Issues in content-based music information retrieval

Aura Lippincott

Glendale, California, USA

Received 7 August 2001

Abstract.

Representing a significant shift from methods that employ bibliographic metadata for music retrieval, modern Music Information Retrieval (MIR) research seeks to utilize characteristics of music to search databases of musical content via musically expressed queries. These investigations hold significant promise for the future development of commercial and research applications such as searchable databases of popular tunes on the web and databases of historical music. However the challenges are diverse, encompassing music cognition, storage and processing, and traditional information retrieval (IR) issues. This paper focuses on IR issues in MIR, discussing various approaches to music representation and matching techniques. Prototype MIR systems Meldex, Semex, Themefinder and Arthur are briefly described.

1. Introduction

In the past, users seeking information about music turned to print sources containing metadata hand-recorded and arranged by title, composer, and other categories. Obviously access methods mirrored print-based retrieval techniques for bibliographic information retrieval of the time and similarly presupposed some prior musical knowledge or access to a librarian. Much current research into automated Music Information Retrieval (MIR) is based upon characterizations of the music itself, rather than information about it. For instance, instead of requiring a search by song title, a user inputs a query in the form of audio and retrieves results similar to that query. The implications for average users of content-based music retrieval systems are significant because prior bibliographic knowledge of a piece of music is not necessary; rather a bit of music running through one's mind suffices for retrieval purposes. This paper will look at some content-based music retrieval research focusing on issues in music representation and matching techniques and will describe some prototype MIR systems.

The complexity of music and of how it is understood and comprehended are important to bear in mind when considering the specifics of MIR research and system design. For example, the layperson and musician may each understand music in a different way, so upon whose comprehension do you base your system? Answers to questions like this one are addressed in terms of the research issues focused on in the paper, although they do constitute important research areas themselves, such as user interface design, presentation and evaluation of results. Other significant areas of research beyond the scope of this paper are database issues and music or audio feature extraction techniques.

2. Historical background

The earliest history of music information retrieval begins with bibliographic publications in print form. Often arranged by title, composer and possibly genre, these reference works are not based on representation of the music itself, rather on metadata describing the music. Early music could also sometimes be identified through the use of thematic catalogues of incipits, which are fragments of melody or themes symbolically represented. Thus manual retrieval systems of this sort

Correspondence to: A. Lippincott, 933 North Glendale Avenue, C, Glendale, CA 91206, USA. E-mail: alippinc@library.ucla.edu

required knowledge of musical notation to use. Thematic catalogues are still in use today both in manual print form and as the basis for some automated MIR systems.

Investigations of automated music retrieval began in earnest in the 1960s. A very early problem faced by researchers was the lack of databases of music data with which to conduct research and test theory. Once technical database storage issues were sufficiently worked out, researchers had decisions to make about what constituted meaningful music representation fit for database storage and computer processing. The outcomes of these initial decisions form an important basis for distinguishing MIR systems, especially as methods of music representation evolve over time and as new representational methods arise. As might be expected, early automated systems were built around searching databases of symbolically notated music fragments, such as thematic catalogues of incipits [1]. However, as Downie [2] notes, there is some choice in how one symbolically represents music in an incipit, such as by sequence of notes or pitch information. So even though the kind of musical representation was decided, the representations consisted of different types and degrees of encoded musical information. This meant, in fact, that each means of representing the incipit included only a portion of the musical information to be found in an incipit, which in turn represented only a fragment of a piece of music. Regardless of their incompleteness, Downie [2] points out that thematic systems gave users the ability to access musical information on its own terms, meaning framed in musical queries, rather than through bibliographic description.

MIR research exploded in the late 1990s, perhaps in response to interest in digital libraries and the potential of networked communication, specifically the Web, to reach a wide commercial audience of music lovers. Modern research into content-based MIR borrows from three fields: traditional IR, musicology, and music perception [3]. In general, musicology's contribution is in the form of research about the physical structure and characteristics of music, thereby contributing ideas about musical representation, similarity and, by extension, matching. Music perception or cognition research contributes theory about how humans perceive and remember music. This is important when thinking about how musical queries are formed and how MIR systems are evaluated. It is also important because the richness of musical performance may be as much a function and outcome of the inherent physical characteristics of music as how the human brain apprehends it. If this is the case then useful musical representation is also intimately related to the study of musical perception. Traditional IR contributes theory about matching algorithms as well as evaluation methods based upon recall and precision and more broadly the design of functional retrieval systems.

3. The problems

MIR shares many of the problems (along with some of the solutions) of traditional IR efforts, albeit with its own significant complexities. The problems addressed in this paper include document and query representation and matching algorithms. This section will examine the problems in each of these areas that are specific to music, as well as some solutions posed.

3.1. Music representation

Illustrating the nebulous nature of music as a concept, consider the following question. Is a musical performance on CD 'music' or is 'music' a series of notes in a certain tempo and pitch that a musician plays? If music is the former, then a musical representation of melody consisting of notes, note duration (i.e. length or tempo) and pitch that suffices for a musician's comprehension is not fully capturing what the average music listener is hearing and probably understanding to be music. However, this representation is capturing the unique melody of a song, which might be sufficient for retrieval regardless of a user's particular musical education or comprehension. Actually, music can be understood as containing both the audio representation of a CD and the symbolic representation of a score. Unfortunately symbolic representations leave out important music information (and are not 'user-friendly') and 'full-text' digital audio is too complex and anyway cannot be stored as 'sound'. Because full-text audio information cannot be utilized, a choice must be made at the outset about which information to represent, consequently affecting the design of the system and the audience. Bearing this in mind, what various MIR system approaches aim to do is represent music in a form that allows a particular piece to be distinguishable from another and in turn, similar to a query. This probably means finding something unique about it (that serves to differentiate it) then measuring how documents (including queries) are similar to one another or perhaps how different they are based on these quantified or symbolically represented attributes.

Content-based MIR systems commonly capture melody information, which constitutes a unique

combination of notes, pitch and note duration. Any or all of these three melody attributes can be encoded in a variety of ways. For example, pitch may be distilled to the point where it reflects only direction (i.e. up or down). Decisions like these hinge both upon music perception research (i.e. the degree to which humans perceive, remember and can reproduce pitch information) and mathematical matching algorithms that account in varying degrees for how approximate or exact the matches to these representations must be.

There are many other means of representing music, including MIDI, digital audio 'long-term structure' and digitized symbolic notation formats such as '**kern', which can in turn be representative of musical fragments or full text. The earlier mentioned incipits are symbolically notated fragments that capture the beginning of a piece of music or important themes. They may be made up of the space between notes (intervals) or pitch changes. Because incipits are theoretically unique they lend themselves well to automated retrieval systems once they are digitally encoded. However, they are not well suited to all types of music or all types of MIR system users. MIDI is a popular choice for MIR researchers because it is a standard-based digital communications protocol (already digitally encoded) that captures musical performance data, such as notes turned on and off, pitch and note duration. Standard MIDI saves this information in a generic file that may be interpreted by any program that supports the Standard MIDI File. Because of this flexibility, MIDI files are widely available on the web, making them available for gathering into databases. It is crucial to note that MIDI files are not digital audio files, therefore do not fully capture the richness of musical performance. Although not representing full-text audio, Foote's [4] 'long-term structure' approach nevertheless utilizes analysis of audio waveforms, rather than symbolic or MIDI representations. Long-term structure is based on representing 'variation of soft and louder passages' resulting in a unique 'energy profile' that may be compared to other energy profiles. Not only does this approach represent a move away from intermediate representations of music in the form of MIDI and symbolic representation, but significantly, also moves away from representing melody. This illustrates an important distinction between different content-based MIR approaches, namely, whether an approach utilizes the structure of sound (audio) or the structure of music (melody, for example). Although Foot's approach uses audio information, similar to the musical (or melody) approaches it does not utilize the full range of information consisting instead of select representations of the information found in sound waves. Although the utilization of 'full-text' audio is a goal of MIR it is beyond current capabilities due to its complexity and processing issues.

3.2. Music queries

The issues involved in music queries are ones of both representation and user creation. Musical query representation is directly related to document representation for the obvious reason that content-based MIR system algorithms are designed to match query and document representations, thereby treating them equally similar to traditional IR.

In terms of human users of MIR systems, the issues of how to create a query are especially important and depend in part on who is expected to use the system. At the most basic level, system designers must keep in mind the musical expertise of the user. Thus some queries take the form of hummed audio input, while others are inputted using keyboards. In either case, the query probably needs translation to match the representation of the music documents in the database. In effect, an audio query is being used to match documents that in the case of MIDI and symbolic approaches do not constitute audio in the first place. Even after translation the query is likely to be inexact both in terms of what the user has in mind and in terms of the actual music document it is meant to retrieve. Furthermore, depending on the input method, it may not contain one or more of the unique attributes of melody. This proves to be one of the chief challenges to overcome, or at least account for, when designing matching algorithms. Another challenge is that 'features that matter most to listeners pertain to a particular performance of a musical work rather than to inherent features of the composition' [1]. It is not clear that this issue has been, or can be, sufficiently addressed.

3.3. Matching techniques for music

Theoretical approaches for matching algorithms in MIR systems have taken many difference paths. Underlying each are the complex assumptions mentioned earlier about what exactly is being compared based on what musical information is being deemed as representative of the music. Algorithms are also often tested or developed on particular databases of uniform music or genres in which the music has similar structural characteristics. Of the approaches examined herein, there is little discussion of scaling to databases of

heterogeneous musical styles and infrequently to different homogeneous genres of music. Thus it is possible and sometimes admitted that the different approaches described in brief below will not work well on music outside of the genres or contexts in which they were developed.

In general, matching algorithms have to contend with maintaining computing efficiency in a highly complex environment, matching inaccurate user input to accurate musical representations and sometimes matching monophonic (one melody) query representations to polyphonic (multiple melodies) documents. Often, where an approach relies on melody similarity in matching queries to documents, approximate string matching techniques are utilized to account for both query inaccuracy and length (where short queries must be matched to long documents). Uitenbogerd [5] notes 'the problem of matching fragments to music is made difficult be the psychology of music perception'. This, he continues, is due to the fact that 'literal matching may have little relation to the perceived melody similarity'. Of course, perceptual issues can be approached in the design of matching algorithms; but they cannot be completely accounted for. Unfortunately, the problem with approximate string matching as a computational technique to account for various inaccuracies, is that it is inefficient and does not necessarily yield better results than using alternative algorithms [6].

ÓMaidín's [7] approach to music matching is based on determining melodic similarity in a database of folk songs stored as scores. He formulates a 'geometrical algorithm for melodic difference' in which melodic segments of equal length are compared based upon three factors of characteristics of the segments. These include 'juxtaposition of notes in the segments, pitch difference, and note duration'. The algorithm is based upon assigning weights to the note length (or duration) then evaluating the melodic difference. Clearly there are limitations to this approach in terms of the highly restrictive conditions placed on the comparisons, such as equality in melody fragment length.

The approach of Crawford *et al.* [8] is based on using string-matching techniques borrowed from biological and technical sciences. They are in search of the 'characteristic signature' of a 'musical entity' in which patterns of notes, for example, are assigned weights and the sum of weights constitutes the characteristic signature. Unlike the previous example where ÓMaidín compares two fragments, Crawford *et al.* are interested in determining the unique characteristics of a particular musical structure. This fingerprint would not only allow it to be distinguished from all other musical structures, but according to the authors, would provide a basis for understanding human recognition of distinct musical entities. Instead of basing their approach on a homogeneous database of musical tunes, the authors' claim (untested in the article) is that the characteristic signature could be extracted from unstructured audio, symbolically encoded source data, or MIDI commands.

Downie's [2] approach treats musical matching as a text-matching problem based upon traditional text-based IR systems. He represents a monophonic melody as a collection of intervals split into 'n-grams' or 'discrete units of melody information taken from anywhere in a melody'. The n-grams, of varying lengths, are then treated as words for purposes of applying text-based IR techniques of matching. The limitations inherent in this approach are similar to others that rely on melody in that they significantly reduce music to characteristics that do not scale well to all types of music.

Foote's [4] approach again deals with audio similarity rather than symbolic or MIDI similarity. He uses a 'Gaussian model' that sets the captured audio representation (via an 'energy profile') in a vector and measures the 'euclidean distance' between the representations. Similar to Crawford's approach and to text-based vector analysis, Foot's system analyzes each document (including queries), calculating a unique variable, then indexes every document based upon the unique variable, and finally measures the distance of the (query document) to each other document to determine levels of similarity.

4. Some MIR systems

Parallel to the discussion up to now, MIR systems are 'widely varying in approach' [2]. This section will review some of the functional prototype systems currently available, including Meldex, Semex, Themefinder and Arthur. Once again, illustrating the breadth of research endeavors in the field of MIR, these systems can easily vary in their specific techniques of music representation and matching from the ones reviewed in previous sections. However, they tend to fall within either the 'music as audio' or 'music as melody' distinction already drawn. Other MIR systems include Tuneserver (see www.ipd.ira.uka.de/tuneserver/) and Muscle Fish (see www.musclefish.com), which each fall into one camp or the other.

The New Zealand Digital Library Meldex system is built around a database of over 9400 folk song scores (http://nzdl2.cs.waikato.ac.nz/cgi-bin/gwmm

?mt=music&c=meldex&a=page&p=query&qaq=0). The system accepts user-specified musical queries in the form of a short sung, hummed or keyboard input. This query is returned to the user in musical notation along with a ranked list of musically notated songs similar to the query. Matching of query and documents is based on similarity of 'melodic contour', intervals or rhythm (tempo) as represented in musical notation. The system uses both exact and approximate string matching techniques similar to those used in traditional IR. The problems inherent in exact matching are the variability of musical performance and imperfections in user input. Therefore Meldex uses less computationally efficient approximate string matching to overcome these issues. The Meldex researchers are especially interested in the 'number of notes required to identify a unique melody' [9]. The obvious limitations of the Meldex system are that it is based on musical scores rather than audio-based musical performances. In addition, the system requires that queries be translated to match the internal representation of the database, thereby introducing user error into the process.

Semex (Search Engine for Melodic Excerpts) is a prototype system that represents music as pitch levels or notes and supports three file formats that contain this musical information, including MIDI (not online at this time [12]). The matching algorithms used by Semex are also based on the general theory of string-matching, but more specifically use a computational technique called 'bit parallelism' to deal with the efficiency problems of long document lengths required to match imperfect queries. Semex claims to improve on systems like Meldex in terms of efficiency and discriminatory power. The approach importantly differs from Meldex in that it allows matching of monophonic queries to polyphonic music.

Themefinder differs from both Meldex and Semex in that it searches a database of themes, specifically seventeenth, eighteenth and nineteenth century fragments, as one would find in a printed thematic catalogue. The data is represented in the '**kern' format, consisting of notes, rests and barlines as found in Western musical notation (from the explanation of **kern at www.lib.virginia.edu/dmmc/Music/Humdrum/kern_hlp.html#kern). Themefinder allows users to combine precise or 'fuzzy' melody contour (pitch direction), note letter name, pitch class with bibliographic limits. In fact, query results are greatly improved as these limits and parameters are used. Themefinder is an example of a system that, although limited in its application outside of thematically represented Western music, nevertheless improves its usefulness within this domain by combining and offering multiple retrieval techniques to users.

Finally, Arthur is a system designed by Foote [4] that utilizes the audio waveform analysis of long-term structures described in an earlier section on a database of 'classical phonograph recordings'. The shortcoming of this system, according to Foote, is that the analysis relies on softness and loudness variations in music, and therefore is limited to music (in this case classical) that is uniformly structured to contain this audio waveform characteristic. Nevertheless, it provides an alternative to retrieval by melody similarity and represents a growing area of MIR research.

5. Significance/success of projects

As Downie notes, 'one thing that unites . . . these approaches is that they have some kind of shortcoming. The more powerful analytic systems can be very difficult to use, incipit and thematic indexes can leave out a very large amount of music that might be of interest, and approximate string matching can be computationally expensive without necessarily giving better results' [10]. A major drawback referred to as problematic throughout the paper is that, although many modern MIR systems are searching musical content of sorts, they are far from searching the music found on a CD. Even those that search audio representations of music are not capturing 'full text' and are surely missing characteristics crucial to human musical comprehension and crucial to the structures that make audio music. The experimental systems discussed above are either based on theory about music structures or audio structures, not both.

The limits of automated music retrieval are analogous to those of automated text retrieval in that neither have the ability to distinguish nuances in perceptual meaning. People are potentially querying a database for music in their heads that is far more complex that its representations. For instance, a user may wish to find all of the performances of a song that has been performed multiple times by different musicians in different genres (some not melodically based) with the potential for very different sounding results. The first problem is how to turn that information need into a query. Perhaps that can be overcome by enabling the system to accept an imperfectly hummed rendition of a fragment that may occur anywhere in the song. Secondly, what do these different renditions have in common and how do you selectively represent that in a database? Finally, how do you match the user's

fragment to this database of selective music information. This said, it might be entirely possible to design a system that is capable of adequately and usefully meeting the demands of most users looking for melody-based genres of music, but what really connects these songs in the example above is the title! In conclusion, given the complexity of music there is a significant amount of research to be done before MIR systems are capable of flexibly searching and retrieving a heterogeneous database of tunes spanning the universe of present and future musical genres.

6. Conclusion: future directions

MIR systems are considered to have a very wide range of possible applications from searchable consumer-oriented web-based databases of popular tunes to research databases of historical music. MIR systems research is also of interest in digital library research. Aside from retrieval systems, research into music information is used for music copyright cases [11] and to determine historical music attribution.

Within the focus area of this paper, research goals appear to center on creating efficient algorithms, dealing with complex musical representations of varying genres all within the ever-present context of music perception and meaning. Processing issues aside, the real challenge for MIR systems is the complexity of music analogous to that found in language and the multiple meaning of words. Like language, music has hidden meaning. However, music is even more complex in that each rendition adds a layer of complexity. Although a sophisticated language of discourse has been built up to describe and talk about music there is really no way to share and convey its personal meaning. At the same time that music is a universal language, the language of music lacks a codified universal meaning. The physical phenomenon of music is best handled by MIR systems, while the meaning of music, although touched by research into music cognition, is beyond MIR systems to master at this time. Perhaps MIR systems of the future will combine bibliographic metadata, audio content and music notation in powerful retrieval systems that provide a solution to the problems of cross-genre musical searching.

References

[1] E. Selfridge-Field, What motivates a musical query? *International Symposium on Music Information Retrieval*, Plymouth, MA, 23–25 October 2000. Available at: http://ciir.cs.umass.edu.music2000/ (accessed April 2001).

[2] S.J. Downie, Access to music information: the state of the art. *Bulletin of the American Society for Information Science* June/July 2000.

[3] A.L.Uitdenbogerd, A. Chattarajam amd J. Zobel. Music IR: past, present and future. *International Symposium on Music Information Retrieval*, Plymouth, MA, 23–25 October 2000. Available at: http://ciir.cs.umass.edu/music2000/ (accessed April 2001).

[4] J. Foote, Arthur: retrieving orchestral music by long-term structure. *Proceedings of the International Symposium on Music Information Retrieval*, Plymouth, MA, 23–25 October 2000.

[5] A.L. Uitdenbogerd and J. Zobel, Manipulation of music for melody matching. *ACM Multimedia 98 – Electronic Proceedings*. Available at: www.acm.org/sigmm/MM98/electronic_proceedings/uitdenbogerd/ (accessed April 2001).

[6] S.J. Downie, *The Exploratory Workshop on Music Information Retrieval. Call for Participation* (ACM SIGIR, Berkeley, CA, 1999).

[7] D. ÓMaidín, A geometrical algorithm for melodic difference, *Computing in Musicology* 11 (1998) 65–72.

[8] T. Crawford, C.S. Iliopoulos and R. Raman, String-matching techniques for musical similarity and melodic recognition, *Computing in Musicology* 11 (1998) 73–101.

[9] R.J. McNab, L.A. Smith, D. Bainbridge and I.H. Witten, The New Zealand Digital Library MELody inDEX, *D-Lib Magazine* May (1997). Available at:www.dlib.org/dlib/may97/meldex/05written.html (accessed April 2001).

[10] S.J. Downie, Music retrieval as text retrieval: simple yet effective. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA, 15–19 August 1999.

[11] C. Cronin, Concepts of melodic similarity in music-copyright infringement suits', *Computing in Musicology* 11 (1998) 187–210.

[12] K. Lemström and S. Perttu, Semex – an efficient music retrieval prototype, *International Symposium on Music Information Retrieval. MUSIC IR 2000*, Plymouth, MA, 23–25 October 2000. Available at: http://ciir.cs.umass.edu/music2000/ (accessed April 2001).