# Tendencies, Perspectives, and Opportunities of Musical Audio-Mining

PACS REFERENCE: SS-MUS-01 Musical Acoustics – Perception and Recognition of Music

Leman M.[1], Clarisse L. P.[2], De Baets B.[3], De Meyer H.[4], Lesaffre M.[1], Martens G.[4], Martens J.P.[2], Van Steelant D.[3]

[1] Dept. of Musicology (IPEM), Ghent University, Blandijnberg 2, 9000-GHENT, Belgium,
Tel: +32 (0) 9 2644125, Fax:+32 (0) 9 2644125, E-mail:Marc.Leman@rug.ac.be
[2] Dept. of Electronics and Information Systems (ELIS), Ghent University
[3] Dept. of Applied Mathematics, Biometrics and Process Control (KERMIT), Ghent University
[4] Dept. of Applied Mathematics and Computer Science (TWI), Ghent University

## ABSTRACT

Content-based music information retrieval and associated data-mining opens a number of perspectives for music industry and related multimedia commercial activities. Due to the great variability of musical audio, its non-verbal basis, and its interconnected levels of description, musical audio-mining is a very complex research domain that involves efforts from musicology, signal processing, and statistical modeling. This paper gives a general critical overview of the state-of-the-art followed by a discussion of musical audio-mining issues which are related to bottom-up processing (feature extraction), top-down processing (taxonomies and knowledge-driven processing), similarity matching, and user analysis and profiling.

## INTRODUCTION

Musical audio-mining deals with the extraction and processing of patterns and knowledge from musical audio. In its proper context of music information retrieval (MIR) this knowledge allows users to search and retrieve music by means of content-based text and audio queries, such as query-by-humming/singing/playing/excerpts, or specification of a list of musical terms, such as 'happy', 'energetic', etc., or by any combination of these. The result will be a ranking of answers, based on similarity ratings, pointing to relevant audio-files. First we present the general architecture of a MIR system and an overview of the characteristics of existing MIR systems. Then we discuss issues of musical audio-mining which are related to bottom-up processing (feature extraction), top-down processing (taxonomies and knowledge-driven processing), similarity matching, and user analysis and profiling.

## GENERAL ARCHITECTURE OF A MIR SYSTEM

Figure 1 depicts the general architecture of a MIR system. It basically consists of a database part (left), a query part (right), and a similarity matching engine with optional parts that account for a taxonomy and users profiling. The task is to retrieve the wanted music files using information provided by the query. Papers related to several aspects of MIR can be found at the ISMIR websites http://ciir.cs.umass.edu/music2000/ and http://ismir2001.indiana.edu/index.html.

## OVERVIEW OF EXISTING SYSTEMS

Most existing MIR systems (Table 1) have limited capabilities in terms of data- and audio-mining. The music collections typically consist of sets of short fragments, mostly incipits of (monophonic)

melodies which are represented as strings of pitches, pitch intervals, contours, and durations in an electronic score format such as MIDI or Humdrum. Information retrieval is based on dynamic pattern matching (string editing), that is, calculation of the cost of editing a symbolic sequence query such that it fits with a set of symbolic sequences of music. The ranking of answers is based on this cost. No taxonomies thus far have been used. Some systems, however, do allow audio-input queries. These are processed into symbolic strings (melodies) that can be used in pattern matching.

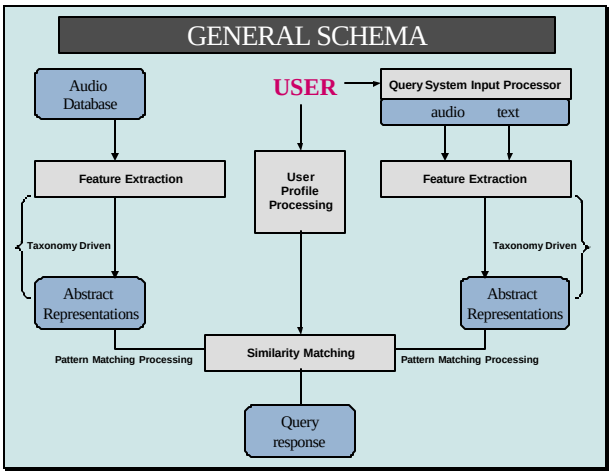


**Figure 1 General schema of a MIR system**

**Tabel 1 Overview of recent MIR Systems**

| SYSTEM | DEVELOPMENT | GENRE, # FRAGM. | ENCODING | ANNOTATION |
|---|---|---|---|---|
| **QBH System**<br>1995 | Cornell Univ.<br>Ghias et Al | classical + pop 183 | MIDI<br>monophone | Parsons encoding (D,U,R) |
| **MELDEX**<br>1996 | Univ. of Waikato<br>New Zealand Dig. Library<br>McNab et Al | folksong: 2 corpora<br>Essen 7000<br>Greenhaus 2354 | MIDI<br>monophone | Parsons (D,U,R)<br>Interval, Duration,Rhythm<br>(Audio query) |
| **MiDiLiB**<br>1997 | Univ. of Bonn<br>Clausen | 20000 | MIDI<br>monophone | polyphone to monophone<br>1. Skyline algorithms<br>2. Melody Lines<br>(Audio query) |
| **Themefinder**<br>1998 | Stanford Univ. &<br>Ohio State Univ.<br>Huron et al | Classical 10000<br>Essen 7000<br>Lincoln 18000 | MIDI & Humdrum<br>monophone | Pitch, Interval, Contour, Scale |
| **Melodic Matching Techniques for Large Databases**<br>1998 –1999 | Univ. of Australia<br>Uitdenbogerd & Zobel | 10466 | MIDI<br>polyphone/<br>monophone | Contour, Interval, Rhythm |
| **Melodiscov**<br>1999 | CNRS-UPMC, Paris<br>Rolland et Al. | Pop small amount<br>200 jazz / folksong | MIDI | Pitch, Duration, FlExPat algoritm: pattern discovery |
| **SEMEX**<br>MIR Prototype<br>2000 | Univ. of Helsinki<br>Lemström & Perttu | notes & chords 2000000 | MIDI<br>IRP<br>MDB | Queries on pitch sequences:<br>1.QPI Classification<br>2.bit-parallelism algorithm<br>3.MonoPoly filtering |
| **Mel. Match. QBH MPEG7-MPEG21**<br>2001 | MIT<br>Kim et Al | pop and folk 8000 | MIDI | Pitch, Contour, Duration<br>ATRAMA (Audio query) |
| **Name-This-Tune "Tuneserver"**<br>2001 | Univ. of Karlsruhe<br>R. Typke | Classical 10000<br>Pop/folk 210 | MIDI<br>monophone | Parsons (D,U,R)<br>Pitch, Duration<br>(Audio query) |

A number of ongoing projects that focus on audio have been described in the literature but are not yet accessible (see ISMIR proceedings). But most of the approaches presented thus far have limited data-mining capabilities and the range of musical data onto which the tools can be applied is restricted.

Going beyond current limitations would imply:

1. An audio database that is representative of the music consumption in our society. This implies the consideration of different musical genres (see Table 2) and the extraction of a wide range of characteristics from polyphonic musical audio.
2. A query system that takes into account content-based audio and text. This implies the development of a taxonomy (network of related concepts and terms) that correlates music descriptions with extracted audio features.
3. User strategies and profiling. This implies the analysis of issues such as long-term memory capacity of users for musical content, fault toleration in sung/hummed queries, as well as the classification of users into typical user communities (profiling).

**Tabel 2 Statistics of music sales based on data provided by IFPI (The International Federation of the Phonographic Industry) (Sales in Belgium)**

| | |
|---|---|
| 46% | Pop/MOR/Easy Listening |
| 24% | Rock/Heavy Metal |
| 7% | R&B/Urban (R&B, Disco, Funk, Fusion, Motown, Reggae, Soul) |
| 5% | Oldies |
| 4% | Soundtracks (filmmusic, musicals) |
| 3% | Dance (Techno, House, Jungle) |
| 3% | Classical |
| 3% | Rap/HipHop |
| 2% | Ethnic/World |
| 2% | Children's/Spoken word/Comedy |
| 2% | New Age |

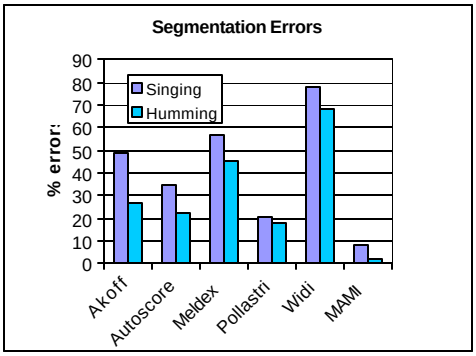


**Figure 2 Data based on Clarisse et al. (2002)**

**BOTTOM-UP PROCESSING (FEATURE EXTRACTION)**

When musical audio is given as input it will be necessary to extract features that make abstraction of differences in key, missing notes, wrong rhythms, and differences in tempo, etc. but preserve the relevant content at a higher level of abstraction.

***QUERY SIDE:*** Most MIR-systems that deal with audio thus far have a focus on the query side but there is clearly a need for more accurate systems. Especially the segmentation was experienced as

being too error prone (20% segmentation errors) (Clarisse et al. 2002). The MAMI transcription system shows segmentation errors that vary between 0 and 7 %, depending on the amount of lyrics that is used by the singer (Fig. 2). An error of less than 10 % is anticipated to be acceptable.

***DATABASE SIDE**:* From the database side it is straightforward to distinguish between polyphonic transcription, source segregation and feature extraction.

**Polyphonic transcription** aims at transcribing the audio database into a score based representation so that similarity matching can be based on the matching of a melodic query with a polyphonic score. Several polyphonic transcription systems are available on the market (see www.intelliscore.net and www.widisoft.com) but these systems are known to perform very poorly (moreover, little or no information is available on how these systems work). Most systems are limited in scope and although a state of the art system is capable of transcribing pieces of at most 3 or 4 voices, most systems work with certain predetermined instruments which have been carefully modeled in advance (Klapuri 1998, 2001). None of these systems may be of direct use but musical audio-mining can benefit from the research efforts that led to the stable and well understood mid-level representations adopted. Also, the research on pitch and rhythm tracking, is extremely useful.

**Source segregation** aims at segregating the audio signal in different complementary audio signals, belonging to separate audio sources. This is generally done by assuming that the source signals are statistical independent. When independent sources are mixed in an arbitrary way, mutual dependencies arise. When these dependent signals are transformed into independent signals (i.e. each signal does not contain any components of the other signals) one obtains the original set of source signals (Casey and Westner 2001). This area is less relevant to musical audio-mining, since a lot of approaches demand that there must be at least as many observable mixture signals as source signals. Also, work in this area has focused especially on speech. If the algorithms are applied to music, the approach is only data-driven (no perceptual or musical information is used).

**Feature extraction** aims at extracting all kinds of sensory/perceptually/cognitive relevant information from the musical signal. In the pitch domain, content about global pitch relationships, such as chord type (instead of exact individual pitches of a chord) are relevant. In the rhythm domain, beat and meter, as well as the dominant rhythmic patterns that occur in the music, provide very useful information. Feature extraction therefore focuses on content that users would tend to deal with when using an audio-based music retrieval system (Aigrain 1999, Cariani and Leman 2001).

**TOP-DOWN PROCESSING (TAXONOMY)**

In musical audio-mining, a taxonomy (or network of musical concepts) is needed for two reasons: (i) to describe and internally represent musical audio features at higher levels of abstraction, using a consistent framework of concepts, and concept relationships and dependencies, (ii) to enable users to specify their queries using content descriptions in addition or apart from singing, humming, or any other input specification. Thus far, however, there exists no commonly agreed upon taxonomy for music description.

MPEG-7 (http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg7.htm#_Toc533998977)
is a worthy attempt at defining a standard in audio content description but it is largely insufficient as standard for audio-based musical content. Unfortunately, most 'taxonomies' in the musicological literature have not been conceived of as operational instruments related to audio. This makes it hard to straightforwardly implement any such musicological theory.

The taxonomy conceptual framework is concerned with the question as to what concepts will be used, how they relate to audio-based features, and how these concepts are defined in terms of the global MIR system. Figure 3 shows the horizontal layers of a taxonomy which is currently under development within the MAMI-project (http://www.ipem.rug.ac.be/MAMI ).

- **Low level** concepts describe content that is close to the acoustical or sensorial properties of the signal. These include features such as frequency, duration, intensity as well as roughness, onset, loudness. The sensorial properties involves processing typically located at the periphery of the human auditory system. They are typically related to temporal features of the signal and local (non-contextual) properties.
- **Mid level** concepts involve time-space transformations and constrained context dependencies within a time-scale of the musical present (the 'now') (< 3 seconds). This is the perceptual level where time-space transformations may allow for the specification of the musical signal content in spatial terms (timbre, pitch, chords…) and temporal terms (beat, meter, rhythmic pattern...).
- **High level** concepts typically involve learning and categorization beyond the representation of the 'now' (> 3 seconds). The concepts deal with structure and interpretation and may be related to cognitive as well as emotional and affective issues. This level is highly determined by the cultural context and processing related to long term memory processes. The concepts may convey defined meanings that are not necessarily directly associated with the signal properties, but perhaps with properties of subjective feelings and interpretations.

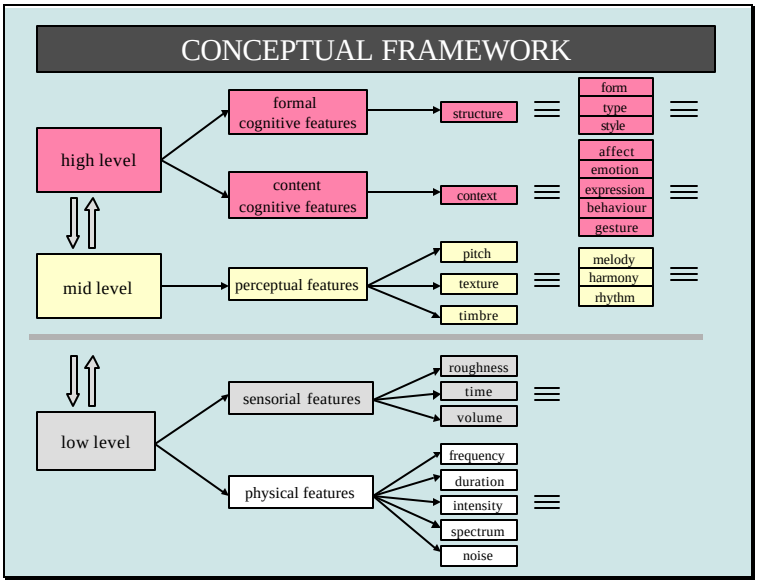


**Figure 3 Horizontal structure of the MAMI-taxonomy**

**SIMILARITY MATCHING**

The complexity and multi-level nature of the musical features, obtained from audio/text processing in a MIR context, may involve a number of different techniques, depending on particular needs and peculiarities. Most systems thus far are based on dynamic pattern matching (calculating the cost of editing a sequence) but a range of different techniques can be taken into account, from graphical models (e.g. classification trees, Hidden Markov Models) to neural networks (e.g. Self-Organizing Maps), and the aggregation of different similarity measurements (global similarity measurement) (Meij 2002).

**USER ANALYSIS AND PROFILING**

Up to now, the role of the user has often been neglected in MIR systems. It is evident, however, that the user's musical memory capabilities, as well as his/her capabilities of music description are determining factors in audio-mining. Little is known about the long-term memory capabilities of

musical recall, nor about the mean performance of an audio-query or the typical sung patterns (e.g. chorus parts) that users tend to focus attention upon. We know very little about differences in audio/verbal query performance between musically-educated and non-educated users. It is evident that this knowledge will add additional constraints to audio-mining. Users may display a typical consumption pattern and therefore may be categorized into particular user groups, whose properties may be of help in audio-mining and information retrieval (collaborative filtering).

## CONCLUSION

Given the current state-of-the-art in electronic content delivery, the technological orientation of the music culture, and the interest of the music industry to provide musical commodities and services via the distributed electronic channels, there is an urgent need to develop advanced tools for music-mining, that is, ways of dealing with content about music and its associated processing. Due to the great variability of musical audio, its non-verbal basis, and its interconnected levels of description, musical audio-mining is a very complex research domain that involves efforts from musicology, signal processing, and statistical modeling. Musical data-mining draws on concept taxonomies that allow users to specify a musical piece in terms of more or less unique rational and emotional descriptors. But these descriptors have their roots in acoustical properties of the musical audio as well, hence, signal processing and statistical modeling are needed to relate audio to the conceptual taxonomy. Similarity measurement plays an important role in finding the appropriate connections between representational structures in the query (which can be sung, or specified in terms of the taxonomy) and representational structures in the database.

## ACKNOWLEDGEMENT

MAMI (www.ipem.rug.ac.be/MAMI) is a musical audio-mining project at Ghent University sponsored by IWT (Institute for the Promotion of Innovation by Science and Technology in Flanders).

## REFERENCES

Aigrain P. (Ed.) (1999). Special Issue on Musical Content Feature Extraction. *Journal of New Music Research* vol. 28 (4).

Cariani P. and Leman M. (Eds.) (2001). Special Issue on Music and Time. *Journal of New Music Research* 30 (2).

Clarisse L. et al. (2002). “An Auditory Model Based Transcriber of Singing Sequences”. *ISMIR 2002* (submitted).

Klapuri A. (1998). *Automatic Transcription of Music*. MSc Thesis, Tampere University of Technology, 1998.

Klapuri A. et al. (2001). “Automatic Transcription of Music”. Symposium on "Stochastic Modeling of Music", *Proceedings of the 14th Meeting of the FWO Research Society on Foundations of Music Research*, 22th of October, Ghent, Belgium.

Casey M.A. & Westner A. (2001). “Separation of mixed audio sources by independent subspace analysis”. *Proceedings of the International Computer Music Conference*, Berlin, August 2001.

Meij J. (Ed.) (2002). *Dealing with the Data Flood: Mining data, text and multimedia*. Rotterdam: STT Netherlands Study Centre for Technology Trends.