



ELSEVIER

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Semantic audio content-based music recommendation and visualization based on user preference examples

Dmitry Bogdanov^{a,*}, Martín Haro^a, Ferdinand Fuhrmann^a, Anna Xambó^b, Emilia Gómez^a, Perfecto Herrera^a

^a Music Technology Group, Universitat Pompeu Fabra, Roc Boronat 138, 08018 Barcelona, Spain

^b Music Computing Lab, The Open University, Walton Hall, MK7 6AA Milton Keynes, UK

ARTICLE INFO

Article history:

Received 26 September 2011

Received in revised form 15 February 2012

Accepted 17 June 2012

Available online 25 July 2012

Keywords:

Music information retrieval

Information systems

User modeling

Recommender system

Preference visualization

Evaluation

ABSTRACT

Preference elicitation is a challenging fundamental problem when designing recommender systems. In the present work we propose a content-based technique to automatically generate a semantic representation of the user's musical preferences directly from audio. Starting from an explicit set of music tracks provided by the user as evidence of his/her preferences, we infer high-level semantic descriptors for each track obtaining a user model. To prove the benefits of our proposal, we present two applications of our technique. In the first one, we consider three approaches to music recommendation, two of them based on a semantic music similarity measure, and one based on a semantic probabilistic model. In the second application, we address the visualization of the user's musical preferences by creating a humanoid cartoon-like character – the *Musical Avatar* – automatically inferred from the semantic representation. We conducted a preliminary evaluation of the proposed technique in the context of these applications with 12 subjects. The results are promising: the recommendations were positively evaluated and close to those coming from state-of-the-art metadata-based systems, and the subjects judged the generated visualizations to capture their core preferences. Finally, we highlight the advantages of the proposed semantic user model for enhancing the user interfaces of information filtering systems.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Over the past decade, we have witnessed a rapid growth of digital technologies, the Internet, and the multimedia industry. Consequently, the overload of generated information has created the current need for effective information filtering systems. Such information systems include tools for browsing and indexing large data catalogs as well as recommendation algorithms to discover unknown but relevant items therein. Their development and related research are usually carried out in the field of information retrieval.

In particular, recommender systems built upon user profiles are currently in the spotlight of the information retrieval community. Since preferences are highly subjective, personalization seems to be a key aspect for optimal recommendation. Ideally, such systems should be able to grasp user preferences and provide, on this basis, the content which is relevant to the user's needs.

Preference elicitation can therefore be regarded as a fundamental part of recommender systems and information filtering systems in general. Several approaches have been proposed in the literature to tackle this problem. In particular, Hanani,

* Corresponding author. Tel.: +34 935422000/935422164; fax: +34 935422517.

E-mail address: dmitry.bogdanov@upf.edu (D. Bogdanov).

Shapira, and Shoval (2001) identified two main strategies – explicit and implicit user preference inference. The former relies on user surveys in order to obtain qualitative statements and ratings about particular items or more general semantic properties of the data. In contrast, the latter relies on the information inferred implicitly from user behavior and, in particular, consumption statistics. In the present work, we focus on music recommender systems and consider explicit strategies to infer musical preferences of a user directly from the music audio data.

When considering digital music libraries, current major Internet stores contain millions of tracks. This situation complicates the user's search, retrieval, and discovery of relevant music. At present, the majority of industrial systems provide means for manual search (Nanopoulos, Rafailidis, Ruxanda, & Manolopoulos, 2009). This type of search is based on metadata¹ information about artist names, album or track titles, and additional semantic² properties which are mostly limited to genres. Music collections are then queried by tags or textual input using this information.

Moreover, current systems also provide basic means for music recommendation and personalization, which are not related to the audio content, i.e., using metadata. Such systems obtain a user's profile by monitoring music consumption and listening statistics, user ratings, or other types of behavioral information, decoupled from the actual music data (Baltrunas & Amatriain, 2009; Celma, 2008; Firan, Nejdil, & Paiu, 2007; Jawaheer, Szomszor, & Kostkova, 2010; Levy & Bosteels, 2010; Shardanand & Maes, 1995). In particular, a user can be simply represented as a vector of ratings or playback counts for different artists, albums, and tracks. Having a database of such user profiles, this allows the use of collaborative filtering to search for similar users or music items (Sarwar, Karypis, Konstan, & Reidl, 2001). Alternatively, semantic tag-based profiles can be built to be matched with music items directly. Firan et al. (2007) proposes to create such a semantic profile using implicit information about a user's listening behavior. To this end, they use the user's listening statistics (artist or track playback counts) and the editorial metadata extracted from the files in the user's personal music collection (artist names and track titles). The tags are obtained for artists, albums, and particular tracks from music services which provide means for social tagging, such as *Last.fm*.³ Tags can also be retrieved from information found on the Web (Celma, 2008; Celma & Serra, 2008; Schedl, Widmer, Knees, & Pohle, 2011) in the form of reviews, biographies, blog posts, music related RSS feeds, etc.

These approaches, notably using collaborative filtering for music recommendation, are found to be effective when considering popular music items. However, it has been shown that they fail in the long tail, i.e., for unpopular items, due to the lack of available user ratings, social tags, and other types of metadata (Celma, 2008). On the other hand, there is evidence (Barrington, Oda, & Lanckriet, 2009; Celma & Herrera, 2008) that content-based⁴ information extracted from the audio can help to overcome this problem.

Existing research in the area of audio content-based music recommendation usually focuses on the related task of measuring music similarity. The Music Information Research (MIR) community has achieved relative success in this task (Casey et al., 2008; Downie, Ehmann, Bay, & Jones, 2010), striving to facilitate both manual search and automatization of music recommendation. In these approaches, music tracks are represented in a given feature space, built upon timbral, temporal, tonal, and/or higher-level semantic dimensions, all extracted from audio content (Barrington, Turnbull, Torres, & Lanckriet, 2007; Bogdanov, Serrà, Wack, & Herrera, 2009; Pampalk, 2006; Pohle, Schnitzer, Schedl, Knees, & Widmer, 2009; West & Lamere, 2007). Such a representation enables the definition of similarity measures (or distances⁵) between tracks, which can be used to search music collections using queries-by-example. Such distance-based approaches are designed and evaluated, in most cases, for the query-by-one-example use-case. Since retrieval based on a single example is just a particular case of using a recommender system, these approaches may not be directly suitable for music recommendation purposes in general. As the users only provide one query, no knowledge about their musical preferences is required. Querying by example implies an active interaction by the user to explicitly define the "direction of search". As a result, such approaches are not suitable when a user does not know her/his exact needs and prefers receiving recommendations from an available music collection without defining an example (seed) item.

In addition to these non-personalized measures, there has only been sparse work on personalized music similarity measures from audio content data (Lu & Tseng, 2009; Sotiropoulos, Lampropoulos, & Tsihrintzis, 2007; Vignoli & Pauws, 2005). These studies introduce metrics, which are adapted according to a user's perception of similarity to measure distances between tracks in a given collection. Nevertheless, these studies are also focused on the query-by-one-example scenario, and, in their majority, do not take musical preferences into account.

Alternatively, there exist few research studies on user preference modeling for music recommendation which include studies of audio content-based (Grimaldi & Cunningham, 2004; Hoashi, Matsumoto, & Inoue, 2003; Logan, 2004; Mandel & Ellis, 2005) and hybrid approaches (Li, Myaeng, Guan, & Kim, 2005; Su, Yeh, & Tseng, 2010; Yoshii, Goto, Komatani, Ogata, & Okuno, 2006). These studies present several shortcomings. Firstly, they operate solely on rough timbral, and sometimes temporal and tonal information. This information is low-level as it does not incorporate higher-level semantics in the description of music. In the case of music similarity, it has been shown that distance measures which operate on semantic descriptors, inferred from low-level features, outperform low-level derived similarities (Barrington et al., 2007; Bogdanov

¹ We pragmatically use the term "metadata" to refer to any information not extracted from the audio signal itself.

² 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, and instrumentation.

³ <http://last.fm>.

⁴ We use the terms "audio content-based" or "audio content" to refer to any information extracted from the raw audio signal.

⁵ For the sake of simplicity we refer to any (dis) similarity estimation with the term "distance".

et al., 2009; West & Lamere, 2007). Recent research suggests that exploiting a semantic domain can be a relevant step to overcome the so-called semantic gap (Aucouturier, 2009; Celma, Herrera, & Serra, 2006), which arises from the weak linking between human concepts related to musical aspects and the low-level feature data extracted from the audio signal. Furthermore, the metadata components of the majority of hybrid approaches solely use information about user ratings, exploiting it in a collaborative filtering manner. This allows to measure relations between different music tracks or between different users, but does not provide insights into the underlying relations between the music and the user himself, i.e., the nature of musical preferences. Moreover, a large amount of users and ratings is usually required for reasonable performance, as such systems are prone to the so-called “cold-start problem” (Maltz & Ehrlich, 1995), i.e., the inability to provide good recommendations at the initial stages of the system.

This all indicates a lack of research on both metadata-based and audio content-based strategies for an effective elicitation of musical preferences, including comprehensive evaluations on large music collections and real listeners. Most existing approaches exploit user ratings as the only source of explicit information. The evaluation of such approaches is often done objectively without the participation of real listeners. Ground truth datasets of user ratings are used instead. However, these ratings can be considered as indirect and even noisy preference statements (Amatriain, Pujol, & Oliver, 2009). They do not necessarily represent real user preferences, as they are biased by the precision of a rating scale, decisions on the design of the recommender interface, etc. (Cosley, Lam, Albert, Konstan, & Riedl, 2003). In turn, implicit listening behavior statistics based on track counts might not represent real preferences in particular since it ignores the difference between track durations or users' activities when listening the music (Jawaheer et al., 2010). Furthermore, these information sources do not guarantee a complete coverage of all kinds of preferred items. Alternative explicit approaches are generally limited to surveying for the names of favorite artists, albums, or preferred genres.

In the present work, we focus on audio content-based user modeling suitable for music recommendation. In contrast to most existing approaches, we propose a novel technique which is based on the automatic inference of a high-level semantic description⁶ of the music audio content, covering different musical facets, such as genre, musical culture, moods, instruments, rhythm, and tempo. These semantic descriptors are computed from an explicit set of music tracks defined by a given user as evidence of her/his musical preferences. To the best of our knowledge this approach for user modeling for music recommendation has never been evaluated before. In particular, our technique relies on two hypotheses. First, we suppose that asking for explicit preference examples is an effective way to infer real user preferences. Second, we assume that high-level semantic description outperforms common low-level feature information in the task of music recommendation. The latter hypothesis is based on similar evidence in the case of music similarity estimation (Bogdanov et al., 2009).

In particular, our focus lies on music discovery as the use-case of a recommender system, where we consider both relevance and novelty aspects, i.e., recommending music liked by, but previously unknown to users. We propose three new recommendation approaches operating on semantic descriptions, based on the proposed user preference modeling technique. To evaluate them, we compare our methods with two baseline approaches working on metadata. First, we employ a simple approach which uses exclusively genre information for a user's preference examples. Second, we apply a state-of-the-art commercial black-box recommender system on the basis of *Last.fm*. This recommender relies on metadata, and partially uses collaborative filtering information (Levy & Bosteels, 2010), operating on a large database of users and their listening statistics. We provide this system with editorial metadata for the preference examples to retrieve recommendations. Moreover, we also consider two audio content-based baseline approaches. In contrast to the proposed semantic methods, these algorithms use the same procedure for recommendation but operate on low-level timbral features. We then evaluate all considered approaches on 12 subjects, for which we use their gathered preference data to generate recommendations and carry out a listening experiment to assess familiarity, liking and further listening intentions of the provided recommendations. The obtained results indicate that our proposed approaches perform close to metadata-based commercial systems. Moreover, we show that the proposed approaches perform comparably to the baseline approach working on metadata which relies exclusively on manually annotated genre information to represent user preferences and a music collection to recommend music from. Furthermore, the proposed approaches significantly outperformed the low-level timbre-based baselines, supporting our hypothesis on the advantage of using semantic descriptors for music recommendation.

In a second step we exploit the proposed user preference model to map its semantic description 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. We propose a novel approach to depict a user's preferences. In our study we 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 the same 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.

In summary, the proposed technique generates a user model from a set of explicitly provided music tracks, which, in turn, are characterized by the computed semantic descriptors. This semantic representation can be useful in different applications, along with music recommendation, to enrich user experience and increase user trust in a final recommender system. The examples of such applications are, among others, user characterization and visualization, and justification of the provided recommendations. To support and evaluate the proposed technique, we focus on two applications, namely music recommen-

⁶ We will use the generic terms “descriptor” and “semantic descriptor” to refer to any high-level semantic description.

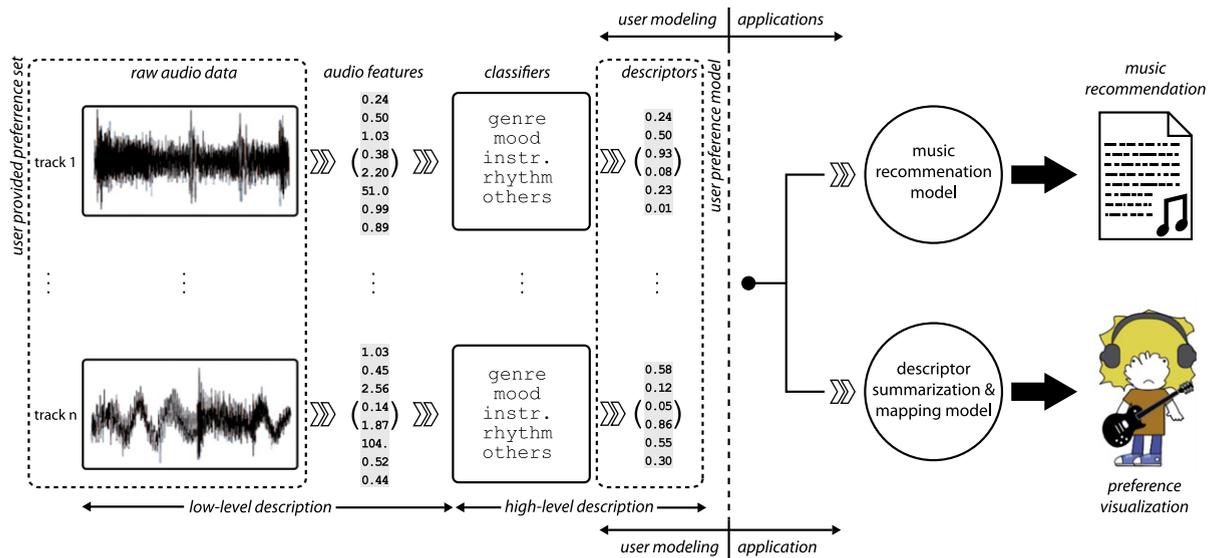


Fig. 1. General scheme of the proposed preference elicitation (user modeling) technique and its applications.

dation and musical preference visualization. A general scheme of the proposed technique and its applications is presented in Fig. 1.

This article is organized as follows: The next section covers related work in the field of audio content-based music recommendation. In Section 3 we describe the proposed preference elicitation technique, including the processes of data gathering Section 3.1 and automatic descriptor extraction Section 3.2. In Section 4 we analyze the evaluation data provided by 12 participants. Section 5 focuses on the first application of the presented preference inference technique – audio content-based music recommendation. In Section 6 we present the second application – audio content-based visualization of musical preferences, starting from the proposed user modeling technique. In Section 7 we provide a general discussion and consider several use-cases of integration of the proposed applications into a final recommender system and their further improvement. Finally, in Section 8 we state general conclusions and highlight future research directions.

2. Related work in music recommendation

In this section we review the most important studies in music recommendation, considering both audio content-based and hybrid approaches. These studies can be divided into three categories: personalized music similarity measures, audio content-based models and hybrid models of user preferences.

A number of studies incorporate perceptual personalization of music similarity measures which can be applied for music recommendation. Sotiropoulos et al. (2007) present an active learning system, which adapts the underlying Euclidean distance measure according to a user's feedback on the perceived music similarity. The system operates on sets of timbral, temporal, and tonal features, employing feature selection based on neural networks. Vignoli and Pauws (2005) present a music recommender system based on a hybrid distance measure defined as a user-weighted combination of timbre, genre, tempo, year, and mood distance components. The weights can be explicitly defined by the user. Moreover, Lu and Tseng (2009) present a personalized hybrid recommender system. They propose to combine a distance working on tonal and rhythmic features together with a distance based on collaborative filtering information about preferred tracks, and a semantic emotion-based distance. In order to train the personalized hybrid distance, the user is given a sample of music tracks and is asked to explicitly supply the system with preference assessments (likes/dislikes) and the underlying reasons (such as preference by tonality, and rhythm) for each track. Based on these assessments, the system searches for the closest tracks to the preferred tracks in a music collection using the personalized distance. 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 mood annotations done by experts.

Regarding the work on audio content-based user modeling for music recommendation, Hoashi et al. (2003) present a system with an underlying classification procedure, which divides tracks into the "good" and "bad" categories according to the genre preferences explicitly given by a user. Tree-based vector quantization is used for classification of the tracks represented in a timbral feature space by mel-frequency cepstral coefficients (MFCCs). A sample of tracks labeled by genre is used for initial training of the algorithm. Additional corrections to the classification algorithm can be done via relevance feedback. Grimaldi and Cunningham (2004) apply similar 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 timbral and temporal features. These classifiers and features were originally suited for the task of genre classification. Due to this fact, the authors found that the

proposed approach fails in the case when the user's preference is not driven by a certain genre. Logan (2004) proposes to generate recommendations based on an explicitly given set of music tracks, which represent a user's preferences. A timbral distance measure is applied to find the tracks similar to the set. As such, the author proposes to use the Earth mover's distance between clusters of MFCCs, which represent music tracks. Unfortunately, no evaluation on real listeners was conducted. Instead, a set of tracks from a randomly chosen album was used to simulate a user's preferences. A track for the same album, not belonging to the user set, is then used as an objective criterion for the evaluation. One of the potential drawbacks of such an evaluation methodology consists in the bias, which leads to the overestimation of real performance, given that timbral distances tend to easily recognize tracks for the same album due to the so-called "album effect" (Mandel & Ellis, 2005). This effect implies that, due to the production process, tracks from the same album share much more timbral characteristics than tracks from different albums of the same artist, and, more so, different artists.

Finally, there are more sophisticated user modeling approaches which use both metadata and audio content information. Yoshii et al. (2006) present a probabilistic user model, which incorporates ratings given by a user and audio content-based "bags-of-timbres". The latter ones represent polyphonic timbre weights, and are obtained from a Gaussian mixture model of MFCCs for each track. The authors use a Bayesian network in the core of their system. A simulation by user ratings obtained from the Amazon Internet store was used to conduct an objective evaluation. Li et al. (2005) and Li, Myaeng, and Kim (2007) propose a track-based probabilistic model, which extends the collaborative filtering approach with audio content-based information. In this model, music tracks are classified into groups based on both available user ratings (by all users in the system) and the extracted set of timbral, rhythmic, and pitch features. The predictions are made based on a user's own ratings, considering their Gaussian distribution on each group of tracks. The authors conducted an objective evaluation using ground truth user ratings. Similarly, Su et al. (2010) present a hybrid recommendation approach, which represents the tracks in a audio content-based feature space. Patterns of temporal evolution of timbral information are computed for each track, represented as frame sequences of clusters of timbral features. Subsequently, given a collaborative filtering information in the form of user ratings, the tracks can be classified into "good" and "bad" according to the ratings of a user and his/her neighbors with similar ratings. To this end, the frequency of the occurrence of "good" and "bad" patterns are computed for each track and are taken as a criterion for classification. The evaluation of the proposed approach is done on ground truth ratings obtained from the Amazon Internet store.

3. Methodology

In this section we explain the proposed audio content-based technique for user modeling. We describe the underlying procedure of gathering user preference examples and the process of descriptor extraction. This technique was partially presented in (Bogdanov, Haro, Fuhrmann, Gómez, & Herrera, 2010; Haro et al., 2010).

3.1. Preference examples gathering

As a first step, we ask users to gather the minimal set of music tracks which is sufficient to grasp or convey their musical preferences (the user's *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 user 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 proposed audio content-based technique and its applications, the music pieces would be informative even without any additional metadata (such as artist names and track titles). Nevertheless, for a considerable amount of users in a real world (industrial) scenario, providing metadata can be easier than uploading audio. In this case, the audio including full tracks or previews can be obtained from the associated digital libraries by the provided metadata.

For our evaluation purposes only, users are obliged to provide audio files and optionally provide metadata. We then, by means of audio fingerprinting,⁷ retrieve and clean metadata for all provided tracks including the ones solely submitted in audio format. Therefore, we will be able to compare our approaches to metadata-based approaches in the case of music recommendation. We also ask the users 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.

3.2. Descriptor extraction

Here we describe the procedure followed to obtain a semantic representation of each music track from the user's preference set. We follow Bogdanov et al. (2009) and Bogdanov, Serrà, Wack, Herrera, and Serra (2011) to obtain such descriptions.

⁷ We use MusicBrainz service: http://musicbrainz.org/doc/MusicBrainz_Picard.

Table 1

Ground truth music collections employed for semantic regression. Source references: (1) Homburg et al. (2005), (2) in-house, (3) Tzanetakis and Cook (2002), (4) Gómez and Herrera (2008), (5) Laurier et al. (2009) + in-house, and (6) 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	Genre & Culture	Classical, dance, hip-hop, jazz, pop, rhythm'n'blues, rock, speech	400 tracks, 50 per genre	(2)
G3	Genre & Culture	Blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock	993 track excerpts, 100 per genre	(3)
CUL	Genre & Culture	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	Moods & Instruments	Sad, non-sad	230 full tracks + excerpts, 96/134 per class	(5)
MAG	Moods & Instruments	Aggressive, non-aggressive	280 full tracks + excerpts, 133/147 per class	(5)
MRE	Moods & Instruments	Relaxed, non-relaxed	446 full tracks + excerpts, 145/301 per class	(5)
MAC	Moods & Instruments	Acoustic, non-acoustic	321 full tracks + excerpts, 193/128 per class	(5)
MEL	Moods & Instruments	Electronic, non-electronic	332 full tracks + excerpts, 164/168 per class	(5)
RPS	Rhythm & Tempo	Perceptual speed: slow, medium, fast	3000 full tracks, 1000 per class	(2)
RBL	Rhythm & Tempo	Chachacha, jive, quickstep, rumba, samba, tango, viennese waltz, waltz	683 track excerpts, 60–110 per class	(6)
ODA	Other	Danceable, non-danceable	306 full tracks, 124/182 per class	(2)
OPA	Other	Party, non-party	349 full tracks + excerpts, 198/151 per class	(2)
OVI	Other	Voice, instrumental	1000 track excerpts, 500 per class	(2)
OTN	Other	Tonal, atonal	345 track excerpts, 200/145 per class	(2)
OTB	Other	Timbre: bright, dark	3000 track excerpts, 1000 per class	(2)

For each music track, we calculate a low-level feature representation using an in-house audio analysis tool.⁸ In total, this tool provides over 60 commonly used low-level audio features, characterizing global properties of the given tracks, related to timbral, temporal, and tonal information. The features include inharmonicity, odd-to-even harmonic energy ratio, tristimuli, spectral centroid, spread, skewness, kurtosis, decrease, flatness, crest, and roll-off factors, MFCCs, spectral energy bands, zero-crossing rate (Peeters, 2004), spectral complexity (Streich, 2007), transposed and untransposed harmonic pitch class profiles, key strength, tuning, chords (Gómez, 2006), pitch, beats per minute (BPM) and onsets (Brossier, 2007). Most of these features are extracted on a frame-by-frame basis and then summarized by their means and variances across all frames. In the case of multidimensional features (e.g., MFCCs), covariances between components are also considered.

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. We opt for multi-class support vector machines (SVMs) with a one-vs.-one voting strategy (Bishop, 2006), and use the libSVM implementation.⁹ 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 17 ground truth music collections (including full tracks and excerpts) presented in Table 1, corresponding to 17 classification tasks.

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 (Hall, 2000) according to the underlying music collection. Moreover, the parameters of each SVM are found by a grid search with 5-fold cross-validation. 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 and RBL show the worst performance, close to 60%,¹⁰ while classifiers for CUL, MAG, MRE, MAC, OVI, and OTB show the best performance, greater than 93%.

With the described procedure we obtain 62 semantic descriptors, shown in Table 1, for each track in the user's preference set. These resulting representations of tracks (i.e., vectors of class probabilities) form our proposed user model, defined as a set U :

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

⁸ <http://mtg.upf.edu/technologies/essentia>.

⁹ <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.

¹⁰ Still, note the amount of classes in G1 and RBL classifiers is 9 and 3, respectively.

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.

As the procedure of the low-level signal analysis and the details of semantic descriptor extraction are out of the scope of this paper, we refer the interested reader to the aforementioned literature on low-level features, and to (Bogdanov et al., 2009, 2011), and references therein, for details on the SVM implementation.

4. User data analysis

In order to evaluate the proposed technique, we worked with a group of 12 participants (8 male and 4 female) selected from the authors' colleagues and acquaintances without disclosing any detail of the targeted research. They were aged between 25 and 45 years old (average $\mu = 33$ and standard deviation $\sigma = 5.35$) and showed a very high interest in music (rating around $\mu = 9.64$, with $\sigma = 0.67$, where 0 means no interest in music and 10 means passionate about music). Ten of the 12 participants play at least one musical instrument, including violin, piano, guitar, synthesizers, and ukulele.

The number of tracks selected by the participants to convey their musical preferences was very varied, ranging from 23 to 178 music pieces ($\mu = 73.58$, $\sigma = 45.66$) with the median being 57 tracks. The time spent for this task also differed a lot, ranging from half an hour to 60 h ($\mu = 11.11$, $\sigma = 22.24$) with the median being 5 h.

It is interesting to analyze the provided verbal descriptions about the strategy followed to select the music tracks. Some of the participants were selecting 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, musical qualities, and chronological order 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 aspects: 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 4 to 94 labels ($\mu = 25.11$, $\sigma = 23.82$). 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 Fig. 2. We observe that most of them were 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). 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.

Finally, the music provided by the participants was very diverse. Fig. 3 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.

5. Music recommendation

The first considered application exploits the computed user model to generate music recommendations based on semantic descriptors. For consistency, we focus on the task of retrieving 20 music tracks from a given music collection as recom-

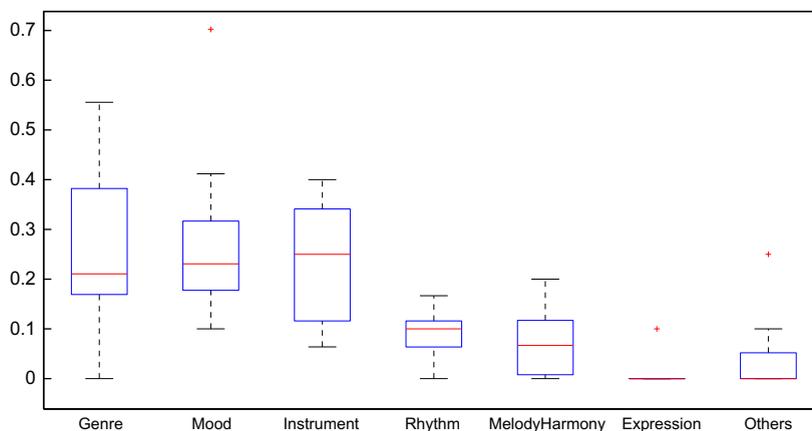


Fig. 2. Box plot of the proportions of provided labels per musical facet, normalized by the total number of labels per participant. Categories from left to right correspond to genre, moods, instruments, rhythm, melody and harmony, musical expression, and other labels respectively. Red crosses stand for extreme outliers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2. *Semantic distance from all tracks (SEM-ALL)*. Alternatively, instead of simplifying the user model to one point we consider all individual tracks. 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. We return the 20 nearest tracks according to this distance as recommendations.
3. *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. We select the number of components 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 values¹¹ and return the 20 most probable tracks as recommendations.

5.2. Evaluation

Here we describe the evaluation of the proposed recommendation approaches against metadata-based and audio content-based baselines.

5.2.1. Metadata-based baselines

We consider two baseline approaches to music recommendation working on metadata. The first baseline is constructed exclusively using information about the user's genre preferences. The second one is based on the information about preferred tracks and artists (taken from the editorial metadata provided by the user for the preference set), and partially employs collaborative filtering information, querying a commercial state-of-the-art music recommender for similar music tracks.

1. *Random tracks from the same genre (GENRE)*. This simple and low-cost approach provides random recommendations relying on genre categories of the user's preference set. We assume that all tracks in the given music collection are manually tagged with a genre category by an expert. We randomly preselect 20 tracks from the preference set and obtain their genre labels. 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. For each of the 20 preselected tracks, we return a random track of the same genre label.
2. *Black-box music similarity from Last.fm (LASTFM)*. As we did not have collaborative filtering data available for our research (and moreover, a large dataset would be required to match with our participants' tracks), we opted to use black box recommendations provided by *Last.fm*.¹² It 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-track¹³ and artist-to-artist¹⁴ similarity computed by the undisclosed algorithm, which is partially based on collaborative filtering, but does not use any audio content. 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. Again, we randomly preselect 20 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 track from the recommended ones with an available preview.¹⁵ 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 with an available preview.

5.2.2. Audio content-based baselines

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 from all tracks (MFCC-ALL)*. This approach is a counterpart to the proposed *SEM-ALL* approach using a common low-level timbral distance (Pampalk, 2006) instead of the semantic one. The tracks are modeled by probability dis-

¹¹ Under the assumption of a uniform distribution of the tracks in the universe within the semantic space.

¹² All experiments were conducted on May 2010.

¹³ For example, http://last.fm/music/Grandmaster+Flash/_/The+Message/+similar.

¹⁴ For example, <http://last.fm/music/Baby+Ford/+similar>.

¹⁵ These previews are downloadable music excerpts (30 s), which are later used in our subjective evaluation for the case of the *LASTFM* approach.

tributions 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.

5.2.3. Evaluation methodology

We performed subjective listening tests on our 12 subjects in order to evaluate the considered approaches. 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 s) by 47,000 artists (approximately 2 tracks per artist).

For each subject, we computed the user model from the provided preference set. According to the considered recommendation approaches we generated 7 playlists (three by the proposed approaches working with the semantic user model, two by the approaches working on metadata, and two by the low-level timbral approaches). Each playlist consisted of 20 music tracks. Following a usual procedure for evaluation of music similarity measures and music recommendations, we applied 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. These playlists were merged into a single list, in which tracks were randomly ordered and anonymized, including filenames and metadata. The tracks offered as recommendations were equally likely to come from each single recommendation approach. This allowed us to 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 were not aware of the amount of recommendation approaches, their names and their rationales.

We designed a questionnaire in order to obtain the different subjective impressions related to the recommended music (see Table 2). For each recommended track the participants were asked to provide a number of ratings:

- *Familiarity* ranged from 0 to 4; with 0 meaning absolute unfamiliarity, 1 feeling familiar with the music, 2 knowing the artist, 3 knowing the title, and 4 the identification of artist and 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 were also asked to provide the track title and artist name for those tracks rated high in the familiarity scale.

5.2.4. Results

First, we manually corrected familiarity ratings when the artist/title provided by a user was incorrect compared to the actual ones. In such situations, a familiarity rating of 3, or, more frequently, 4 or 2, was lowered to 1 (in the case of incorrect

Table 2

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, etc.)
	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 does not 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
Give-me-more	0	I will skip it in any occasion I find in a playlist
	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

artist and track title) or 2 (in the case of correct artist, but incorrect track title). These corrections represented just 3% of the total familiarity judgments.

Considering the subjective scales used, a good recommender system should provide high-liking/listening intentions/request for the greater part of retrieved tracks and in particular for low-familiarity tracks. Therefore, we recoded the user's ratings into 3 main categories, referring to the type of the recommendation: *hits*, *fails*, and *trusts*. Hits were those tracks having a low familiarity rating (<2), high (>2) liking and intentions ratings, and a positive (>0) "give-me-more" request. Fails were those tracks having low (<3) liking and intentions ratings, and null "give-me-more" request. Trusts were those tracks which got a high familiarity (>1), high (>2) liking and intentions ratings, and a positive (>0) "give-me-more" request. 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). A user could be satisfied by getting a trust track from time to time, but annoyed if every other track is a trust, especially in the use-case of music discovery (the main focus of the present work). 18.3% of all the recommendations were considered as "unclear" (e.g., a case when a track received a high liking, but a low intentions rating and a null "give-me-more" request). 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 3 and Fig. 5a. An inspection of it reveals that the approach which yields the largest amount of hits (41.2%) and trusts (25.4%) is *LASTFM*. The trusts found with other approaches were scarce, all below 4%. The approaches based on the proposed semantic user model (*SEM-ALL*, *SEM-MEAN* and *SEM-GMM*) yielded more than 30% of hits, and the remaining ones did not surpass 25%. The existence of an association between recommendation approach and the outcome of the recommendation was statistically significant, according to the result of the Pearson chi-square test ($\chi^2(18) = 351.7, p < 0.001$).

Additionally, we performed three separate between-subjects ANOVA tests in order to test the effects of the recommendation approaches on the liking, intentions, and "give-me-more" subjective ratings. 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 three tested indexes. This pattern highlights the *LASTFM* approach as the one getting the highest overall ratings. It also groups together the timbral *MFCC-GMM* and *MFCC-ALL* approaches (those getting the lowest ratings), and the remaining approaches (*SEM-ALL*, *SEM-MEAN*, *SEM-GMM*, and *GENRE*) are grouped in-between. The mean values of the obtained liking, listening intentions, and "give-me-more" ratings per each approach are presented in Fig. 5b.

Finally, 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. 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.

5.3. Discussion

We presented an application of the considered user model for music recommendation. Based on this computed model, we proposed three approaches operating on a subset of the retrieved semantic descriptors. Two of these approaches recommend tracks similar to the preference set using a semantic distance. The third approach creates a probabilistic model using GMM to estimate the density of the user's preferences within the semantic domain. We evaluated these approaches against two metadata-based and two audio content-based baselines in a subjective evaluation on 12 participants. Specifically, we employed a simple metadata-based approach which recommends random tracks, selected from the genres preferred by the

Table 3

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

Approach	Fail	Hit	Trust	Unclear
<i>SEM-MEAN</i>	49.167	31.250	2.500	17.083
<i>SEM-ALL</i>	42.500	34.583	3.333	19.583
<i>SEM-GMM</i>	48.750	30.000	2.500	18.750
<i>MFCC-ALL</i>	64.167	15.000	2.083	18.750
<i>MFCC-GMM</i>	69.583	11.667	1.250	17.500
<i>LASTFM</i>	16.667	41.250	25.417	16.667
<i>GENRE</i>	53.750	25.000	1.250	20.000

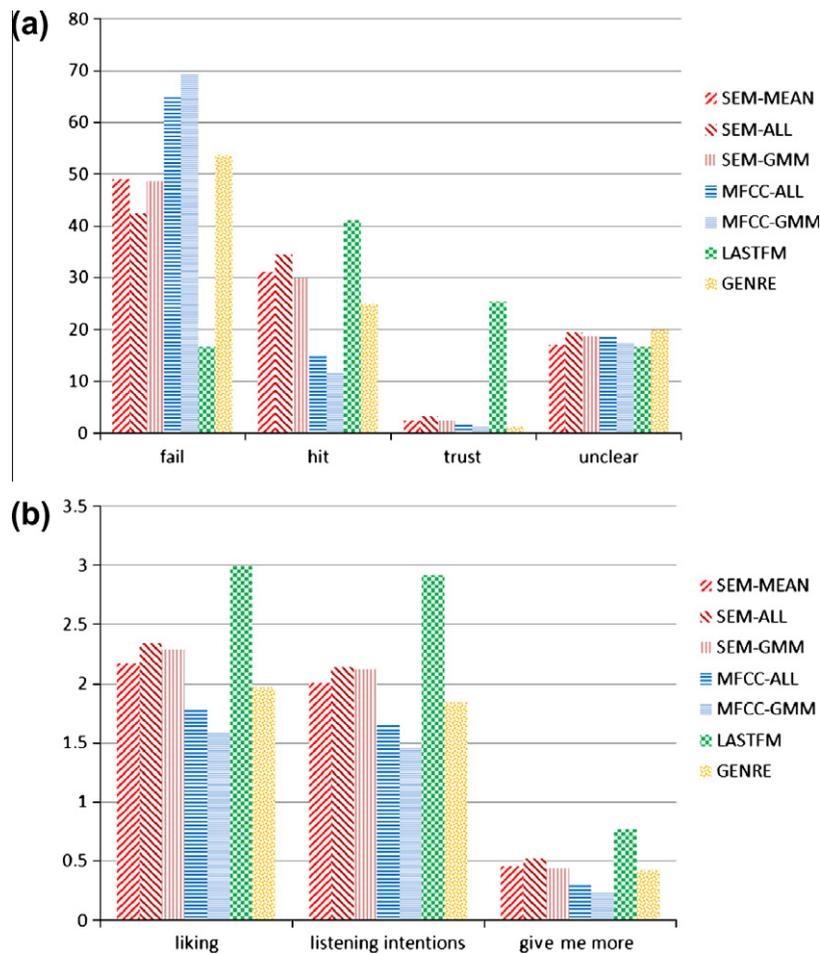


Fig. 5. The percent of fail, trust, hits, and clear categories per recommendation approach (a); the liking, listening intentions, and “give-me-more” mean ratings for recommendation approach (b). The results for the *LASTFM* approach were obtained on a different underlying music collection. The “give-me-more” rating varies in the $[0,1]$ interval.

user. Alternatively, given the editorial metadata for the user’s preference set, we employed a state-of-the-art black-box recommender working on collaborative filtering information – *Last.fm*, to retrieve similar music. Among the audio content-based baselines, we employed two approaches operating on low-level timbral features (MFCCs) instead of the semantic descriptors. These approaches are counterparts to our semantic distance-based and probabilistic approaches, working with a timbral user model.

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 semantic description for music recommendation. Moreover, it complements the outcomes from the previous research on semantic music similarity (Bogdanov et al., 2009). We may conclude that the high-level semantic description outperforms the low-level timbral description in the task of music recommendation and, in particular, musical preference elicitation.

Comparing with the baselines working on metadata, we found that the proposed approaches perform better than the simple genre-based recommender (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 (see Fig. 2), and that manually annotated genre and sub-genre 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*) in terms of the number of successful novel recommendations (hits). Still, this metadata-based approach using collaborative filtering yielded only 7 absolute percentage points more hits than one of our proposed semantic methods (*SEM-ALL*). Considering trusted recommendations, the *LASTFM* baseline provided about 22% more recommendations already known by the participants. Interestingly, one track out of four recommended by this baseline was already familiar to the participants, which might be considered an excessive amount considering the music discovery use-case. In particular, the larger amount

of both hits and trusts provided by the *LASTFM* 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 the obtained performance of this approach to be an upper bound in both hits and trusts and expect a lower performance on our smaller in-house collection. Taking all this into account, we expect the proposed semantic approaches, and the underlying semantic user model, to be suitable for music discovery in the long tail which can suffer from insufficient, incorrect, or incomplete metadata information.

6. Visualization of musical preferences

The second application exploits the computed user model to generate a visualization of the user's musical 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.

6.1. Descriptor summarization

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 operate on a reduced subset of descriptors for simplicity reasons in the mapping process. To this end, we select this subset considering the classifiers' accuracy against ground truth values provided by a subset of five 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 4.

Having refined the semantic descriptors for the computed user model, 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* representation proposed for music recommendation (Section 5.1). 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 (Section 5.2.3). Moreover, we range-normalize the aforementioned standardized descriptor values according to the following equation:

$$N_i = \frac{d_i - \min}{\max - \min}, \quad (2)$$

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 6.2). We then summarize the descriptor values across tracks by computing the mean for every normalized descriptor ($\mu_{N,i}$).

Table 4

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)	

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:

$$\text{Ternary}_i = \begin{cases} 1 & \text{if } \mu_i > (\mu_{R,i} + th_i), \\ 0.5 & \text{if } (\mu_{R,i} - th_i) \leq \mu_i \leq (\mu_{R,i} + th_i), \\ 0 & \text{if } \mu_i < (\mu_{R,i} - th_i), \end{cases} \quad (3)$$

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 17 semantic descriptors.

6.2. Visualization

In order to generate the *Musical Avatar*, we convert the summarized semantic descriptors to a set of visual features. According to MacDONALD, HARGREAVES, and MIELL (2002), individual, cultural and sub-cultural musical identities emerge through social groups concerning different types of moods, behaviors, values or attitudes. 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, de AGUIAR, THEOBALT, MAGNOR, & SEIDEL, 2005; PETAJAN, 2005; SAUER & YANG, 2009). Secondly, it is a graphical medium which, by eliminating superfluous features, amplifies the remaining characteristics of a personality (MCCLLOUD, 2009). Thirdly, there are examples of existing popular avatar collections of this kind such as Meegos¹⁶ or Yahoo Avatars.¹⁷

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, AALTONEN, & SIIRTOLA, 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 a neutral character,¹⁸ 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 5. 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). Apart from 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 gray background color that ranges from RGB (100,100,100) to RGB (200,200,200). The relation between graphic groups and categories of the semantic descriptors is presented in Table 5. We note that our decisions on the design, and in particular on the descriptor mapping, are arbitrary, being a matter of choice, of visual and graphic sense, and common sense according to many urban styles of self-imaging.

¹⁶ <http://meegos.com>.

¹⁷ <http://avatars.yahoo.com>.

¹⁸ 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 to any descriptor at all.

Table 5
Mapping of the descriptor categories to the graphic groups.

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

Table 6

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 6 for an example of the vector representation of these elements). This subset is defined according to the mapping criteria depicted in Table 5. As a result, a particular *Musical Avatar* is generated for the user's musical preferences. All graphics are done in vector format for rescalability and implemented using Processing¹⁹ (Reas & Fry, 2007).

According to the summarization methods considered in Section 6.1, 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. Fig. 6 shows a graphical example of our visualization strategy where, given the summarized binary user model, the best graphic symbol for each graphic group is chosen. Fig. 7 shows a sample of *Musical Avatars* generated by the three summarization methods and Fig. 8 shows a random sample of different *Musical Avatars*.

6.3. Evaluation

6.3.1. Evaluation methodology

We carried out a subjective evaluation on our 12 subjects. 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 4). We requested a real number between 0 and 1 to rate the degree of preference for

¹⁹ <http://processing.org>.

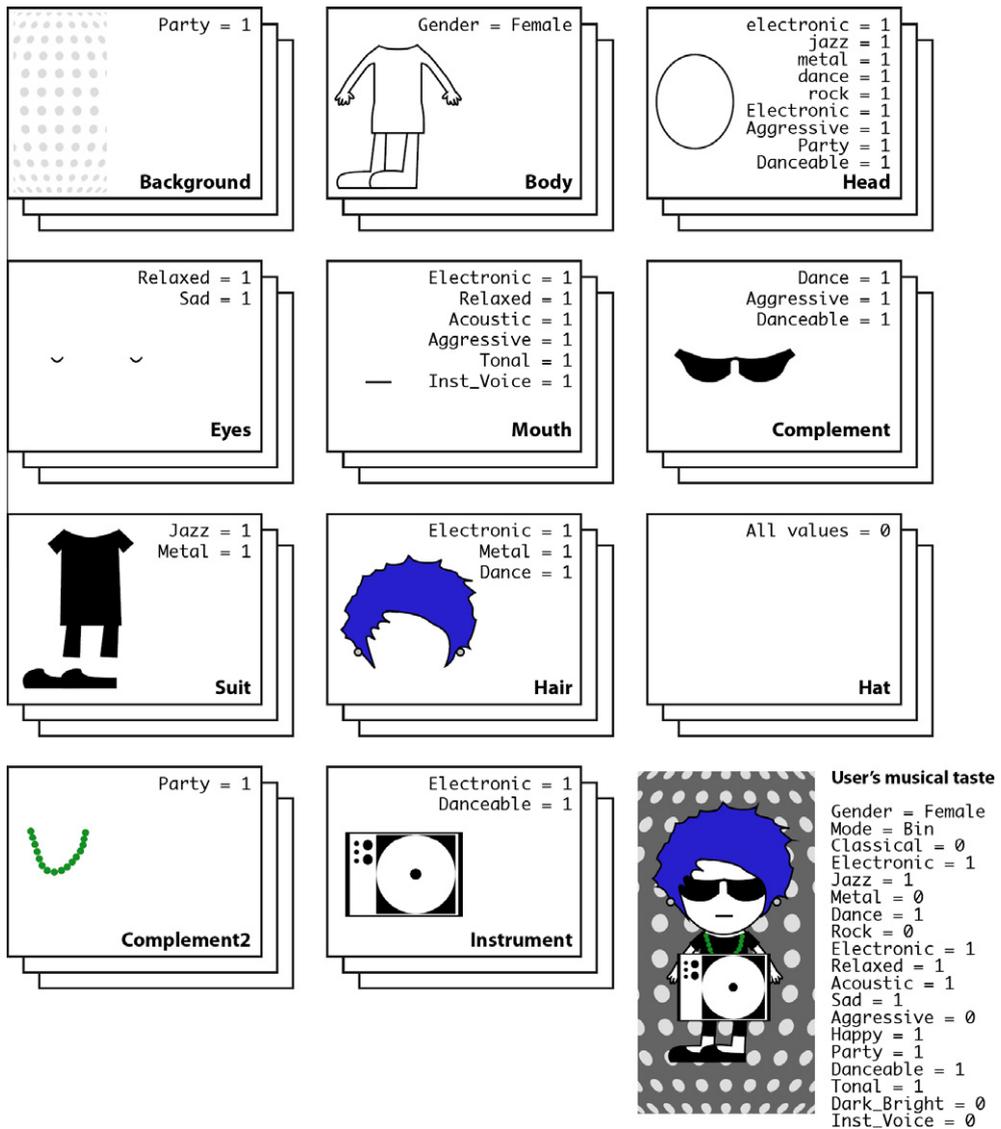


Fig. 6. Example of the visualization approach. It can be seen how the descriptor values influence the selection of the different graphic elements used to construct the avatar. The values inside the graphic element boxes represent all possible descriptor values that can generate the presented element.



Fig. 7. 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.

each descriptor (e.g., 0 meaning “I do not 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



Fig. 8. A random sample of Musical Avatars.

Table 7

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

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.²⁰

6.3.2. 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. We tested the effect of the method on the ratings obtained from the subjects using a within-subjects ANOVA. 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 considered 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 then introduced a dissimilarity measure to assess the significance of the summarized description of musical preferences. In particular, we estimated 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 ($t(11) = -5.11$, $p < 0.001$).

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 clearly shows a user preference for all three proposed

²⁰ A screenshot of the evaluation and more Musical Avatars are available online <http://mtg.upf.edu/project/musicalavatar>.

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 until a larger sample of user evaluations is available. Specifically, the conducted ANOVA did not reveal a clear particular preference for any of the three considered methods.

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 user model). Finally, some participants could not decode the meaningfulness of some visual features (e.g., glasses, head shape). 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. General discussion and possible enhancements for recommender systems

Let us shortly recapitulate the major contents of the current work. We proposed a novel technique for semantic preference elicitation suitable for various applications within music recommender systems and information filtering systems in general. Two such applications – music recommendation and musical preference visualization, were presented and evaluated from a user modeling point-of-view. Moreover, the proposed user modeling approach and the considered applications 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.

Considering the limitations of our study, we would like to note that we employed a small sample of subjects (12 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 tradeoff between the number of participants (Hoashi et al., 2003) and the number of evaluated tracks per approach by a particular user Barrington et al. (2009) and Lu and Tseng (2009). Furthermore, no studies on human evaluation of visualization approaches considering musical preferences are known to the authors.

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 applications, recommendation and visualization, into a single, interactive music recommender interface. For both considered applications a user-defined preference set served as a starting point. This preference set is automatically converted into a fully semantic user model. The preference set offers a compact description of presumably multi-faceted preferences explicitly in terms of multiple music tracks. Therefore it is not limited to a single seed item or semantic term to draw recommendations from. In contrast, the preference set manually provided by the user assures a noise-free representation of the user's preference with a maximum possible coverage. Moreover, the chosen preference elicitation technique – namely inferring the dimensions of the user's preferences in a fully audio content-based manner – provides the system with the flexibility to overcome the so-called cold-start problem, which audio content-unaware systems are typically faced with (see Section 1). It also guarantees recommendations of non-popular items, which may be preferred by specialized or long-tail seeking listeners. Finally, having semantic descriptions for both the user and the recommended items allows to automatically generate justifications for the recommended music (e.g., “This track was recommended because you like *jazz* music with *acoustic* instrumentation and *relaxed* mood”)., which is a highly desirable feature for a recommendation system (Tintarev & Masthoff, 2007).

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 6. 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. 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).

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.²¹ Nevertheless, such an implicit inference of a preference set can lead to noisy representations or to the lack of coverage of all possible facets of the user’s musical preferences (see also Section 1).

8. Conclusions

In the present work we considered audio content-based user modeling approaches suitable for music recommendation and visualization of musical preferences. We proposed a novel technique for preference elicitation, which operates on an explicitly given set of music tracks defined by a user as evidence of her/his musical preferences and builds a user model by automatically extracting a semantic description of the audio content for each track in the set. To demonstrate the effectiveness of the proposed technique we considered (and evaluated) two applications: music recommendation and visualization of musical preferences. The results obtained from the subjective evaluations, conducted on 12 subjects, are promising.

In the case of music recommendation, we demonstrated that the approaches based on the proposed semantic model, inferred from low-level timbral, temporal, and tonal features, outperform state-of-the-art audio content-based algorithms exploiting only low-level timbral features. Although these approaches perform worse than the considered commercial black-box system, which exploit collaborative filtering, the difference in performance is greatly diminished when using the semantic descriptors computed in our model. It is important to notice that one of the main advantages of our model is the fact that it does not suffer from the long-tail and cold-start problems, which are inherent to collaborative filtering approaches.

In the case of musical preferences visualization, we presented an approach to automatically create a visual avatar of the user, capturing their musical preferences, from the proposed user model. To the best of our knowledge, such a task has not been previously explored in the literature, and we have developed an appropriate procedure and an evaluation methodology. The subjective evaluation showed that the provided visualization is able to reliably depict musical preferences, albeit in a coarse way.

In addition to the demonstrated applications, we also described a number of possible enhancements of music recommender systems based on the proposed user model. Specifically, we discuss justification of recommendations, interactive interfaces based on visual clues, playlist description and visualization, tracking the evolution of a user’s musical preferences, and social applications.

As future work, we plan to focus our research on performance improvements, enriching the current model with more semantic descriptors (e.g., instrument information), and improving the accuracy of the underlying classifiers. We also plan to expand the present prototypical study and conduct a large scale Web-based user evaluation in order to better assess the representativeness of the obtained user models for their further refinement. In particular, as the proposed technique requires some effort from the user to gather preference examples, a comparison with implicit methods to obtain information about preferences would be of interest.

Acknowledgements

The authors thank all participants involved in the evaluation and Justin Salamon for proofreading. This research has been partially funded by the FI Grant of Generalitat de Catalunya (AGAUR) and the Buscamedia (CEN-20091026), Classical Planet (TSI-070100-2009-407, MITYC), DRIMS (TIN2009-14247-C02-01, MICINN), and MIRES (EC-FP7 ICT-2011.1.5 Networked Media and Search Systems, grant agreement No. 287711) Projects.

²¹ 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 <http://mtg.upf.edu/project/musicalavatar>.

References

- Abdullah, M. B. (1990). On a robust correlation coefficient. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 39(4), 455–460.
- Ahmed, N., de Aguiar, E., Theobalt, C., Magnor, M., & Seidel, H. (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).
- 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.
- 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.
- Baltrunas, L., & Amatriain, X. (2009). Towards time-dependant recommendation based on implicit feedback. In *Workshop on context-aware recommender systems (CARS'09)*.
- 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. (2007). Semantic similarity for music retrieval. In *Music information retrieval evaluation exchange (MIREX'07)*. <http://www.music-ir.org/mirex/abstracts/2007/AS_barrington.pdf>.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Bogdanov, D., Haro, M., Fuhrmann, F., Gómez, E., & Herrera, P. (2010). Content-based music recommendation based on user preference examples. In *ACM conf. on recommender systems. Workshop on music recommendation and discovery (Womrad 2010)*.
- Bogdanov, D., Serrà, J., Wack, N., & Herrera, P. (2009). From low-level to high-level: Comparative study of music similarity measures. In *IEEE int. symp. on multimedia (ISM'09)*. *Int. workshop on advances in music information research (AdMIRE'09)* (pp. 453–458).
- 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.
- Brossier, P.M. (2007). *Automatic annotation of musical audio for interactive applications*. Ph.D. thesis, QMUL, London, UK.
- Cano, P., Gómez, E., Gouyon, F., Herrera, P., Koppenberger, M., Ong, B. et al. (2006). *ISMIR 2004 audio description contest*. Tech. rep. <<http://mtg.upf.edu/node/461>>.
- 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., & 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*. <<http://mtg.upf.edu/node/874>>.
- 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.
- Chang, C., & Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 27:1–27:27.
- 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., et al (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.
- 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).
- Firan, C. S., Nejdil, W., & Paiu, R. (2007). The benefit of using tag-based profiles. In *Latin American web conf.* (pp. 32–41).
- 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* (3rd ed.). Oxford University Press.
- Gómez, E. (2006). *Tonal description of music audio signals*. Ph.D. thesis, UPF, Barcelona, Spain.
- 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(3), 140–156.
- 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.
- 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. In *Audio Mostly (AM'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).
- 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).
- 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)*. *HetRec'10* (pp. 47–51). New York, NY, USA: ACM. ACM ID: 1869453.
- Jones, N., & Pu, P. (2007). User technology adoption issues in recommender systems. In *Networking and electronic commerce research conf.*
- Laurier, C., Meyers, O., Serrà, J., Blech, M., & Herrera, P. (2009). Music mood annotator design and integration. In *Int. workshop on content-based multimedia indexing (CBMI'2009)*.
- 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)*.
- 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 and Management*, 43(2), 473–487.
- Logan, B. (2004). Music recommendation from song sets. In *Int. conf. on music information retrieval (ISMIR'04)* (pp. 425–428).
- Lu, 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.
- Maltz, D., & Ehrlich, K. (1995). Pointing the way: active collaborative filtering. In *SIGCHI conf. on human factors in computing systems (CHI'95)* (pp. 202–209).
- 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).
- McCloud, S. (2009). *Understanding comics: The invisible art* (36th ed.). HarperPerennial.
- Nanopoulos, A., Rafailidis, D., Ruxanda, M., & Manolopoulos, Y. (2009). Music search engines: Specifications and challenges. *Information Processing and Management*, 45(3), 392–396.

- Pampalk, E. (2006). *Computational models of music similarity and their application in music information retrieval*. Ph.D. thesis, Vienna University of Technology.
- Peeters, G. (2004). *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. CUIDADO Project Report.
- Petajan, E. (2005). Face and body animation coding applied to HCI. In *Real-time vision for human-computer interaction*. US: Springer.
- 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). Cambridge, MA: MIT Press.
- 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).
- Reas, C., & Fry, B. (2007). *Processing: A programming handbook for visual designers and artists*. MIT Press.
- Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620.
- 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. (2009). Music-driven character animation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 5(4), 1–16.
- Schedl, M., Widmer, G., Knees, P., & Pohle, T. (2011). A music information system automatically generated via web content mining techniques. *Information Processing and Management*, 47(3), 426–439.
- 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* (pp. 210–217).
- Sotiropoulos, D. N., Lampropoulos, A. S., & Tsihrintzis, G. A. (2007). MUSIPER: A system for modeling music similarity perception based on objective feature subset selection. *User Modeling and User-Adapted Interaction*, 18(4), 315–348.
- Streich, S. (2007). *Music complexity: a multi-faceted description of audio content*. Ph.D. thesis, UPF, Barcelona, Spain.
- Su, J. H., Yeh, H. H., & Tseng, V. S. (2010). A novel music recommender by discovering preferable perceptual-patterns from music pieces. In *ACM symp. on applied computing (SAC'10)* (pp. 1924–1928).
- Tintarev, N., & Masthoff, J. (2007). Effective explanations of recommendations: user-centered design. In *Proceedings of the 2007 ACM conference on recommender systems* (pp. 153–156).
- Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), 293–302.
- Vignoli, F., & Pauws, S. (2005). A music retrieval system based on user-driven similarity and its evaluation. In *Int. conf. on music information retrieval (ISMIR'05)* (pp. 272–279).
- Voong, M., & Beale, R. (2007). Music organisation using colour synaesthesia. In *CHI'07 extended abstracts on Human Factors in Computing Systems* (pp. 1869–1874).
- West, K., & Lamere, P. (2007). A model-based approach to constructing music similarity functions. *EURASIP Journal on Advances in Signal Processing*, 2007, 149.
- 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)*.