Tracks and Topics: Ideas for Structuring Music Retrieval Test Collections and Avoiding Balkanization

Jeremy Pickens
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst
jeremy@cs.umass.edu

ABSTRACT

This paper examines a number of ideas related to the construction of test collections for evaluation of music information retrieval algorithms. The ideas contained herein are not so much new as they are a synthesis of existing proposals. The goal is to create retrieval techniques which are as broadly applicable as possible, and the proposed manner for creating test collections supports this goal.[1]

1. INTRODUCTION

One of the fundamental problems encountered by music information retrieval system designers is that the representations for and sources of music are incredibly diverse. Music may be monophonic or polyphonic. It may be represented as digital audio, (digitized) analog audio (for example from old scratchy record or hissy tape collections), conventional music notation in symbolic/computer-readable format, conventional music notation as scanned images (sheet music), and event-level music such as MIDI, to name a few. One may have access to a full piece of music, or only to a snippet, such as a chorus or an incipit. Pieces of music may occur in the same key, or they might exist in numerous different keys. Pieces might be played or otherwise represented in a wide variety of tempos, or they might all be normalized to a single tempo. Depending on the source of the piece, there might be different types of errors in the final representation: users humming a piece will produce one type of error, automated transcriptions of audio could produce another type of error, and automated transcriptions of digitized sheet music could produce yet another type of error.

The combinatorial possibilities of these music sources are enormous. One such combination might be "incipits of polyphonic, MIDI music, normalized to C-Major but left in their original tempos," Another combination might be "monophonic full tunes, in audio format, unnormalized in any way, hummed by first year university music students". A rough combinatorial enumeration yields a total of 240 different varieties or types of music chunks. If we add to that the fact that the query could be of a different type than the music source collection (for example, the query might be audio and monophonic, while the collection has only sheet music, polyphonic pieces), then there are $\binom{240}{2} = 28,680$ possible experimental configurations.

Clearly the music retrieval community does not have the resources to build even a couple dozen different test collections, much less twenty-eight thousand. Nevertheless, the varied types of systems being built by the community continue to proliferate. Systems are being built which work only for monophonic music, or only polyphonic music, or only audio music, or only key-transposed music, and so on. While these are necessary first stages in such a new research community, the goal should be to produce retrieval algorithms which are robust to the various music sources and representations. Otherwise, the community risks balkanization of the retrieval process and the creation of algorithms myopic in scope and unable to function outside of their narrow, specialized situations.

2. BACKGROUND

In this paper the focus is on ad hoc retrieval experiments. There are certainly many other important MIR-related tasks, such as automated audio transcription, automatic clustering and heirarchy creation for user browsing, and so on. For the sake of this discussion, we are focus on the ad hoc task, defined as new queries on a static (or nearly static) collection of documents. The collection is known, a priori, but the query which will be asked is not. The Cranfield model is the standard evaluation paradigm for this sort of task and was outlined in the 1960s by Cleverdon et al [1]. Along with many others in the MIR community, we support this model for music information retrieval evaluation.

TREC (Text REtrieval Conference) expanded upon these ideas, providing a centralized forum in which large, standardized collections could be assembled [4]. Rueger [6], Sodring [7], and Voorhees [8] have spent a good deal of effort explaining how TREC-style experiments may be applied to the MIR context. We wish to synthesize the lessons learned in these discussions with additional requirements suggested by other researchers.

In particular, Futrelle explained that "The goal of any basic research in MIR should be to develop techniques that can fit into a broad and comprehensive set of techniques. An

[1]This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #IIS-9905842. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

evaluation of an MIR technique should situate itself in this larger context, and should acknowledge the implications the results have for the technique's role in a broader and more comprehensive set of techniques.[3]" We wish to integrate this goal with the advantages of TREC-style evaluation by emphasizing the notion of a *track*, as we will explain in section 3.

Perhaps the most similar proposal to ours is the Melucci and Orio task-oriented approach [5]. In their paper, they propose identifying and separating queries into separate tasks by their "information requirements", or the broad category by which similar pieces of music will be found. For example, a certain query might be identified as a "melody" query, meaning that relevant documents will be melodically similar. Or another query might be identified as a "rhythm" query, meaning that relevant documents are going to be rhythmically similar. This is a very important distinction to make, as the same exact piece of music may be used as a query, but depending on a user's infomation requirement (need) a different algorithm will have to be built. At the same time, systems should not be so specific that you need a different type of system for every single subspecies of information requirement. For example, there should not have to be one information retrieval system for jazz melodies, another for classical melodies, and yet another for folk melodies. Knowing the broad information requirement, the fact that melodic similarity is desired, should be enough. Otherwise, balkanization increases as too many narrowly defined retrieval systems proliferate. There is a balance between homogeneity and variety that must be struck.

3. TRACKS AND TOPICS

We feel that striking this balance between homogeneity and variety is important, and we wish to carry it a step further. By so doing, we also believe we will meet the Futrelle goal of being able to develop techniques which fit into a broader context of music information retrieval research. The manner in which we propose balancing homogeneity and variety this is to divide test collections into TREC-style *tracks* and *topics*.

A track is a broad statement about the type of task which will be done. A topic is an individual query, an expression of a user information need and other supporting information. A single track contains multiple topics, multiple variegated types of information need expressions. In this sense, a *track* is no different from the Melucci *task* in the previous section. Homogeneity is achieved in the sense that all the topics within a track have the same basic information need as their foundation. Variety is achieved in the sense that many topics within a track are slightly different expressions of that type of information need.

Returning to the example above, a "melodic" track is homogenous because all the topics within that track have as their core need melodic similarity. A melodic track simultaneously has variety because there are not only folk melody queries, but jazz melody queries and classical melody queries as well. Furthermore, tracks allow us to meet the Futrelle requirement that algorithms developed for searching be as broadly applicable as possible, because in order to score well across all the topics in a track, a retrieval system developer cannot optimize only toward jazz queries, or only toward classical queries. In order to perform well on the task, more powerful, more broadly applicable algorithms will be developed. "The ideal MIR technique could be effectively applied to a wide variety of music, regardless of its cultural origin [3]."

Up to this point, our proposals are in alignment with most of the other white papers detailing TREC-style evaluation. Again, our goal is not to replace these ideas, but to expand on them. The main idea of this paper is simply to carry the notion of tracks and topics one step further, into the realm of representation and complexity. In section 1 we spoke of the huge number of combinatorial possibilities that arose when systems were built and specifically tailored only toward monophonic music, polyphonic music, symbolic music, audio music, full pieces, incipits, choruses, and so on. Yet it should not matter if the music is in symbolic format, or scanned sheet music format, or audio, or if it is monophonic or polyphonic. In all cases, a user with the information need of finding pieces of music with the same *tune* as his query will have that need met no matter what the format of the retrieved piece.

Therefore, we wish to expand the "melody" track to include topics (and source collections) which contain not only jazz, classical, and folk pieces, but which also contain monophonic, polyphonic, audio, and symbolic pieces. The topics should also contain "full-text" pieces as well as incipit-only pieces, and chorus-only pieces. In short, homogeneity is preserved because all of the information needs expressed are thematically equivalent; users want pieces of music that contain the same "tune" as their query, no matter the form of the query or the source collection. At the same time, variety is also preserved because all the topics are slightly different in not only genre, but representation and complexity. It therefore becomes a worthy research goal to find algorithms which can deal with melodic similarity across all these boundaries.

4. TRACK SELECTION

Melodic similarity, increasingly misnamed because polyphony is an ingredient in the mixture, is only one possible track. Some users are not actually interested in melodic similarity, and thus the algorithms developed by systems using this track would not work. Tracks should work to maximize homogeneity, to a point; when the information need of a particular topic is too disparate from an existing track, a different track is needed.

I propose the following three major tracks for consideration in music information retrieval test collection construction:

1. `‘‘Melody/Tune’’ Track` – Contains topics in which information needs (and thus relevance) is determined primarily from note pitch features. This does not mean that other features, such as duration and timbre to name just a few, cannot be used to aid the retrieval process. Indeed, durations of notes might better inform some sort of rhythmic structure, which could be used in determining melodic boundaries or significant changes. Timbral features in an audio piece might offer clues about which notes or chords are or are not part of a "tune". But the point is that, no matter what features are used, the similarity sought by this type of information need related to the "tune" of the piece in question.

2. ‘‘Rhythm’’ Track – Contains topics in which information needs (and thus relevance) is determined primarily from note onset and duration features. Again, this does not mean that other features are unimportant. Suppose someone lays down a salsa beat, as a query, and the goal is to find other songs with a similar rhythm. Then being able to determine the timbre of the high-pitched clave, and using that timbre to determine when this instrument is struck, might give a good indication of where the main or important beats in a particular piece of music lie, thus better educating a rhtymic similarity matching algorithm. Or, if you saw in some symbolic piece of music that the note pitches returned to the tonic at some regular interval, that might help better identify measure boundaries or phrase/passage boundaries, which also could be useful for creating better rhythmic mathing algorithms. Once again, no matter what features are used, the similarity algorithms associated with this track, with this type of information need, relate to rhythmic patterns.

3. ‘‘Genre’’ track – Contains topics in which information needs (and thus relevance) is determined primarily from human-based genric judgements. This might be the hardest track to define, as genres include everything from heavy metal/country/rap in the popular audio domain, to mazurkas and cha-chas in the dance domain, to distinctions such as baroque, classical, and romantic in a period-based domain. Features used could include anything: pitch, harmony, duration, timbre, rhythm, and so on. A cha-cha might be similar to other cha-chas because of rhythmic clues, a baroque piece might be similar to another baroque piece because of certain harmonic progression clues (not the actual harmonic progressions, but the patterns inherent in those progressions), and a country song might be similar to another country song because of certain timbral clues (such as that characteristic “twang”). But in all cases, topics in this track have as their information need a similarity of genric type.

As with the tune track, the rhythmic and genre tracks also contain music of sources of representation and complexity: monophonic, polyphonic, audio, symbolic, full-text, incipits only, and so on. As such, homogeneity is best preserved across tracks, while variety is expanded within a track.

These are not the only possible tracks, nor do I feel that these existing proposals are set in stone. The community might feel that the “tune” track is too broad, too homogeneizing, and that, for the time being, there should be both an *audio* tune track, and a *symbolic* tune track. Whatever the final decisions, however, we would like to reemphasize the notion of having only a small number of tracks, and a large number of topics within a track. If there are too many tracks, the community risks balkanization. If there are too few topics within a track, the statistical significance of retrieval evaluation and system comparison will be low. More tracks may be added in a few years, as community interest and size continues to grow. But in these beginning stages, a small number of tracks is preferable.

5. THE ROLE OF MUSICGRID

One more piece is needed to make the proposals in this paper possible. Dovey has recently proposed a WebServices-related framework for distributed MIR collaboration and evaluation: MusicGrid [2]. This architecture allows a community to share not only resources such as topics (queries) and source collections, but also algorithms which operate on this data. Not only can these algorithms be migrated to the data, rather than the other way around, but components may be pieced together like a puzzle, mixed and matched.

This has important consequences for the track and topic based evaluation we propose in this paper. In particular, one of the difficulties associated with collections of multifarious music, from monophonic to polyphonic, from audio to symbolic, from jazz to classical, is that not every research group in the community has the expertise, let alone the resources, to work with every type of music and representation.

Thus, if I am trying to work with some sort of pitch-based feature, and the data in the collection is piecewise audio, I will have to write my own transcription algorithm before I can even begin to examine those pieces. This can be prohibitively expensive, and leads many research groups to focus only on symbolic data. Yet with a GRID architecture, if one member of the research community has implemented a transcription algorithm, no matter how good or bad, that algorithm may be taken and plugged in to someone else’s system as a front end.

As long as a “parser” exists for a particular music format, there is no need to develop music collection and/or queries in a standardized format. Research groups may bring collections of music, whether 50 pieces or 10,000 pieces, to the community, and as long as they also provide a parser which can read and “take apart” data in their format, the data will be accessible to all within the community. Thus more researchers can get up to speed quicker, designing algorithms which do better matching, rather than spending their time trying to parse various formats.

Therefore, with a MusicGrid architecture, a large number of topics may quickly be assembled, which topics may be tested against a large collection of music. For a given track, research groups need only submit a set of music pieces (the background collection), a set of topics (queries) which are intended to be run on this collection, and a parser which handles the data format of this collection. MusicGrid lets us simply take the union of all these topics and music to form a larger test collection for everyone to share.

Suppose I am building a retrieval algorithm which uses a pitch-based feature in some manner. Now, suppose two research groups provide access to their collections, and it turns out that the same piece of music is found in both collections. However, in the first collection, this piece of music exists in symbolic format, and in the second collection it is audio. The sequence of pitches gleaned from the symbolic parser on the symbolic piece will undoubtedly be slightly different than the sequence gleaned from the audio parser/transcriber on the audio piece. This is actually the whole point of amalgamating symbolic and audio pieces into the same track; the algorithms that will need to be developed to function on both perfect symbolic data as well as imperfect transcribed data should yield better insights into the nature of the problem than algorithms specifically tailored to a particular representation.

6. CONCLUSION

Evaluation drives research. Benchmarks help define research goals. Possession of a valid evaluation metric allows researchers to develop techniques which push the envelope of existing technologies and successfully meet the task at hand. By dividing music information retrieval evaluation into tracks and topics, we insure that the techniques which will be developed in the future are sufficiently broad and powerful enough to handle a variety of different music sources, representations, and complexities, while at the same time are focused enough to meet a user's information need.

By employing the MusicGrid architecture in support of this evaluation paradigm, it will become much easier to bootstrap large, varied test collections together. Not only do larger collection and topic sets increase the communities confidence in the results of an evaluation metric, but the very manner in which the test collections are assembled helps prevent the balkanization of algorithms that might otherwise occur. Furthermore, this same architecture lets research groups, who otherwise would not have the resources, participate in the algorithm-crafting arena.

The tracks proposed in this paper are not set in stone. Further discussion is necessary to agree within the community which tasks are the most interesting, the most widely applicable. But whatever the outcome of such discussions, the very process of spanning together numerous topics with the same core information need, no matter what the representation format or music piece length or complexity, will help create robust and powerful music information retrieval systems.

7. REFERENCES

[1] C. W. Cleverdon, J. Mills, and M. Keen. *Factors Determining the Performance of Indexing Systems, Volume I - Design, Volume II - Test Results*. ASLIB Cranfield Project, Cranfield, 1966.

[2] M. J. Dovey. Music grid – a collaborative virtual organization for music information retrieval collaboration and evaluation. In J. S. Downie, editor, *The MIR/MDL Evaluation Project White Paper Collection (Edition #2)*, pages 50–52, http://music-ir.org/evaluation/wp.html, 2002.

[3] J. Futrelle. Three criteria for the evaluation of music information retrieval techniques against collections of musical material. In J. S. Downie, editor, *The MIR/MDL Evaluation Project White Paper Collection (Edition #2)*, pages 20–22, http://music-ir.org/evaluation/wp.html, 2002.

[4] D. Harman. The trec conferences. In R. Kuhlen and M. Rittberger, editors, *Hypertext - Information Retrieval - Multimedia; Synergieeffekte Elektronischer Informationssysteme, Proceedings of HIM '95*, pages 9–28. Universitaetsforlag Konstanz, 1995.

[5] M. Melucci and N. Orio. A task-oriented approach for the development of a test collection for music information retrieval. In J. S. Downie, editor, *The MIR/MDL Evaluation Project White Paper Collection (Edition #2)*, pages 29–31, http://music-ir.org/evaluation/wp.html, 2002.

[6] S. Rueger. A framework for the evaluation of content-based music information retrieval using the trec paradigm. In J. S. Downie, editor, *The MIR/MDL Evaluation Project White Paper Collection (Edition #2)*, pages 68–70, http://music-ir.org/evaluation/wp.html, 2002.

[7] T. Sodring and A. F. Smeaton. Evaluating a music information retrieval system, trec style. In J. S. Downie, editor, *The MIR/MDL Evaluation Project White Paper Collection (Edition #2)*, pages 71–78, http://music-ir.org/evaluation/wp.html, 2002.

[8] E. M. Voorhees. Whither music ir evaluation infrastructure: Lessons to be learned from trec. In J. S. Downie, editor, *The MIR/MDL Evaluation Project White Paper Collection (Edition #2)*, pages 7–13, http://music-ir.org/evaluation/wp.html, 2002.