USER-DEPENDENT TAXONOMY OF MUSICAL FEATURES AS A CONCEPTUAL FRAMEWORK FOR MUSICAL AUDIO-MINING TECHNOLOGY

Micheline Lesaffre[1], Marc Leman[1], Koen Tanghe[1], Bernard De Baets[3], Hans De Meyer[4] and Jean-Pierre Martens[2]

1 Department of Musicology (IPEM), Ghent University, Blandijnberg 2, 9000-GHENT, Belgium,
2 Department of Electronics and Information Systems (ELIS), Ghent University
3 Department of Applied mathematics, Biometrics and Process Control, Ghent University
4 Department of Applied Mathematics and Computer Science, Ghent University
Micheline.Lesaffre@rug.ac.be

ABSTRACT

Musical audio-mining technology allows users to search and retrieve music by means of content-based text and audio queries. Though these systems are promising, there is a need for a better understanding of the role of user preferences and user profiles. The development of taxonomies for different aspects of a musical audio-mining system aims at bridging the gap between system development and user interactive interfacing. In the first part of this paper, the need for user-dependent taxonomy development is addressed and an experiment in spontaneous user behavior is described, based on 72 users and 1148 vocal queries. Statistical analysis provides insight into the characteristics of vocal querying and possible useful concepts. In the second part of the paper, it is described how categories and concepts from the statistical analysis have been used for the refinement of taxonomies that address user interactive interfacing and feature extraction.

1. INTRODUCTION

Search and retrieval of information is a core activity of the information society and music is a product of interest to many members of this society. Hence the need for advanced music information retrieval (MIR) systems, which allow users to specify musical content interactively [1]. For this aim, reliable automatic annotation and processing of musical content is needed (see the proceedings of ISMIR 2000-2).

The extraction and processing of musical content from musical audio is called musical audio mining. Figure 1 shows the general architecture of an audio-based music retrieval system. It consists of a target audio-database (left), a query interface (right), and a similarity-matching engine with optional parts that account for users profiling. The task is to retrieve the wanted musical audio files using information provided by the query.

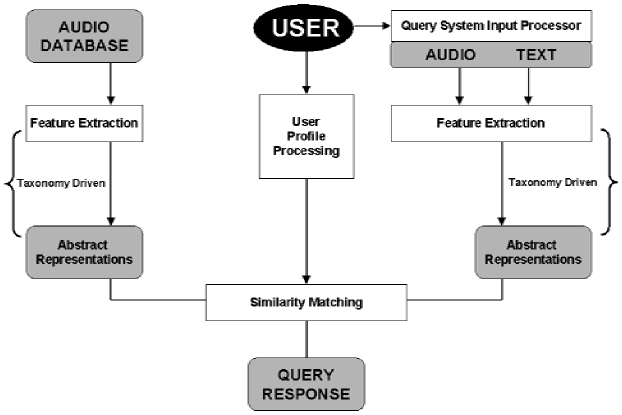


Figure 1: *General architecture of an audio-based MIR system. Left: the target structures with audio database and abstract representation. Right: the query structures, with query input and abstract representations. In the middle, user with particular preferences and profile. Bottom: query response (e.g. in the form of a list of titles)*

A main problem, however, is how users want to interact with those systems. Although there is general agreement that the user-centered approach is a key area of music information retrieval, few studies have been undertaken that concentrate on user behavior [2]. Thus far, a variety of systems have been developed that can automatically extract information from an audio signal [3, 4] but less attention has been paid to the user-friendly characteristics of this interaction.

Progress in developing user-friendly musical audio-mining depends on a number of factors, among which: (i) careful analysis of spontaneous user query behavior, (ii) the development of taxonomies, that is, sets of meaningful concepts and relationships between concepts dealing with musical content, (iii) the implementation of a system that uses this network of concepts and interrelations for automatic annotation and processing of musical content, (iv) the design of a user-friendly query interface that deals with the spontaneous ways in which users tend to interact with music.

This paper highlights the approach of the MAMI-project in several of these domains. In the first part of this paper, the need for user-dependent taxonomy development is addressed and an experiment in spontaneous user behavior is described, based on a large-scale experiment with 72 users and 1148 vocal queries. Statistical analysis provides insight into the characteristics of vocal querying and possible useful concepts. In the second part of the paper, it is described how categories and concepts from the statistical analysis have been used for the refinement of taxonomies that address user interactive interfacing and feature extraction.

2. A TAXONOMY OF MUSIC DESCRIPTORS

2.1. Why a Taxonomy?

Music is characterized by multiple levels of content descriptors, from low-level acoustical features up to high-level structural descriptions and qualitative specifications. Taxonomies aim at providing descriptor labels for these many levels, which is useful for two reasons: (i) provide a common language of descriptors to developers of audio-based MIR systems (they tend to come from very different disciplines), (ii) provide a common language to users that interact with these systems (also users tend to have different backgrounds). The development of a taxonomy is challenging and dynamic: refinement is needed in view of application results. The role of taxonomies is to make bridges between system development and user interactive interfaces.

Until now, taxonomies for music description have not been conceived of as operational instruments related to audio mining. The music industry, Internet music retailers, copyright companies and many others have designed music taxonomies for genre classification [5], but most of these taxonomies are created to meet the particular aims of the industries rather than the users, or guide their consumers, without having any foundation within a network of musical relevant concepts. As a result of this particularity, these taxonomies show many constraints and inconsistencies. They are not very useful in audio mining where instrumental and user-friendly taxonomies are needed.

Instrumental user-friendly taxonomies are meant to define content, structure and boundaries as a framework for the systematic development of musical audio mining. They deal with the interdisciplinary background of researchers working in audio-mining as well as with user behavior, and they have to cope with the multi-facetted ways in which music can be described.

2.2. What kind of Taxonomy?

The taxonomies worked out within the context of the MAMI project are multi-leveled, and multi-dimensional. Distinctions have been made, for example, between description levels, concept categories, features and methods [6]. Low (physical and sensorial), mid (perceptual) and high description levels (formal and expressive) point to the degree of abstraction in which content is related to the audio signal. Concept categories are classes of items with common attributes (e.g. genres, styles, sound sources, and structural entities). Features are properties extracted from the musical signal (e.g. note duration, pitch). Methods are related to how users are dealing with content specification (e.g. vocal: singing, humming, textual).

It is not our intention to work on a single taxonomy for all problems, rather to develop different taxonomies dedicated to particular aspects of the MIR system.

As mentioned before, a taxonomy serves both the developers and the users, although users may be the driving factor in taxonomy design. A user-oriented taxonomy aims at specifying concept categories that can deal with the broad diversity of how users deal with music. In view of Figure 1, a user-friendly taxonomy is related to the query-interface part. Such a taxonomy should draw on the different query statements which users tend to formulate spontaneously.

In what follows, a coarse taxonomy, defined in a previous phase of the MAMI-project [7], has been refined on the basis of a study of how users tend to deal with audio-based MIR systems. We first describe the results of that study, and then describe its effect on the development of our taxonomy.

3. SPONTANEOUS USER BEHAVIOR

A query-by-voice (QBV) experiment for studying spontaneous user behavior has been set up [8] with 72 subjects. Their task was to reproduce a tune from memory. The experiment resulted in a large database of 1148 vocal queries that was annotated and statistically analyzed [2]. The main results are briefly reviewed.

3.1. Characteristics of vocal querying

Vocal querying has defined characteristics. Results of the statistical analysis indicate that a user-friendly MIR system should be able to deal with people starting up to 2 seconds after the start of the recording and reproducing a few seconds of a tune up to 30 seconds or more. The mean query length however is around 14 seconds.

The ideal system should also cope with changes in query method such as singing syllables, singing text, humming and whistling. Sometimes subjects prefer to tap along with the drum or make spoken comments while performing a query. More than half of the queries is performed using only one method, the rest varies between 2 and 6 methods. Most common query methods are singing lyrics and singing syllables.

The results suggest that any system that concentrates on one single query method will confine the spontaneous way in which the user tends to describe the musical content. Our experiment showed that less than half of the participants produce whistled queries and a particular group even tends to provide long whistled queries of high quality. In contrast to this, most existing systems require the use of specific syllables for the ease of query processing [9]. Yet the kind of syllables requested from users are often the wrong ones. Syllables like [ba] or [fa], for example, hardly occur in spontaneous vocal querying.

Three quarter of the query collection is performed in a melodic way, which supports the known wisdom that melodic content is to be regarded as a major salient feature of vocal queries [10].

3.2. User groups

The analysis of query methods suggests the categorization of users into different groups. This categorization depends on age, gender, musical background and memory type. Five user profiles were found. The first relies on singing text, the second prefers singing syllables, the third divides its query time between two methods, the fourth mixes several methods and the fifth (small group) prefers whistling.

Significant effects on the methods used were found between musicians and non-musicians. For musicians, text occupies a significantly smaller position than for non-musicians. As to gender, men tend to use a larger variety of syllables than women do. There is also clear influence of the memory type on the query method that changes from textual dominance to syllabic dominance with a growing importance of short-term memory. Remarkable is the sudden increase in percussive queries when long-term memory of the song is no longer present.

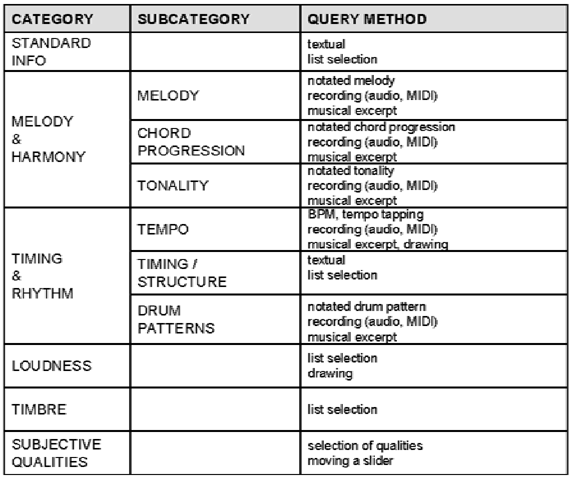
Conclusion: In this study of spontaneous querying, a large variety of strategies was observed but the experiment generated guidelines that provided information for the refinement of user-oriented taxonomies.

4. TAXONOMY REFINEMENT

4.1. User-Oriented Taxonomy

The experiment on spontaneous user behavior provides guidelines for the development of user-oriented taxonomies, which serve as bridges between user preferences and needs, and system development.

The taxonomy depicted in Figure 2 has been used for developing an interactive user interface for querying. Query methods are categorized into four approaches called standard info, melody and harmony, timing and rhythm, loudness, timbre and subjective qualities. Standard info requires only textual input about standard meta-data such as the title of the piece, name of the composer, orchestra director, or record company. Apart from that, content specific information can be given about melody, harmony, rhythm, loudness and timbre. These categories allow the specification of structural descriptions. Inputs may be based on specifications of structural descriptors (e.g. selecting pre-defined verbal descriptors related to pitch and rhythm), audio queries and audio examples (e.g. recording of a voice or an instrument, example), graphisms (e.g. make a drawing of tempo evolution) or motorical actions (e.g. tapping the beat, handle a slider). Subjective qualities, related to perceived affective qualities of the music, up to now, are described by a limited set of bipolar adjectives (e.g. sad-gay, tender-brutal, boring-exiting). Input can be given by moving a slider, for example.



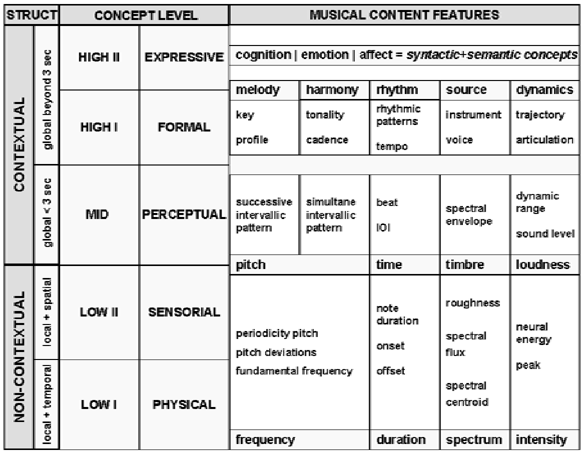
CATEGORY	SUBCATEGORY	QUERY METHOD
STANDARD INFO		textual list selection
MELODY & HARMONY	MELODY	notated melody recording (audio, MIDI) musical excerpt
	CHORD PROGRESSION	notated chord progression recording (audio, MIDI) musical excerpt
	TONALITY	notated tonality recording (audio, MIDI) musical excerpt
TIMING & RHYTHM	TEMPO	BPM, tempo tapping recording (audio, MIDI) musical excerpt, drawing
	TIMING / STRUCTURE	textual list selection
	DRUM PATTERNS	notated drum pattern recording (audio, MIDI) musical excerpt
LOUDNESS		list selection drawing
TIMBRE		list selection
SUBJECTIVE QUALITIES		selection of qualities moving a slider

Figure 2: *Categories and subcategories of a user interface, with related query methods that are facilities for vocal input, verbal specifications in terms of low to high level structural descriptions, and qualitative descriptions.*

4.2. Constituent music-related concepts

A major part of taxonomy development is devoted to musical feature extraction and analysis. This taxonomy (see Figure 3) includes six basic categories: melody, harmony, rhythm and timbre, dynamics and expression. Pitch, as a unit in time and space, is constituent for melody and harmony. Melody, harmony and rhythm attach to spatio-temporal structural properties. Timbre is part of the concept class that describes sound sources or sonic material. The Dynamics category relates to aspects of loudness and the many aspects of movement and musical gesture. Dynamics levels and accents are natural indicators for emotional mood as well. As shown in the figure, the categories relating to audio-structural properties are also distinguished according to physical, sensorial, perceptual and formal descriptors. Many researchers are currently studying the automated extraction of these audio-structural features.

Expression, on the other hand, is a high level inter-subjective category which cannot be directly extracted from the audio, but which can be modeled on the basis relationships between extracted structural (cognitive) features and perceived qualities (affects, emotions). Investigation on expressive qualities reveals interesting relationships between structural and semantic descriptions [6]. It has been shown that the listener's perception of qualitative musical content tends to be inter-subjectively consistent and that semantics can connect with automatically extracted low and mid level features.



<table>
<tr><th colspan="2">STRUCT</th><th colspan="2">CONCEPT LEVEL</th><th colspan="5">MUSICAL CONTENT FEATURES</th></tr>
<tr><td rowspan="5">CONTEXTUAL</td><td rowspan="3">global beyond 3 sec</td><td>HIGH II</td><td>EXPRESSIVE</td><td colspan="5">cognition | emotion | affect = <i>syntactic+semantic concepts</i></td></tr>
<tr><td></td><td></td><td>melody</td><td>harmony</td><td>rhythm</td><td>source</td><td>dynamics</td></tr>
<tr><td>HIGH I</td><td>FORMAL</td><td>key
profile</td><td>tonality
cadence</td><td>rhythmic patterns
tempo</td><td>instrument
voice</td><td>trajectory
articulation</td></tr>
<tr><td rowspan="2">global &lt; 3 sec</td><td>MID</td><td>PERCEPTUAL</td><td>successive intervallic pattern</td><td>simultane intervallic pattern</td><td>beat
IOI</td><td>spectral envelope</td><td>dynamic range
sound level</td></tr>
<tr><td></td><td></td><td colspan="2">pitch</td><td>time</td><td>timbre</td><td>loudness</td></tr>
<tr><td rowspan="3">NON-CONTEXTUAL</td><td>local + spatial</td><td>LOW II</td><td>SENSORIAL</td><td colspan="2" rowspan="2">periodicity pitch
pitch deviations
fundamental frequency</td><td rowspan="2">note duration
onset
offset</td><td rowspan="2">roughness
spectral flux
spectral centroid</td><td rowspan="2">neural energy
peak</td></tr>
<tr><td rowspan="2">local + temporal</td><td>LOW I</td><td>PHYSICAL</td></tr>
<tr><td></td><td></td><td colspan="2">frequency</td><td>duration</td><td>spectrum</td><td>intensity</td></tr>
</table>

Figure 3: schematic overview of a taxonomy for feature extraction

5. TAXONOMY AND SYSTEM IMPLEMETATION

The development of the MAMI audio-based MIR system draws on the development of taxonomies such as the ones described above. As to system implementation, two distinctions can be made. (i) A query interface is currently under construction which allows users to specify musical content in terms of vocal melody inputs, musical excerpts and verbal descriptors, based on the taxonomy of Figure 2. The main task is to transform the query input in a representation that is appropriate for search and similarity matching. Search results are reported as a hierarchical list ranked in order of decreasing relevance. Audio feedback allows the user to check whether the engine returned the required music. (ii) Underlying this interface is a library of modules (C++ classes) designed in a clear relationship to the basic concepts of the taxonomy.

5. CONCLUSIONS

Structures found in spontaneous vocal queries were helpful for the refinement of a taxonomy that addresses the design of MIR systems. Taxonomies offer a list of well-defined concepts and relationships between concepts. They provide structure to developers and to users whose backgrounds may be very different. The main requirement for a MIR system is that it should offer the user as much flexibility as possible in performing a search on a music database. Concepts in the taxonomy should allow designers and users to deal with the broad diversity of preferences, needs, and expectations.

Although the work thus far has been focusing on the query part of the MIR system, definitions of the triangular dependencies between query, target and low-level features derived from the signal are urgently needed. The MAMI taxonomies provide steps towards MIR system development that focus on the user, making audio mining more efficient and attractive.

ACKNOWLEDGEMENTS

This research and the QBV experiment were conducted in the framework of the MAMI project, which is funded by the Flemish Institute for the Promotion of Scientific and Technical Research in Industry. The authors gratefully acknowledge Dirk Moelants and Gaëtan Martens for their assistance.

6. REFERENCES

[1] Downie, S. J., “Music information retrieval” in Annual Review of Information Science and Technology 37, ed. Cronin Blaise (Medford, NJ: Information Today, 2003), pp. 295-340

[2] Lesaffre, M., Moelants, D. and Leman, M., “Spontaneous user behavior in vocal queries for audio-mining.” submitted, 2003.

[3] McNab, R. J., Smith L. A., Bainbridge D. and Witten I. H., “The New Zealand Digital Library MELody inDEX”, D-Lib Magazine (1997), pp. 11-18

[4] Prechelt, L. and Typke R. “An interface for melody input” in ACM Transactions on Computer-Human Interaction (2001), pp. 133-149

[5] Pachet, F. and Cazaly, D. “A Classification of Musical Genre.” Proceedings of Content-Based Multimedia Information Access (RIAO) Conference, Paris, France, 2000

[6] Leman, M., Vermeulen, V., De voogdt, L., Taelman J., Moelants D., and Lesaffre M. “Correlation of gestural musical audio cues and perceived expressive qualities” submitted, 2003

[7] Leman, M., Clarisse, L., De Baets, B., De Meyer, H., Lesaffre, M., Martens, G., Martens, J., and Van Steelant, D. “Tendencies, perspectives, and opportunities of musical audio-mining.” In A. Calvo-Manzano, A. Pérez-López, & J. S. Santiago (Eds.), Forum Acusticum Sevilla 2002, 16-20 september, 2002. Madrid: Sociedad Española de Acustica - SEA. (Special issue of Journal Revista de Acustica Vol XXXIII, no. 3-4), 2002.

[8] Lesaffre, M., Tanghe, K., Martens, G., Moelants, D., Leman, M., De Baets B., De Meyer H. and Martens J.P. “ The MAMI Query-By-Voice Experiment: Collecting and annotating vocal queries for music information retrieval” proc. ISMIR 2003

[9] Pauws, S. “CubyHum: A fully operational Query by Humming System” in Proceedings of the Third International Conference on Music Information Retrieval, ed. Michael Fingerhut (Paris: IRCAM – Centre Pompidou, 2002), pp.187-196

[10] Chai, Wei. “Melody Retrieval on the Web” M.S. Thesis, MIT Media Lab, 2001.