How Potential Users of Music Search and Retrieval Systems Describe the Semantic Quality of Music

Micheline Lesaffre, Liesbeth De Voogdt and Marc Leman
IPEM, Department of Musicology, Ghent University, Blandijnberg 2, B-9000 Ghent, Belgium.
E-mail: {Micheline.Lesaffre, Liesbeth.Devoogdt, Marc.Leman}@UGent.be

Bernard De Baets
Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, B-9000 Ghent, Belgium. E-mail: Bernard.DeBaets@UGent.be

Hans De Meyer
Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Ghent, Belgium. E-mail: Hans.DeMeyer@UGent.be

Jean-Pierre Martens
Department of Electronics and Information Systems (ELIS), Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium. E-mail: Jean-Pierre.Martens@elis.UGent.be

A large-scale study was set up aiming at the clarification of the influence of demographic and musical background on the semantic description of music. Our model for rating high-level music qualities distinguishes between affective/emotive, structural and kinaesthetic descriptors. The focus was on the understanding of the most important attributes of music in view of the development of efficient search and retrieval systems. We emphasized who the users of such systems are and how they describe their favorite music. Particular interest went to inter-subjective similarities among listeners. The results from our study suggest that gender, age, musical expertise, active musicianship, broadness of taste and familiarity with the music have an influence on the semantic description of music.

Introduction

This article investigates how potential users of digital audio-libraries describe the semantic quality of music. Within this frame, this article asks the following questions: What are the effects of demographic and musical background? How is the semantic description of music structured? These questions are important for the development of future semantic-based music search and retrieval systems.

Received April 11, 2006; revised April 2, 2007; accepted April 17, 2007

 DOI: 10.1002/asi.20731

To retrieve music from a digital audio-library, users have to express a search intention. Current technology allows only meta-data descriptions, such as title, name, and year. Future technologies however may also allow content-based descriptions (Leman, 2002a; Leman et al., 2002b). For example, users could employ previously recorded audio fragments as search examples. They could further specify which audio cues are relevant for the search, for instance, that the melody should be similar to that of a given audio fragment or that color, mood, or emotional flavor should be the same. To indicate these intentions, they could rely on graphical navigation tools or just describe them in words. All these technologies call for an extension of the traditional meta-data description to a more advanced content description of music.

An important subset of such a content description is the semantic description. This is a verbal description of intrinsic musical qualities or, more specifically, what *intrinsic qualities* mean to a user. Semantic descriptions are sometimes associated with cultural, historical, or other nonmusical significations. However, these so-called *extra-musical* meanings are not taken into account in the present article nor are any other types of nonlinguistic or gesture-based descriptions, such as tapping, singing, or moving.

There are several reasons to believe that descriptions of intrinsic musical qualities form a self-contained description

category for music search and retrieval. First of all, they have a linguistic aspect and are meant to function in a social context. This requires that they are effective in communication as well as explanatory in terms of their particular intention, revealing what is perceived and experienced while avoiding cultural interpretations or associations. Second, semantic descriptions are high-level in that they rely on a cognitive assessment of the sensory and corporeal experiences induced by musical audio. Yet, for this reason, they can be elusive and influenced by a number of subjective factors that introduce uncertainty into the communication pattern. Finally, in the context of music search and retrieval systems, semantic descriptions are used to mediate between the user's verbally described search intention and the audio contained in a music library. Machines need to be able to deconstruct their meaning into formal entities that the machine can deal with. This process may not be a concern of the user, but it is nevertheless of interest when considering semantic descriptions. Deconstruction of semantic descriptors is of central importance in building semantic-based music search and retrieval systems capable of dealing with large music collections. It is safe to assume that the particular linguistic nature of semantic descriptions, their function in social context, and their link with private experience and the musical audio stream put severe constraints on building music search and retrieval systems.

Despite their appeal as a description category, there is still a lack of knowledge about the degree to which users can successfully deal with semantic descriptions. The regularities of the relations between semantic description and subjective background are neither clearly understood nor easy to identify. Music consists of many aspects that interact at multiple levels of description. Descriptors may focus on cues that relate to physical or sensory features (such as articulation, roughness) or they may focus on higher level properties whose semantics may range from structural to synaesthetic/kinaesthetic to affective/emotive qualities (Leman, 2007; Leman et al. 2004; Lesaffre, 2005). Many functions of music are connected to social and psychological factors (Hargreaves & North, 1997; Huron & Aarden, 2002), so that it is not easy to select or define the population of subjects that should be taken into account. Many studies tend to rely on a population of university students or on some other population whose representativeness is not identified. Consequently, the regularities found may be biased. It is very likely that user groups such as musical novices and musicians or adolescents and middle-aged people may understand music in different ways and that there may therefore be a strong effect of subjective background on how music is described. For this reason, the present study focuses on intersubjective similarities among users, taking into account the proper music environment and description context.

This study addresses the relationships between semantic descriptions and user backgrounds. In particular, it is intended as a means of exploring the following questions: What is a representative sample of the people that are likely to use future content-based music search and retrieval systems? What is a representative sample of the music that will need to be described in this context? To what extents are listeners consistent in their use of semantic descriptions and to what extent do they rely on an intersubjective semantics? What type of descriptors may work better than other descriptors? Finally, how do all the above relate to factors such as musical education, genre preferences, taste, and familiarity with music?

This article comprises four parts. The first part gives a critical overview of related work, showing that this work is sparse. The second part describes a large-scale survey and explorative statistical analysis of demographic and musical backgrounds in order to define the population of potential users of music audio-libraries. The third part focuses on an experiment in which a representative subset of the population described pieces music by means of semantic descriptors. It is shown that demographic and musical background affect these descriptions. In the final part, the results are discussed and conclusions are drawn with a view to future investigations.

Related Work

User Studies

To the best of our knowledge, no large-scale studies investigating the effects of demographic and musical background on semantic description, or at least none using a population that is representative for the music information retrieval context and musical stimuli that are representative of its music consumption pattern have yet been published. Kim and Belkin (2002) report an experiment in which 11 students describe seven pieces of classical music, but the small number of subjects and the nature of the stimuli do not allow generalization. A more rigorous approach is needed to categorize the perception of musical features.

Studies that focus on the social use of music (e.g., Hargreaves & North, 1997; Williams, 2001) have the potential to reveal important insights into demographic and musical backgrounds. However, so far such studies have not addressed the potential users of music information retrieval systems (Droetboom, 2004). Obviously, the social context of semantic-based music information retrieval is not yet fully functional and thus it does not yet form part of the current social context. However, in view of the development of music information retrieval test beds, Cunningham (2002) emphasizes the need to identify the potential users of a digital music library. It is indeed likely that the potential users form a varied population open to new technology. The Human Use of Music Information Retrieval Systems (HUMIRS) project (Downie, 2004) aims at collecting and analyzing data concerning real-world users. Thus far, however, results have been reported only for a population of users from a university community (Lee & Downie, 2004).

Semantic Description

In the context of music search and retrieval, the use of semantic descriptors has been explored through both linking

DOI: 10.1002/asi

and annotation approaches. *Linking approaches* aim at collecting users' music ratings for recommendation applications. Kalbach (2002) praises these approaches for being innovative in that they are based on a large population of users dedicated to search and retrieval of music. However, the semantic description often relies on an ad hoc taxonomy that is not grounded in empirical studies (e.g., MoodLogic). *Annotation approaches* collect the user's music ratings in pursuit of system evaluations and algorithm testing (Lesaffre, Leman, De Baets, & Martens, 2004; Tzanetakis & Cook, 2000; Yang & Lee, 2004). Unfortunately, despite requests for more input from psychologists and musicologists (Futrelle & Downie 2002), most such studies provide scarce reference material regarding how ratings were obtained and the representativeness of the population of users. In general, music annotation is often hindered by a lack of representative audio databases and studies often tend to underestimate the importance of a representative set of music.

A number of studies have explored the relationship between music and emotion, more specifically between descriptions of musical structure and descriptions of emotional appraisal (Gabrielson & Juslin, 2003; Gabrielsson & Lindström, 2001; Juslin & Laukka, 2003; Juslin & Sloboda, 2001). The latter form an important subcategory of the category of semantic descriptions. Most studies reveal that semantic/emotive descriptors rely on a number of subjective factors. However, they are not related to music information retrieval and for that reason suffer from a lack of representative population and music. Leman, Vermeulen, De Voogdt, Moelants, and Lesaffre (2005) investigated the hypothesis that the intersubjective basis of semantic descriptors can be predicted by a combination of acoustic cues. This study suggests that valence and activity factors can be better predicted by these means than descriptors that refer to interest factors but that both automated extraction of acoustic cues and statistical approaches remain in need of improvement. It calls for more large-scale studies that, using a representative population of subjects, address the relationship between semantic descriptions and respondent backgrounds.

The present study expands on Leman et al. (2004, 2005), in terms of both scale and approach. The selection of musical excerpts has been improved, the number of excerpts has been expanded, and the population is more representative. Unlike previous research in which the selected stimuli were assumed to be unknown to the subjects, this study works with participants who have a high degree of familiarity with the music they are requested to annotate. This was assumed to be a more realistic and reliable approach, especially when users with little or no musical background are involved.

To summarize, from a literature review and previous experiments conducted in-house, it is clear that in music information retrieval research, opinion and attitude questions are the most challenging and still need a great deal of investigation. So far, there exist no annotated music databases to support such investigation. User dependency with regard to semantic description of music has been acknowledged but also largely neglected, as most experimental research recruited subjects from selected populations of university students and relies on scores obtained for a limited musical database.

User Study

The study presented here contains (a) a survey of the potential users of interactive music systems, their demographic and musical background, as well as their favorite music and (b) an annotation experiment involving a representative set of respondents. A self-administering Web-based questionnaire was used. This generated the main dataset, which contains a list of respondents and information about their background, including the titles of their favorite music. From this main dataset, subjects and musical excerpts were selected for the annotation experiment. This experiment provided the annotation dataset that contains quality ratings of semantic descriptions of the selected music excerpts. All data are incorporated in a relational database.

Survey on Demographic and Musical Background

Aim and approach. The survey aimed at identifying potential users of content-based search and retrieval of music. Our recruitment strategy guaranteed that a valid cross section of potential end users of music information retrieval systems was attracted. The survey was announced on mailing lists and postings to music and ICT newsgroups. In addition, a media campaign was launched and interviews with the researchers about the goals of the survey were broadcast on the radio and published in newspapers.

Survey design and procedure. Respondents could fill in the questionnaire—on the Internet at home or in a booth installed at the music library of the city of Ghent—any time from April 2004 to January 2005. Completing the questionnaire took about 1 hour on average. An introductory text explained the goal of the study and guaranteed privacy.

First, information was gathered about global background, then the musical background was assessed. The questionnaire began with a number of questions addressing respondents' sociodemographic and cultural backgrounds, and their familiarity with the Internet. Respondents were then asked about their music education background, their music level and ability to play an instrument, and if so, which type of instrument they played. They also had to report on their singing and dancing skills, the way they interact with music, and their preferred medium for listening to music. Music genre preferences were addressed in multiple ways: (a) participants were asked which *genres* they listen to; (b) their personal evolution of musical taste was checked; (c) they were requested to provide up to 10 titles of their favorite music. The genre items on the questionnaire used a genre taxonomy that covered 14 major categories with which most subjects were assumed to be familiar, such as classical, world, jazz, and rock. Participants could make multiple choices from a checklist of all 14 genre categories. Additionally, they were offered the possibility of naming a genre that was not in the list. With the aim

DOI: 10.1002/asi

of generating a music audio database that reflected both genre and title preferences, an open-ended section was included at the end of the questionnaire.

There follows a brief summary of the statistics produced by this procedure (for full detailed tables please visit www.IPEM.UGENT.be/staff/Mich/PhDMLesaffre). The focus is especially on participants' music background and on the results relevant to semantic description.

Background. 774 people responded to the questionnaire. Statistical analysis was performed on the 663 (86%) who provided enough data for relevancy. The age of respondents ranged from 15 to 75 years with a mean of 29.4. Males (55.8%) slightly outnumbered females (44.2%). As could be expected from a set-up targeting a population open to new technology, 92.6% of the respondents claimed familiarity with the Internet, and 50% had Internet access at home. They spent an average of 9.6 hours a week using the Internet, and 3.1 hours a week on activities related to music. Almost two-thirds (63.2%) of the subjects had had some form of musical education (i.e., self-study, private education, music school, conservatory or university); half (50.7%) of the subjects claimed to play one or more instruments. Our questions allowed us to make a distinction between more active listeners (32%), active-and-passive listeners (43.3%), and more passive listeners (24.7%). Active listening was defined as conscious listening, whereas passive listening was defined as listening without really paying attention. Among instruments, keyboards were the most popular instrument group (25.6%) followed by plucked strings (21.1%) and woodwind (19%). Respondents evaluated themselves as better dancers than singers and nearly all (94.9%) said they spontaneously move along with the music they hear. Concerning the musical medium they use, subjects could choose between CD/minidisc, radio, television, and the Internet.[1] The most preferred medium was the CD (62.9%) and the second most popular medium was the radio (47%).

Genre preferences and musical taste. The 663 participants made a total of 3,096 genre selections. This represents an average of 4.7 selections from a checklist of 14 predetermined classes. These were classical, world/ethnic, light/oldies, pop, rock/metal/noise, jazz, blues/soul/reggae, folk/country, rap/hip-hop, new age, dance/house/techno, children's songs, and "other". This genre taxonomy is based on a preliminary study of genre classes as provided by the International Federation of Phonographic Industry and All Music Guide in which the top-ranked genres are pop (13.0%), classical (12.0%), and rock (10.8%).

In the survey, musical taste was tested by linking the genre set (as used in the genre questionnaire) with six age categories. According to the sum of counts, it was found that pop (19%), classical (12.9%), and rock/metal/noise (11.2%) are the most popular genres. The importance of classical music is remarkably high and increases with age. Most extremes appear in the youngest age category (1–12) in which popular music gets the highest score (29.5%), followed by children's music (28.7%). The interest in pop music remains rather stable between 18 and 35 years and then gradually decreases. Interest in rock/metal/noise music is most prominent around the age of 12–18, after which it slowly decreases.

Favorite titles. 523 of 663 (79%) participants provided from 1 up to 10 titles of their favorite music together with the name of the composer or performer of each piece. Regarding favorites, the questionnaire provided a list of 3,021 titles with presumed long-term familiarity to the participants. These titles were taken as a starting point for building up the musical database of the annotation experiment (see infra). Participants were also asked to describe the genres to which they think their chosen pieces belong. Their genre descriptions were subsequently labeled by music experts. A clear preference for pop and rock styles was found, but classical music is strongly represented as well. Moreover, the favorites represented higher percentages of pop (20.4%), rock (23.5%), and classical (17.7%) than one would have expected from the genre preferences. This could be explained by the fact that for the latter, participants made multiple choices of genres they listen to in general, whereas for favorites, they provided titles of music they listen to very often. This picture suggests that the music types of most interest to a music information retrieval system user would be pop, rock, and classical music.

Relationships. Relationships between binary distinctions within the variables of age, age category, gender, musical education, musical expertise, and breadth of taste were investigated. For each variable, a two-way contingency table analysis was performed to evaluate statistically significant relationships. A chi-square statistic guaranteed that the differences between the observed and expected values were sufficiently different. Multiple significant relationships were found between these categorical variables. It is, for example, likely that the potential users of interactive music systems have the following characteristics: (a) among those older than 25, two out of three (63%) are men; (b) of those who do not listen to classical music, two out of three (66%) are younger than 25; (c) two-thirds of the women are musical novices (65%); (d) of classical music listeners, more than two-thirds (70%) are musical experts; (d) most (95%) people who play an instrument are musically educated; (e) of those who listen to classical music, two out of three (63%) play an instrument; and (f) of those who do not listen to classical music, two out of three (63%) have a narrow taste range.

Conclusion. With 774 participants representing a broad distribution of music lovers and technology-minded people, we have reached a sample size that is large enough to permit quantitative and qualitative deductions. Although participants were not asked if they might use music information retrieval systems, we developed a recruitment strategy that

[1] At the time of this study, the iPod version for Windows was newly available and had not yet reached many users.

DOI: 10.1002/asi

made sure we attracted the targeted population. Among potential users of music information retrieval systems, it was found likely that music plays an active role in their lives. According to the findings in the survey, a global profile of envisaged users can be outlined. The average music information retrieval system users are likely to have the following characteristics: (a) be young people (three-quarters of them younger than 35 years); (b) use the Internet regularly (92.6%); (c) spend one-third of their Internet time on music-related activities; (d) be actively involved with music (two out of three in our survey had had a musical education and one out of five had a high-level of musical expertise); (e) have the broadest musical taste in the 12–25 age category; (f) have pop, rock, and classical music as preferred genres; (g) be good at genre description; (h) have difficulties assigning qualities to classical music; and (i) assign most variability to classical music.

Annotation Experiment on Semantic Description

The annotation experiment aimed at finding out how potential users of music information retrieval systems would describe their search intention using semantic descriptions of music. It was an in-depth study of feelings, judgments, and appraisal by a selected target group. The focus was on unveiling relationships that could support linking between musical structure and musical expressiveness, taking into account intersubjective similarities and subjective differences. The description of perceived musical qualities involved five aspects: (a) the attribution of appraisal and interest (e.g., affective/emotive descriptors), (b) the description of structural features (e.g., structure descriptors), (c) the description of involved activity (e.g., kinaesthetic descriptors), (d) the impact of memory, and (e) personal judgment.

Subjects and stimuli. Invitations were sent to the 490 participants who at a certain point in time had completed the questionnaire. Of the 115 who responded, 94 finally participated in the experiment. All subjects that participated in the annotation experiment were paid.

From 3,021 favorite titles provided by the participants of the survey (see the favorite titles section), 160 were selected for the experiment's music database. This database was divided into three categories, namely *classical*, *pop* (with a balanced distribution of pop music, rock/metal/noise, electronic music, dance/house/techno, rap/hip hop, soul/reggae, light music, new age, and soundtracks), and *roots* (jazz, blues, world/ethnic and folk/country). The distribution of titles within these three categories reflected the distribution of the related classes in the genre questionnaire of the survey (see the genre preferences and musical taste section). We assumed that subjects' familiarity with particular pieces of music would increase their competence in semantic annotation. Therefore, only titles with high popularity, as measured by their repeated occurrence in the survey, were retained for selection. For each title that was selected, one excerpt of exactly 30 seconds in length was recorded. The selection of the particular excerpt was made on the basis of musical homogeneity and attention to a natural beginning and ending. The 160 excerpts were stored in 16 bit 44100 kHz PCM wav format. A table with references to all the selected musical excerpts is available at http://www.ipem.ugent.be/mami/public/MuziekVoorMAMI/MAMIexp2Refs.xls.

Design. A music information retrieval system that offers users the opportunity to formulate a query by using a set of affective/emotive, structural, and kinaesthetic descriptors (Leman, 2007) that would allow the retrieval of intended music does not yet exist. Therefore, we could not make use of existing systems. Web-based forms were used to gather subjects' ratings of a set of semantic adjectives that describe expressive and structural qualities of music. For each rated musical excerpt, subjects were also asked to give additional information on how familiar they were with it, their overall judgment (beautiful/awful), how complicated they found it (easy/difficult), and their physical response (move along, sing along, etc.). In Table 1, a global overview is given of the model for rating high-level qualities of music.[2] To avoid ambiguity, questions that assess a qualitative measure were phrased carefully, for example, "What feeling goes out of this music?" In the response form, adjectives were not offered as single words but in predications such as "This music is cheerful."

Affective/emotive descriptors were subdivided into two sets of adjectives related to *appraisal* and *interest* (Leman et al., 2005; Scherer & Zentner, 2001). *Appraisal* assumes a cognitive evaluation of the affective/emotive value of the music, that is, whether its affective value can be communicated in a social context. Ratings were made on a five-point scale (ranked from *not* to *very*) for a set of eight adjectives that appeared in random order on a list. The list of adjectives contained potentially bipolar terms, but a unipolar scale was used to avoid weak choices. By offering the possibility of "no opinion" on a particular affect, people were not forced to make a rating. Additionally, respondents had to indicate which of the eight attributes they found the most important. *Interest* focused on the subjective state in response to music, that is, whether the music was of interest to the subject. Participants rated four adjectives presented in the same way as for *appraisal*.

The section dedicated to structural description had a focus on *sonic* properties and *pattern-related* properties. Sonic properties were rated on a nine-point scale, a bipolar version of the five-point scale used for *appraisal* and *interest* for it is obvious that when music is soft, it is not hard.[3] Sonic properties include adjectives that relate to the perception of tempo, loudness, timbre, and dynamics. Concerning

[2] The experiment was conducted in Dutch. In this paper, the adjectives have been translated into English. We tried to find the correct English adjectives whilst staying reasonably faithful to the Dutch model.

[3] For affective/emotive features, on the contrary, it is difficult to maintain bipolarity. One can easily imagine a piece of music that at the same time expresses some degree of cheerfulness and some degree of sadness.

DOI: 10.1002/asi

TABLE 1. Model for semantic description of music used in the annotation experiment.

Semantic Descriptors				
Affective/Emotive		Structural		Kinaesthetic
Appraisal	Cheerful	**Sonic**	Soft/hard	Gesture
	Sad		Clear/dull	Melody imitation
	Carefree		Rough/harmonious	
	Anxious		Void/compact	Memory
	Tender		Slow/quick	No recognition
	Aggressive		Flowing/stuttering	Style recognition
	Passionate		Dynamic/static	Vaguely known
	Restrained			Well known
Interest	Annoying	**Pattern**	Timbre	Judgement
	Pleasing		Rhythm	Beautiful/awful
	Touching		Melody	Difficult/easy
	Indifferent		None	

pattern-related properties, multiple choices had to be made between timbre, rhythm, and melody.

The kinaesthetic description probed the extent of subjects' *bodily action* as a response to the experience of music. This aspect was limited to two phrases: "With this music I start to move spontaneously" and "I could imitate the melody." Both phrases were rated on a five-point scale. The familiarity with the excerpts was checked in the part dedicated to memory. A forced choice had to be made between four options: (a) not knowing either the piece of music or the genre, (b) not knowing the piece of music but being familiar with the genre, (c) having heard the piece before, and (d) having heard the piece a lot. Finally, in judgment, subjects had to indicate whether they personally liked or disliked the music. The bipolar adjectives *beautiful-awful* and *difficult-easy* were rated on a nine-point scale.

Procedure. The annotation experiment took place over four sessions, each covering 40 excerpts of music. As each fragment had to be annotated for all the features in our model for semantic description, 40 excerpts per session were considered the maximum. The experiment was conducted in groups of maximum 10 participants, who performed the test under guidance. The sessions took place in a computer classroom where the subjects sat in front of a PC while the music was played through headphones. The order of the music excerpts was randomized. Only the number of the excerpt was presented on the screen.

The experimenter explained the aims and procedure of the study, using examples that clarified the meaning of the semantic descriptors. After that, participants received written guidelines and were asked to fill out seven forms for each fragment (that is, 280 forms per session). They rated each music excerpt using the phrased semantic descriptors. While giving their ratings, the participants could listen to the music excerpts as often as they wanted. Each session took an average of about 3 hours. Although the task was demanding, most participants enjoyed doing the tests and some of them even asked for more sessions than the four they had already done. Out of 94 participants, 79 (84 %) judged the whole set of 160 musical fragments, one person stopped after three sessions, three after two sessions, and eleven after one session.

Statistics. Statistical procedures were performed to examine the means and standard deviations of the bipolar pairs as well as to check for possible outliers or entry errors. No outliers were found. The annotations of the two persons who did not succeed in finishing even one set were omitted. Although they were sufficiently motivated, the task seemed too complicated for them. Analysis was conducted on the 12,640 responses generated by the 79 subjects who judged all 160 fragments.

First, a comparison was made between the survey population ($N = 663$) and the population of the annotation experiment ($N = 79$). Recall that the latter group of participants (the sample) formed a subset of the former group (the whole group). The aim was to investigate whether the outcome of the sample was typical of the whole group. Then, semantic descriptions of music were related to demographic and musical background information (see the Survey on demographic and musical background section), taking into account the effect of familiarity. Factor analysis was run in order to investigate underlying dimensions of the cognitive assessments. Standard deviations were used as a measure of unanimity among subjects. Finally, correlation analysis was conducted in order to establish relationships among semantic descriptor groups.

Results

Comparison of populations. Comparison between the demographic and musical background of the participants in the survey and those who took part in the experiment showed that population and sample were quite similar. Frequencies for downloading music, being musical educated, being an amateur, and being an active listener, for example, were approximately equal (less than 2% difference). The general lack of differences between the two populations supports the hypothesis that this was a representative sample of the targeted population. The only noticeable difference was that there was

DOI: 10.1002/asi

a better spread of age categories in the subset, with a smaller group younger than 35 (58% instead of 75%).

Influence of subject-related factors. Using the profile information of the 79 subjects, the data were divided into six binary categories as follows: (a) gender, (b) musical expertise (based on answers about music education and music skills), (c) age category, (d) breadth of taste, (e) familiarity with classical music, and (f) active musicianship. For each of these six categories, nonparametric Mann-Whitney tests were performed on data of each of the judged adjectives and adjective pairs describing the *affective/emotive*, *structural*, and *kinaesthetic* qualities of the music examples. Calculations were performed with summed ratings of variables per participant (79 cases). Split up for the different parameters, an overview of the statistically significant effects is included in Table 2. Although they are not numerous (of the 164 tests, 18 appear as significant), some effects of subject-related factors revealed by the Mann-Whitney tests warrant discussion.

In view of the fact that the largest number of significant values (5) was found for the category *gender*, it is likely that gender has a significant influence on the perception of music. Men rated the musical excerpts significantly more restrained, more harmonious, and more static than women did. The latter judged them more beautiful and more difficult than their male counterparts. For *age categories*, some significant values were found as well. These showed that the age of listeners is a contributing factor with regard to differences in perception of

TABLE 2. Summary of Mann-Whitney test results for six participant categories and adjectives and adjective pairs describing several measured structural and more subjective qualities of the 160 musical excerpts.

Affective/Emotive descriptors: Appraisal

Categories	Cheerful	Sad	Carefree	Anxious	Tender	Aggressive	Passionate	Restrained
(a) Gender								−1,97*
(b) Expertise	−2,06*						−2,2*	
(c) Age							−2,33*	
(f) Musician	−2,22*							
(g) Familiarity	−9***	−5,81***	−3,70***	−5,11***	−18,45***	−7,69***	−18,13***	−17,6***

Affective/Emotive descriptors: Interest

Categories	Annoying	Pleasing	Touching	Indifferent
(d) Taste		−2,32*		
(g) Familiarity	−28,71***	−36,31***	−28,15***	−28,51***

Structural descriptors: Sonic

Categories	Soft Hard	Clear Dull	Rough Harmonious	Void Compact	Slow Quick	Flowing Stuttering	Dynamic Static
(a) Gender			−2,27*				−2,16*
(b) Expertise		−2,06*					
(c) Age							−2,80**
(f) Musician		−2,62**					−1,98*
(g) Familiarity	−5,19***	−8,35***	−14,8***			−12,90***	−10,36***

Structural descriptors: Pattern

Categories	Timbre	Rhythm	Melody
(g) Familiarity	−9,75***	−7,54***	−10,47***

	Activity		Memory	Judgment	
Categories	Moving	Imitation	Memory	Beatiful Awful	Difficult Easy
(a) Gender				−2,03*	−2,87**
(b) Expertise		−3,78***			
(d) Taste				−2,25*	
(e) Classical			−2,47*		
(f) Musician		−3,87***			
(g) Familiarity	−24,46***	−36, 31***		−39,59***	−16,73***

Note. The resulting values Z are given in those cases where a high significant effect was found.
* = $p < 0,05$. ** = $p < 0,01$.*** = $p < 0,001$

DOI: 10.1002/asi

features: People older than 35 found the music more "passionate" and less "static" than younger listeners did. Listeners with no (advanced) *music education* judged the pieces as being more "cheerful," more "passionate," and more "dull" than the experts did. As could be expected, the experts had less difficulty with imitating the melodies they heard in the musical pieces. The tests also revealed that people with a *broad musical taste* judged the music to be more "pleasing" and more "beautiful" than subjects with a narrow musical taste. The tests on the influence of *musicianship* revealed almost identical results, as did the test on musical expertise. An additional significant effect is that a *musician* often considered the music as being more "static" than a nonmusician did. Finally, there is a significant influence of being a *listener to classical music* and being *familiar* with the music excerpts in the stimuli set (see next subsection).

Effect of familiarity with the musical excerpts. When the participants indicated the extent of their familiarity with the music excerpts, the most frequent answer was "I do not know this piece but I am familiar with the genre" (34.3%), followed by "I have heard this piece before" (24.8%), "I do not know this piece and I am not familiar with its genre" (21.8%), and finally "I have heard this piece a lot" (19.1%). From these percentages it can be deduced that in 78,2% of cases people were familiar with the style of the excerpts and in 43,9% they even knew the specific piece of music. These findings comply with the aim to minimize the number of unfamiliar pieces in the stimuli set.

Familiarity with the excerpts was used as an additional binary categorization factor. The music excerpts were divided into two groups (known and unknown fragments, 43.9% and 56.1%, respectively). Because there were no 100% unknown excerpts, it is obvious that working with summarized data was not an option for this test. As a consequence, all 12,640 cases (79 subjects*160 music excerpts) were retained for analysis. The results are displayed in Table 2, category g.

Familiarity with the music excerpts is highly significant for all affective/emotive descriptors. In descending order of significance, known music was perceived as being more "tender," more "passionate," less "restrained," more "cheerful," less "aggressive," less "anxious," more "sad," and more "carefree" than unknown music. Familiarity also affects how subjects feel about music: Known excerpts were judged as being more "pleasing," less "annoying," less "indifferent," and more" touching." Even for structural descriptors, except for "void-compact" and "slow-quick," there is a very significant effect of familiarity. The effect on kinaesthetic descriptors is very significant as well; more participants said that they spontaneously start to move and that they can imitate the melody of music they know. And finally, highly subjective adjectives concerning appreciation judgments (beautiful-awful) and difficulty level (easy-difficult) of the music excerpts were influenced in the same way: Known music was judged as being more "beautiful" and "easier" than unknown music.

Factor analysis of affect/emotion descriptors. A nonparametric analysis of all affective/emotive descriptors (12 adjectives) revealed that several adjectives are correlated. Highly significant correlation coefficients were found for "touching" and "pleasing" (0,631**, $p < 0,01$), "cheerful" and "carefree" (0,565**), "touching" and "tender" (0,467**),

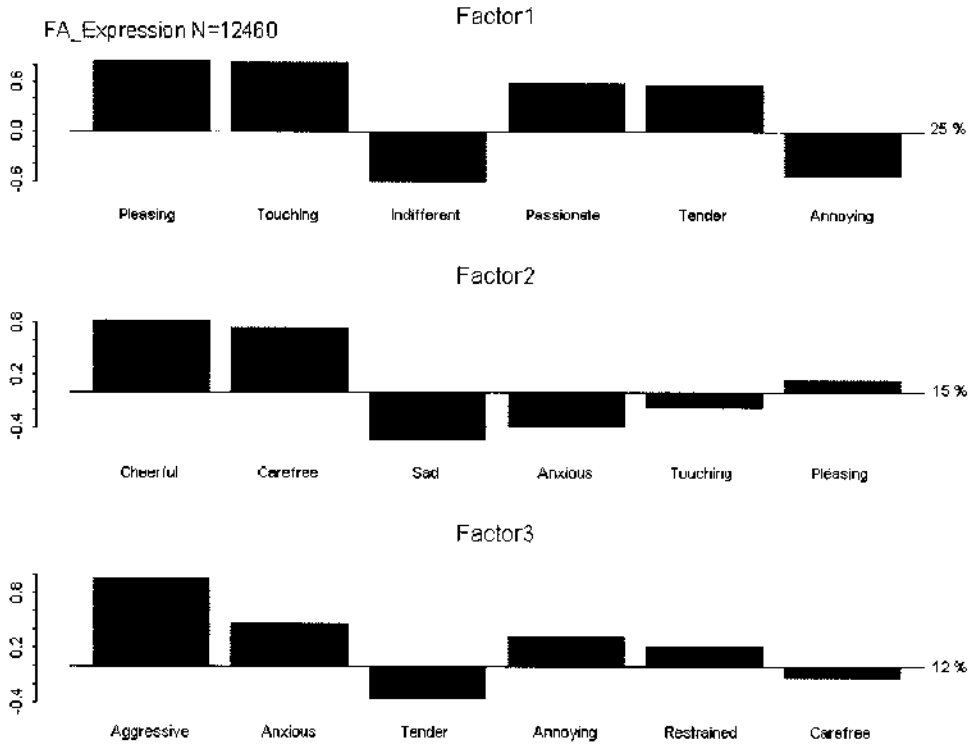

FIG. 1. Dimensions of the affective/emotive space: Three factor loadings together explain 52% of the variance in the data.

DOI: 10.1002/asi

and for "pleasing" and "annoying" (−0,476**). Factor analysis was used to better understand the nature of these correlations in terms of underlying factors. The question investigated here is whether the 12-dimensional description of the perceived affective/emotive qualities in music can be explained to a large extent in a lower dimensional space. Using maximum likelihood estimation and varimax rotation, three factors were obtained (see Figure 1), which together explain 52% of the variance in the data.

The first dimension (first factor) shows a contrast in judging the excerpts between the positive adjectives "pleasing," "touching," "passionate," and "tender" and the negative adjectives "indifferent" and "annoying." This dimension explains 25% of the total variance in the data and includes terms that express rather *high intensity experiences* of music. The second dimension (second factor) pertains to the adjectives "cheerful," "carefree," and "pleasing" (positive loadings) versus the adjectives "sad," "anxious," and "touching" (negative loadings). This dimension explains 15% of the differences in the data and accounts for the *diffuse affective state*. The third dimension (third factor) pertains to the adjectives "aggressive," "anxious," "annoying," and "restrained" (positive loadings) versus "tender" and "carefree" (negative loadings). This dimension explains 12% of the total variance. These adjectives relate to *physical involvement*.

Factor analysis of structural descriptors. A nonparametric correlation analysis of all structural descriptors (seven adjectives) revealed that several adjective pairs were correlated. Highly significant correlation coefficients were found between "soft-hard" and "void-compact" (0,459**, $p < 0,01$), "soft-hard" and "slow-quick" (0,529**), "soft-hard" and "flowing-stuttering"(0,458**), and "slow-quick" and "dynamic-static" (−0,429**). Factor analysis revealed three dimensions that can explain 60% of the variance (see Figure 2).

The first dimension accounts for "soft-hard," "void-compact," "flowing-stuttering," "bright-dull," and "slow-quick" (positive loadings) versus "rough-harmonious" (negative loading). This dimension explains 26% of the total variance. Loudness, spectral density, articulation, timbre, and tempo are located on the positive pole of the axis. The negative pole has only one adjective pair related to timbre.

The second dimension refers to "slow-quick" and "soft-hard" (positive loadings) versus "dynamic-static" (negative loading). This dimension explains 19% of the differences in the data. It juxtaposes tempo and loudness versus articulation.

The third dimension refers to "rough-harmonious" (positive loading) versus "flowing-stuttering" and "clear-dull" (negative loadings). Timbre gets a very high loading (95%) and is not clearly correlated with other features. This dimension explains 15% of the total variance.

Unanimity among subjects. To what extent do people have the same perceptions about, for example, tempo, loudness, and brightness in a particular piece of music? Moreover, do some musical features engender more unanimity in described perception than others? For the seven structural descriptors, the experiment provided enough data to investigate this issue. With 79 responses for each music excerpt, standard deviations were calculated for the 160 excerpts for the structural features. The mean standard deviation for a structural feature gives a rough idea of the objective level of unanimity among subjects

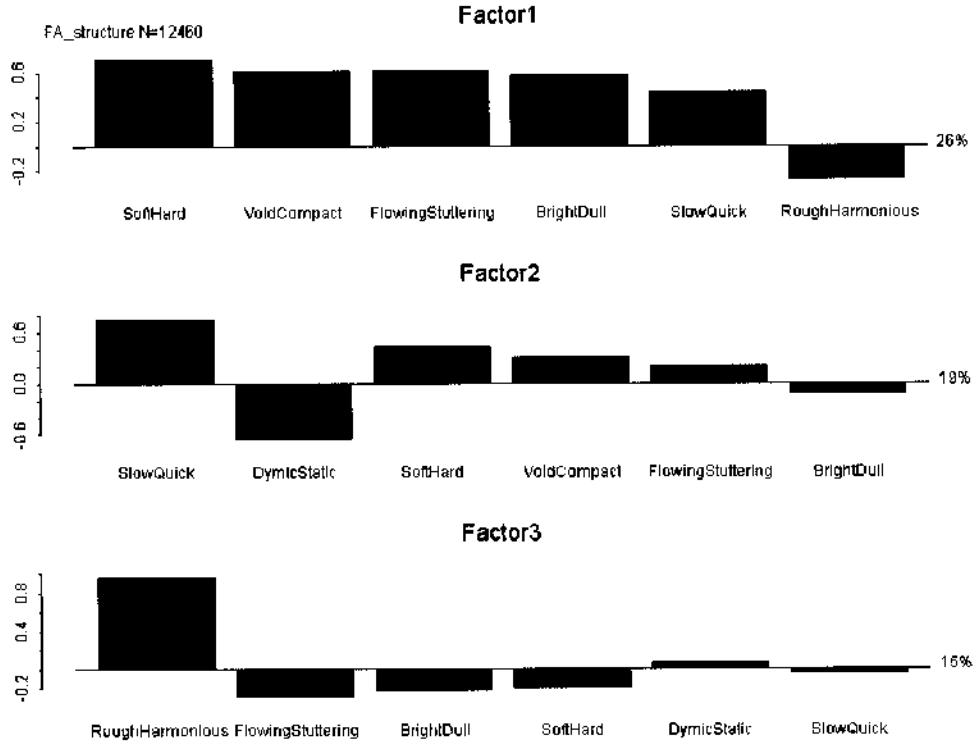

FIG. 2. Dimensions of the structure space: Three factor loadings together explain 60% of the variance in the data.

DOI: 10.1002/asi

for this feature. Brightness ("clear-dull") and roughness ("rough-harmonious") display the lowest degree of unanimity among responses (stdev. 1.9), closely followed (stdev. 1.8) by density ("void-compact"), articulation ("flowing-stuttering"), and movement ("dynamic-static"). Each with a standard deviation of 1.4, loudness ("soft-hard") and tempo ("slow-quick") are perceived with the strongest unanimity.

Relationships among semantic descriptor groups. As a useful guide for further statistical exploration of the relationship between affective/emotion descriptors and structural descriptors, two approaches were explored. In a first analysis all 12,640 (79 subjects*160 musical excerpts) scores were considered independently and in a second analysis the scores per excerpt (sum over subjects).

Given the nonparametric character of the full dataset (data consists of points on an ordinal rating scale), the chosen correlation coefficient for the first analysis method was Kendall's tau. Nonparametric correlations are summarized in Table 3.

Although most relationships between variables are highly significant ($p < 0,01$), high correlation coefficients are rare (most coefficients stay under 0.3). They were found between "tender" and "soft," the affect "tender" also correlating with "harmonious," "void," "slow," and "flowing," and between "aggressive" and "hard," the affect "aggressive" also correlating with "rough," "compact," "quick," and "stuttering." Furthermore, a correlation was found between the experience adjective "touching" and "soft". The only two relationships in the table that are not significant are those between "anxious" and "slow-quick" and between "carefree" and "flowing-stuttering".

In a second approach, mean ratings over all variables assigned to each excerpt were computed and correlated. The results of this analysis are presented in Table 4.

The affective/emotive adjectives "sad," "tender," "aggressive," "annoying," "pleasing," and "touching" correlate most strongly with almost all structural features. Other strong correlations can be found between the following adjectives: "cheerful" and "quick," "cheerful" and "dynamic," "anxious" and "dull," "anxious" and "rough," "restrained" and "dull," "restrained" and "stuttering," "indifferent" and "hard," "indifferent" and "dull," "indifferent" and "stuttering".

The affective/emotive adjective with the lowest significance is "carefree." Relationships in the table that are not significant are between the majority of affect/emotive adjectives and the structural adjective pair "dynamic-static".

To sum up, descriptors such as "tender" and "pleasing" (positive valence or pleasant) are negatively correlated with "bright-dull," "void-compact," "flowing-stuttering," and "soft-hard," and positively correlated with "rough-harmonious". These descriptors are thus correlated with "bright," "void," "flowing," "soft," and "harmonious." Descriptors such as "aggressive" and "annoying" (negative valence or unpleasant) are correlated with "dull," "compact," "stuttering," and "hard". The correlations between the adjective "indifferent" and the structural descriptors correspond with the kind of relationships found for descriptors of negative valence. These findings contradict what one would expect from a descriptor that expresses an uninterested experience. This discrepancy is due to the fact that subjects had difficulties in rating the phrase *"this music leaves me indifferent"*. Many subjects rated "not indifferent" (0) while they meant "indifferent".

TABLE 3. Relationships between affective/emotive and structural descriptors.

	Soft-hard	Clear-dull	Rough-harmonious	Void-compact	Slow-quick	Flowing-stuttering	Dynamic-static
Cheerful	.101(**)	−.136(**)	.076(**)	.133(**)	.286(**)	.049(**)	−.318(**)
	0	0	0	0	0	0	0
Sad	−.309(**)	−.048(**)	.102(**)	−.248(**)	−.366(**)	−.206(**)	.216(**)
	0	0	0	0	0	0	0
Carefree	−0.012	−.121(**)	.125(**)	.050(**)	.150(**)	−0.002	−.161(**)
	0.085	0	0	0	0	0.828	0
Anxious	.131(**)	.177(**)	−.216(**)	.046(**)	0.003	.146(**)	.031(**)
	0	0	0	0	0.695	0	0
Tender	−.507(**)	−.283(**)	.326(**)	−.347(**)	−.371(**)	−.392(**)	.116(**)
	0	0	0	0	0	0	0
Aggressive	.527(**)	.269(**)	−.298(**)	.352(**)	.354(**)	.384(**)	−.175(**)
	0	0	0	0	0	0	0
Passionate	−.087(**)	−.103(**)	.088(**)	−.043(**)	−.063(**)	−.128(**)	−.127(**)
	0	0	0	0	0	0	0
Restrained	.133(**)	.122(**)	−.148(**)	.065(**)	.055(**)	.185(**)	.093(**)
	0	0	0	0	0	0	0
Annoying	.264(**)	.214(**)	−.243(**)	.161(**)	.126(**)	.258(**)	.057(**)
	0	0	0	0	0	0	0
Pleasing	−.256(**)	−.239(**)	.257(**)	−.121(**)	−.092(**)	−.253(**)	−.120(**)
	0	0	0	0	0	0	0
Touching	−.300(**)	−.220(**)	.218(**)	−.191(**)	−.207(**)	−.286(**)	−.026(**)
	0	0	0	0	0	0	0
Indifferent	.130(**)	.156(**)	−.135(**)	.100(**)	.060(**)	.153(**)	.116(**)
	0	0	0	0	0	0	0

Note. Summary of nonparametric Kendall tau correlations ($N = 12640$) showing for each relation the correlation coefficient (first row) and significance (second row).

DOI: 10.1002/asi

TABLE 4. Relationships between affective/emotive and structural descriptors.

	Soft-hard	Clear-dull	Rough-harmonious	Void-compact	Slow-quick	Flowing-stuttering	Dynamic-static
Cheerful	.298(**)	−.196(*)	0.093	.426(**)	.621(**)	.238(**)	−.699(**)
	0	0.013	0.241	0	0	0.002	0
Sad	−.718(**)	−.331(**)	.400(**)	−.753(**)	−.838(**)	−.694(**)	.710(**)
	0	0	0	0	0	0	0
Carefree	0.076	−.266(**)	.257(**)	.219(**)	.403(**)	0.061	−.488(**)
	0.342	0.001	0.001	0.005	0	0.443	0
Anxious	.338(**)	.521(**)	−.598(**)	.173(*)	0.047	.331(**)	0.078
	0	0	0	0.029	0.553	0	0.327
Tender	−.942(**)	−.695(**)	.770(**)	−.858(**)	−.795(**)	−.885(**)	.542(**)
	0	0	0	0	0	0	0
Aggressive	.887(**)	.740(**)	−.757(**)	.766(**)	.648(**)	.776(**)	−.430(**)
	0	0	0	0	0	0	0
Passionate	−.256(**)	−.261(**)	.176(*)	−.251(**)	−.246(**)	−.364(**)	−0.056
	0.001	0.001	0.026	0.001	0.002	0	0.484
Restrained	.466(**)	.512(**)	−.497(**)	.330(**)	.297(**)	.555(**)	0.033
	0	0	0	0	0	0	0.68
Annoying	.756(**)	.699(**)	−.704(**)	.604(**)	.509(**)	.722(**)	−.215(**)
	0	0	0	0	0	0	0.006
Pleasing	−.701(**)	−.707(**)	.703(**)	−.561(**)	−.447(**)	−.708(**)	0.128
	0	0	0	0	0	0	0.108
Touching	−.737(**)	−.593(**)	.574(**)	−.696(**)	−.665(**)	−.742(**)	.407(**)
	0	0	0	0	0	0	0
Indifferent	.548(**)	.526(**)	−.413(**)	.485(**)	.413(**)	.549(**)	−0.127
	0	0	0	0	0	0	0.11

Note. Summary of Pearson's correlations ($N = 160$) showing for each relation the Pearson's coefficient (first row) and significance (second row).

The adjectives "carefree" and "passionate," on the other hand, are not, or only very slightly, correlated with the rated structural features.

Consistency Tests

Can we claim objectivity for these test results? The subjectivity of listener evaluations is a well-known problem. For quality judgments, we need to be concerned about the consistency of measurement. A study on the reliability of the annotations was done through consistency tests that were carried out over time.

Two months after the annotation experiment, the subjects were requested to participate in a follow-up experiment. A set of six music excerpts was selected by means of a randomizer, using a distribution by which each genre had an equal probability of selection: two classical, two pop, and two roots excerpts. Subjects were asked to describe these excerpts using the semantic descriptors from the annotation experiment over multiple sessions, with an interval of at least a week between each. Responses were quite satisfactory.

The data of two follow-up experiments in which 37 subjects participated, together with the corresponding data extracted from the original experiment, made it possible to compare judgments of identical musical excerpts at different points in time. 37 subjects evaluating six music excerpts each gives 222 observations per judged variable. To compare the three different samples (i.e., the original plus the two follow-up experiments) with each other, the Marginal Homogeneity Test (MHT) was used. Consistency has been checked for affective/emotive and structural descriptors.

Five of the 24 comparisons of *affective/emotive descriptors* turn out to be significant. It is remarkable that p-values lower than 0.05 all pertain to comparison with the original annotations. This is probably a result of the longer time interval between the start of the follow-up experiments and the original experiment (2 months) than between the follow-up experiments themselves (approximately a week). It appears that descriptors such as "anxious," "cheerful," "aggressive," and "passionate" can be ambiguous when measurements are repeated.

Eight out of 21 comparisons of *structural descriptors* are significantly different. As with affective/emotive descriptors, significant p-values are the outcome of a comparison with the original experiment, except for the adjective pair "clear-dull," which shows different judgments for the two follow-up experiments. The pairs "rough-harmonious" and "slow-quick" are ambiguous over both follow-up experiments. It was found in the original experiment (see unanimity section above) that there is little inter-subjective agreement for brightness ("clear-dull") and roughness ("rough-harmonious").

To summarize, to check the reliability of a semantic description, the outcome of two repeated follow-up experiments was compared with the result of the original experiments. Although there were inconsistencies, evaluations for the majority of the variables in our annotation experiment seem to be rather reliable. It was found that the ambiguous terms already showed a rather high standard deviation when consistency among subjects was checked.

DOI: 10.1002/asi

Discussion

Up to now, studies in semantic description of music have not focused on potential users of music search and retrieval systems. Typically, they are confined to student populations. In our study, we have used a population that was a representative sample of the population of subjects participating in a large-scale survey.

The present study also has a focus on the musical stimuli that reflect the musical taste of the target population. It assumes that users will formulate a search intention using music they know or music which is related in terms of a known style or mood. This is in contrast with previous studies (e.g., Juslin, 1997; Peretz, 1998; Wedin, 1972) in which stimuli sets were chosen in function of their expressive variance. Those stimuli are less valid in terms of the target population.

In the present study, the number of excerpts ($N = 160$) greatly exceeds the quantity of stimuli used in previous studies. With a few exceptions (e.g., Leman et al., 2005), the number of stimuli generally used in studies on semantic descriptors varies from 2 to 50 (Gabrielson & Juslin, 2003). Furthermore, in the present study, the main criteria for the selection of excerpts were acoustic (the homogeneous character of distinguishing music features, such as a slow tempo, a repeating rhythm pattern, a staccato performance, and absence of lyrics), and not, as in previous research, emotional expressivity.

In contrast with studies that present a simple list of terms, this study uses adjectives presented in a verbal context in order to give subjects a better understanding of how to interpret them. In addition, our annotation experiment involved two response formats, namely, (a) choices between semantic descriptors and (b) semantic ratings using bipolar and unipolar scales. As a consequence, a variety of statistical methods could be applied in data analysis.

A novel element in our investigation of relationships among semantic descriptors (e.g., between affective/emotive and structural descriptors) is that a detailed background on demography and musicality was taken into account. This approach revealed significant correlations with gender, age, music expertise, musicianship, breadth of taste, and familiarity with classical music. The results suggest that developers of music information retrieval systems should take into account the fact that the semantic description of music is strongly affected by a number of subjective factors.

Two approaches were taken to the investigation of affective/emotive descriptors, namely, a categorical approach and a dimensional approach. The categorical approach divided the descriptors into two sets, those of appraisal (e.g., cheerful, sad) and those of interest (e.g., pleasing, boring). The dimensional approach was based on previous research reported by Leman et al. (2005), which distinguishes the dimensions of "Valence," "Activity," and "Interest." For each of these three dimensions, adjective pairs were included in the present study. Although some differences were found, the dimensions reported in Leman et al. (2005) are closely related to the three factors found in the present study. The "Interest" dimension (moving, exciting, pleasing, and passionate, versus indifferent, boring, annoying, and restrained) is quite similar to the factor denoting "intense experience of music" (pleasing, touching, passionate, and tender versus indifferent and annoying). The "Valence" dimension (carefree, gay, hopeful, and positive versus anxious, sad, desperate, and negative) corresponds with the factor denoting a "diffuse affective state" (cheerful, carefree, and pleasing versus sad, anxious, and touching). The "Activity" dimension (bold, restless, and powerful versus tender, calm, and fragile) resembles the factor denoting "physical involvement" (aggressive, anxious, annoying, and restrained versus tender and carefree). Note that in the present experiment, the semantic space was reduced from fifteen to eight dimensions. Unlike Leman et al. (2005), where the subjects were university students, the present study draws on a much broader population, thus providing an interesting addition to the previous study, namely the finding that the semantics are more or less similar for different groups of users.

The finding that subjects agree on structural descriptors such as tempo and loudness confirm previous research (e.g., Juslin & Laukka, 2003 ; Scherer & Oshinsky, 1977). In the annotation experiment, slow tempo is associated with various appraisal descriptions such as "sad" and "tender" and with the interest descriptors "pleasing" and "touching." Quick tempo is associated with the appraisal descriptor "aggressive" and the interest descriptor "annoying." The findings that loud music may be determinant for the perception of aggression and the experience of annoyance and soft music for that of tenderness and a pleasing experience, is also in agreement with previous results (e.g., Baroni & Finarelli, 1994 ; Juslin, 1997).

Conclusion

The present study shows that the semantic description of music may offer an appropriate way to access music in an electronic music library. Users are able to give cognitive assessments of music in terms of linguistic-based semantic descriptors. It can be assumed that this linguistic-based semantic framework can easily be used to formulate a search intention. The study reveals that the semantic framework has an intersubjective validity, even if demographic and musical background also has an impact. This is shown for gender, age, expertise, musicianship, breadth of taste, and familiarity with a particular musical piece. The latter has the highest significant effect on all semantic descriptors. Music search and retrieval systems should therefore distinguish between different categories of users. Semantic descriptors can then be mediated by taking these categories into account.

In this study, a distinction is made between three types of semantic descriptors, namely, affective/emotive, structural, and kinaesthetic. Affective/emotive and kinaesthetic descriptors are genuine *second-person descriptors*, expressing how intrinsic qualities of music are felt. In contrast, structural semantic descriptors reflect the sonic properties of music. One could argue that these descriptors are more closely related to acoustic descriptors (Leman, 2007). Factor analysis on affective/emotive and kinaesthetic descriptors revealed three dimensions. These were labeled *high*

DOI: 10.1002/asi

intense experience, diffuse affective state, and *physical involvement*. These factors are closely related to the dimensions of "Interest," "Valence," and "Activity" found in previous research (Leman et al., 2005). With regard to unanimity among semantic descriptors, adjectives were tested that relate to loudness, timbre, tempo, and articulation. Subjects agreed most on loudness and tempo, less on timbre and articulation. Interesting relationships were found between affective/emotive and structural descriptors and a strong correlation was found between the appraisal descriptor (tender-aggressive) and the structural descriptor (soft-hard). This result suggests that it may be possible to relate semantic descriptors to structural descriptors, so that the latter can mediate between the former and acoustic descriptors.

In order to check the reliability of semantic descriptions, the outcome of two follow-up experiments was compared with the original experiment. Apart from a few inconsistencies, evaluations for the semantic descriptors used in this study seem to be reliable, the ambiguous ones having already showed a rather high intersubject variability in the original experiment. In general, the results from the original experiment were confirmed, which suggests that semantic description of music may provide a stable basis for the future development of content-based access to music.

Acknowledgements

This research was funded by the Flemish Institute for the Promotion of Scientific and Technical Research in Industry (project "Musical Audio Mining" 010035-GBOU). The authors would like to thank V. Vermeulen and F. Desmet for their assistance in setting up the experiment and processing the data.

References

Baroni, M. & Finarelli, L. (1994). Emotions im spoken language and in vocal music. Proceedings of the 3th International Conference of Music Perception and Cognition. Université de Liége, Belgium, 129–133.

Cunningham, S.J. (2002). User studies: A first step in designing a MIR testbed. The MIR/MDL Evaluation Project White Paper Collection (2nd ed.), 17–19.

Downie, J.S. (2004). The creation of music query documents: Framework and implications of the HUMIRS project. Proceedings of the Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH),Götcborg. Retrieved November 8, 2004, from http://www.hum.gu.se/allcach2004/AP/html/prop134.html

Droetboom, M. (2004). Report of the international conference on music information retrieval, October 10–14, Universitat Pompeu Fabra, Barcelona. Retrieved September 10, 2005, from: http://www.dlib.org/dlib/december04/droettboom/12droettboom.html

Futrelle, J., & Downie, J.S. (2002). Interdisciplinary communities and research issues in music information retrieval. In M. Fingerhut. (Ed.) Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR02), Paris, 215–221.

Gabrielsson, A., & Juslin, P.N. (2003). Emotional expression in music. In R.J. Davidson, H.H. Goldsmith, & K.H. Scherer (Eds.), Handbook of affective sciences (pp. 503–534). New York: Oxford University press.

Gabrielsson, A., & Lindström, S. (2001). The influence of musical expression on emotion in music. In P. Juslin & J. Sloboda (Eds.), Music and emotion: Theory and research (pp. 223–248). New York: Oxford University Press.

Hargreaves, A.C. & North, D.J. (1997). The musical milieu: studies of listening in everyday life. The Psychologist, 10, 309–312.

Huron, D., & Aarden, B. (2002). Cognitive issues and approaches in music information retrieval. Unpublished document. Retrieved January 9, 2003, from http://dactyl.som.ohio-state.edu/Huron/Publications/huron.aarden.MIR.html

Juslin, P.N. (1997). Can results from studies of perceived expression in musical performances be generalized across response formats?, Psychomusicology 16, 77–101.

Juslin, P.N., & Laukka, P (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bullettin, 129(5), 770–814.

Juslin, P.N., & Sloboda, J.A. (Eds.) (2001). Music and emotion: Theory and research. New York, Oxford: Oxford University Press.

Kalbach, J. (2002). Classifying emotion for information retrieval: Three websites. Notes, 59(2), 408–411.

Kim, J.-Y., & Belkin, N. (2002). Categories of music description and search terms and phrases used by non-music experts. Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR02), Paris, 164–171.

Lee, J.H., & Downie, J.S. (2004). Survey of music information needs, uses and seeking behaviours: Preliminary findings. In M. Fingerhut (Ed.) Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR04), Barcelona, 441–448.

Leman, M. (2002a). Musical audio mining. In J. Meij (Ed.), Dealing with the data flood: Mining data, text and multimedia. Rotterdam: SST Netherlands Study centre for Technology Trends.

Leman, M., Clarisse, L.P., De Baets, B., De Meyer, H., Lesaffre, M., Martens, G. et al. (2002b). Tendencies, perspectives, and opportunities of musical audio-mining. In A. Calvo-Manzano, A.Pérez-Lopez, & J. Salvador Santiago (Eds.), Forum acusticum sevilla (FAS), Special Issue of the Journal Revista de Acústica XXXIII (3–4), Sevilla, CD-Rom.

Leman, M., Vermeulen, V., De Voogdt, L., Moelants, D., & Lesaffre, M. (2005). Prediction of musical affect attribution using a combination of structural cues extracted from musical audio. Journal of New Music Research, 34(1), 39–67.

Leman, M., Vermeulen, V., De Voogdt, L., Taelman, J., Moelants, D., & Lesaffre, M. (2004). Correlation of gestural musical audio cues and perceived expressive qualities. In A. Camurri & G. Volpe (Eds.), Gesture-based communication in human-computer interaction (40–54). Berlin Heidelberg: Springer-Verlag.

Leman, M. (2007). Embodied music cognition and mediation technology. MIT Press, Cambridge.

Lesaffre, M., (2005). Music information retrieval: Conceptual framework, annotation and user behaviour. Doctoral dissertation. Available at http://www.ipem.ugent.be/staff/Mich/PhDMLesaffre_Download.html

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

Peretz, I., Gagnon, L. & Bouchard, B. (1998). Music and emotion: Perceptual determinants, immediacy and isolation after brain damage. Cognition, vol. 68, 111–141.

Scherer, K.R., & Oshinsky, J.S. (1977). Cue utilization in emotion attribution from auditory stimuli. Motivation and Emotion, 1, 331–346.

Scherer, K.H., & Zentner, M.R. (2001). Emotional affects of music: Production rules. In P.N. Juslin & J.A. Sloboda (Eds.), Music and emotion: Theory and research (pp. 361–392). New York: Oxford University Press.

Tzanetakis, G., & Cook, P. (2000). Experiments in computer-assisted annotation of audio. Paper presented at the International Conference on Auditory Display (ICAD), Atlanta, Georgia, USA.

Wedin, L. (1972). A multidimensional study of perceptual—emotional qualities in music, scandinavian. Journal of Psychology 13, 241–256.

Williams, C. (2001). Does it really matter? Young people and popular music. Popular Music, 20(2), 232–242.

Yang, D., & Lee, W. (2004). Disambiguating music emotion using software agents. In M. Baroni (Ed.) Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR04), Barcelona, 52–58.

DOI: 10.1002/asi