Comparing Aural Music-Information Retrieval Systems

Bryan Pardo
EECS Dept., University of Michigan
1101 Beal Avenue, Ann Arbor, MI 48109 USA
+1-(734)369-3207
wpb@eecs.umich.edu

Colin Meek
EECS Dept., University of Michigan
1101 Beal Avenue, Ann Arbor, MI 48109 USA
+1-(734)763-1043
meek@umich.edu

William Birmingham
EECS Dept., University of Michigan
1101 Beal Avenue, Ann Arbor, MI 48109 USA
+1-(734)936-1590
wpb@eecs.umich.edu

ABSTRACT

Aural queries, often called "query by humming", are a popular input to music information-retrieval systems. These systems perform a complex series of operations, from transcribing input to retrieving pieces. We suggest a test-bed needs to represent each data-processing stage, from initial transcription of an audio query, to the relevancy ranking of each piece in the database compared to a particular query. Such a multi-stage dataset would let each research group concentrate on the portion of the task most interesting to that group, while at the same time providing training and testing data for each component of the system.

1. INTRODUCTION

We are interested in creating systems that retrieve music based on an aural query provided by a user. The query is assumed to be a theme, hook, or riff from the piece of music the user wants to find. It is assumed that lyric information is either not provided or not used by the system. This has been called "query by humming" in the literature [1], although singing, whistling, etc., are all forms of aural queries. A system must transcribe the aural query and search for related pieces of music in a database, returning "similar" pieces by some similarity measure. Figure 1 outlines the steps in the aural query task.

In recent years, music retrieval from an aural query has been investigated by a number of groups [2][3][4][5]. Unfortunately, the lack of a standard test-bed and method of performance evaluation makes it impossible to directly compare the performance of the various systems described in the literature (we call all these systems aural music-information retrieval systems, or AMIR, for the remainder of this paper).

In Figure 1, data are represented by rectangles and operations are represented as ovals. We propose a standard data set for each rectangle in the diagram: a standard set of audio recordings of queries, as well as a standard set of transcriptions of those queries into some kind of (possibly) note-based representation.

Similarly, the song corpus would be represented both in an unprocessed format (audio files or bitmaps of sheet music) and as some kind of agreed-upon structured data (MIDI, GUIDO). Ideally, the corpus would include a complete ordering of the pieces in the corpus for each query, based on similarity between query and song. This ordering would be assumed "correct" and be a measure against which systems could be tested. We suggest that the ordering be done by a group of experts familiar with the corpus. Moreover, we suggest that there may be multiple criteria for ranking the songs, and ordering for each criterion may be necessary (e.g., rhythm, pitch contour, etc.)

There are considerable difficulties with this tack, as it is clear in many cases that a particular song is the (sole) intended target for a search: it may be difficult to establish a corollary concept of musical query relevance to the more established relevance criteria for text documents.

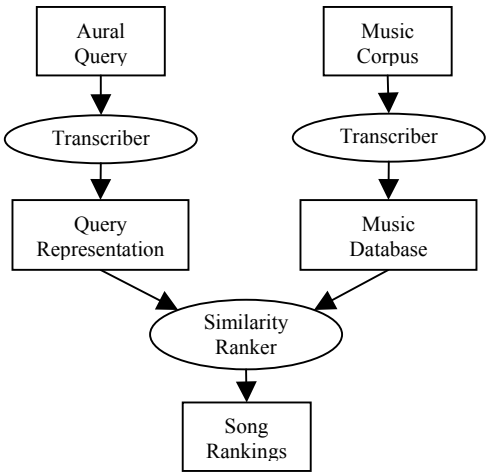

Figure 1.

A multi-stage dataset, such as we describe, would let each research group concentrate on the portion of the task most interesting to that group, while at the same time providing training and testing data for each component of the system. For instance, a group working on query transcription would benefit from "correct" reference query transcriptions for testing and training purposes, and a group working on similarity ranking could use the same dataset as input to their component of the system. Also, with clear boundaries between subtasks, it would be possible to "mix and match" the work of several research groups working on different aspects of the problem, finding the best combination of (for example) OCR transcriber from sheet music, pitch-tracker for hummed queries and similarity ranker.

New representations for queries (or pieces) could be compared to the standard representation to see how a change in the query representation affects retrieval performance. An example would be a comparison of representing pitch contours using "same, up, down" versus representing pitch contour by MIDI not number. This would let researchers begin to tease apart the interaction between how the data is represented and the kind of similarity ranker used to find matches.

The remainder of this paper outlines details we feel are important in designing a standard test-bed.

2. QUERIES

One of the most common music information needs involves finding a piece of music without knowing the song title or the artist name. Typically, a person will give some genre/instrumentation information ("it is sung by a Gospel choir") and then sing a portion of the piece ("la la la doo bee Oh Lord, Oh Lord"), as best recalled. Since most people cannot read or write music, nor can they play a song "by ear" on any given instrument, it is likely that the typical person can best transmit musical content information by singing, whistling, or humming.

We suggest that an initial set of queries should consist of the sung/whistled/hummed information. It may be assumed, for the time being, that other information provided about genre, instrumentation, etc, can be provided by having the person enter such information in response to questioning by the information retrieval system ("Click the box next to the genre of music you wish to search"). Provided genre information could then be supplied as metadata, along with the sung query.

It might eventually be useful to support an interactive query process, wherein the user refines their query and the search space based on the results returned. However, the test-bed we envision would not easily support such a paradigm and we do not see a clear way to do so without requiring the test-bed to cover an unwieldy number of queries and query formats.

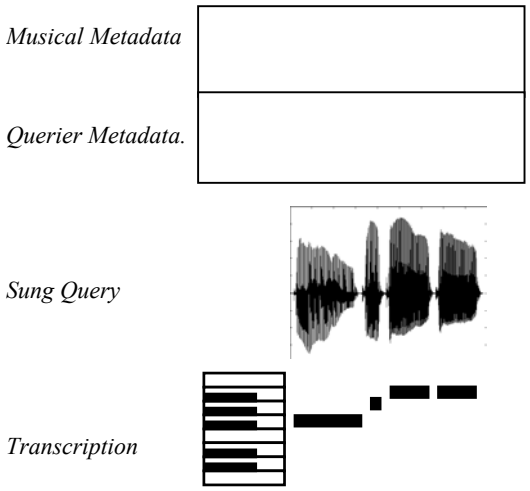

Figure 2.

Figure 2 illustrates a query format that would be useful for a test-bed. Here, the sung query is stored as a .wav file, musical genre, instrumentation and some information about the person performing the query are stored as metadata. A transcription of the sung query into pitch, timing and syllable information is also shown. Ideally, such a transcription would be done by an expert human from the sung query and provided for each query. This would allow researchers interested in honing their transcribers a "gold standard" to measure their systems against. Those uninterested in transcription could simply something that has already been transcribed.

One key piece of metadata is that of the ability of the person performing the query to duplicate a melody by singing. People interested in finding pieces of music range widely in this ability, as anyone who has heard karaoke or a group of people singing "Happy Birthday" can attest. Thus, the population of users will have a strong influence on the difficulty of the retrieval task. In our current research, we perform a simple singing test on each user, to determine his or her ability to reproduce the pitch contour of a melody. Users are then divided into groups, based on ability. The query representation and retrieval method can then be tailored to the musical skill of the user, and performance of the system may be measured given user skill sets. For example, a person may have good ability to duplicate rhythm, but poor ability to sing pitch intervals correctly. A matching algorithm could, thus, give more weight to a rhythmic match than the pitch contour. We recommend that each query in a standard query set be categorized by the ability of the person performing the query, as measured by some agreed-upon set of standards.

3. THE MUSIC CORPUS

While we are neutral as to the pieces selected for the database for a test-bed, we argue against a homogeneous database. Genre, (rockabilly vs. madrigal, vs. house music), instrumentation (string quartet vs. gospel choir, vs. synthesizer), musical approach (whimsical vs. dramatic) and extra-musical considerations (on the soundtrack to "The Big Chill" vs. on the sound track to "Dirty Dancing") can all strongly determine how a listener perceives the similarity of two pieces of music. Other dimensions are not hard to come up with (length of piece, lyrical content).

Given that we are interested in query by humming, we would like music collections to consist of "singable" pieces. Thus, Broadway show tunes would form a much better corpus than would twelve-tone piano pieces by Webern.

As to the initial format of the corpus set, we are again neutral as long as the format can be automatically generated from the formats used by humans in their typical interaction with music. Obvious candidates include audio files, (.wav, .mp3) and bitmap scans of sheet music (.bmp, .jpg).

As with queries, metadata (such as the title and genre of the piece), along with a transcription into an MIR-friendly format would be stored along with the original .wav file. Also, since queries are likely to be based on some catchy hook or theme, it would be useful for people engaged in automated meta-data creation [6] to also annotate songs with information about motif and theme.

4. SONG RANKINGS

Ideally, the distance between each query and each piece in the database would be measured by humans and stored in the benchmark set of score rankings (see Figure 1). Distance would be a vector in a space based on rhythm, pitch contour and syllabic content (lyrics). Each element in the vector would represent distance between the query and the piece along a particular dimension. Example dimension are pitch contour, rhythm, and syllables sung. It is unreasonable, however, to assume a non-toy database with a complete ordering of the pieces in the database according to their distance from the query along each dimension

As a simplification, we suggest simply listing the correct piece for each query.

5. EVALUATION

At each level of processing, there should be standard performance evaluations. For example, a transcriber from the audio query to a representational system would be measured against a hand-transcription that is presumed correct. Multiple evaluative measures will likely be useful, since various retrieval systems will likely require different performance specifications (for instance, one method may be tolerant of pitch error, while another is tolerant of rhythm error). That said, for a given task, a standard evaluation measure should be specified, so that systems performing the same task may be measured by the same yardstick. To avoid bias, evaluation should be performed automatically, without human intervention.

6. CONCLUSIONS

We suggest a test-bed be created that is specific to aural music information retrieval. This test-bed would have a standard set of audio queries (stored as audio files), with associated transcriptions, a standard set of music pieces (audio files or score bitmaps) and associated transcriptions, and a mapping between each query and the right member of the set of music pieces. Such a multi-stage dataset would let each research group concentrate on the portion of the task most interesting to that group, while at the same time providing training and testing data for each component of the system. Given this standard set of data, automatically performed standard performance measures should be agreed upon. This would let research groups in aural music information retrieval compare experimental results in a meaningful, relatively unbiased, way.

7. REFERENCES

[1] Ellen Vorhees. *WhitherMusic IR Evaluation Infrastructure: Lessons to be Learned from TREC*. Workshop on the Creation of Standardized Test Collections, Tasks, and Metrics for Music Information Retrieval (MIR) and Music Digital Library (MDL) Evaluation at the 2002 Joint Conference on Digital Libraries (JCDL2002), Portland, OR, 2002.

[2] Birmingham, W.P., Dannenberg, R.D., Wakefield, G.H., Bartsch, M., Bykowski, D., Mazzoni, D., Meek, C., Mellody, M., Rand, W. Musart: Music Retrieval Via Aural Queries, in Proceedings of ISMIR 2001 (Bloomington, IN, October 2001), 73-81

[3] McNab, R. J., Smith, L. A. et al. Towards the digital music library: tune retrieval from acoustic input. Digital Libraries, ACM. 1996.

[4] Clausen, M., Englebrecht, R. et al. Proms: A web-based tool for searching in polyphonic music. Proceedings of the International Symposium on Music Information Retrieval, 2000.

[5] Tseng, Y. H. (1999). Content-based retrieval for music collections. SIGIR, ACM. 1999.

[6] Meek, C., Birmingham, W. Thematic Extractor, in Proceedings of ISMIR 2001 (Bloomington, IN, October 2001), 119-128.