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

5.5.2 Musical audio mining

Marc Leman[1]

Introduction

Musical audio mining can be defined as data mining on musical audio. It will allow users to search and retrieve music, not only by means of text queries (such as title, composer, text song, conductor, orchestra), but also by means of content-based musical queries, such as query-by-humming/singing/playing, by specification of a list of musical variables (such as 'happy', 'energetic', etc.), or by means of given sound excerpts, or any combination of audio and text. Musical audio mining opens a number of perspectives for the music industry and related multimedia commercial activities. In this paper we introduce the reader to some basic concepts of musical audio mining and discuss its relevance for commerce and industry.

Musical audio mining in the information society

Music plays an important role in our high-tech information society. Music is nowadays present in every aspect of life, public and private, and thus has a very noticeable influence on people. It allows the expression of affect and emotion which is necessary for the formation of identity and the development of the well-being of the individual in a social environment. This involvement with music is reflected in an elaborate network of music commodities and services, called 'the music industry'. The core of this industry is dominated by a few major companies, followed by a large number of smaller companies [Pichevin, 1997].

Just to give a rough idea of the size of the market: in 1999, the global music retail market (music sales of singles, LPs, MCs, CDs) was worth about 38,506 million US$ with a total unit sales of 3,459 million units showing a growth by 2,6% a year in real value since 1991 [IFPI, 2000]. If, next to that, one considers the social and economical impact of the music recording sector, copyright associations, organizations for concerts and other live performances, musical instrument and consumer audio markets, music in broadcasting, and education, one quickly understands that the music 'business' relies on enormous human and financial resources. Data from Laing [1996] suggest that the music industry had a greater turnover than both the cinema and video industries, providing work in Europe alone to over 600,000 people, and having an impact on millions of people involved in education and business.

[1]Prof Dr M. Leman, Marc.Leman@rug.ac.be, IPEM - Department of Musicology, Ghent University, Ghent, Belgium, http://www.ipem.rug.ac.be/

Despite its consumer market and large network of interconnected industrial activities, musical content has long been underestimated as a potential source of economical activity. But this is rapidly changing. Phenomena such as Napster and Freenet have demonstrated the potential of distributed decentralized information storage and retrieval systems. Given the large quantity of mp3 audio-files available, many people have become aware of the need for content-based search and retrieval.

The classical approach based on text as supplier of meta-data of musical audio, such as the name of the song-writer, the title and the text of the song, the musician, the director, the orchestra, recording date , etc. can be extended with descriptions based on the characteristics of audio. These descriptions can be generated partly manually, partly automatically. In combination with sophisticated profiling techniques, they offer the basis for future intelligent and flexible music mining systems. Given the nature of the problem, it goes without saying that the musical audio mining community is an interdisciplinary research community where musicologists, engineers, mathematicians, sociologists, and others have joined their forces [Leman, 2002].

Techniques for musical audio mining and retrieval

Music search and retrieval is often based not on audio, but on symbolic representations of music. A simple but powerful idea has been to conceive those symbolic representations in terms of the now familiar 'electronic scores'. MIDI (Music Instrument Digital Interface) is an example. This industrial standard for communication between electronic musical instruments and its encoding of music can be conceived of as a score, or note notation, in electronic format. But there are other formats for electronic scores as well, see [Selfridge-Field, 1997]. Many approaches in music information retrieval aim at reducing music first to a notated melody (as electronic score) and associate the proper audio files to this. Search is done on the melody, and retrieval can deliver the necessary audio files. It is often admitted that melodies may not be sufficient as an encoding of all kinds of music, but at least they provide an interesting paradigm for content-based music search and retrieval.

How, then are melodies related to musical content? A melody may be notated as a sequence of pitches or intervals (distances between pitches), and durations for each pitch (or distances between durations). That's fairly simple, and therefore appealing. Search in audio is then reduced to search on symbolic strings that handle pitch and duration. Yet, this simple encoding method already shows that musical content is not always easy to define.

Think of defining the notion of melodic similarity, which is a necessary component in any search and retrieval system. It is evident that proper accounts of musical perception, and perception-based modeling should be taken into account, when melodies are compared. The similarity of two strings is typically calculated as the 'edit'-cost it takes to make the two strings identical by performing a deletion, insertion, and replacement operation on the characters. Characters, in this case, stand for musical notes, or durations, and sequences of characters define melodies. [Rolland, 1999] developed algorithms for discovering patterns in musical sequences, taking into account insertion, deletion and replacement of notes, as well as multiple notes. The operations are furthermore weighted depending on the level of abstraction that is used. This is indeed very typical for music: a melody is not just a

concatenation of notes, but notes constitute relationships at different levels of abstraction. One may consider exact pitches, pitch intervals, or general melodic contour as three different abstraction levels related to pitch. But similar abstraction levels exist for rhythm (such as pulse, beat, meter, rhythm units). Furthermore, the meaning of a melody often depends on the tempo, some notes that are 'edited' are less important than others, depending on their position in the metric structure, and so on. In other words, similarity measurement is not just a matter of string comparison, but one may prefer to introduce high level musical knowledge to make comparisons relevant from the viewpoint of human perception. Musical content processing, therefore, is not that simple. Furthermore, systems may take into account user profiles [Rolland, 2001]. Music education or none? It may make a difference to how one deals with musical content.

Melody representations thus far have been very important, because they offer a major paradigm for music search and retrieval. The paradigm relies on both a music database and a query system. The retrieval will often pass via a symbolic encoding system, such as melody notation, although the representation is not necessarily limited to notated melodies. It can be based on different encoding techniques such as Hidden Markov Models (see Section 6.2.9) or neural networks (see Section 6.2.7). Nevertheless, the procedure is often similar: queries based on audio input (humming/sound examples) will first be transformed into a more abstract representation, making search and similarity measurement feasible. The data entries found can then be linked to the proper audio-file which is retrieved.

In order to become acquainted with some of the techniques, we introduce two approaches based on electronic scores. In the next section, we discuss music mining based on audio-files. A general observation is that many research groups focus on electronic scores, rather than on audio. This seems to indicate that audio research is in an initial phase, where basic concepts and techniques for audio need further development. Speech processing offers some inspiring paradigms, although one has to take into account that musical audio and speech are different in many aspects.

Music search and retrieval based on melodies

Some basic techniques of musical search and retrieval can be illustrated by means of practical applications.
The MELDEX[2] system is designed to retrieve melodies from a database on the basis of a few notes sung into a microphone, hummed, or otherwise entered. The acoustic input is transcribed into music notation and then a database is searched for melodies that contain similar patterns [Ghias, Logan, 1995; McNab, 1997]. The audio to melody transformation is based on the extraction of the fundamental pitch. Users are requested to sing with 'da' and 'ta' so that beginnings and endings of notes can be easily determined by segmentation on amplitude. Retrieving the music from the collection of musical scores is then essentially a matter of matching input strings against a database. To facilitate the task, users can constrain the search. For example, they can compare their input with the beginning or with the chorus of the music. Search is then based on interval directions (contour), rather than pitch ratios, or musical intervals, but the latter is an option. The technique comes down to the classical calculation of the 'edit'-cost by means of deletion, insertion, and replacement. The cost of a sequence of edit operations is the sum of the costs of the individual operations, and the aim is to find the lowest-cost sequence that accomplishes the desired transformation. Using dynamic programming, the optimal solution can be found in a time proportional to the product of the

[2] http://www.nzdl.org/cgi-bin/library

lengths of the sequences. As mentioned before, the search algorithms are improved, when knowledge of musical content processing can be included. Tunes with a certain percentage of matching are returned and the user can then download the files. Themefinder[3] is a similar search and retrieval system based on melodies, but it does not involve audio. Related work on melodies can be found in [Cambouropoulos, 2001].

An entirely different approach is based on so-called content-addressed memories. [Toiviainen, 2001][4] applies this idea to melodies, but it can be extended to any kind of symbolic encoding. The method draws on the findings that listeners judge the similarity of melodies on the basis of frequently occurring features of music. Common statistical measures of music, such as distributions of pitches, intervals and durations, pitch transitions, interval transitions, and duration transitions are extracted from each melody separately. Each melody is thus encoded by a set of variables defining a particular quality of the melodic content. Similarities among melodies can be studied on the basis of a comparison of their probability distributions. For example, when you are interested in a qualitatively high match of pitch sequences, you may focus on the variable (and its corresponding probability distribution) that is based on counting the occurrences of any 3 pitch successions. You may then proceed by performing a proper clustering of these distributions over all melodies. Similarity measurement will then typically be based on distance calculations in a multi-dimensional space. Constrained search can be realized by combinations of the appropriate distributions. Toiviainen and Eerola, for example, use the SOM [Self-Organizing Map, Kohonen, 1995] to cluster the melodies, and combinations of the variables of a melody were realized through the clustering of the position coordinates of the melody on different maps, each representing a single feature (see also Section 6.5.1, Visualization tools and [Kaski, 1997] on the cd-rom).

The statistical approach shows similarities with statistical techniques in speech processing [Jelenik, 1999] and is certainly worth further investigation. The technique is robust and could allow the representation of musical style.

Musical search and retrieval based on audio

In the previous section, we discussed approaches that reduce music to a notated melody. Of course, there is a price to be paid for this simplicity. First of all, not all music is notated by means of a score (think about ethnic music, electronic music, improvisations) and, secondly, the melody approach is not suited for polyphonic music or music with no explicit melodic character. As a consequence the technique is often restricted to the encoding of particular and well-documented folksongs and popular songs. It seems straightforward to expand the search techniques to polyphonic music notation, rather than melodic notation. This work is just starting, see [Dovey, 2001] who describes techniques for searching in polyphonic music notations, and [Doraisamy, 2001] who develops a search method based on all combinations of monophonic musical sequences. However, in dealing with much of the musical reality we need more advanced techniques to deal with audio.

Are there any techniques available which could handle any kind of musical audio? Transforming musical audio into some kind of symbolic or abstract representation is a necessary step for most musical search and retrieval systems. A straightforward paradigm for musical encoding is inspired by speech recognition research and runs as follows. The musical signal is first windowed into short (e.g. 30 ms) overlapping frames. For each of the frames, a short-time spectrum is computed and transformed into Mel Cepstrum

[3] http://www.themefinder.com
[4] http://www.jyu.fi/musica/essen/

Coefficients[5]. Alternatively, one may use an auditory model or wavelet transform. Then a Hidden Markov Model (HMM) can be used to classify the frames in an unsupervised way. The outcome will be an alphabet obtained by self-organization. In the study by [Aucouturier, 2001], the HMM learns different textures occurring in the music in terms of mixtures of Gaussian distributions[6] over the space of spectral envelopes. Learning is done with the Baum-Welsh algorithm, while the labeling is done using Viterbi decoding [Jelenik, 1999](see Section 6.2.10). Each state of the HMM then accounts for one texture. A particular musical piece will be encoded by means of state transitions, hence by a sequence of labels (corresponding to the states), which we could call a 'state-score'. This is the classical technique borrowed from speech recognition (Figure 1).

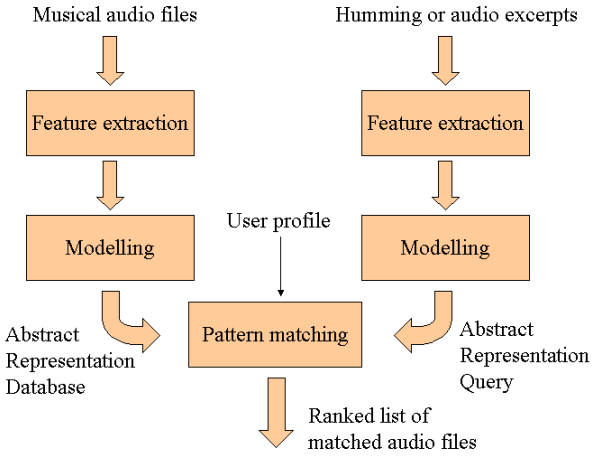

Figure 1

General diagram of musical audio mining based on audio queries.

Rather than comparing audio with audio directly, more abstract representations of musical audio are needed in order to take into account issues such as pitch transposition and tempo differences between the query and the stored music file. In the left part of the figure, musical audio files are transformed into abstract representations through feature extraction[7] and modeling. Feature extraction may produce pitch frames which are modeled, using HMM. The result is a kind of 'score' for each musical audio file with labels for pitches and some (flexible) representation for duration. The 'scores' make up a database which represents the musical audio files at a level of abstraction where pattern matching is possible. The query part has as input a musical audio file, which could be an excerpt of an existing piece of music, or a piece sung by the user in a microphone. A similar feature extraction and modeling can be done, which results in an abstract representation of the query. The pattern matching will be typically based on a similarity matching between the abstract representation of the query and the abstract representation of the music in the database. Pattern matching can be refined by a user profile. Many variations of this general schema are possible. Audio

[5] A transformation with an algorithm which incorporates compensation for the variations in sensitivity of the human ear to different audible frequencies.
[6] indicates the probability density of values
[7] extracting specific features from the original data

queries can be extended by a text-based music information retrieval system, different levels of search, and so on.

It goes without saying, however, that the definition of the acoustic events in music is not straightforward. Speech events are typically described in terms of a limited set of phonemes, and allow the statistical structures to be learned in a supervised way (i.e. on the basis of manual annotation). However, the acoustical events in music are much more numerous and should be learned in an unsupervised way, for example, by training HMM's on a large audio database. Once the HMM's have been derived, musical audio can be decomposed in terms of the states of the HMM. Instead of comparing raw audio material, similarity can then be based on the sequences of labels. In other words: appropriate string-matching can then be done by comparing 'state-scores'. Conceptually, they are on the same level as the notated melodies and therefore, similarity costs can be based on 'edit'-operations such as deletion, insertion, and replacement, using dynamic programming algorithms. Further research is needed, however, to demonstrate the feasibility of this paradigm.

Meanwhile, it is clear that variants of this schema can be explored. For example, one could first cluster the acoustical features using a neural network, and then represent the music as a trajectory in this clustering space. Trajectories can be learned and clustered so that similar music is projected at close distance in this space. Temporal characteristics can furthermore be dealt with by means of time-warping methods. [Mazzoni and Dannenberg, 2001], for example, extract a continuous pitch contour from a recording of the audio query and use then a time-warping algorithm to match this string of pitches against songs in a database of MIDI files.

Musart[8] [Birmingham, 2001] is an example of a query by a humming system based on the extraction of melodic contour and HMM's. Concerning sound effects search and query by example systems, see Musclefish[9]. [Blum, 1997] develops a variety of algorithms for analyzing the content of an audio signal and producing meta-data describing the signal. The first official beta release of the SoundFisher[10] program became available in August 2000. Follow-up versions are posted on a regular basis. The user can build complicated search requests based on any of the data, including the audio qualities or similarity to audio examples.

Quite a lot of other projects are related to query-by-humming. Just to mention a few: QBHAgent[11] is a system that employs the accepted notion of melodic pitch contours to support content-based navigation around a body of multimedia documents including MIDI and digital audio files. The system adopts an open hyper-media model, which enables the user to find available links from an arbitrary fragment of a piece of music, based on the content or location of that fragment. ECHO[12] is a software system that takes a sung melody as a search query. Bad singing could be adjusted for by calculating the difference in pitch and intervals between adjacent notes. Those calculations are then compared with music sequences in a database until a match is found. SoundCompass [13] is a query-by-humming system for a large music database. All processing of musical information is done on the basis of

[8] http://musen.engin.umich.edu
[9] http://www.musclefish.com
[10] http://www.soundfisher.com/
[11] http://audio.ecs.soton.ac.uk/sbh
[12] http://www.sonoda.net/
[13] http://woodworm.cs.uml.edu/~rprice/ep/kosugi/

beats instead of notes. Tuneserver[14] attempts to match a wav-file containing a user's whistling to a classical music piece. Audio Logger[15] has an interactive query which interfaces with multiple modes such as a query by example (audio clip or music example), query-by-humming, query by concept, query by user drawn frequency and amplitude shape, query by instrument. Query by emotional conception of music: 'sorrow', 'sad', 'joy', query by pitch, query by frequency or amplitude shape, and query by keyword are also possible.

The University of Milan[16] has a project related to archiving, in collaboration with the Teatro alla Scala and the Bolshoi Theatre. Part of the project is devoted to the development of a content-based system [Haus, 2001]. Related groups in Europe are situated in different countries, among which IRCAM (France), and Pompeu Fabra University (Spain). CUIDADO[17] is an example of a EU-supported project, which aims at developing content-based technologies using MPEG 7 standard, building reusable modules for audio feature extraction, statistical indexing, database management, networking and constraint based navigation. A project called MAMI (Musical Audio mining) has recently been established at Ghent University[18].

This list is not exhaustive and the reader may consult the recent proceedings of ISMIR[19] 2000 and 2001 for further details. In general, most audio-based systems so far deal with very limited databases. Large databases are typically encoded in melody notation.

MPEG-7 audio

Some words should be said about MPEG-7 audio. The background for this is that musical search and retrieval will be based on both automated and manual descriptions and that all this information needs to be represented in one way or another. Therefore, a recurrent idea has been to develop a standard for representing musical content, also called 'meta-data' in this context. This is the goal of MPEG-7 Audio, a multimedia content description interface, whose first version is available[20], but its stability has not yet been proven, and some of its representation types may raise questions from a musical point of view. Nevertheless, MPEG-7 Audio is an interesting initiative and a first attempt to develop a common representation for meta-data on music.
MPEG-7 Audio has been developing a description scheme for melodic information in terms of pitched, monophonic melodies. Another description scheme deals with musical instrument timbre, which is, according to this standard, either 'harmonic' or 'percussive'.
It is important, however, to understand that apart from some reference tools MPEG-7 does not come up with algorithms for, say, fundamental pitch extraction. Rather the aim is to provide a framework for the representation within a larger context of feature representations and schemes. The MPEG home page[21] and the MPEG-7 Industry Focus Group website[22] contain links to

[14] http://wwwipd.ira.uka.de/tuneserver/
[15] http://www.inficad.com/~roz/audio.htm
[16] http://woodworm.cs.uml.edu/~rprice/ep/haus/index.html
[17] http://www.cuidado.mu/
[18] http://www.ipem.rug.ac.be/mami/index.html
[19] http://music-ir.org/
[20] http://www.lis.e-technik.tu-muenchen.de/research/bv/topics/mmdb/e_mpeg7.html
[21] http://mpeg.telecomitalialab.com/
[22] http://mpeg-industry.com/

information about MPEG, links to other MPEG-7 web pages and many publicly available documents (the MPEG-7 standard is included on the cd-rom).

Understanding music for musical audio mining

Many audio-based approaches to music are characterized by a straightforward bottom-up approach and do not take music models into account in their processing. Most music researchers, however, argue that music needs to be described at different levels [Todd, 1991; Poli, 1991; Marsden, 1992; Balaban, 1992; Bigand, 1993; Haus, 1993; Roads, 1997; Leman, 1997; Zannos, 1998; Godoy, 2001], and that music like language obeys rules which are grounded in human biology and in cultural context. Progress in the acoustical analysis of music may be expected from the development of good music models, a better understanding of the musical structures underlying particular musical styles, and models of music perception. The following levels of description of music attempt at to give an idea of the different possible approaches.

Signal representation

This is just how music is represented at the most basic level: as waveforms.

Frame-based representations

Frame based representations are derived from the waveform using shifted frames, as is nowadays done in many applications. Different types of frame-based representations can be related to different musical parameters such as energy, pitch, roughness, rhythm, spectrum, etc. The Mel Cepstrum Coefficients representation is popular for acoustical encoding, because it is a fast way to incorporate aspects of the human auditory system. In auditory modeling, which is often used as the acoustical front end in speech processing, the frame-based representations reflect neuronal activity at different time steps. [Leman, 2001] offers a toolbox for perception-based music analysis.

Parameter-based representations

These representations rely on modeling techniques over frame-based representations. An example is tonal tension, which can be calculated by comparing pitch images that reflect the development of musical pitch on a local scale on the one hand with pitch patterns of a global scale on the other hand [Leman, 2000]. Another straightforward technique is to use inter-onset-intervals as constrained time-intervals from which parameters may be extracted. [Carreras, 1999] apply this technique to extract chord patterns from pitch images. In vibrato, the mean pitch pattern will be the pitch heard, but the pitch modulation may be as great as 60 cents (half a tone). [Rossigno, 1999] have worked out a feature extraction and temporal segmentation device. Characterizing the expressive content of musical signals is another 'hot' topic of research [Dannenberg, 1998]. The work of [Canazza, 1997] aims at extracting performance features from musical signals. These are just a few examples of a large range of features that can be extracted from musical signals.

Event and gesture-based representations

Events can be linked to the underlying parametric and frame-based representations. [Klapuri, 2001] has developed a system for the automatic transcription of music with the main emphasis on finding the multiple pitches of concurrent musical sounds. These techniques may be very useful for encoding, because they allow for the representation of musical content in terms of symbols: objects having attributes such as a particular beginning,

end, duration, pitch, loudness, timbre, vibration frequency, vibration index, etc. Larger musical structures can be taken into consideration as well. An example is a repeating rhythm pattern, where the energy evolution over time represents the deployment of gestural characteristics in time. A musical content, therefore, may point to a melodic figure representing a particular musical thought or Gestalt. [Bresin, 2000] uses performance variables such as pitch number, inter-onset duration, off-time duration, sound level, pitch deviation, vibrato amplitude, vibrato frequency, etc. to define expressive performances in event-based music representations.

The concept-based representations

The next level then is purely conceptual. Concept based representations provide descriptions for musical features in terms of the natural language such as 'rough', 'expressive', 'dissonant', 'graceful', etc. Primary concepts are those such as 'high', 'low', 'far', 'near', 'accellerando' or 'descellerando', which describe physical features of a musical pattern or event. At a second level, concepts may have a metaphorical origin rooted in space, movement, and effect. These are the so-called kinaesthetic [Laban, 1963], synaesthetic [Cytowic, 1989], and cenaesthetic concepts [Broeckx, 1981], which actually belong to another domain of experience, but which are often used to characterize music. Examples are 'press', 'glide', 'flick' or concepts such as 'labile', 'conflict', 'extinction'. Often these concepts can be connected to lower representational levels. Much more research is needed to understand how people use these concepts for retrieving music. Some related work has been done by [Pachet, 2001] who studied a method for parsing music file names.

To summarize: different representational levels for music exist, and they provide cues for meta-data representation of musical audio. What is needed is a conceptual architecture for music representation in which these different levels can be handled in a flexible way. As far as I know, such a representational schema does not yet exist, though steps in the direction of multi-level architectures for musical representation are undertaken.

Economical and societal impacts

The economical and societal impacts of musical audio mining are primarily situated in the music industry, but they may prove to have an even wider impact as well. Results may be relevant to an entire domain of multimedia applications concerned with the automatic retrieval of information from databases incorporating files of different natures — such as text, speech, music and images — coexisting in the same software environment. In particular, the industries which have thus far been involved in text-based and speech-based data mining may learn how to extend their methodologies so as to include musical audio files in their applications as well.
In the present stage of development, musical audio mining research projects are upstream industrial research projects with a high potential in follow-up technological trajectories for diverse industrial players in the multimedia industry, in particular in the large and growing field of the data mining and content business (content creation, distribution, delivery, retrieval, consumption). Many of the tools envisioned in musical audio mining will prove valuable in the general domains of signal processing and pattern matching. In particular, we think of 'intelligent' interactive multimedia systems that work on the processing of musical content. Musical audio mining can further be used to retrieve songs from a database in a Karaoke machine, to shop for specific songs on CDs in a music store, or to request certain songs via a mobile phone. Research and industrial communities nowadays have a strong interest in non-verbal multi-modal communication both with man and

machine. The focus on non-verbal multi-modal communication is of particular relevance to many areas in man-machine communication.

Related industry and commerce

Following is a more detailed view of the absorption of musical audio mining by the aimed sector in different application domains.

Music industry

The primary sectors of the music industry are: (a) the production and sale of sound recordings (b) the administration of authors rights and neighboring rights (c) the organization of concerts and other manifestations (d) the manufacture and sale of musical instruments. Musical audio mining research will have an impact on each of these:

- Musical audio mining is relevant for the production of sound recordings, because it has a focus on content. The current production techniques can be improved so that more efficient content creation based on automated audio-analysis can be achieved. Search and retrieval of music on the basis of audio mining techniques is of course a core application.
- Techniques for feature extraction, statistical modeling, and similarity measurement can prove to be useful for companies such as SABAM, EMO and IFPI' which are involved in the protection of author's rights. Nobody questions the importance of an efficient copyright protection system, and good content extraction algorithms may be an indispensable means for the efficient tracing of eventual plagiarism.
- Concerts and performances can be advertised using content description techniques. The core technology of musical audio mining is furthermore highly relevant to the entertainment business. The (future) interactive discotheque is one example of a system in which groups of dancers interact and produce their music by dance in response to music machines. By dancing they can also influence the music which is produced by the system. Karaoke is another very popular field in which the musical audio mining techniques are central. In the Karaoke paradigm, a singer sings a popular song whose score is stored in the computer. The machine plays the accompaniment, and so as with the interactive discotheque, techniques of feature extraction and similarity measurement are of central importance in selecting the right music.
- As to the manufacture and sale of electronically enhanced music instruments, musical audio mining will be relevant in that it provides core technologies for the new generation of interactive multi-sensory audio/dance/video systems. The musical audio mining technology may serve to extract the appropriate content that is needed for man-machine interaction based on expressiveness [Camurri, 2001]. Applications are conceivable in the artistic cultural sector, for example in theater and opera houses, or in the commercial cultural sector such as discotheques, large multimedia manifestations, and very popular dance music environments. Applications which envision the interaction of large dancing masses with technology on the basis of content extraction in movement and audio, such as singing offers interesting opportunities.

Multimedia industry

- Apart from the specific musical applications mentioned in the previous section, there is a vast domain of possible applications where music is considered part of a larger information context.
- Musical audio mining is relevant for the search in multimedia archives, because audio is indeed an important aspect of many multimedia documents. Musical audio mining technology will provide the technology for

separating for example songs from speech and from purely instrumental musical passages. Access to sound sources of all kinds, including speech and music are very useful for researchers, for TV-production houses (e.g. finding appropriate musical sequences, or spoken or sung film/video fragments), as well as for individual consumers who want to get direct access to particular forms of music. Such systems will allow users to fast-forward through the news broadcast, stopping at all the occurrences of a given word or phrase, or avoiding musical intermezzos. By eliminating the need to listen to or transcribe lengthy recordings, this breakthrough capability can save enormous amounts of time and increase productivity. The capabilities of such systems can be greatly enhanced by more sophisticated tools for musical audio mining.

- Libraries, museums, foundations, universities make up some of the group of users that are focused on cultural heritage archives. Until now, rapidly growing multimedia databases offered only author/title/edition-like search and indexing capacities. This very general and crude method of archiving does not comply with modern needs of library users and modern technological possibilities. Musical audio mining technology will provide core technology for the development of new types of indexing and searching in multimedia archives, thus giving easier access to larger user groups. In connection to a film archive, for example, users may be interested in retrieving film fragments where similar musical sequences have been used.
- Other important multimedia users are recording studio's, composers and new media artists, TV, radio, Internet-TV, radio and TV on-demand, and film production companies. The use of multi-purpose multimedia archives is a core aspect of their functioning. This category of users is dependent on easy access to custom multimedia databases, and their daily work produces hours of material, which has to be archived in a modern, efficient fashion. It is important to mention that the ever-growing role of Internet pushes technologies which will probably force a replacement of traditional TV and radio transmission with TV and radio on-demand. The success of these technologies will depend on the effectiveness of the embedded storage and retrieval systems. Again we claim that musical audio mining technology is a core technology in this field.

E-commerce applications

On-line shopping and E-commerce is an important new domain which will dominate the future retail market. As in the case of TV and radio-on-demand, musical audio mining offers core technologies to facilitate faster implementations of efficient and reliable archiving systems. Future promotion of record companies will depend on the accuracy with which the buyer can find the recording he is looking for. Present title/number catalogues for videos, for example, will be replaced by more human-friendly catalogues which comply with ‘natural language’ search systems and ‘query-by-humming' or ‘sounds like this song’.

Hardware manufacturers

If the musical audio mining research were to awaken a genuine need in the multimedia users, hardware manufacturers would follow with implementing compatible features on their hardware. In the audio/video home entertainment equipment field, one can imagine the video camera or DVD recorder automatically indexing recorded data with MPEG-7 descriptors. The same changes may be implemented in the fields of computer hardware, and electronic instruments.

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 in providing musical commodities and services via the distributed electronic channels, there is an urgent need to develop advanced tools for music mining, that is, ways to deal with content concerning music and associated processing.

Musical audio mining is a young, but promising research field. However, due to the great variability of musical audio, its non-verbal basis, and its interconnected levels of description, musical audio mining is a rather difficult field and despite a growing group of researchers, it is still in its initial stage of development. Musical audio mining is rooted in musicology, where it draws on concept taxonomies that allow users to specify a musical piece in terms of more or less unique descriptors. These descriptors have their roots in acoustical properties of the musical audio, hence signal processing and statistical modeling are core disciplines as well because, they relate the audio to the conceptual taxonomy. As in text-based data mining, similarity measurement plays an important role in finding the appropriate connections between representational structures in the query and representational structures in the database.
The intrinsic interdisciplinarity as well as its foundation in musicology, sound and music computing, and similarity measurement give musical audio mining the status of a core research domain in multimedia. Given its relevance for the music industry and related commercial applications in multimedia, audio mining can be considered a very promising research field.

Acknowledgements
Thanks to M. Lesaffre and F. Carreras for their useful comments.

References

- Aucouturier, J., M. Sandler. (2001). Using Long-Term Structure to Retrieve Music: Representation and Matching. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA
- Balaban, M., K. Ebcioglu, O. Laske. (eds.). (1992). Understanding Music with AI: Perspectives on Music Cognition. The MIT Press, Cambridge, MA, USA
- Bigand, E., S. McAdams. (1993). Thinking in Sound: the Cognitive Psychology of Human Audition. Oxford University Press, Oxford, UK
- Birmingham, W., R. Dannenberg, G. Wakefield, M. Bartsch, D. Bykowski, D. Mazzoni, C. Meek, M. Mellody, W. Rand. (2001). Musart: Music Retrieval via Aural Queries. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA, pp73-81
- Blum, T., D. Keisler, J. Wheaton, E. Wold. (1997). Audio Databases with Content-Based Retrieval. In: M. Maybury (ed.). Intelligent Multimedia Information Retrieval. MIT/AAAI Press, Cambridge, MA, USA
- Bresin, R., A. Friberg. (2000). Emotional Coloring of Computer Controlled Music Performance. Computer Music Journal. Computer Music Journal **24** (4):44-63
- Broeckx, J. (1981). Muziek, ratio en affect - over de wisselwerking van rationeel denken en affectief beleven bij voortbrengst en ontvangst van muziek. Metropolis, Antwerp, Belgium
- Cambouropoulos, E. (2001). Melodic Cue Abstraction, Similarity, and Category Formation: A Formal Model. Music Perception **18** (3):347-370

- Camurri, A., G. De Poli, M. Leman. (2001). A Multi-Layered Conceptual Framework for Expressive Gesture Applications, Proceedings of the Workshop on Current Research Directions in Computer Music, Audiovisual Institute, Pompea Fabra University, Barcelona, Spain
- Canazza, S., G. De Poli, S. Rinaldin, A. Vidolin. (1997). Sonological Analysis of Clarinet Expressivity. In: M. Leman. (ed.). (1997). Music, Gestalt, and Computing — Studies in Cognitive and Systematic Musicology. Springer Verlag, Berlin, Germany. pp431-440
- Carreras, F., M. Leman, M. Lesaffre. (1999). Automatic Description of Musical Sgnals Using Schema-Based Chord Decomposition. Journal of New Music Research **28** (4):310-331
- Cytowic, R. (1989). Synesthesia. Springer Verlag, Berlin, Germany
- Dannenberg, R., G. De Poli. (eds.). (1998). Synthesis of Performance Nuance. Special Issue of Journal of New Music Research, Swets & Zeitlinger, Lisse, The Netherlands
- Doraisammy, S. S. Rüger. (2001). An Approach towards a Polyphonic Music Retrieval System. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA. pp187-193
- Dovey, M. (2001), A Technique for 'Regular Expression' Style Searching in Polyphonic Music. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA. pp179-185
- Ghias, A., J. Logan, D. Chamberlin, B. Smith. (1995). Query-by-Humming – Large Musical Information Retrieval in an Audio Database. In Proceedings of the Third ACM International Conference on Multimedia. ACM-95, San Francisco, California, USA
- Godoy, R. H. Jorgensen. (eds.) (2001). Musical Imagery. Swets & Zeitlinger, Lisse, The Netherlands
- Haus, G. (ed.). (1993). Music Processing. Oxford University Press, Madison, A-R Editions, Wisconsin, USA
- Haus, G., E. Pollastri. (2001). An Audio Front End for Query-by-Humming Systems. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA. pp65-72
- IFPI (2000). The Recording Industry in Numbers 2000. IFPI, London
- Jelenik, F. (1999). Statistical Methods for Speech Recognition. The MIT Press, Cambridge, MA, USA
- Klapuri, A. (2001). Automatic Transcription of Music. Proceedings of the Fourteenth Meeting of the FWO Research Society on Foundations of Music Research – Stochastic Modeling of Music. IPEM, Ghent University, Ghent, Belgium
- Kohonen, T.(1995). Self-organizing Maps. Springer Verlag, Heidelberg
- Laban, R. (1963). Modern Educational Dance. MacDonald & Evans, London, UK
- Laing, D. (1996). The Economic Importance of Music in the European Union. (http://www.icce.rug.nl/~soundscapes/DATABASES/MIE/Part1_introduction.html)
- Leman, M. (2000). An Auditory Model of the Role of Short-Term Memory in Probe-Tone Ratings. Music Perception **17** (4):481-509
- Leman, M. (2002). Systematic Musicology in the Information Society — Tendencies, Perspectives and Opportunities for Musical Content Processing. Manuscript in press
- Leman, M. (ed.). (1997). Music, Gestalt, and Computing — Studies in Cognitive and Systematic Musicology. Springer Verlag, Berlin, Germany

- Leman, M., M. Lesaffre, K. Tanghe. (2001). A Toolbox for Perception-Based Music Analysis. Ghent: IPEM - Department of Musicology, Ghent University, Ghent, Belgium. http://www.ipem.rug.ac.be/Toolbox/index.html
- Marsden, A., A. Pople. (eds.). (1992). Representations and Models in Music. Academic Press, London, UK
- Mazzoni, D., B. Dannenberg. (2001). Melody Matching Directly from Audio. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA. pp17-18
- McNab, R.J., L.A. Smith, D. Bainbridge, I.H. Witten. (1997). The New Zealand Digital Library. D-Lib Magazine. May
- Pachet, F., D. Laigre. (2001). A Naturalist Approach to Music File Name Analysis. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA. pp51-58
- Pichevin, A. (1997). Le Disque à l'Heure d'Internet: l'Industrie de la Musique et les Nouvelles Technologies de Diffusion. L'Harmattan, Paris, France
- Poli, G. De, A. Piccialli, C. Roads. (eds.). (1991). Representations of Musical Signals. The MIT Press, Cambridge, MA, USA
- Roads, C., G. De Poli, S. Pope. (eds.). (1997). Musical Signal Processing. Swets & Zeitlinger, Lisse, The Netherlands
- Rolland, P. (1999). Discovering Patterns in Musical Sequences. Journal of New Music Research **28** (4):334-350
- Rolland, P. (2001). Adaptive User Modeling in a Content-Based Music Retrieval System. Proceedings of the 2nd Annual International Symposium on Music Information Retrieval. Indiana University Press, Bloomington, USA. pp27-30
- Rossignol, S., X. Rodet, J. Soumagne, J.-L. Collette, P. Depalle. (1999). Automatic Characterization of Musical Signals: Feature Extraction and Temporal Segmentation. Journal of New Music Research **28** (4):281-295
- Selfridge-Field, E. (ed.). (1997). Beyond MIDI. The Handbook of Musical Codes. The MIT Press, Cambridge, MA, USA
- Todd, P., D. Loy. (eds.). (1991). Music and Connectionism. The MIT Press, Cambridge, MA, USA
- Toiviainen, P., T. Eerola. (2001). A Method for Comparative Analysis of Folk Music Based on Musical Feature Extraction and Neural Networks. Proceedings of the International Symposium on Systematic and Comparative Musicology III International Conference on Cognitive Musicology 2001. Department of Musicology, Jyväskylä, Finland
- Zannos, I. (ed.). (1998). Music and Signs. ASKO Art & Science, Bratislava, Slovakia