according to many urban styles of self-imaging.

Table 7.2: Mapping of the descriptor categories to the graphic groups.

	Descriptor categories		
Graphic Group	Genre	Moods & Inst.	Others
Background			•
Head	•	•	•
Eyes		•	•
Mouth		•	•
Complement	•	•	•
Suit	•		•
Hair	•		
Hat	•	•	
Complement2			•
Instrument	•	•	

[4] A neutral character corresponds to an empty avatar. It should be noted that the same representation can be achieved if all normalized descriptor values are set to 0.5 meaning no preference for any descriptor at all.

Table 7.3: Vector representation example: user profile vs. the instrument graphic group (continuous summarization). A visual element with the minimum distance to the user profile is selected (in this case, the turntable).

Category	Descriptor	User Profile				
Genre	Classical (G3)	0.0	0.0	0.0	0.0	0.0
Genre	Electronic (G1)	1.0	0.0	0.0	0.0	1.0
Genre	Jazz (G3)	0.0	0.0	1.0	0.0	0.0
Genre	Metal (G3)	0.0	0.0	0.0	0.0	0.0
Genre	Dance (G2)	1.0	0.0	0.0	0.0	0.0
Genre	Rock (G2)	0.5	1.0	0.0	0.0	0.0
Moods & Inst.	Electronic (MEL)	1.0	0.0	0.0	0.0	1.0
Moods & Inst.	Relaxed (MRE)	0.0	0.0	0.0	0.0	0.0
Moods & Inst.	Acoustic (MAC)	0.8	0.0	0.0	1.0	0.0
Moods & Inst.	Sad (MSA)	0.0	0.0	0.0	0.0	0.0
Moods & Inst.	Aggressive (MAG)	0.0	1.0	0.0	0.0	0.0
Moods & Inst.	Happy (MHA)	1.0	0.0	0.0	0.0	0.0
Distance to user profile			2.43	2.62	2.07	**1.70**

We construct a vector space model and use a Euclidean distance as a measure of dissimilarity to represent the user's musical preferences in terms of graphic elements. For each graphic group we choose the best graphic symbol among the set of all available candidates, i.e., the closest to the corresponding subset of the user's vector model (see Table 7.3 for an example of the vector representation of these elements). This subset is defined according to the mapping criteria depicted in Table 7.2. As a result, a particular *Musical Avatar* is generated for the user's musical preferences. All graphics are done in vector format for scalability and implemented using Processing[5] (Reas & Fry, 2007).

According to the summarization methods considered in Section 7.3, the mapping is done from either a discrete or continuous space resulting in different data interpretations and visual outputs. These differences imply that in some cases the graphic symbols have to be defined differently. For instance, the "vocal" descriptor set to 0.5 in the case of *continuous* method means "she likes both instrumental and vocal music", while this neutrality is not present in the case of the *binary* method. Furthermore, in the *continuous* method, properties such as size or chromatic gamma of the graphic symbols are exploited while this is not possible within the discrete vector spaces. Figure 7.1 shows a graphical example of our visualization strategy where, given the summarized binary user model, the best (i.e., the closest by a Euclidean distance) graphic symbol for each graphic group is chosen. Figure 7.2 shows a sample of *Musical Avatars* generated by the three summarization methods and Figure 7.3 shows a random sample of different *Musical Avatars*.

[5] http://processing.org

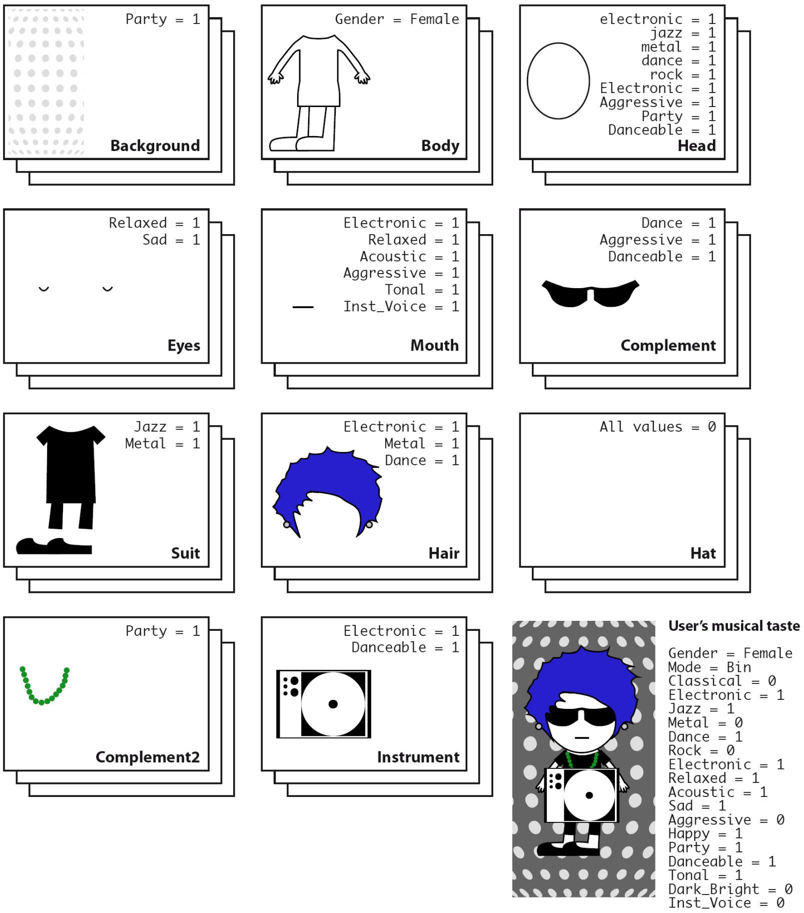


Figure 7.1: Example of the visualization approach. It can be seen how the descriptor values influence the selection of the different graphic elements used to construct the avatar. The values inside the graphic element boxes represent all possible descriptor values that can generate the presented element.

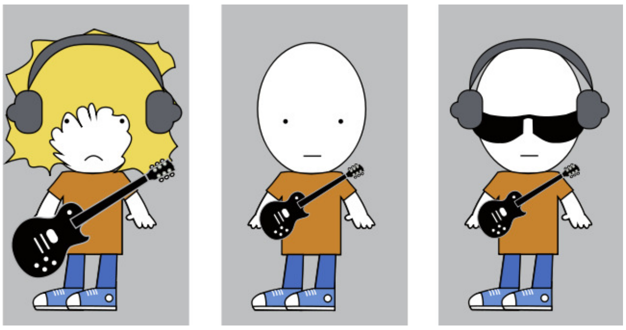


Figure 7.2: Sample *Musical Avatars* generated by the three summarization methods (i.e., from left to right, *binary*, *ternary*, and *continuous*) for the same underlying user model. Notice the differences in guitar and headphones sizes among the generated avatars.

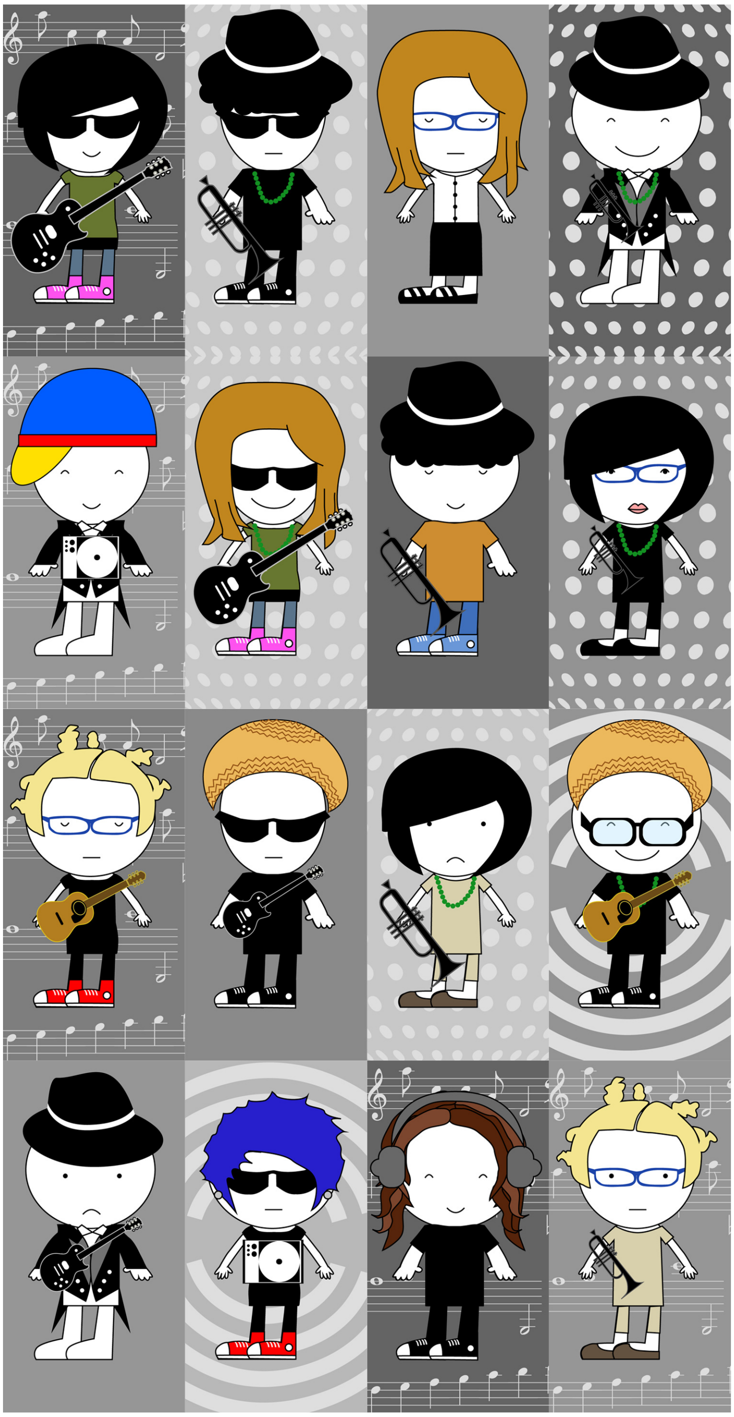


Figure 7.3: A random sample of *Musical Avatars*.

7.5 Evaluation methodology

We carried out a subjective evaluation on our 12 subjects being a subset of the population described in Section 3.3.2. They included 8 males and 4 females with the average age of 34 years ($\mu = 33.83, \sigma = 5.2$) and a high interest in music ($\mu = 9.58, \sigma = 0.67$). For each participant, we generated three *Musical Avatars* corresponding to the three considered summarization methods. We then asked the participants to answer a brief evaluation questionnaire. The evaluation consisted in performing the following two tasks.

In the first task, we asked the participants to manually assign values for the 17 semantic descriptors used to summarize their musical preferences (see Table 7.1). We requested a real number between 0 and 1 to rate the degree of preference for each descriptor (e.g., 0 meaning "I don't like classical music at all" up to 1 meaning "I like classical music a lot" in the case of the "classical" descriptor). For the second task, we first showed 20 randomly generated examples of the *Musical Avatars* in order to introduce their visual nature. We then presented to each participant six avatars: namely, the three images generated from her/his own preference set, two randomly generated avatars, and one neutral avatar. We asked the participants to rank these images assigning the image that best express their musical preferences to the first position in the rank (i.e., rank = 1). Finally, we asked for a written feedback regarding the images, the evaluation procedure, or any other comments.[6] A screenshot of the evaluation is presented in Appendix B.

7.6 Evaluation results

From the obtained data we first analyzed the provided rankings to estimate the accuracy of the visualization methods examined in the questionnaire. To this end, we computed the mean rank for each method. The resulting means and standard deviations are reported in Table 7.4. We tested the effect of the method on the ratings obtained from the subjects using a within-subjects ANOVA. A prerequisite for this type of experimental design is to check the sphericity assumption (i.e., all the variances of the differences in the sampled population are equal) using the Mauchly's test, which indicated that the assumption could be trusted (Huberty & Olejnik, 2006). The effect of the visualization method was found to be significant (Wilks Lambda = 0.032, $F(4,7) = 52,794$, $p < 0.001$). Pairwise comparisons (a least significant differences t-test with Bonferroni correction, which conservatively adjusts the observed significance level based on the fact that multiple comparisons are made) revealed significant differences between two groups of avatars: on one side, the random and the neutral avatars (getting ratings that cannot be con-

[6] More *Musical Avatars* are available online: http://mtg.upf.edu/project/musicalavatar.

Table 7.4: Mean ranks and standard deviations for the different visualization methods obtained in the user evaluation. The random column corresponds to the average values of the individual random results (see text for details).

	Continuous	Binary	Ternary	Random	Neutral
μ	1.73	2.27	2.91	4.28	5.18
σ	0.79	1.49	1.45	1.16	0.98

sidered different from each other) and, on the other side, the *binary*, *ternary*, and *continuous* avatars (which get ratings that are statistically different from the random and the neutral ones, but without any significant difference between the three). The differences between those two groups of avatars are clearly significant ($p < 0.005$) except for the differences between random and *ternary*, and between *binary* and neutral, which are only marginally significant ($p \leq 0.01$).

We also assessed the significance of the summarized description of musical preferences by estimating how the computed representation performs against a randomly generated baseline. Therefore, we first computed the Euclidean distance between the obtained descriptor vector representing the user profile (standardized and range-normalized) and the vector containing the participants' self-assessments provided in the first task of the evaluation. We then generated a baseline by averaging the Euclidean distances between the self-assessments and 10 randomly generated vectors. Finally, a t-test between the algorithm's output ($\mu = 0.99, \sigma = 0.32$) and the baseline ($\mu = 1.59, \sigma = 0.25$) showed a significant difference in the sample's means ($t(11) = -5.11, p < 0.001$). Additionally, Figure 7.4 shows box plots of the obtained dissimilarities.

From the obtained results, we first observe that the generated description based on audio content analysis shows significant differences when compared to a random assignment. The mean distance to the user-provided values is remarkably smaller for the generated data than for the random baseline; i.e., the provided representations reasonably approximate the users' self-assessments in terms of similarity. Furthermore, Table 7.4 clearly shows a user preference for all three proposed quantization methods over the randomly generated and the neutral *Musical Avatars*. In particular, the *continuous* summarization method has been found top-ranked, followed by the *binary* and *ternary* quantization methods. This ranking, given the ANOVA results, should be taken just as approximative. Specifically, the conducted ANOVA did not reveal a clear particular preference for any of the three considered methods; i.e., no statistically significant difference between simple avatars with discrete visual elements and more complex ones with continuous elements were found. On the other hand, we can see that the neutral avatar is less preferred than the random avatars.

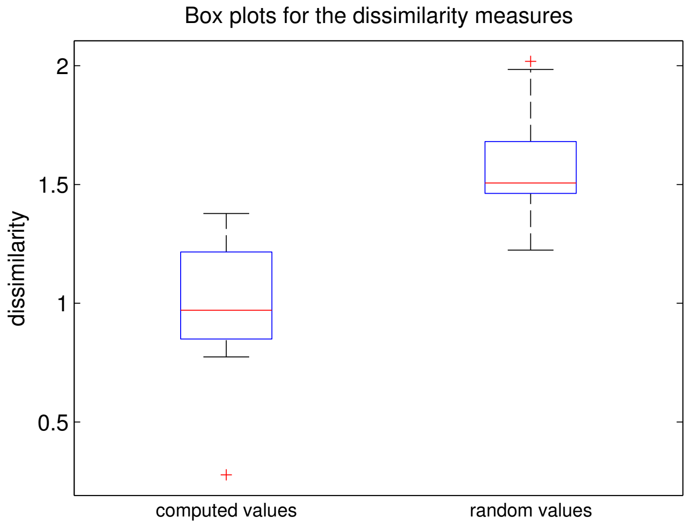


Figure 7.4: Box plots of the dissimilarity estimation. The Euclidean distance between the ground truth labels and the computed descriptors shows a significantly lower mean than the one obtained using 10 randomly generated descriptor vectors. Red crosses stand for extreme outliers.

This suggests that the users prefer images that carry some information (even if it does not match the users' preferences) rather than avatars lacking of visual features. This poses the problem of visualizing users with varied musical preferences (i.e., mean values of a majority of the descriptors close to 0.5), especially in the case of the *ternary* quantization. We have expected to observe the difference between the random and neutral avatars, however it was not statistically significant, probably due to the small number of participants in the study.

Evaluation of the participants' comments can be summarized as follows. First, we can observe a general tendency towards an agreement on the representativeness of the *Musical Avatar*. As expected, some subjects reported missing categories to fully describe their musical preferences (e.g., country music, musical instruments). This suggests that the provided semantic descriptors seem to grasp the essence of the user's musical preference, but fail to describe subtle nuances in detail. This could be explained by the fact that we use a reduced set of semantic descriptors in our prototype (17 descriptors out of the 62 initially extracted for the proposed semantic user profile). Indeed, by providing better semantic descriptions of the musical content under consideration (i.e. better classifiers and descriptors), the algorithm's accuracy in describing these aspects would benefit to a great extent. In consequence, since we are working with state-of-the-art algorithms, the available tools are only able to solve the problem on a very coarse level. Finally, some participants could not decode the meaningfulness of some visual features (e.g., glasses, head shape) because of the arbitrarienss of the mappings. This information will be considered in our future work for refining the mapping strategy. According to the obtained results, we observed participants' preference for all three summarization methods based on the proposed user model over the baselines. In general, we conclude that the *Musical Avatar* provides a reliable, albeit coarse, visual representation of the user's musical preferences.

7.7 General discussion and possible enhancements for recommender systems

In what follows we comment on the implications of the presented approaches for the user's interaction as well as future implementations of "final systems", which unite both recommendation and preference visualization approaches into a single, interactive music recommender interface. The mapping of the semantic dimensions to visual features, resulting in the *Musical Avatar*, enables an intuitive, yet still arbitrary, depiction of musical preferences. This by itself enriches and facilitates the user's interaction process, an appealing feature for any recommender system. Furthermore, allowing the user to interact and manipulate graphical representations offers a straightforward path towards user adaptive models. One possible extension here is the filtering of music recommendations according to the presence or absence of certain visual features of the *Musical Avatar*. This allows users to actively control the output of the music recommender by selecting certain visual attributes which are connected to acoustic properties via the mapping described in Section 7.4. Also, the iconic *Musical Avatar* may serve as a badge, reflecting a quick statement of one's musical preferences, with possible applications in online social interaction. This use-case is highly consistent with finding of research on social factors of music preferences (see Section 2.2.2). Moreover, users can share preferences related to the generated avatars or group together according to similar musical preferences represented by the underlying user models.

Both aforementioned applications can be easily united into a single interactive recommender system. In addition to the already discussed music recommendation and static preference visualization, the concepts introduced in the present work can be extended to reflect time-varying preferences. For example, an underlying user model can be computed considering different time periods (e.g., yesterday, last week, last month). Also, tracking preferences over time enables the generation of "preference time-lines", where *Musical Avatars* morph from one period to the next, while users can ask for recommendations from different periods of their musical preferences.

Moreover, in the visualization application, exploiting multiple instances of preference sets can alleviate the limitations introduced by a single preference set. Multiple graphical instances can be used to visually describe different subsets of a music collection, thus serving as high-level tools for media organization and browsing. Hence, recommendations can be directed by those avatars, introducing one additional semantic visual layer in the recommendation process. Using multiple representations can help to better visually depict preferences of certain users, where a single avatar is not sufficient for describing all facets of their musical preferences. Moreover, users may want to generate context-dependent avatars, which can be used for both re-playing preference items or listening to recommendations depending on the context at hand (e.g., one may use his avatar for happy music at a party or listen to recommendations from the “car” avatar while driving).

CHAPTER 8

Conclusions and future work

8.1 Introduction

Let us shortly recapitulate the major contents of the present thesis. We focused our work on music recommendation, and addressed the related problems of music preference elicitation and music similarity measurement. Since 2008, when this thesis was started till the present date, there was a limited amount of research on music recommendation and there was a lack of meta-studies systematizing the existing approaches. In Chapter 2 we introduced the state-of-the-art of music preferences from the perspective of music cognition, psychology, and sociology. We then conducted an extensive review of literature on music recommendation, and provided a systematization by the sources of information used, underlying algorithms, and evaluation methodologies, which was missing in the existing literature. We have found that the existing studies on music recommendation employed objective metrics in user simulation tests, but very rarely conducted real user-centered A/B listening tests. The absolute majority of them did not consider measuring actual usefulness of the recommendations and their novelty for the listener. The existing approaches proposed to employ audio and metadata. However, the studies proposing audio content-based approaches to recommendation lacked proper comparison with the state-of-the-art metadata-based approaches, and the subjective quality of both types of recommenders was rarely assessed. Content-based approaches worked on limited amounts of low-level audio features, while we believe it is advisable to incorporate more musical knowledge that is already computable with our state-of-the-art algorithms. Therefore, we aimed to improving content-based recommender systems by introducing high-level semantic descriptors. In Chapter 3 we proposed to build a semantic user model from explicitly given preference examples by applying automatic inference of high-level concepts from the audio content. Recommendation approaches can be based on music similarity measurement between the user model and tracks in music collection. Therefore, in Chapter 4 we focused on the problem of content-based music similarity. We proposed and evaluated a novel semantic similarity measure together with a hybrid low-level/semantic approach, which allowed for an improved music similarity estimation, according to the conducted objective and subjective evaluations. In Chapter 5 we focused on different content-based, metadata-based and hybrid recommendation approaches. We designed a new evaluation methodology, which makes possible to assess subjective usefulness and novelty of recommendations for the listener in subjective listening tests. We employed the proposed similarity measures in the context of recommendation, and studied how their simple filtering by genre metadata can improve the performance. We used the state-of-the-art approaches, working by means of collaborative filtering and social tags, as our baselines, and we proposed our own approach working with editorial metadata. In Chapter 6 we studied how audio content information can be exploited to provide quantitative insights on the factors of music preferences from both acoustical and semantic perspectives. Finally, we demonstrated a novel application of the proposed semantic user model for music preference visualization in Chapter 7.

8.2 Summary of contributions

- We proposed novel audio content-based measures of music similarity. In particular, we proposed to use high-level semantic description of music tracks inferred by SVMs from the low-level timbral, temporal, and tonal features, and use their hybrid combination. The measures were evaluated both objectively and subjectively in A/B listening tests, and were ranked among the best results within the MIREX evaluations. Our approaches performed comparably to the current state of the art, providing satisfactory music similarity estimation. Furthermore, semantic categorization used in our measures allows justification to users of the provided similarity estimations, and thereby it can increase the transparency of the final systems.

- We proposed a number of content-based distance-based approaches to music recommendation based on these similarity measures. We demonstrated their advantage over baseline timbral methods, and revealed the fact, that simple genre/style tags can be a reasonable source of information to provide recommendations superior to the common low-level timbral music similarity based on MFCCs. We demonstrated how filtering of the proposed approaches by simple genre metadata can significantly improve performance up to the level of the state-of-the-art metadata-based systems in terms of user satisfaction ratings. Although the overall amount of relevant recommendations was smaller than for metadata-based systems, the proposed approaches provided a larger amount of novel relevant recommendations, therefore, being well-suited for music discovery.

- We proposed a novel lightweight approach to music recommendation based on publicly available editorial metadata as an alternative to systems working with cumbersome and commercially withhold user ratings and social tags. We observed, that it is also comparable to the state-of-the-art metadata-based approaches and is well-suited especially for the case of playlist generation.

- We have conducted a comprehensive evaluation of the proposed approaches against a number of state-of-the-art recommenders in subjective A/B listening tests. To this end, we employed a novel evaluation methodology for music recommendation, which takes into account novelty factors, together with different behavioral aspects of satisfaction with the provided recommendations. Our evaluations provide reliable insights on the nature of the considered content-based and metadata-based methods, including commercial recommenders.

- In general, we evidenced the advantage of adding semantic description to the low-level audio feature representations for both music similarity and music recommendation tasks. Our proposed content-based preference elicitation strategy takes advantage of automatic semantic description of the preference examples explicitly provided by the listener and is suited for various applications, such as music recommendation and music preference visualization. The proposed semantic user model is able to shrink the gap between low-level signal features and human-level judgments about music, and is able to provide insights on music preferences in an understandable form for humans.

- We provided computational insights on the important factors of music preferences by analyzing audio content. The results correlate with the existing research on music cognition: we have evidenced the importance of rhythm, tonality and timbre descriptors as well as semantic categories on the prediction of preference.

The outcomes of the research carried out in this thesis have been published in several papers in international conferences and journals (see Appendix D). Besides, a part of conducted research related to music similarity measures was incorporated into an existing commercial music recommendation engine, meanwhile the proposed preference elicitation and visualization approach was presented online at an international conference and on a public exhibition (see Appendix A), and was featured in public media.

8.3 Open issues and future perspectives

The experiments conducted in this thesis revealed that state-of-the-art content-based algorithms for music recommendation are still inferior in their performance compared to the state-of-the-art metadata-based approaches. Nevertheless, this difference can be greatly diminished employing a minimum amount of cheap genre metadata. Furthermore, we demonstrated that alternative metadata sources can be effectively used to achieve the same performance on the example of our approach working exclusively on editorial metadata. Given that, we still have evidenced that all considered approaches reached only moderate (above-average) levels of user satisfaction. This means that still there is no method, metadata-based nor content-based, which could cogently address the problem of music recommendation.

Considering the limitations of our study, we would like to note that we employed small samples of subjects (up to 27 music enthusiasts) that might not represent the general population. We nevertheless observed statistical significant differences which, in this context, mean that the detected trends are strong enough to override the individual differences or potentially large variability that might be observed in small-size samples of listeners. We also believe that users of music recommender systems, at least to date, are mainly music enthusiasts, and hence we have properly and sufficiently sampled that population. More importantly, to the best of our knowledge, the few existing research studies on music recommendation involving evaluations with real participants are significantly limited in the trade-off between the number of participants and the number of evaluated tracks per method by a particular user, as we discussed in Section 2.3.2. Furthermore, no studies on human evaluation of visualization approaches considering musical preferences are known to the authors, and we believe this can be a fruitful direction for further research.

The proposed content-based semantic user model and the approach to its visualization can be used as basic tools for Human Computer Interaction to enrich the experience with music recommender systems. A number of innovative personalized interfaces for understanding, discovering, and manipulating music recommendations can be built on top of our developed methodologies. In what follows, we highlight a number of open issues and future perspectives regarding the work presented in this thesis.

8.3.1 Role of user-context in music recommendation

Can the achieved above-average level of performance signify that both types of approaches, metadata-based and content-based, are possibly reaching a glass ceiling? We consciously left out of this thesis' scope a question of how user-context can be integrated into a recommender system. However, we hypothesize that the reason might be due to the absence of user context information within the considered systems. In Section 2.2.2 we highlighted the evidence of importance of the "use-of-music" factor and the listener's context in general for understanding music preference. We believe the ability to understand the user's current needs (i.e., the required use of music) to be a very important facet of recommender systems of the future. Fortunately, this emerging topic of research gains importance within recommender system and music information retrieval community (Schedl & Flexer, 2012).

8.3.2 Improvements of proposed approaches

We may also hypothesize that finding better audio features will lead to higher performance of content-based approaches: some the high-level and mid-level aspects of music are still missing in our research. For example, incorporating better rhythmic features is very promising, as we have evidenced great importance of the rhythm in the description of music preferences. Fulfilling musical dimensions of percussiveness, smoothness, noisiness, and rhythmic complexity, suggested by ongoing research in Last.fm (Sutton, 2012), can greatly improve preference models and performance of content-based music recommenders. Melodic features have been shown to be complementary to timbral features (Salamon et al., 2012) and can be used to expand timbral descriptions. As well, with recent advances in auto-tagging, more semantic descriptors can be implemented with reliability comparable to the inconsistency between humans' annotations themselves. Recognition of specific music styles from audio is challenging, and we have demonstrated the advantage of genre metadata. Improving the accuracy of semantic descriptors is another important task. Their quality can be significantly conditioned by the underlying ground truths and the choice of the classifier.

Furthermore, we hypothesize that designing personalized music similarity measures instead of "universal" measures, identical for all users, can bring further advantage for distance-based recommendation approaches (Stober & Nürnberger, 2009). Better hybrid combinations of audio and metadata information, in particular, exploiting the proposed editorial metadata, are to be considered as well.

8.3.3 Non-obtrusive user feedback

Finally, alternative methods for gathering the preference set can be employed. Since selecting representative music tracks may be a boring and exhausting task for certain users, data-driven approaches can be applied. Audio content-based methods may be used to infer preference items from the user's personal collection by, for instance, clustering the collection according to certain musical facets to find central elements within each cluster (i.e., centroids). Additionally, listening statistics or personal ratings of particular items can be used to infer musical preferences without actually processing a full music collection.[1] Nevertheless, such an implicit inference of a preference set can lead to noisy

[1] A demo of such a music recommender/visualization system working on the proposed principles, but taking listening statistics instead of explicitly given preference set, is available online at http://mtg.upf.edu/project/musicalavatar

representations or to the lack of coverage of all possible facets of the user's musical preferences (see also Section 2.3.1).

8.3.4 Sociological and psychological issues

To the best of our knowledge, a problem of how to address sociological and psychological issues related to music recommendation remains unexplored. This problem is hard to address computationally and we believe it to be a crucial problem for music recommender systems. In particular, determining whether particular music pieces are preferred by a listener due to psychological and sociological factors rather than acoustical ones would be of great interest. These factors may include recall of associated social relations, life events of the past, or pure conditioning, i.e., just because some music item has been presented in association with some rewarding stimulus (Lamont & Webb, 2009). Nevertheless, solving this problem would require explicit user feedback of such psychological and sociological factors which could significantly complicate the design of recommender systems.

Dmitry Bogdanov, Barcelona, June 17, 2013.

Bibliography

Abdullah, M. B. (1990). On a robust correlation coefficient. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *39*(4), 455–460.

Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality. In *Int. Conf. on Very Large Data Bases (VLDB'05)*, pp. 901–909.

Ahmed, N., de Aguiar, E., Theobalt, C., Magnor, M., & Seidel, H.-P. (2005). Automatic generation of personalized human avatars from multi-view video. In *ACM Symp. on Virtual Reality Software and Technology (VRST'05)*, pp. 257–260.

Alpaydin, E. (2004). *Introduction to Machine Learning.* MIT Press.

Aman, P. & Liikkanen, L. (2010). A survey of music recommendation aids. In *ACM Conf. on Recommender Systems. Workshop on Music Recommendation and Discovery (Womrad 2010).*

Amatriain, X. (2005). *An Object-Oriented Metamodel for Digital Signal Processing with a focus on Audio and Music.* Ph.D. thesis, UPF, Barcelona, Spain.

Amatriain, X., Pujol, J., & Oliver, N. (2009). I like it... i like it not: Evaluating user ratings noise in recommender systems. *User Modeling, Adaptation, and Personalization*, *5535/2009*, 247–258.

Arevalillo-Herráez, M., Domingo, J., & Ferri, F. J. (2008). Combining similarity measures in content-based image retrieval. *Pattern Recognition Letters*, *29*(16), 2174–2181.

Aucouturier, J. J. (2009). Sounds like teen spirit: Computational insights into the grounding of everyday musical terms. In J. Minett & W. Wang (Eds.) *Language, Evolution and the Brain*, Frontiers in Linguistics, pp. 35–64. Taipei: Academia Sinica Press.

Aucouturier, J. J. & Pachet, F. (2002). Music similarity measures: What's the use. In *Int. Conf. of Music Information Retrieval (ISMIR'02)*, p. 157–163.

Aucouturier, J. J., Pachet, F., & Sandler, M. (2005). "The way it sounds": timbre models for analysis and retrieval of music signals. *IEEE Trans. on Multimedia*, *7*(6), 1028–1035.

Baltrunas, L. & Amatriain, X. (2009). Towards time-dependant recommendation based on implicit feedback. In *Workshop on Context-aware Recommender Systems (CARS'09)*.

Barrington, L., Chan, A., Turnbull, D., & Lanckriet, G. (2007a). Audio information retrieval using semantic similarity. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'07)*, vol. 2, pp. 725–728.

Barrington, L., Oda, R., & Lanckriet, G. (2009). Smarter than genius? human evaluation of music recommender systems. In *Int. Society for Music Information Retrieval Conf. (ISMIR'09)*, pp. 357–362.

Barrington, L., Turnbull, D., Torres, D., & Lanckriet, G. (2007b). Semantic similarity for music retrieval. In *Music Information Retrieval Evaluation Exchange (MIREX'07)*. Available online: http://www.music-ir.org/mirex/abstracts/2007/AS_barrington.pdf.

Baur, D. & Butz, A. (2009). Pulling strings from a tangle: visualizing a personal music listening history. In *Int. Conf. on Intelligent User Interfaces (IUI'09)*, pp. 439–444.

Berenzweig, A., Ellis, D. P. W., & Lawrence, S. (2003). Anchor space for classification and similarity measurement of music. In *Int. Conf. on Multimedia and Expo (ICME'03)*, vol. 1, pp. 29–32.

Berlyne, D. E. (1974). *Studies in the new experimental aesthetics: steps toward an objective psychology of aesthetic appreciation*. Hemisphere Pub. Corp.

Bertin-Mahieux, T., Eck, D., Maillet, F., & Lamere, P. (2008). Autotagger: A model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research*, *37*(2), 115–135.

Bertin-Mahieux, T., Eck, D., & Mandel, M. (2010). Automatic tagging of audio: The state-of-the-art. In W. Wang (Ed.) *Machine Audition: Principles, Algorithms and Systems*. IGI Publishing.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "Nearest neighbor" meaningful? In *Database Theory — ICDT'99*, no. 1540 in Lecture Notes in Computer Science, pp. 217–235. Springer Berlin Heidelberg.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bosnić, Z. & Kononenko, I. (2008). Comparison of approaches for estimating reliability of individual regression predictions. *Data & Knowledge Engineering*, *67*(3), 504–516.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Brossier, P. M. (2007). *Automatic Annotation of Musical Audio for Interactive Applications*. Ph.D. thesis, QMUL, London, UK.

Bu, J., Tan, S., Chen, C., Wang, C., Wu, H., Zhang, L., & He, X. (2010). Music recommendation by unified hypergraph: combining social media information and music content. In *ACM Int. Conf. on Multimedia (MM'10)*, p. 391–400.

Bühlmann, P. u. (2002). Analyzing bagging. *The Annals of Statistics*, *30*(4), 927–961.

Cano, P., Gómez, E., Gouyon, F., Herrera, P., Koppenberger, M., Ong, B., Serra, X., Streich, S., & Wack, N. (2006). ISMIR 2004 audio description contest. Tech. rep. Available online: http://mtg.upf.edu/node/461.

Cano, P., Koppenberger, M., & Wack, N. (2005). Content-based music audio recommendation. In *ACM Int. Conf. on Multimedia (MM'05)*, pp. 211–212.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, *96*(4), 668–696.

Celma, O. (2008). *Music recommendation and discovery in the long tail*. Ph.D. thesis, UPF, Barcelona, Spain.

Celma, O. (2010). *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer.

Celma, O. & Herrera, P. (2008). A new approach to evaluating novel recommendations. In *ACM Conf. on Recommender Systems (RecSys'08)*, pp. 179–186.

Celma, O., Herrera, P., & Serra, X. (2006). Bridging the music semantic gap. In *ESWC 2006 Workshop on Mastering the Gap: From Information Extraction to Semantic Representation*. Available online: http://mtg.upf.edu/node/874.

Celma, O., Ramırez, M., & Herrera, P. (2005). FOAFing the music: A music recommendation system based on RSS feeds and user preferences. In *Int. Conf. on Music Information Retrieval (ISMIR'05)*.

Celma, O. & Serra, X. (2008). FOAFing the music: Bridging the semantic gap in music recommendation. *Web Semantics: Science, Services and Agents on the World Wide Web*, *6*(4), 250–256.

Chamorro-Premuzic, T., Swami, V., & Cermakova, B. (2010). Individual differences in music consumption are predicted by uses of music and age rather than emotional intelligence, neuroticism, extraversion or openness. *Psychology of Music*, *40*(3), 285–300.

Chang, C.-C. & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. on Intelligent Systems and Technology*, *2*(3), 27:1–27:27.

Charbuillet, C., Tardieu, D., Cornu, F., & Peeters, G. (2011a). 2011 IRCAM audio music similarity system #2. In *Music Information Retrieval Evaluation Exchange (MIREX'11)*. Available online: http://www.music-ir.org/mirex/abstracts/2011/CTCP2.pdf.

Charbuillet, C., Tardieu, D., & Peeters, G. (2011b). GMM supervector for content based music similarity. In *Int. Conf. on Digital Audio Effects (DAFx'11)*.

Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., & Riedl, J. (2003). Is seeing believing? how recommender system interfaces affect users' opinions. In *Conf. on Human factors in Computing Systems (CHI'03)*, pp. 585–592.

Cramer, H., Evers, V., Ramlal, S., Someren, M., Rutledge, L., Stash, N., Aroyo, L., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*(5), 455–496.

Cripps, A., Pettey, C., & Nguyen, N. (2006). Improving the performance of FLN by using similarity measures and evolutionary algorithms. In *IEEE Int. Conf. on Fuzzy Systems*, p. 323–330.

Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: information retrieval in practice.* Addison-Wesley.

Cupchik, G. C., Rickert, M., & Mendelson, J. (1982). Similarity and preference judgments of musical stimuli. *Scandinavian Journal of Psychology*, *23*(4), 273–282.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

D'Elia, A. & Piccolo, D. (2005). A mixture model for preferences data analysis. *Computational Statistics & Data Analysis*, *49*(3), 917–934.

Downie, J., Ehmann, A., Bay, M., & Jones, M. (2010). The music information retrieval evaluation eXchange: some observations and insights. In *Advances in Music Information Retrieval*, pp. 93–115.

Downie, J. S. (2008). The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, *29*(4), 247–255.

Dror, G., Koenigstein, N., Koren, Y., & Weimer, M. (2011). The yahoo! music dataset and KDD-Cup'11. In *KDD-Cup Workshop*, vol. 2011.

Dunn, P. G., de Ruyter, B., & Bouwhuis, D. G. (2011). Toward a better understanding of the relation between music preference, listening behavior, and personality. *Psychology of Music*, *40*(4), 411–428.

Ellis, D. P. & Poliner, G. E. (2007). Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'07)*, pp. IV–1429–1432.

Ertoz, L., Steinbach, M., & Kumar, V. (2003). Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *SIAM Int. Conf. on Data Mining*, vol. 47.

Fernández, M., Vallet, D., & Castells, P. (2006). Probabilistic score normalization for rank aggregation. In *Advances in Information Retrieval*, pp. 553–556.

Ferrer, R. & Eerola, T. (2011). AMP: artist-based muscial preferences derived from free verbal responses and social tags. In *IEEE Internation Conf. on Multimedia & Expo (ICME'11). Int. Workshop on Advances in Music Information Research (AdMIRe'11).*

Finnäs, L. (1989). How can musical preferences be modified? a research review. *Bulletin of the Council for Research in Music Education*, (102), 1–58.

Firan, C. S., Nejdl, W., & Paiu, R. (2007). The benefit of using tag-based profiles. In *Latin American Web Conf.*, pp. 32–41.

Flexer, A., Schnitzer, D., Gasser, M., & Widmer, G. (2008). Playlist generation using start and end songs. In *Int. Symp. on Music Information Retrieval (ISMIR'08)*, pp. 173–178.

Flexer, A., Schnitzer, D., & Schlueter, J. (2012). A MIREX meta-analysis of hubness in audio music similarity. In *Int. Society for Music Information Retrieval Conf. (ISMIR'12).*

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1–22.

Garcıa-Dıez, S., Saerens, M., Senelle, M., Fouss, F., & universitaires de Mons, F. (2011). A simple-cycles weighted kernel based on harmony structure for similarity retrieval. In *Int. Society for Music Information Retrieval Conf. (ISMIR'11)*, pp. 61–66.

Gibbons, J. D. & Chakraborti, S. (2003). *Nonparametric Statistical Inference.* CRC Press.

Glasgow, M. R., Cartier, A. M., & Wilson, G. D. (1985). Conservatism, sensation-seeking and music preferences. *Personality and Individual Differences*, *6*(3), 395–396.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality*, *37*(6), 504–528.

Gouyon, F. (2005). *A computational approach to rhythm description: Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing.* Ph.D. thesis, UPF, Barcelona, Spain.

Gouyon, F., Klapuri, A., Dixon, S., Alonso, M., Tzanetakis, G., Uhle, C., & Cano, P. (2006). An experimental comparison of audio tempo induction algorithms. *IEEE Trans. on Speech and Audio Processing*, *14*(5), 1832–1844.

Green, S. J., Lamere, P., Alexander, J., Maillet, F., Kirk, S., Holt, J., Bourque, J., & Mak, X. W. (2009). Generating transparent, steerable recommendations from textual descriptions of items. In *ACM Conf. on Recommender Systems (RecSys'09)*, p. 281–284.

Grimaldi, M. & Cunningham, P. (2004). Experimenting with music taste prediction by user profiling. In *ACM SIGMM Int. Workshop on Multimedia Information Retrieval (MIR'04)*, pp. 173–180.

Grimmett, G. & Stirzaker, D. (2001). *Probability and random processes.* Oxford University Press, 3rd edn.

Gruzd, A. A., Downie, J. S., Jones, M. C., & Lee, J. H. (2007). Evalutron 6000: collecting music relevance judgments. In *ACM/IEEE-CS Joint Conf. on Digital Libraries (JCDL'07)*, pp. 507–507.

Gómez, E. (2006). Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing*, *18*(3), 294–304.

Gómez, E. & Herrera, P. (2008). Comparative analysis of music recordings from western and non-western traditions by automatic tonal feature extraction. *Empirical Musicology Review*, *3*, 140–156.

Gómez, E., Herrera, P., Cano, P., Janer, J., Serrà, J., Bonada, J., El-Hajj, S., Aussenac, T., & Holmberg, G. (2009). Music similarity systems and methods using descriptors. WIPO Patent No. 2009001202.

Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Int. Conf. on Machine Learning*, pp. 359–366.

Hanani, U., Shapira, B., & Shoval, P. (2001). Information filtering: Overview of issues, research and systems. *User Modeling and User-Adapted Interaction*, *11*(3), 203–259.

Hargreaves, D., MacDonald, R., & Miell, D. (2005). How do people communicate using music? In D. E. Miell, R. A. R. MacDonald, & D. J. Hargreaves (Eds.) *Musical communication*, pp. 1–25. Oxford University Press.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. on Information Systems*, *22*(1), 5–53.

Herrera, P., Resa, Z., & Sordo, M. (2010). Rocking around the clock eight days a week: an exploration of temporal patterns of music listening. In *ACM Conf. on Recommender Systems. Workshop on Music Recommendation and Discovery (Womrad 2010)*, p. 7–10.

Hoashi, K., Matsumoto, K., & Inoue, N. (2003). Personalization of user profiles for content-based music retrieval based on relevance feedback. In *ACM Int. Conf. on Multimedia (MULTIMEDIA'03)*, pp. 110–119.

Hoashi, K., Matsumoto, K., Sugaya, F., Ishizaki, H., & Katto, J. (2006). Feature space modification for content-based music retrieval based on user preferences. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'06)*, vol. 5, pp. 517–520.

Hoerl, A. (1962). Application of ridge analysis to regression problems. *Chemical Engineering Progress*, *58*(3), 54–59.

Hoffman, M., Blei, D., & Cook, P. (2008). Content-based musical similarity computation using the hierarchical dirichlet process. In *Int. Symp. on Music Information Retrieval (ISMIR'08).*

Holbrook, M. B. & Schindler, R. M. (1989). Some exploratory findings on the development of musical tastes. *Journal of Consumer Research*, *16*(1), 119–124.

Holm, J., Aaltonen, A., & Siirtola, H. (2009). Associating colours with musical genres. *Journal of New Music Research*, *38*(1), 87–100.

Homburg, H., Mierswa, I., Möller, B., Morik, K., & Wurst, M. (2005). A benchmark dataset for audio classification and clustering. In *Int. Conf. on Music Information Retrieval (ISMIR'05)*, pp. 528–531.

Hu, X. & Downie, J. S. (2007). Exploring mood metadata: Relationships with genre, artist and usage metadata. In *Int. Conf. on Music Information Retrieval (ISMIR'07)*.

Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *IEEE Int. Conf. on Data Mining (ICDM'08)*, p. 263–272.

Hu, Y. & Ogihara, M. (2011). Nextone player: A music recommendation system based on user behaviour. In *Int. Society for Music Information Retrieval Conf. (ISMIR'11)*.

Huberty, C. J. & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis*. Wiley-Interscience, 2nd edn.

Huq, A., Bello, J. P., & Rowe, R. (2010). Automated music emotion recognition: A systematic evaluation. *Journal of New Music Research*, *39*(3), 227–244.

Jawaheer, G., Szomszor, M., & Kostkova, P. (2010). Comparison of implicit and explicit feedback from an online music recommendation service. In *Int. Workshop on Information Heterogeneity and Fusion in Recommender Systems (HetRec'10)*, p. 47–51.

Jensen, J. H., Christensen, M. G., Ellis, D. P. W., & Jensen, S. H. (2009). Quantitative analysis of a common audio similarity measure. *IEEE Trans. on Audio, Speech, and Language Processing*, *17*, 693–703.

Jensen, J. H., Ellis, D. P., Christensen, M. G., & Jensen, S. H. (2007). Evaluation of distance measures between gaussian mixture models of MFCCs. In *Int. Conf. on Music Information Retrieval (ISMIR'07)*, p. 107–108.

Jones, N. & Pu, P. (2007). User technology adoption issues in recommender systems. In *Networking and Electronic Commerce Research Conf.*

Kemp, A. E. (1996). *The Musical Temperament: Psychology and Personality of Musicians*. Oxford University Press.

Kim, J.-H., Jung, K.-Y., Ryu, J.-K., Kang, U.-G., & Lee, J.-H. (2008a). Design of ubiquitous music recommendation system using MHMM. In *Int. Conf. on Networked Computing and Advanced Information Management (NCM '08)*, vol. 2, pp. 369–374.

Kim, J.-H., Song, C.-W., Lim, K.-W., & Lee, J.-H. (2006). Design of music recommendation system using context information. In *Agent Computing and Multi-Agent Systems*, pp. 708–713.

Kim, K., Lee, D., Yoon, T. B., & Lee, J. H. (2008b). A music recommendation system based on personal preference analysis. In *Int. Conf. on the Applications of Digital Information and Web Technologies (ICADIWT'08)*, p. 102–106.

Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., & Friederici, A. D. (2004). Music, language and meaning: brain signatures of semantic processing. *Nature Neuroscience*, *7*(3), 302–307.

Koenigstein, N., Shavitt, Y., Weinsberg, E., & Weinsberg, U. (2010). On the applicability of peer-to-peer data in music information retrieval research. In *Int. Society for Music Information Retrieval Conf. (ISMIR'10)*.

Korn, F., Pagel, B.-U., & Faloutsos, C. (2001). On the 'dimensionality curse' and the 'self-similarity blessing'. *IEEE Trans. on Knowledge and Data Engineering*, *13*(1), 96–111.

Lamont, A. & Webb, R. (2009). Short- and long-term musical preferences: what makes a favourite piece of music? *Psychology of Music*, *38*(2), 222–241.

Laurier, C. (2011). *Automatic Classification of Musical Mood by Content-Based Analysis.* Ph.D. thesis, UPF, Barcelona, Spain.

Laurier, C., Meyers, O., Serrà, J., Blech, M., & Herrera, P. (2009a). Music mood annotator design and integration. In *Int. Workshop on Content-Based Multimedia Indexing (CBMI'2009)*.

Laurier, C., Meyers, O., Serrà, J., Blech, M., Herrera, P., & Serra, X. (2009b). Indexing music by mood: design and integration of an automatic content-based annotator. *Multimedia Tools and Applications*, *48*(1), 161–184.

Leblanc, A. (1982). An interactive theory of music preference. *Journal of Music Therapy*, *19*, 28–45.

Lee, J. (2011). How similar is too similar?: Exploring users' perceptions of similarity in playlist evaluation. In *Int. Society for Music Information Retrieval Conf. (ISMIR'11)*.

Lee, J. & Lee, J. (2008). Context awareness by case-based reasoning in a music recommendation system. In *Ubiquitous Computing Systems*, pp. 45–58.

Levy, M. & Bosteels, K. (2010). Music recommendation and the long tail. In *ACM Conf. on Recommender Systems. Workshop on Music Recommendation and Discovery (Womrad 2010)*.

Levy, M. & Sandler, M. (2008). Learning latent semantic models for music from social tags. *Journal of New Music Research*, *37*(2), 137–150.

Levy, M. & Sandler, M. (2009). Music information retrieval using social tags and audio. *IEEE Trans. on Multimedia*, *11*(3), 383–395.

Li, Q., Myaeng, S., Guan, D., & Kim, B. (2005). A probabilistic model for music recommendation considering audio features. In *Information Retrieval Technology*, pp. 72–83.

Li, Q., Myaeng, S. H., & Kim, B. M. (2007). A probabilistic music recommender considering user opinions and audio features. *Information Processing & Management*, *43*(2), 473–487.

Li, T. & Ogihara, M. (2006). Toward intelligent music information retrieval. *IEEE Trans. on Multimedia*, *8*(3), 564–574.

Lidy, T. & Rauber, A. (2009). MIREX 2009 spectral and rhythm audio features for music similarity retrieval. In *Music Information Retrieval Evaluation Exchange (MIREX'09).*

Lin, Y.-C., Yang, Y.-H., & Chen, H. H. (2011). Exploiting online music tags for music emotion classification. *ACM Trans. on Multimedia Computing, Communications and Applications*, *7S*(1), 26:1–26:16.

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *Int. Symp. on Music Information Retrieval (ISMIR'00).*

Logan, B. (2004). Music recommendation from song sets. In *Int. Conf. on Music Information Retrieval (ISMIR'04)*, pp. 425–428.

Logan, B. & Salomon, A. (2001). A music similarity function based on signal analysis. In *IEEE Int. Conf. on Multimedia and Expo (ICME'01)*, p. 190.

Lonsdale, A. J. & North, A. C. (2011). Why do we listen to music? a uses and gratifications analysis. *British journal of psychology (London, England: 1953)*, *102*(1), 108–134.

Lu, C.-C. & Tseng, V. S. (2009). A novel method for personalized music recommendation. *Expert Systems with Applications*, *36*(6), 10035–10044.

MacDonald, R. A. R., Hargreaves, D. J., & Miell, D. (2002). *Musical identities.* Oxford University Press.

Maffesoli, M. (1996). *The time of the tribes: the decline of individualism in mass society.* SAGE.

Magno, T. & Sable, C. (2008). A comparison of signal-based music recommendation to genre labels, collaborative filtering, musicological analysis, human recommendation, and random baseline. In *Int. Conf. on Music Information Retrieval (ISMIR'08)*, pp. 161–166.

Maillet, F., Eck, D., Desjardins, G., & Lamere, P. (2009). Steerable playlist generation by learning song similarity from radio station playlists. In *Int. Conf. on Music Information Retrieval (ISMIR'09)*.

Mandel, M. I. & Ellis, D. P. (2005). Song-level features and support vector machines for music classification. In *Int. Conf. on Music Information Retrieval (ISMIR'05)*, pp. 594–599.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Marolt, M. (2008). A mid-level representation for melody-based retrieval in audio collections. *IEEE Trans. on Multimedia*, *10*(8), 1617–1625.

McCloud, S. (2009). *Understanding comics: the invisible art*. HarperPerennial, 36 edn.

McDermott, J. H. (2012). Auditory preferences and aesthetics. In *Neuroscience of Preference and Choice*, pp. 227–256.

McFee, B., Barrington, L., & Lanckriet, G. (2012a). Learning content similarity for music recommendation. *IEEE Trans. on Audio, Speech, and Language Processing*, *20*(8), 2207–2218.

McFee, B., Bertin-Mahieux, T., Ellis, D. P. W., & Lanckriet, G. R. G. (2012b). The million song dataset challenge. In *World Wide Web Conf. (WWW'12)*, p. 909–916.

McFee, B. & Lanckriet, G. (2009). Heterogeneous embedding for subjective artist similarity. In *Int. Conf. on Music Information Retrieval (ISMIR'09)*.

McKinney, M. F. & Moelants, D. (2006). Ambiguity in tempo perception: What draws listeners to different metrical levels? *Music Perception*, *24*(2), 155–166.

McNee, S., Riedl, J., & Konstan, J. (2006). Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human Factors in Computing Systems*, p. 1101.

Menon, V. & Levitin, D. J. (2005). The rewards of music listening: response and physiological connectivity of the mesolimbic system. *Neuroimage*, *28*(1), 175–184.

Moh, Y. & Buhmann, J. M. (2008). Kernel expansion for online preference tracking. *Int. Symp. on Music Information Retrieval (ISMIR'08)*, p. 167–172.

Moh, Y., Orbanz, P., & Buhmann, J. M. (2008). Music preference learning with partial information. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'08)*, pp. 2021–2024.

Morozov, E. (2012). *The Net Delusion: The Dark Side of Internet Freedom.* PublicAffairs.

Nanopoulos, A., Rafailidis, D., Ruxanda, M., & Manolopoulos, Y. (2009). Music search engines: Specifications and challenges. *Information Processing & Management*, *45*(3), 392–396.

North, A. C. (2004). Uses of music in everyday life. *Music Perception*, *22*(1), 41–77.

North, A. C. & Hargreaves, D. J. (1995). Subjective complexity, familiarity, and liking for popular music. *Psychomusicology: Music, Mind and Brain*, *14*(1-2), 77–93.

North, A. C. & Hargreaves, D. J. (1999). Music and adolescent identity. *Music Education Research*, *1*(1), 75–92.

North, A. C. & Hargreaves, D. J. (2007a). Lifestyle correlates of musical preference: 1. relationships, living arrangements, beliefs, and crime. *Psychology of Music*, *35*(1), 58–87.

North, A. C. & Hargreaves, D. J. (2007b). Lifestyle correlates of musical preference: 2. media, leisure time and music. *Psychology of music*, *35*(2), 179.

North, A. C. & Hargreaves, D. J. (2007c). Lifestyle correlates of musical preference: 3. travel, money, education, employment and health. *Psychology of Music*, *35*(3), 473–497.

North, A. C. & Hargreaves, D. J. (2008). *The social and applied psychology of music.* Oxford University Press.

Novello, A., McKinney, M. F., & Kohlrausch, A. (2006). Perceptual evaluation of music similarity. In *Int. Conf. on Music Information Retrieval (ISMIR'06).*

OMR (2011). Orpheus media research consumer survey executive summary. Tech. rep. Available online: http://www.cliomusic.com/wp-content/uploads/2011/04/omr.executivesummary.consumer_110324-1500.pdf.

Pampalk, E. (2006). *Computational models of music similarity and their application in music information retrieval.* Ph.D. thesis, Vienna University of Technology.

Pampalk, E., Dixon, S., & Widmer, G. (2003). On the evaluation of perceptual similarity measures for music. In *Int. Conf. on Digital Audio Effects (DAFx'03)*, p. 7–12. London, UK.

Pampalk, E., Flexer, A., & Widmer, G. (2005a). Improvements of audio-based music similarity and genre classification. In *Int. Conf. on Music Information Retrieval (ISMIR'05)*, pp. 628–633.

Pampalk, E., Pohle, T., & Widmer, G. (2005b). Dynamic playlist generation based on skipping behavior. In *Int. Conf. on Music Information Retrieval (ISMIR'05)*, p. 634–637.

Pariser, E. (2012). *The Filter Bubble: What the Internet Is Hiding from You.* Penguin Books, Limited.

Park, H. S., Yoo, J. O., & Cho, S. B. (2006). A context-aware music recommendation system using fuzzy bayesian networks with utility theory. *Fuzzy Systems and Knowledge Discovery*, p. 970–979.

Parra, D. & Amatriain, X. (2011). Walk the talk: analyzing the relation between implicit and explicit feedback for preference elicitation. In *Int. Conf. on User Modeling, Adaption, and Personalization (UMAP'11)*, p. 255–268.

Pazzani, M. & Billsus, D. (1997). Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, *27*(3), 313–331.

Pearson, J. L. & Dollinger, S. J. (2004). Music preference correlates of jungian types. *Personality and individual differences*, *36*(5), 1005–1008.

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *CUIDADO Project Report.* Available online: http://recherche.ircam.fr/equipes/analyse-synthese/peeters/ARTICLES/.

Perkins, S. (2008). Personality and music: An examination of the five-factor model in conjunction with musical preference. Thesis, Wheaton College, USA.

Petajan, E. (2005). MPEG-4 face and body animation coding applied to HCI. In *Real-Time Vision for Human-Computer Interaction*, pp. 249–268.

Platt, J. C. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In P. J. Bartlett, B. Schölkopf, D. Schuurmans, & A. J. Smola (Eds.) *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, Cambridge, MA.

Pohle, T., Knees, P., Schedl, M., & Widmer, G. (2006). Automatically adapting the structure of audio similarity spaces. In *Workshop on Learning the Semantics of Audio Signals (LSAS'06)*, pp. 66–75.

Pohle, T. & Schnitzer, D. (2007). Striving for an improved audio similarity measure. In *Music Information Retrieval Evaluation Exchange (MIREX'07)*. Available online: http://www.music-ir.org/mirex/2007/abs/AS_pohle.pdf.

Pohle, T. & Schnitzer, D. (2009). Submission to MIREX 2009 audio similarity task. In *Music Information Retrieval Evaluation Exchange (MIREX'09)*. Available online: http://music-ir.org/mirex/2009/results/abs/PS.pdf.

Pohle, T., Schnitzer, D., Schedl, M., Knees, P., & Widmer, G. (2009). On rhythm and general music similarity. In *Int. Society for Music Information Retrieval Conf. (ISMIR'09)*, pp. 525–530.

Radlinski, F. & Craswell, N. (2010). Comparing the sensitivity of information retrieval metrics. In *Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'10)*, p. 667–674.

Reas, C. & Fry, B. (2007). *Processing: a programming handbook for visual designers and artists*. MIT Press.

Reed, J. & Lee, C. (2011). Preference music ratings prediction using tokenization and minimum classification error training. *IEEE Trans. on Audio, Speech, and Language Processing*, *19*(8), 2294–2303.

Rentfrow, P. J., Goldberg, L. R., & Levitin, D. J. (2011). The structure of musical preferences: A five-factor model. *Journal of Personality and Social Psychology*, *100*(6), 1139–1157.

Rentfrow, P. J. & Gosling, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, *84*, 1236–1256.

Rentfrow, P. J. & Gosling, S. D. (2006). Message in a ballad. *Psychological Science*, *17*(3), 236–242.

Salamon, J., Rocha, B., & Gómez, E. (2012). Musical genre classification using melody features extracted from polyphonic music signals. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'12)*, p. 81–84.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Saris, W. E. & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Wiley-Interscience.

Sarwar, B., Karypis, G., Konstan, J., & Reidl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Int. Conf. on World Wide Web (WWW'01)*, pp. 285–295.

Sauer, D. & Yang, Y.-H. (2009). Music-driven character animation. *ACM Trans. on Multimedia Computing, Communications, and Applications*, *5*(4), 1–16.

Schedl, M. & Flexer, A. (2012). Putting the user in the center of music information retrieval. In *Int. Society for Music Information Retrieval Conf. (ISMIR'12)*.

Schedl, M. & Knees, P. (2009). Context-based music similarity estimation. In *Int. Workshop on Learning Semantics of Audio Signals (LSAS'09)*.

Schedl, M., Pohle, T., Knees, P., & Widmer, G. (2011a). Exploring the music similarity space on the web. *ACM Trans. on Information Systems*, *29*(3), 1–24.

Schedl, M., Widmer, G., Knees, P., & Pohle, T. (2011b). A music information system automatically generated via web content mining techniques. *Information Processing & Management*, *47*(3), 426–439.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social Science Information*, *44*(4), 695–729.

Schindler, A., Mayer, R., & Rauber, A. (2012). Facilitating comprehensive benchmarking experiments on the million song dataset. In *Int. Society for Music Information Retrieval Conf. (ISMIR'12)*.

Schnitzer, D., Flexer, A., Schedl, M., & Widmer, G. (2011). Using mutual proximity to improve content-based audio similarity. In *Int. Society for Music Information Retrieval Conf. (ISMIR'11)*, pp. 79–84.

Schwartz, K. D. & Fouts, G. T. (2003). Music preferences, personality style, and developmental issues of adolescents. *Journal of Youth and Adolescence*, *32*(3), 205–213.

Schäfer, T. (2008). *Determinants of Music Preference.* Ph.D. thesis, Technischen Universität Chemnitz.

Schäfer, T. & Sedlmeier, P. (2009). From the functions of music to music preference. *Psychology of Music*, *37*(3), 279–300.

Serrà, J., Serra, X., & Andrzejak, R. G. (2009). Cross recurrence quantification for cover song identification. *New Journal of Physics*, *11*(9), 093017.

Sethares, W. A. (2005). *Tuning, timbre, spectrum, scale.* Springer Verlag.

Seyerlehner, K., Schedl, M., Pohle, T., & Knees, P. (2010). Using block-level features for genre classification, tag classification and music similarity estimation. In *Music Information Retrieval Evaluation Exchange (MIREX'10)*.

Shani, G. & Gunawardana, A. (2009). Evaluating recommender systems. *Recommender Systems Handbook*, p. 257–298.

Shardanand, U. & Maes, P. (1995). Social information filtering: algorithms for automating "word of mouth". In *Proceedings of the SIGCHI conference on Human factors in computing systems*, p. 210–217.

Shental, N., Hertz, T., Weinshall, D., & Pavel, M. (2002). Adjustment learning and relevant component analysis. *Lecture Notes In Computer Science*, pp. 776–792.

Sigurdsson, S., Petersen, K. B., & Lehn-Schiøler, T. (2006). Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music. In *Int. Conf. on Music Information Retrieval (ISMIR'07)*, p. 286–289.

Sinha, R. & Swearingen, K. (2002). The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human Factors in Computing Systems*, p. 831.

Slaney, M. (2011). Web-scale multimedia analysis: Does content matter? *IEEE Multimedia*, *18*(2), 12–15.

Slaney, M., Weinberger, K., & White, W. (2008). Learning a metric for music similarity. In *Int. Symp. on Music Information Retrieval (ISMIR'08)*, pp. 313–318.

Slaney, M. & White, W. (2007). Similarity based on rating data. In *Int. Symp. on Music Information Retrieval (ISMIR'07)*.

Smith, L. M. (2010). Beat critic: Beat tracking octave error identification by metrical profile analysis. In *Int. Society for Music Information Retrieval Conf. (ISMIR'10)*.

Song, S., Kim, M., Rho, S., & Hwang, E. (2009). Music ontology for mood and situation reasoning to support music retrieval and recommendation. In *Int. Conf. on Digital Society (ICDS'09)*, pp. 304–309.

Song, Y. & Zhang, C. (2008). Content-based information fusion for semi-supervised music genre classification. *IEEE Trans. on Multimedia*, *10*(1), 145–152.

Sordo, M. (2012). *Semantic Annotation of Music Collections: A Computational Approach*. Ph.D. thesis, UPF, Barcelona, Spain.

Sordo, M., Celma, O., Blech, M., & Guaus, E. (2008). The quest for musical genres: Do the experts and the wisdom of crowds agree? In *Int. Conf. of Music Information Retrieval (ISMIR'08)*, pp. 255–260.

Stober, S. & Nürnberger, A. (2009). User-adaptive music information retrieval. *Künstliche Intelligenz*, *23*(2), 54–57.

Sturm, B. L. (2012). A survey of evaluation in music genre recognition. *Proc. Adaptive Multimedia Retrieval, Copenhagen, Denmark*.

Su, J. H., Yeh, H. H., & Tseng, V. S. (2010a). A novel music recommender by discovering preferable perceptual-patterns from music pieces. In *ACM Symp. on Applied Computing (SAC'10)*, pp. 1924–1928.

Su, J.-H., Yeh, H.-H., Yu, P. S., & Tseng, V. S. (2010b). Music recommendation using content and context information mining. *IEEE Intelligent Systems*, *25*(1), 16–26.

Sutton, C. (2012). Last.fm – the blog · advanced robotics. Available online: http://blog.last.fm/2012/07/24/advanced-robotics.

Swearingen, K. & Sinha, R. (2001). Beyond algorithms: An HCI perspective on recommender systems. In *ACM SIGIR 2001 Workshop on Recommender Systems*.

Szymanski, G. (2009). Pandora, or, a never-ending box of musical delights. *Music Reference Services Quarterly*, *12*(1), 21–22.

Teo, T. (2003). Relationship of selected musical characteristics and musical preference (a review of literature). *Visions of Research in Music Education*, *3*.

Ter Bogt, T. F. M., Mulder, J., Raaijmakers, Q. A. W., & Nic Gabhainn, S. (2010). Moved by music: A typology of music listeners. *Psychology of Music*, *39*(2), 147–163.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tiemann, M. & Pauws, S. (2007). Towards ensemble learning for hybrid music recommendation. In *ACM Conf. on Recommender Systems (RecSys'07)*, p. 177–178.

Tintarev, N. & Masthoff, J. (2007). Effective explanations of recommendations: user-centered design. In *ACM Conf. on Recommender Systems (RecSys'07)*, p. 153–156.

Tullis, T. & Albert, W. (2008). *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Morgan Kaufmann.

Turpin, A. & Scholer, F. (2006). User performance versus precision measures for simple search tasks. In *Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'06)*, p. 11–18.

Tzanetakis, G. (2009). Marsyas submissions to MIREX 2009. In *Music Information Retrieval Evaluation Exchange (MIREX'09).*

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

Uitdenbogerd, A. L. & van Schyndel, R. (2002). A review of factors affecting music recommender success. In *Int. Conf. of Music Information Retrieval (ISMIR'02)*, pp. 204–208.

Urbano, J., Martín, D., Marrero, M., & Morato, J. (2011). Audio music similarity and retrieval: Evaluation power and stability. In *Int. Society for Music Information Retrieval Conf. (ISMIR'11).*

Voong, M. & Beale, R. (2007). Music organisation using colour synaesthesia. In *CHI'07 extended abstracts on Human Factors in Computing Systems*, pp. 1869–1874.

Wack, N., Cano, P., de Jong, B., & Marxer, R. (2006). A comparative study of dimensionality reduction methods: The case of music similarity. Tech. rep., Music Technology Group. Available online: http://mtg.upf.edu/files/publications/NWack_2006.pdf.

Weinberger, K. Q. & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, *10*, 207–244.

West, K. & Lamere, P. (2007). A model-based approach to constructing music similarity functions. *EURASIP Journal on Advances in Signal Processing*, *2007*, 149–149.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann.

Xu, C., Maddage, N. C., Shao, X., Cao, F., & Tian, Q. (2003). Musical genre classification using support vector machines. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'03)*, pp. 429–432.

Yang, Y.-H. & Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Trans. Intell. Syst. Technol.*, *3*(3), 40:1–40:30.

Yoshii, K. (2008). *Studies on Hybrid Music Recommendation Using Timbral and Rhythmic Features.* Ph.D. thesis, Kyoto University, Japan.

Yoshii, K., Goto, M., Komatani, K., Ogata, T., & Okuno, H. G. (2006). Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences. In *Int. Conf. on Music Information Retrieval (ISMIR'06).*

Yoshii, K., Goto, M., Komatani, K., Ogata, T., & Okuno, H. G. (2008). An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model. *IEEE Trans. on Audio, Speech, and Language Processing*, *16*(2), 435–447.

Zheleva, E., Guiver, J., Mendes Rodrigues, E., & Milić-Frayling, N. (2010). Statistical models of music-listening sessions in social media. In *Int. Conf. on World Wide Web (WWW'10)*, p. 1019–1028.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

Zweigenhaft, R. L. (2008). A do re mi encore. *Journal of Individual Differences*, *29*(1), 45–55.

Çataltepe, Z. & Altinel, B. (2007). Hybrid music recommendation based on different dimensions of audio content and an entropy measure. In *Eusipco 2007 Conf.*

Çataltepe, Z. & Altinel, B. (2009). Music recommendation by modeling user's preferred perspectives of content, Singer/Genre and popularity. *Collaborative and social information retrieval and access: techniques for improved user modeling*, pp. 203–221.

Appendix A: Demonstrations

PHAROS: Music similarity in a search engine

We have integrated the proposed hybrid music similarity measure (Section 4.6) within the framework of a search-engine project PHAROS.[2] PHAROS was an Integrated Project funded by the European Union under the Information Society Technologies Programme (6th Framework Programme). Its strategic objective was defined as “Search Engines for Audiovisual Content”, and it was aimed to advance search of audiovisual content from a point-solution search engine paradigm to an integrated search platform paradigm. One of the main goals of the project was to develop a scalable and open search framework that lets users search, explore, discover, and analyze contextually relevant data.

PHAROS uses automatic content annotation to index audiovisual content. In particular, it provides music search using the music similarity measures proposed in this thesis. The required audio features and semantic descriptors are computed by the C++ implementation based on an audio analysis tool Essentia[3] developed at Music Technology Group.

MyMusicalAvatar

Presentation

MyMusicalAvatar is a demonstration of the proposed approach to visualization of music preferences and music recommendation. Our original proposed approach is grounded on a set of tracks explicitly provided by the user. In this demonstration system, we explore an alternative way to obtain the user’s preference set taking information from her/his accounts on popular online music services *Last.fm* and *Soundcloud*. The system analyzes audio content and generates music recommendations, using a semantic music similarity measure, and the user’s preference visualization, mapping semantic descriptors to visual elements.

The user interface is designed in the form of a web page.[4] The user specifies her/his account name on *Last.fm* and/or *SoundCloud* services, from which the preferred tracks should be retrieved. *SoundCloud* is a platform which allows users (mostly musicians) to collaborate, promote, and distribute their music.

[2] Platform for searcHing of Audiovisual Resources across Online Spaces. `http://www.pharos-audiovisual-search.eu`, `http://mtg.upf.edu/research/projects/pharos`

[3] `http://essentia.upf.edu`

[4] A demo of the system is available online: `http://mtg.upf.edu/project/musicalavatar`

Specifically, it allows users to upload their own tracks or mark tracks as their favorites. This information is available via the *SoundCloud* API.[5] Our system is currently limited to the users of these musical services, but will be further extended with an option to upload preferred tracks to our server.

Different types of tracks can be used to infer the user's preference set:

- Tracks marked as favorites by the user on *Last.fm*.
- Tracks listened the most by the user according to their *Last.fm*'s statistics.
- Tracks marked as favorites by the user on *SoundCloud*.
- Tracks uploaded by the user on *SoundCloud*.

The type of tracks to use and their amount can be specified by the user. The system retrieves the URLs of the tracks to be included in the preference set using the *Last.fm* and *SoundCloud* APIs. Using these URLs, audio fragments (30 sec.) of the track previews are downloaded.[6] By means of audio analysis we infer the semantic descriptions for each track from the user's preference set.

To generate recommendations, we employ an in-house music collection of 50.000 music excerpts, covering a wide range of musical genres. This collection was analyzed to retrieve the same semantic descriptions as those used for the preference set. We follow the methodology presented in Section 5 to search for the tracks from our in-house music collections which are similar to the user's preference set and use the proposed semantic distance (Section 4.3.2). The recommendation outcomes are presented to the user, including metadata of the tracks, audio previews, and the reason why a particular track was recommended (i.e., recommendation sources).

We follow the proposed visualization approach (Section 7.4) to form the *Musical Avatar*. Selected descriptors of each track are summarized across all tracks in the preference set. The resulting descriptors represent degrees of the user preference for different genres, moods, and instrumentation types. These descriptors are then mapped to visual elements of the avatar, which are implemented using Processing. An example of the output provided to a user is presented in Figure 1.

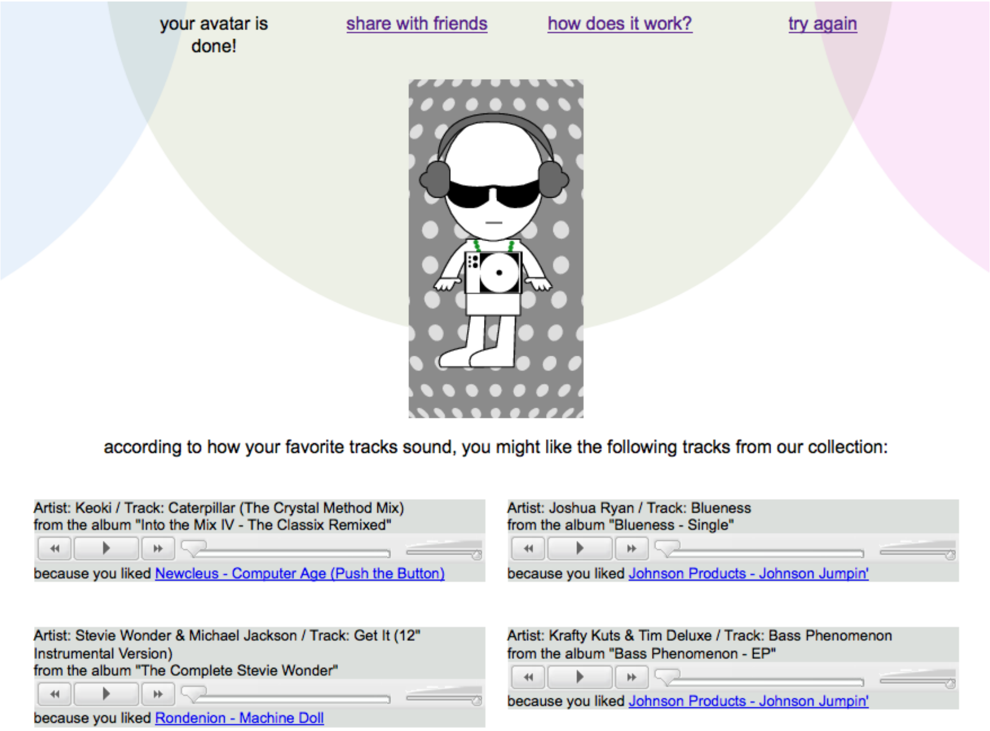
Interactive demo

In addition, we have developed a simple interactive demonstration of how avatar visualization is conditioned by different types of music for a public exhibition for children and adolescents.[7] A sample of 16 tracks was employed for this purpose. A user can select to preview them and select the preferred

[5] `http://soundcloud.com/developers`

[6] The system considers solely the tracks with available previews.

[7] This demonstration is available online: `http://mtg.upf.edu/project/musicalavatar`



your avatar is done!

according to how your favorite tracks sound, you might like the following tracks from our collection:

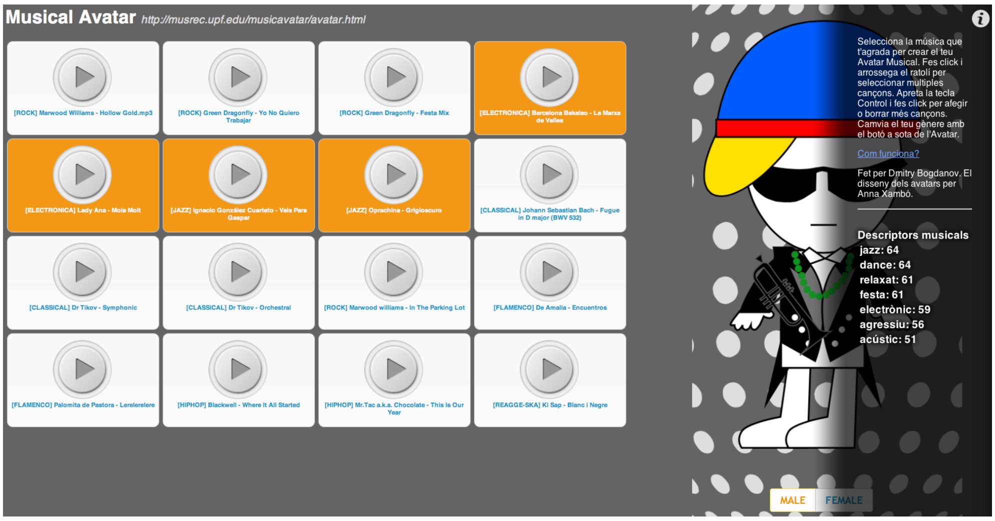
Artist: Keoki / Track: Caterpillar (The Crystal Method Mix)
from the album "Into the Mix IV - The Classix Remixed"
because you liked Newcleus - Computer Age (Push the Button)

Artist: Joshua Ryan / Track: Blueness
from the album "Blueness - Single"
because you liked Johnson Products - Johnson Jumpin'

Artist: Stevie Wonder & Michael Jackson / Track: Get It (12" Instrumental Version)
from the album "The Complete Stevie Wonder"
because you liked Rondenion - Machine Doll

Artist: Krafty Kuts & Tim Deluxe / Track: Bass Phenomenon
from the album "Bass Phenomenon - EP"
because you liked Johnson Products - Johnson Jumpin'

Figure 1: Screenshot of the system output returned to a user including the generated avatar and music recommendations.



Musical Avatar *http://musrec.upf.edu/musicavatar/avatar.html*

[ROCK] Marwood Williams - Hollow Gold.mp3

[ROCK} Green Dragonfly - Yo No Quiero Trabajar

[ROCK] Green Dragonfly - Festa Mix

[ELECTRONICA] Barcelona Bakalao - La Marxa de Valles

[ELECTRONICA] Lady Ana - Mola Molt

[JAZZ] Ignacio González Cuarteto - Vals Para Gaspar

[JAZZ] Oprachina - Grigioscuro

[CLASSICAL] Johann Sebastian Bach - Fugue in D major (BWV 532)

[CLASSICAL] Dr Tikov - Symphonic

[CLASSICAL] Dr Tikov - Orchestral

[ROCK] Marwood williams - In The Parking Lot

[FLAMENCO] De Amalia - Encuentros

[FLAMENCO] Palomita de Pastora - Lerelerelere

[HIPHOP] Blackwell - Where It All Started

[HIPHOP] Mr.Tac a.k.a. Chocolate - This Is Our Year

[REAGGE-SKA] Ki Sap - Blanc i Negre

Selecciona la música que t'agrada per crear el teu Avatar Musical. Fes click i arrossega el ratolí per seleccionar multiples cançons. Apreta la tecla Control i fes click per afegir o borrar més cançons. Camvia el teu gènere amb el botó a sota de l'Avatar.

Com funciona?

Fet per Dmitry Bogdanov. El disseny dels avatars per Anna Xambó.

Descriptors musicals
jazz: 64
dance: 64
relaxat: 61
festa: 61
electrònic: 59
agressiu: 56
acústic: 51

Figure 2: Screenshot of the interactive demonstration.

ones. The visualization changes accordingly to the selection. A screenshot of the interactive demonstration is presented in Figure 2.

Technical details

The back-end of the demo system is coded in C++ and python. It uses libraries of the Music Technology Group (Essentia and Gaia) for audio analysis and music similarity. The front-end GUI is web-based. It is made using HTML5/CSS, javascript, jQuery, and Processingjs for interactive avatar visualization.

Appendix B: Supplementary Material

Chapter 3: Preference elicitation strategies

We present the exact text of the instructions and the questionnaire presented to the participants in our experiments: “Your goal is to gather the minimal set of songs needed to graps or convey your music preferences. It’s important to note that these are not artists, these are single musical pieces which are informative by themselves (without any additional context). Ideally we would like to have a folder with the selected songs in mp3 format (or any format you may use). If you cannot get the mp3 files, please provide a list of the songs, indicating the artist, name of the piece and if needed the edition. Please answer the following questionnaire.

Personal data:

- Age
- Sex
- Interest for music [0-10] (e.g. 0 =no interest - 10 =passionate)
- Do you play any instrument? which one?
- How many songs did you select?

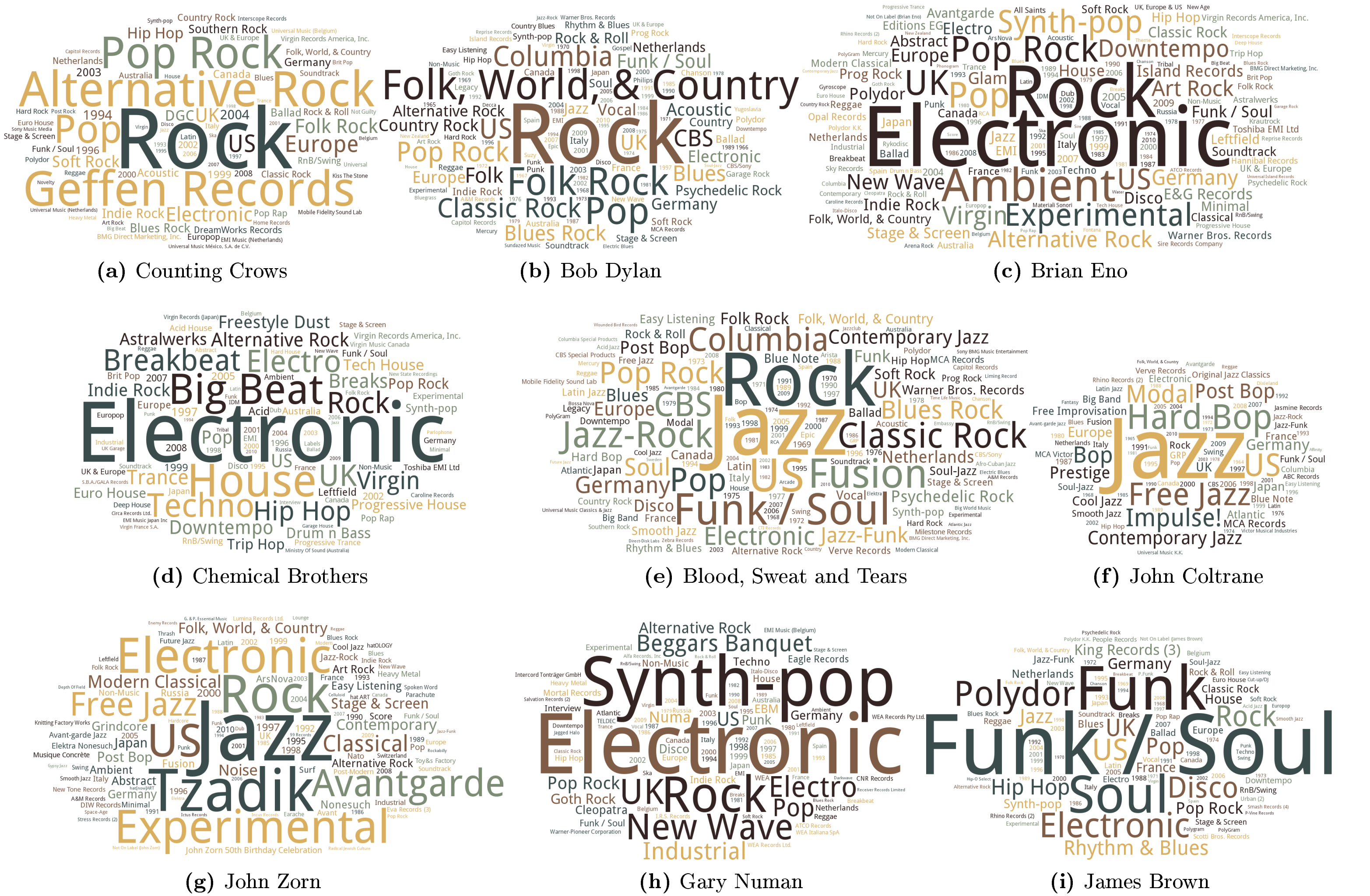
1. Please describe the strategy that you used to make the musical selection.
2. Please specify how long did it take to build it.
3. Give a set of labels that you think define your musical preferences and then the music you selected. These are some examples or aspects you can use:
    - Genre (e.g. rock, heavy metal)
    - Mood (e.g. calm, happy, party)
    - Instrumentation (e.g. orchestral, big band, female singers, mellow voice)
    - Rhythm (e.g. fast, regular)
    - Melody / harmony (e.g. atonal)
4. Other comments you want to make.”

Chapter 5: Music recommendation based on preference examples

Figure 3 presents examples of artist tag-clouds generated following the proposed recommendation approach working with editorial metadata found on *Discogs* (DISCOGS-ALL-1).

Chapter 6: Content-based analysis of music preferences

Figure 4 presents a screenshot of the user evaluation conducted to assess the quality of computed semantic user profiles and the user satisfaction with the generated avatar.



(a) Counting Crows

(b) Bob Dylan

(c) Brian Eno

(d) Chemical Brothers

(e) Blood, Sweat and Tears

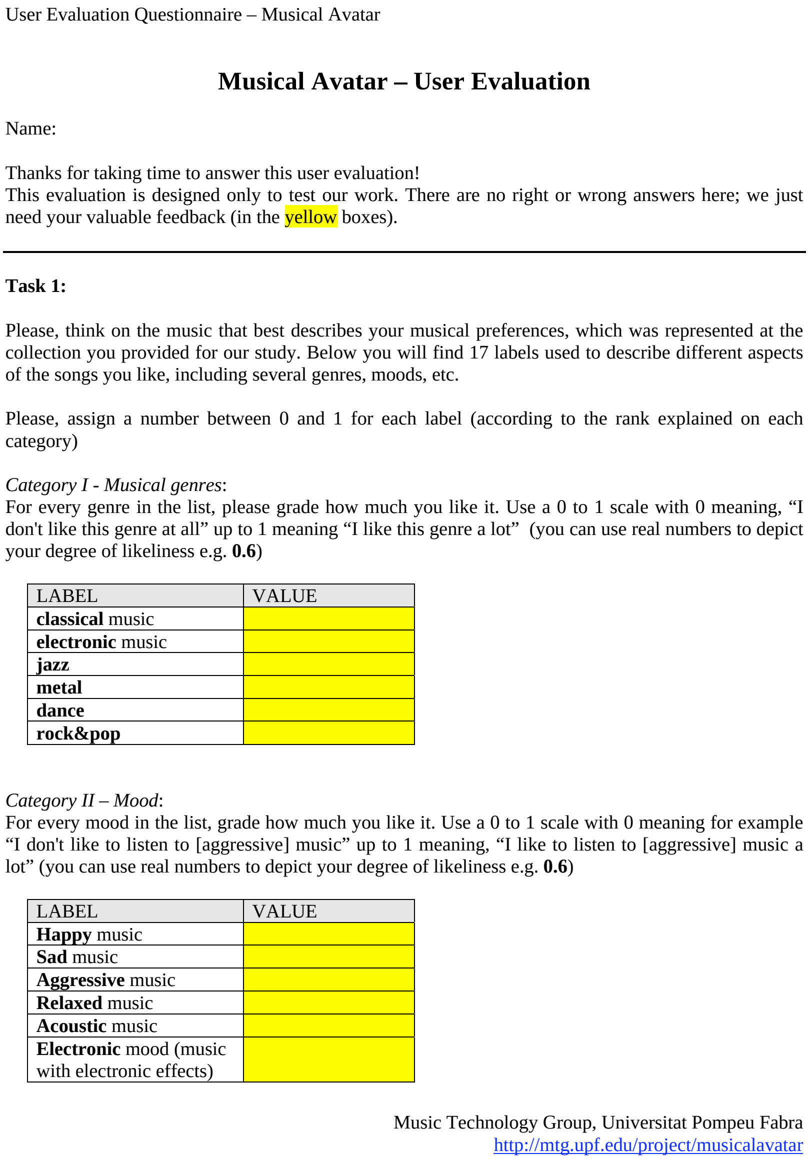
(f) John Coltrane

(g) John Zorn

(h) Gary Numan

(i) James Brown

Figure 3: Artist tag-clouds generated using the editorial information found on *Discogs* following the proposed recommendation approach.



User Evaluation Questionnaire – Musical Avatar

Musical Avatar – User Evaluation

Name:

Thanks for taking time to answer this user evaluation!
This evaluation is designed only to test our work. There are no right or wrong answers here; we just need your valuable feedback (in the yellow boxes).

Task 1:

Please, think on the music that best describes your musical preferences, which was represented at the collection you provided for our study. Below you will find 17 labels used to describe different aspects of the songs you like, including several genres, moods, etc.

Please, assign a number between 0 and 1 for each label (according to the rank explained on each category)

Category I - Musical genres:
For every genre in the list, please grade how much you like it. Use a 0 to 1 scale with 0 meaning, “I don't like this genre at all” up to 1 meaning “I like this genre a lot” (you can use real numbers to depict your degree of likeliness e.g. **0.6**)

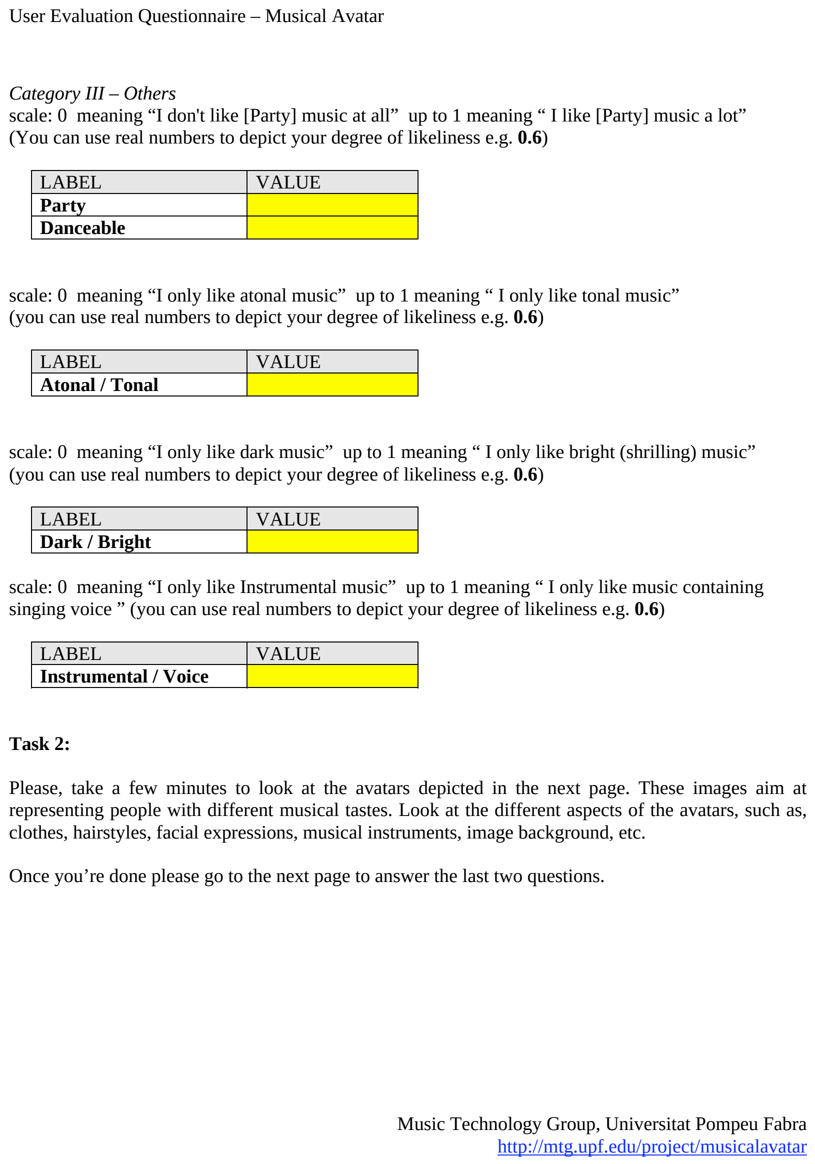
LABEL	VALUE
classical music	
electronic music	
jazz	
metal	
dance	
rock&pop	

Category II – Mood:
For every mood in the list, grade how much you like it. Use a 0 to 1 scale with 0 meaning for example “I don't like to listen to [aggressive] music” up to 1 meaning, “I like to listen to [aggressive] music a lot” (you can use real numbers to depict your degree of likeliness e.g. **0.6**)

LABEL	VALUE
Happy music	
Sad music	
Aggressive music	
Relaxed music	
Acoustic music	
Electronic mood (music with electronic effects)	

Music Technology Group, Universitat Pompeu Fabra
http://mtg.upf.edu/project/musicalavatar

Figure 4: A screenshot of the evaluation questionnaire given to the participants. (Continued on next page.)



User Evaluation Questionnaire – Musical Avatar

Category III – Others
scale: 0 meaning “I don't like [Party] music at all” up to 1 meaning “ I like [Party] music a lot”
(You can use real numbers to depict your degree of likeliness e.g. **0.6**)

LABEL	VALUE
Party	
Danceable	

scale: 0 meaning “I only like atonal music” up to 1 meaning “ I only like tonal music”
(you can use real numbers to depict your degree of likeliness e.g. **0.6**)

LABEL	VALUE
Atonal / Tonal	

scale: 0 meaning “I only like dark music” up to 1 meaning “ I only like bright (shrilling) music”
(you can use real numbers to depict your degree of likeliness e.g. **0.6**)

LABEL	VALUE
Dark / Bright	

scale: 0 meaning “I only like Instrumental music” up to 1 meaning “ I only like music containing singing voice ” (you can use real numbers to depict your degree of likeliness e.g. **0.6**)

LABEL	VALUE
Instrumental / Voice	

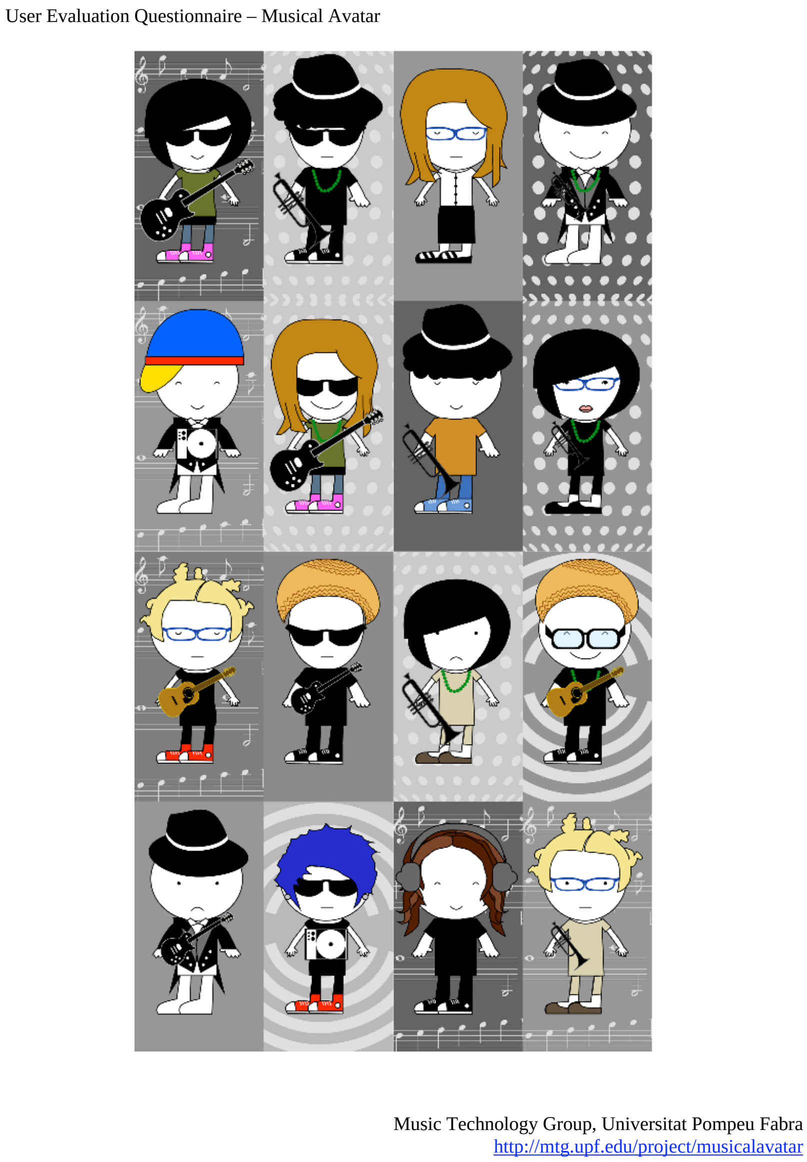
Task 2:

Please, take a few minutes to look at the avatars depicted in the next page. These images aim at representing people with different musical tastes. Look at the different aspects of the avatars, such as, clothes, hairstyles, facial expressions, musical instruments, image background, etc.

Once you’re done please go to the next page to answer the last two questions.

Music Technology Group, Universitat Pompeu Fabra
http://mtg.upf.edu/project/musicalavatar

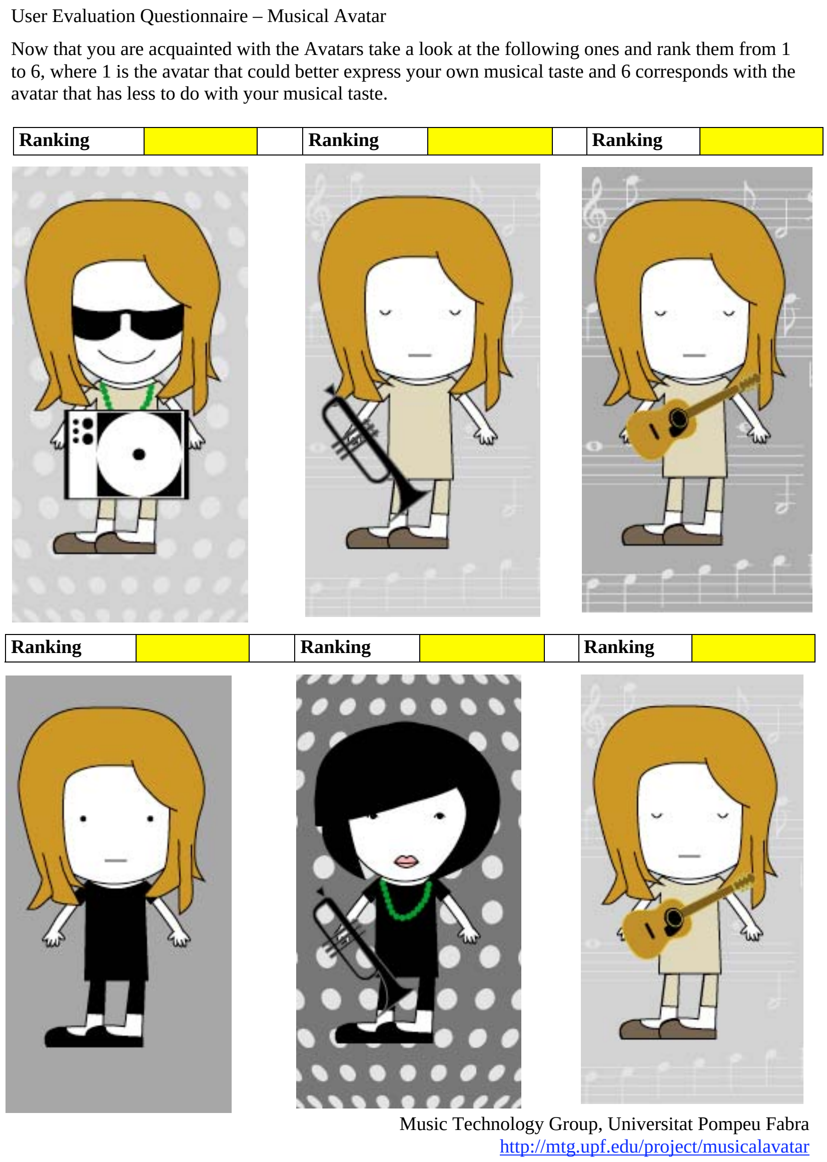
Figure 4: Continued (caption shown on previous page.)



User Evaluation Questionnaire – Musical Avatar

Music Technology Group, Universitat Pompeu Fabra
http://mtg.upf.edu/project/musicalavatar

Figure 4: Continued.



User Evaluation Questionnaire – Musical Avatar

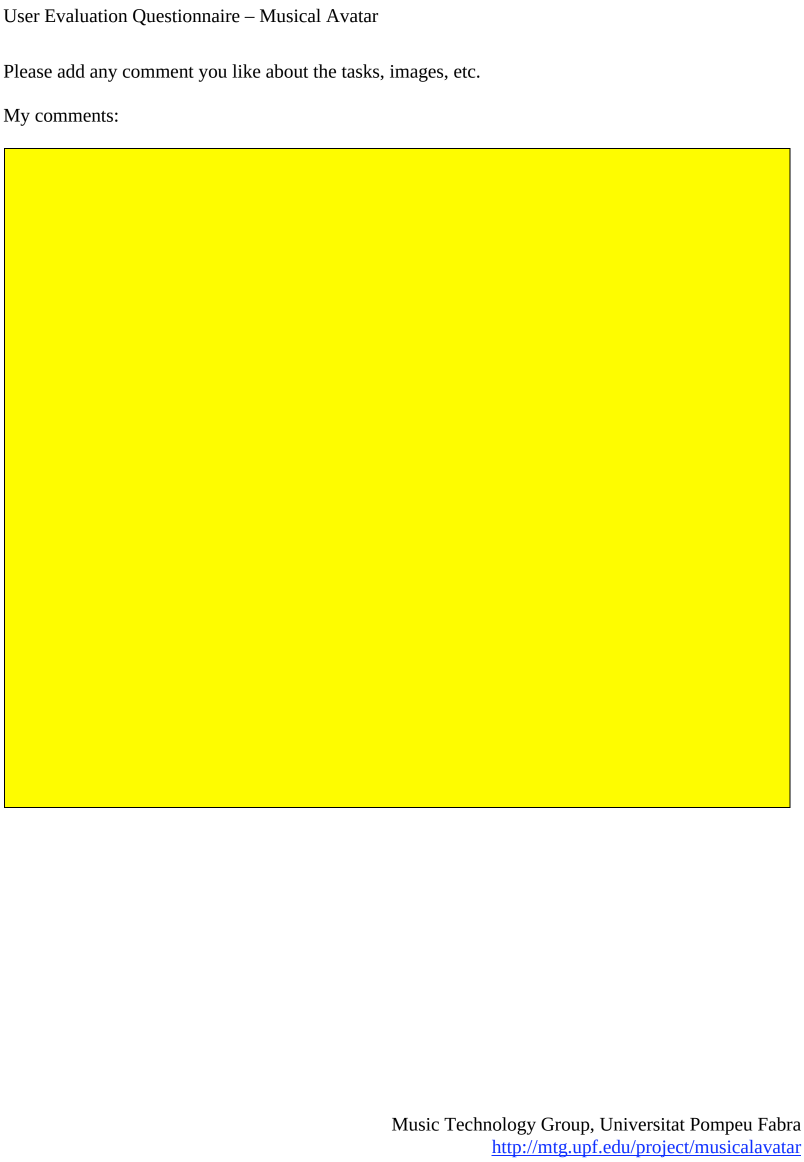
Now that you are acquainted with the Avatars take a look at the following ones and rank them from 1 to 6, where 1 is the avatar that could better express your own musical taste and 6 corresponds with the avatar that has less to do with your musical taste.

Ranking			**Ranking**			**Ranking**	

Ranking			**Ranking**			**Ranking**	

Music Technology Group, Universitat Pompeu Fabra
http://mtg.upf.edu/project/musicalavatar

Figure 4: Continued.



User Evaluation Questionnaire – Musical Avatar

Please add any comment you like about the tasks, images, etc.

My comments:

Music Technology Group, Universitat Pompeu Fabra
http://mtg.upf.edu/project/musicalavatar

Figure 4: Continued.

Appendix D: Publications by the author

ISI-indexed peer-reviewed journals

Bogdanov, D., Haro M., Fuhrmann F., Xambó A., Gómez E., & Herrera P. (2013). Semantic audio content-based music recommendation and visualization based on user preference examples. Information Processing & Management. 49(1), 13-33.

Bogdanov, D., Serrà J., Wack N., Herrera P., & Serra X. (2011). Unifying Low-level and High-level Music Similarity Measures. IEEE Transactions on Multimedia. 13(4), 687-701.

Full-article contributions to peer-reviewed conferences

Bogdanov, D., Wack N., Gómez E., Gulati S., Herrera P., Mayor O., Roma G., Salamon J., Zapata J., & Serra X. (Submitted) ESSENTIA: an audio analysis library for music information retrieval. International Society for Music Information Retrieval Conference (ISMIR).

Bogdanov, D. & Herrera P. (2012). Taking advantage of editorial metadata to recommend music. 9th International Symposium on Computer Music Modeling and Retrieval (CMMR 2012), pp. 618-632.

Yang, Y., Bogdanov D., Herrera P., & Sordo M. (2012). Music retagging using label propagation and robust principal component analysis. 21st International World Wide Web Conference (WWW 2012). 4th International Workshop on Advances in Music Information Research (AdMIRe 2012), pp. 869-876.

Bogdanov, D. & Herrera P. (2011). How much metadata do we need in music recommendation? A subjective evaluation using preference sets. International Society for Music Information Retrieval Conference (ISMIR), pp. 97-102.

Bogdanov, D., Haro M., Fuhrmann F., Xambó A., Gómez E., & Herrera P. (2011). A Content-based System for Music Recommendation and Visualization of User Preferences Working on Semantic Notions. 9th International Workshop on Content-based Multimedia Indexing (CBMI'11), pp. 249-252.

Bogdanov, D., Haro M., Fuhrmann F., Gómez E., & Herrera P. (2010). Content-based music recommendation based on user preference examples. The 4th

ACM Conference on Recommender Systems. Workshop on Music Recommendation and Discovery (Womrad 2010).

Haro, M., Xambó A., Fuhrmann F., Bogdanov D., Gómez E., & Herrera P. (2010). The Musical Avatar – A visualization of musical preferences by means of audio content description. 5th Audio Mostly Conference: A Conference on Interaction with Sound (AM'10).

Bogdanov, D., Serrà, J., Wack, N., & Herrera, P. (2009). From Low-level to High-level: Comparative study of music similarity measures. IEEE International Symposium on Multimedia. International Workshop on Advances in Music Information Research (AdMIRe), pp. 453–458.

Other contributions to conferences

Sordo, M., Celma Ò., & Bogdanov D. (2011). Audio Tag Classification using Weighted-Vote Nearest Neighbor Classification. Music Information Retrieval Evaluation eXchange (MIREX) extended abstract.

Wack, N., Laurier, C., Meyers, O., Marxer, R., Bogdanov, D., Serrà, J., Gómez, E., & Herrera, P. (2010). Music classification using high-level models. Music Information Retrieval Evaluation eXchange (MIREX) extended abstract.

Wack, N., Guaus, E., Laurier, C., Meyers, O., Marxer, R., Bogdanov, D., Serrà, J., & Herrera, P. (2009). Music type groupers (MTG): generic music classification algorithms. Music Information Retrieval Evaluation eXchange (MIREX) extended abstract.

Bogdanov, D., Serrà, J., Wack, N., & Herrera, P. (2009). Hybrid similarity measures for music recommendation. Music Information Retrieval Evaluation eXchange (MIREX) extended abstract.