A Score-driven Approach to Music Information Retrieval

Goffredo Haus, Maurizio Longari, Emanuele Pollastri

Laboratorio di Informatica Musicale - L.I.M.,
Dipartimento di Scienze dell’Informazione,
Università degli Studi di Milano,
via Comelico 39,
20135 Milan, Italy

{haus,longari,pollastri}@dsi.unimi.it
Abstract

As the dimension and the number of digital music archives grow, the problem of storing and accessing multimedia data is no longer confined to the database area. Specific approaches for music information retrieval are necessary to establish a connection between textual and content-based meta data. This paper addresses such issues with the intent of surveying our own perspective on music information retrieval. In particular, we stress the use of symbolic information as a central element in a complex musical environment. Musical themes, harmonies and styles are automatically extracted from electronic music scores and employed as access keys to data. The database schema is extended to handle audio recordings. A score/audio matching module provides a temporal relationship between a music performance and the score played. Besides standard free-text search capabilities, three level of retrieval strategies are employed. Moreover, the introduction of a hierarchy of input modalities assures to meet the needs and the expertise of a wide group of users. Singing, playing and notating melodic excerpts are combined with more advanced musicological queries, such as querying by a sequence of chords. Finally, we present some experimental results and our future research directions.
1 Introduction

In the last few years, the contents of a growing number of music archives have been digitally encoded and stored into various formats. Not only music sheets, but even videos and audio recordings have been translated into their digital counterpart. The preservation of cultural heritage and the possibility of employing the powerful instruments of information technology are the main motivations for this transformation. However, the process of digitisation has brought to the fore the problem of organizing and accessing huge amounts of data. This being the case, multimedia information is typically unstructured for the lack of identifiable entities comparable, for example, with words. Further, the high storage requirements and the fuzziness with which people are used to manage multimedia content make that information definitely difficult to handle. Following what was happening all over the world, in 1997 LIM –Laboratorio di Informatica Musicale– of the Computer Science Department of the University of Milan started a project whose ultimate goal was the preservation of the musical heritage of Teatro alla Scala and the creation of an electronic database for audio recordings, music scores and the relevant documents (Haus, 1998b). Starting from this project, Music Information Retrieval has been established as one of the main area of original research and experimentation at LIM. There are some peculiar features that originate from those previous experiences and still exist in our studies. First of all, we lay special stress on western music and, in particular, on tonal music. This choice is not only geographical, but even cultural; in fact, the support of musical notations is assumed leaving out oral traditions and musical performances with improvisation, like Jazz. Moreover, we are aware of the fact that there should be so many approaches to MIR as the number of possible users. In this perspective, grouping users by similar skills and expectations is both necessary and desirable. A rough but useful simplification consists in dividing the intended users into three groups: musically untrained users, music amateurs and music professionals. While we refer the reader to the following for a more formal definition, we anticipate that
Pollastri, 3

this simplified representation can be recognized throughout the whole article, guiding our vision of the MIR problem.

2 An integrated environment for audio and music scores

Given the heterogeneous nature of the documents in a music archive, the first big issue to be tackled is the choice of what should be addressed by content. Accessing music materials through image processing of a video-clip could be a novelty, nonetheless it may not be the most natural way to do it! As we said, we rely on the assumption that, for each document, a music notated representation actually exists in some form. From this point of view, the central role is played by symbolic music sources; audio, textual and eventually video documents could be linked together by this common source of information. A database architecture could be based on music-objects obtained by carving up the symbolic representations into somehow meaningful elements semantically related to their musical contents. This approach introduces a \textit{conceptual} structure that allows a semantic-oriented description of the document (Bertino, Catania & Ferrari, 1999), where the semantic is constituted by musical elements like, for instance, a sequence of notes or chords. In this “score-centric” vision, there are two important facets to be considered. First, the effectiveness of the database is subject to the ability of setting out a process for the extraction of musical features from the symbolic representations of music. We should be able to include all those dimensions of music which can be roughly identified with melody, harmony, rhythm and structure. Second, all the different media contained in a multimedia database and the related textual documents must be connected to the symbolic information. Notwithstanding these difficulties, some practical methods can be automatically employed. The audio recording of a performance, for instance, can be segmented out of the music score played. When an automatic procedure can not be applied, a list of meta-data can be acquired at the time of data entering. This is the case of textual attributes like the title and the composer of a piece that will necessarily be introduced by an operator. Besides the aspects peculiar to our approach, a music database has to face the
characteristic problems of multimedia information retrieval, that is looking for similarity measures in order to handle fuzzy querying-by-content, effective database schemas and human-machine interfaces for data-browsing and retrieval. The ultimate goal of our researches is to devise a cross score/audio integrated database environment in which one can find any kind of information by singing, playing and writing score excerpts. The overall architecture is illustrated in figure{architecture}, where the fundamental unit is the music score both for indexing and retrieval purposes. Musical themes, harmonies and styles could be automatically extracted and employed as access keys to data. A hierarchy of input modalities is assured since combining multiple inputs will lead to more accurate results. Audio data is linked to symbolic information, providing an integrated system for music browsing/retrieval. The actual media storage is put offline. Audio CDs and DVDs can be seen as objects linked to the database through proper dedicated device drivers. Further, software tools dedicated to sound browsing for music professionals must be explored; such tools facilities will extend the basic database environment through a set of packages, each one devoted to a group of users. In the next sections, a brief survey of each component will be given in order to clarify their specific goals and methods. In particular, the topics that will be covered in some details are integration of audio and score sources, segmentation of symbolic music code for indexing purposes, retrieval strategies and user interfaces issues.

3 Audio Indexing from Symbolic Information

The extraction of musical attributes from audio recordings of any complexity has proved to be a daunting task. There is a large number of previous studies about automatic transcription of audio sources, but none of them has lead to very promising results in the case of real performances. So far, the translation of simple monophonic sources (i.e. one sound and one note at a time) is the only actual application. If we have to deal with the recording of a symphony, for example, there is no practical way of transcribing the melodic lines played by the
violins. Some researches have dealt this difficulty using some mid-level representation, like the introduction of a set of descriptors for the underlying acoustic events. However, these features are valid only insofar as we are concerned with acoustic properties of the signal. This approach has been successfully adopted by some of the authors of the present paper for timbre instrument classification (Agostini, Longari & Pollastri, 2001); other important applications involve the concept of audio fingerprinting, for which some commercial products are still on the market. In our instance of the music information retrieval problem, musical attributes have to be captured by the system to support querying by score excerpts, such as short tunes and harmonies. Moreover, the indexing schema followed in our studies is extracted by the symbolic music sources, so that a connection between audio and score sources must be found. Thus, an intermediate solution is needed between the “blind” translation of audio sources and the use of acoustic features. We propose to use the information contained in the music scores as an indexing template for the process of audio segmentation (Frazzini, Haus & Pollastri, 1999). The goal is to identify a temporal relationship between a music performance and the score played or, in other words, to associate each note of a score with its position in the audio file. Under the hypotheses presented in section {introduction}, it is possible to design a software tool which is able to extract such bindings between a symbolic representation and each real performance. We are currently employing a bank of notch-filters centered at frequencies specified by the pitches in the music score. The recognition of acceptable events is estimated from the energy output of each filter. Basically, we proceed with an analysis-and-refinement process as it is illustrated in the block diagram of figure {audio\score_block_diagram}. First, the beginning of each bar is recursively looked for, until a stable estimation of the bar duration is obtained. This estimated time could show a maximum variation of 30% between adjacent bars, so that musical \italic{crescendo} or \italic{rallentando} can be anyhow followed. Then, each audio segment between the beginnings of two adjacent bars is analyzed for localizing all
the notes on strong accents. The tempo of the piece of music helps us, because we can choose to estimate the time elapsed between two beats (e.g. one beat is a quarter note if tempo is 4/4). The algorithm iteratively looks into the energy curves at the output of the filters, it guesses the most likely positions of each note-event by inspecting the points above a variable threshold and it updates the estimated time for each time-unit.

Thus, for each performance we can compile a list of cues in the audio source that binds every event in a score. This approach enables us to store into the database only textual information (composer, title, date, musicians, conductor,…), the indexed music scores and a table of pointers to the CDs containing the digitalized performance. The table of pointers is extended with the list of cues from the symbolic representation to each audio source. Thus, each score fragment in the database points to all its instances recorded on CDs (see figure{query_audio_score}). Further, it is possible to design an application that allows browsing an audio file and a score file at the same time and at different levels of temporal granularity, namely notes, bars, pages and themes.

4 Data indexing and database schema

In a music archive, the sources of information are essentially constituted by text, symbolic music, audio and other multimedia resources. The indexing strategy followed in our work is built on symbolic representations of music; thus, the process of score acquisition and processing is our primary concern. We start from a graphical representation of scores that is obtained by a digitization process and it is encoded in standard bitmapped format files. The graphical formats are translated into symbolic notations by means of Optical Music Recognition (OMR) techniques. Because of the great number of errors made by the commercial OMR software, the bitmapped files are pre-processed in order to improve the results of the symbolic conversion by automatically deskewing the staves, normalizing the number of staves per system and widening the closest staves. Then, the remaining
errors are manually adjusted by means of notation editing software. (Frazzini &
The choice of a suitable electronic standard for music scores is another critical
task. Each of the existing notation formats focuses on different aspects of musical
information. NIFF format embraces the graphical representation of the score
subdividing the information in pages, systems, staves and so on; MIDI represents
performance music information that is actually an incomplete score representation;
Enigma is a proprietary notation format and it is published only partially. Since
such differences are notable, different file formats cannot be straightforwardly
put together but must be integrated at a higher level of abstraction (Haus &
Longari, 1998). This process is not only necessary for symbolic notations acquired
from music sheets, but also for the music files already in one of the cited formats
(for example, for the huge amount of midifiles available on the net). Accordingly,
we construct an abstract of the musical information defining a high level data
structure called \emph{spine}. A spine is a timed list whose nodes are made of
time representation of notes (pitch and duration). An extension of the standard
NIFF has been defined to represent this information (Haus, 1998a). The idea behind
these data structures is similar to the logical domain of SMDL (Standard Music
Description Language) (Newcomb, 1995).
Once the information contained in the symbolic representations has been structured,
a suitable form for the representation in the database environment must be assured.
For the sake of performance, we cannot afford to store into the database all the
information contained in the complete music scores. In our approach, the symbolic
information is interpreted from a musicological point of view. The resulting
semantic is the core of our system, upon which we have built all the content-
representation strategies. The idea is to find the most memorable fragments by
means of some theoretical rules. This results in a music summarization process by
which we extract the indexes to be stored into the database.
Following a method introduced by Haus (Bertoni, Haus, Mauri & Torelli, 1978), a symbolic representation of music is analyzed through melodic-operators. We apply iteratively the transposition, mirror inversion, retrogradation operators and mixtures thereof to extract the generative segments of a piece. The algorithm starts with a group of two notes, then the length is incremented and the algorithm is iterated until no other segments can be found. These segments are the most representative and they are called themes. Unfortunately, extraction of the generative elements is a time consuming task and requires human intervention in setting several complicated parameters. An improvement of this approach is achieved working at the harmonic level. We recognize the harmonic structure of a piece analyzing notes vertically. We start from the assumption that it is most likely to find a generative segment in sections where a repetition of a harmonic sequence occurs. Melodic search is then applied only to those sections. The harmonic movement is also stored as part of the semantic information, allowing content-based harmonic queries.

Another high level notion that can be exploited for accessing music is the stylistic information, for example the music genre and/or period (Pollastri & Simoncelli, 2001). This feature is by all means important in a music library and it is normally entered by an operator into an electronic database in the form of meta-data. Pursuing the goal of automating this process, some experiments have been carried out. The idea is to consider a melodic segment as a stochastic process; the style of a composer is represented by a Hidden Markov Model (HMM) trained with his/her best known tunes. An unknown segment is then classified as written by the composer for which the corresponding HMM gives the highest score of similarity.

The methodologies discussed so far allow the definition of new content-based features describing musical pieces by musicological point-of-views (Ferrari & Haus, 1999). Traditional textual attributes must be stored in the database besides this content-based information. Textual information comprises general catalogue information (e.g. title, composer) and meta-information specific for each
representation of a particular piece of music. In the latter category, we include performance information (conductor, players, ballet cast, ..), external storage information (where audio files, scanned images, notation files and videos are stored, and whether they are on-line or off-line) and other information like the quality of a recording and the state of preservation. All the attributes are linked to the composition which is the main conceptual object in our database schema. Each composition may be represented by several performances and graphic formats (e.g. different music editions). The resulting schema contains a complex structure made of a number of entities and relationships. When the information is completely specified and exact matching processing can be applied, the traditional relational database model can be employed. On the contrary, content-based features need approximate matching strategies to be retrieved. Object-oriented is a promising technology for unstructured information modeling. However, performances of Object-Oriented Database Models (OODBM) in terms of storage requirements and query processing are not comparable to those of relational database models. Moreover, OODBMs are still a topic of research and are rarely exploited in commercial database systems. In our experience with the La Scala Musical Database (Ferrari & Haus, 1999), we focused our efforts on integrating textual and multimedia data by means of object-relational database models. This technology extends the relational model to the ability of representing complex data types by maintaining, at the same time, its performance and simplicity.

5 Retrieval Strategies

Our aim is both to support a widespread group of users and to access a huge amount of data. The introduction of querying-by-content may be seen as an extension of a basic free-text search. In fact, an alphanumeric query should delimit the scope of a musical query (for example, "find all Mozart works in which \italic(this) melody occurs") or it should be restricted by it (for example, "find all composers that have used \italic(this) melody") (Ferrari & Haus, 1999). The database schema presented in section{database} assures that music objects are completely integrated
with textual attributes, so that all pieces of information regarding a particular piece of music are interconnected. For computational efficiency, given a query, we prefer to filter information starting from the textual attributes, if any, and then applying one of the following three retrieval-by-content strategies:

1- approximate string matching on melodic queries
2- calculation of music metrics on melodic queries
3- approximate string matching of a sequence of melodic and harmonic objects extracted from a musical query

At a first level, we consider the comparison between the melodic fragment proposed by the user and the stored themes as a simple string-matching problem (Pollastri, 1998). A distance measure is computed on both melodic and rhythm countour, where the former is given by the sequence of music intervals and the latter by a series of relative durations quantized by a logarithmic function. We employed the Levenshtein distance and an algorithm by Chang-Lampe (Chang & Lampe, 1992). The second level is introduced for the comparison of the melodic fragment and the themes using the distance metrics introduced by Polansky (Polanski, 1996). The comparison can be done in three modalities. In the first, the distance measures are computed only on the incipits of the themes. In the second modality, the melodic query slides over each of the themes and only the best matches are retained. The last method is an extension of the second one; the input melodic segment is shifted over the strong stresses of the themes and the distance metrics are then computed in the usual way.

The last level among our retrieval strategies enables to achieve a real musicological environment in which the musical semantics expounded in the above are exploited. By means of the melodic/harmonic information stored into the database, it is possible to design a retrieval/browsing environment in which the interconnected melodic and harmonic objects can be queried. Melodic objects can be melodic fragments (themes or tunes) or application of melodic operators to fragments. Harmonic objects can be sequences of chords or application of harmonic
operators like tonal functions and harmonic cadenzas. Thus, a user can enter a sequence of chords, obtaining a list of pieces of music with similar harmonic movement by the system. This musicological retrieval level needs some extra efforts to be implemented, for it requires a special graphic interface to manage the great number of possibilities that can be offered to the expert user. Such matter will be dealt with in the near future.

6 User Interface Issues

Designing interfaces between humans and machines is always a critical mission. If anything, it is even more difficult in a musical environment. Users of a music information retrieval system show motivations and behaviors that are hard to characterize and change rapidly. Moreover, their backgrounds are very heterogeneous, ranging from a friendly use of information technologies to a cold attitude towards them. Music professionals like composers and conductors are more likely to ask for musical structures, harmonies and thematic repetitions among different lines, while musically untrained users of a pop-music service will be more interested in the title of a piece or in the name of an artist. At the same time, users have very different skills; a musician is able to enter information via a musical keyboard or writing a score excerpt, while the layman would prefer her/his voice or a standard textual query. Since the usefulness of a multimedia system is due to the way it matches users’ expectations and skills, a system allowing different input modes must be conceived. The interface for a music information retrieval system is multimodal because it relies on audio, text and symbolic music (Haus & Pollastri, 2000). We established a hierarchy of music-interfaces, each one reflecting a particular level of expertise of the user. The textual input is placed at the top, as the easiest mode of interaction; the use of complex queries through notated music or other means lays at the bottom of the hierarchy. In the middle, there are audio and electronic inputs; employing voice is an easy and natural interaction for everyone, while musical keyboards are harder to use by non-professionals.
To better understand their needs, expected users have been divided into following three categories:

-Music professional: users that can notate music and play a musical instrument; they are able to formulate musicological queries, such as finding a sequence of harmonies or looking for a particular structure in the music.

-Music Amateurs: they can play a musical keyboard and write score excerpt, even if only at a basic level (i.e. notation of tunes)

-Musically untrained users: it is the widest category because it includes people that passively enjoy music, for example listening to the radio or looking for mp3s on the net. Nonetheless, they are interested in seeking information by standard search methods and by content.

The hierarchy we have introduced provides alternative interfaces for each group of users. Novice users may prefer to access information by textual attributes through a standard free-text search strategy or by content through a sung tune. These interfaces are easier to learn, sometimes at the expense of less effective results.

On the other hand, experts may enter some textual information and a more precise query by means of a keyboard. The tradeoff between simplicity and power is evident; we suggest that human/computer interaction and retrieval effectiveness can be greatly improved exploiting the possibilities of singing, playing and notating music in addition to standard alphanumeric searches.

Amongst such range of inputs, audio sources are the most difficult to handle. Since our basic unit of knowledge is the symbolic representation of music, audio inputs have to be translated into note-like attributes. This process of transcription can be reliably accomplished only for monophonic inputs, although we are experimenting with algorithms for polyphonic sources too. Currently, two pitch-tracking algorithms for musical instruments and human voice have been fully implemented. For technical details, we refer the interested reader to some recent papers (Haus & Pollastri, 2000). If the input is entered using a keyboard this translation is not necessary and a sequence of notes can be directly built.
7 Data-set and experiments

Given the complexity of the whole architecture, the testing phase has been accomplished through separate experiments. Each segmentation and retrieval algorithm went through a specific evaluation procedure. In the case of thematic indexing of music scores, the indicators of success/failure have been established in comparison to the existing themes in musical dictionaries (Barlow & Morgenstern, 1949). Some meaningful case studies have been deeply investigated from a qualitative point of view; for example, Symphony N.36 by Mozart, Symphony N.8 by Beethoven, English Suite N.1 by Bach and some Beatles’ songs. Figure \figure{theme} shows the first of the three themes extracted from the second movement of Mozart’s symphony; the corresponding theme in the music dictionary is shown in red notes. Notice that, at the charge of few extra notes, the theme was recognized. Since we use themes as indexing schema, it is preferable to have an excess of information than lack thereof. In these tests, we have isolated a set of special parameters for some music periods (e.g. baroque, romantic, pop-songs) that guarantees to find themes equal or a few notes longer than the ones in the dictionary. The algorithm for thematic search has been evaluated with the harmonic extension, providing a less computationally expensive implementation. In the direction of gaining some insight into the process of automatically extracting information on musical style, we have our recent work about classification of melodies by composer. A set of 720 musical themes from five different composers has been analyzed with a success rate comparable to human performances on the same task (Pollastri & Simoncelli, 2001).

Within the presented retrieval strategies, the string matching approach and the music metric have been tested with a small benchmark constituted by the themes and arias from the “Nozze di Figaro” by Mozart. The results were similar for both algorithms and the indicator of success was the usual precision measure. A test session with a database of nearly 4000 midi files is currently under way. In this experiment, we are employing precision, recall, the number of hits in the first n-matches and the coverage ratio, which are standard performance measures in
Information Retrieval. Notice that a standard set of queries to make our results comparable with other studies actually does not exist. Finally, we refer the reader to the already cited works about the tests on the other algorithms.

8 Conclusion

Far from being complete, this paper was intended to give an overview of the most peculiar aspects of our vision of MIR. Nonetheless, there are some topics currently under study that are worth mentioning. The architecture presented in section \section{integrated environment} can not handle complex audio queries like, for example, an audio excerpt from an mp3. To overcome this limitation, we are developing a package dedicated to sound browsing (Agostini, Longari & Pollastri, 2001); a natural extension of those algorithms of pattern matching and features extraction could be employed for audio comparison. Moreover, most of our acoustic features are contained into the low-level descriptors of MPEG-7 (MPEG, 2001) and we will investigate the use of standard meta-data into the database schema presented here. In the current literature, there are some significant directions of research in which we are rather interested; for example, we plan to explore the separation of well-defined audio sources within musical signals, like the singing voice (Berenzweig & Ellis, 2001) and the retrieval of orchestral music even in the presence of different performances of the same piece of music (Foote, 2000).

In the context of symbolic representation, we aim to find a unique suitable format for all the aspects of musical information. The purpose is to integrate the various standards (graphical, notational, performance and audio) and the exchanging of data with external entities. Following the guidelines defined in the similar SMDL, we are considering the XML format as a candidate tool (Haus & Longari, 2001).

Acknowledgment

This project has been partially supported by the Italian National Research Council in the frame of the “Methodologies, techniques, and computer tools for the preservation, the structural organization, and the intelligent query of musical
audio archives stored on heterogeneous magnetic media” research, Finalized Project `Cultural Heritage” (Subproject 3, Topic 3.2, Subtopic 3.2.2, Target 3.2.1). The authors gratefully acknowledge Prof. Elena Ferrari and Massimiliano Pancini for their support and the helpful discussions. A special thank to Vincenzo Marra for having carefully read this paper. The authors are mainly indebted to the current staff at LIM and to researchers and undergraduate students that made this work possible.

Bibliography


Figure(Architecture): architecture of an integrated environment for music storage and retrieval. The central role is played by features extracted by music scores (theme, harmony and style); both standard and content-based queries are supported. The audio sources are kept offline (Juke-Box for audio CDs).
Figure {Audio\score_block_diagram}: data flow of the algorithm for audio indexing from symbolic information. The number and the value of center frequencies in the band-pass filterbank are extracted from score information. Positions of the analysis windows are updated by the bar/beat estimation blocks. See text for details.
The Audio/Score match module supplies a table of pointers for each performance of a music score. This table of pointers binds score events to locations in the audio recording. Since several recordings of different performances may exist for a music score, the process must be applied to each audio recording.
figure(interface): Graphical representation of a hierarchy of musical interfaces. Five different groups of users are related to five modes of interaction. Level of expertise and degree of expectation for each group of user grow from the top to the bottom of the figure.
figure(theme): Example of theme extracted from the second movement of Symphony N.36 by Mozart. This is the first of three fragments extracted by the employed algorithm. The theme as it appears in the Barlow Morgenstern’s dictionary is indicated in red notes (gray rectangle).