

Three Criteria for the Evaluation of Music Information Retrieval Techniques Against Collections of Musical Material

Joe Futrelle

Graduate School of Library and Information Science

University of Illinois at Urbana-Champaign

501 E. Daniel, Champaign, IL 61820, USA

+1-(217) 265-0296

futrelle@uiuc.edu

ABSTRACT

Evaluation of MIR systems requires honesty and skepticism with respect to the selection of test collections and the interpretation of results. We describe three minimal criteria which help to ensure this honesty and skepticism through appropriate selection of test collections, elimination of bias, and objective analysis of results.

1. INTRODUCTION

In order to evaluate a Music Information Retrieval (MIR) technique, it is necessary to apply the technique to one or more collections of musical material. The validity of this process rests on the researchers' honesty and skepticism in selecting and using the collections. We believe that three important criteria illustrate the kind of honesty and skepticism required of any MIR evaluation, and that MIR evaluation should be judged, minimally, against these criteria. They are:

1. Does the study use a large, comprehensive music collection? If not, why not?
2. Has the study eliminated or acknowledged collection / researcher bias?
3. Does the study use objective criteria to evaluate its results?

In this paper we will briefly examine each of these criteria.

2. DOES THE STUDY USE A LARGE, COMPREHENSIVE MUSIC COLLECTION? IF NOT, WHY NOT?

The ideal MIR technique could be effectively applied to a wide variety of music, regardless of its cultural origin. At the same time, this hypothetical technique would also allow us to probe deeply into the content of musical material in order to locate, analyze, or otherwise use the material. In practice, these two goals often contradict one another and require trade-offs. For instance, harmonic analysis may provide a powerful way of interacting with collections of common-practice European music (as in [1]), but may be of relatively little use on modal music such as Indian classical music or the blues. Evaluations of MIR techniques must explicitly identify these trade-offs. For instance, a study of an MIR technique based on harmonic analysis which evaluates it against a collection of common-practice European music must remain skeptical about the technique's applicability to other musics, and scale back its

claims accordingly.

Furthermore, it is the responsibility of an MIR researcher to evaluate a technique against the largest and most heterogeneous collection that is practical given the scope of the technique's applicability. If a technique is hypothesized to work for a particular kind of music, that kind of music must be included in the test collection, or the claims of the evaluation must be appropriately qualified. Finally, 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 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. For instance, query-by-humming (QBH) systems such as [2-9] are designed to support a single mode of user interaction with a collection indexed by melodic fragments. Given that scope, evaluations of QBH should produce hypotheses about the usefulness of QBH systems as a component of a system including other modes of user interaction and collections not indexed by melodic fragments.

3. HAS THE STUDY ELIMINATED OR ACKNOWLEDGED COLLECTION / RESEARCHER BIAS?

Researcher selection of music collections and queries can introduce bias, so pains should be taken to eliminate this bias. For example, many studies such as [10] evaluate MIR techniques with a hand-selected subset of the researchers' personal music collections. To eliminate selection bias, studies should select test collections at random from large, comprehensive collections of musical material. Where this is not possible, researchers should refrain from making any claims which could be impacted by selection bias.

In addition, the representation of music used for evaluating an MIR technique may introduce simplifying assumptions that may impact the claims. For instance, most QBH studies represent queries as sequences of notes, and therefore assume that query errors consist of equivalence classes of mismatches between those sequences of notes and themes extracted from the collection [3]. This eliminates classes of errors which cannot be expressed in the sequence-of-notes representation, such as various kinds of tuning problems. Such studies should (and often do) adjust their claims about the techniques' ability to cope with errors. In a larger sense, ways of representing music embody claims about the nature of music which must be

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

considered provisional until some universal theory of music may be developed and accepted [11]. A choice of representations therefore introduces a kind of bias towards a particular set of claims about the nature of music, and this bias should be acknowledged.

Style and genre present similar problems for MIR evaluation. If the technique is evaluated against a particular style or genre, or set of styles and genres (as in [12]), why is it not evaluated against others? Do the styles and genres differ in some way which is critical to the claims of the study? For instance, a study of tempo detection which evaluated it with a collection of rock music would beg the question of whether the technique might apply to a genre of music in which the beat is presumably harder to discern, such as Korean opera. Test collections should not be genre-limited unless the technique requires it, for instance as in [13] which investigates a technique designed to work primarily on classical music.

Another important area in which researcher bias can be introduced is query selection. It is often impractical or difficult for researchers to gather “real” queries from “real” users, since the experimental systems being evaluated lack the context in which real users can issue such queries. Therefore queries must be elicited from volunteers or synthesized according to a set of assumptions. In classic IR research, queries are synthesized by extracting fragments from target documents, and this approach has been extended to MIR systems as well as in [14]. Since evaluating the performance of an IR system for a particular query hinges on knowing *a priori* which items in the collection are relevant, the process of selecting the query typically involves selecting a relevant item from the collection and using it to elicit or synthesize a query. This process must be as transparent and repeatable as possible. For instance, the item must be selected randomly.

4. DOES THE STUDY USE OBJECTIVE CRITERIA TO EVALUATE ITS RESULTS?

MIR systems are typically evaluated for the relevance judgments they make. In order to do this evaluation, it is necessary to compare the relevance techniques against some external measure of relevance. The selection of this external measure of relevance may introduce bias. For instance, studies of theme-based MIR techniques such as [15] often evaluate their relevance judgments against published incipit indexes. But the results may indicate that the technique has been over-fitted to the kind of music for which incipit indexes are available.

Another subjective consideration which may impact the results of an MIR study is the choice of data analysis techniques. For instance, an MIR study about timbre classification might use a particular clustering algorithm to group feature vectors, and draw some conclusion from the results. But what if the clustering algorithm is not appropriate for the data? Perhaps the result is an artifact of the algorithm, and the algorithm should therefore be discarded in favor of a more appropriate one. Conclusions should not be drawn from the analysis of results unless the analysis techniques used accurately identify the patterns that are present in the data, considered independently of the hypotheses of the study.

Speaking of hypotheses, MIR studies should explicitly identify and reject the null hypothesis. This is a fundamental principle of

statistical analysis, but it has larger implications than simply reporting a measure of significance with every result. It means accounting not just for error, but also for a variety of plausible sources of bias, and either rejecting them or acknowledging them by adjusting the claims of the study. This practice embodies the honesty and skepticism required of MIR researchers, and indeed of researchers in any discipline.

5. CONCLUSION

The three criteria described above can serve as minimal guidelines to design evaluations of MIR systems. By choosing collections appropriately, eliminating or acknowledging bias, and assessing results objectively, MIR researchers can ensure that their evaluations meet a minimal standard for honesty and skepticism. And this minimal standard can serve as a common point of departure for the discussion of significant broader questions about MIR, such as the nature of music and the purpose of MIR systems.

6. REFERENCES

- [1] Cope, D. Computer Analysis of Musical Allusions. In *Proceedings of International Symposium on Music Information Retrieval* (Bloomington, IN, USA, 2001).
- [2] McPherson, J.R. and Bainbridge, D. Usage of the MELDEX Digital Music Library. In *Proceedings of International Symposium on Music Information Retrieval* (Bloomington, IN, USA, 2001).
- [3] Rand, W. and Birmingham, W. Statistical Analysis in Music Information Retrieval. In *Proceedings of International Symposium on Music Information Retrieval* (Bloomington, IN, USA, 2001).
- [4] Rolland, P.-Y. Adaptive User Modeling in a Content-Based Music Retrieval System. In *Proceedings of International Symposium on Music Information Retrieval* (Bloomington, IN, USA, 2001).
- [5] Sorsa, T. and Huopaniemi, J. Melodic Resolution in Music Retrieval. In *Proceedings of International Symposium on Music Information Retrieval* (Bloomington, IN, USA, 2001).
- [6] Haus, G. and Pollastri, E. An Audio Front End for Query-by-Humming Systems. In *Proceedings of International Symposium on Music Information Retrieval* (Bloomington, IN, USA, 2001).
- [7] Birmingham, W., et al. MUSART: Music Retrieval Via Aural Queries. In *Proceedings of International Symposium on Music Information Retrieval* (Bloomington, IN, USA, 2001).
- [8] Nishimura, T., et al. Music Signal Spotting Retrieval by a Humming Query Using Start Frame Feature Dependent Continuous Dynamic Programming. In *Proceedings of International Symposium on Music Information Retrieval* (Bloomington, IN, USA, 2001).
- [9] Jang, J.-S.R., Chen, J.-C., and Kao, M.-Y. MIRACLE: A Music Information Retrieval System with Clustered Computing Engine. In *Proceedings of International Symposium on Music Information Retrieval* (Bloomington, IN, USA, 2001).
- [10] Whitman, B., Flake, G., and Lawrence, S. Artist Detection in Music with Minnowmatch. In *Proceedings of IEEE*

Workshop on Neural Networks for Signal Processing
(Falmouth, MA, 2001).

- [11] Smiraglia, R.P. Musical Works as Information Retrieval Entities: Epistemological Perspectives. In *Proceedings of International Symposium on Music Information Retrieval* (Bloomington, IN, USA, 2001).
- [12] Aucouturier, J.-J. and Sandler, M. Using Long-Term Structure to Retrieve Music: Representation and Matching. In *Proceedings of International Symposium on Music Information Retrieval* (Bloomington, IN, USA, 2001).
- [13] Foote, J. ARTHUR: Retrieving Orchestral Music by Long-Term Structure. In *Proceedings of International Symposium on Music Information Retrieval* (2000).
- [14] Downie, J.S., *Evaluating a Simple Approach to Music Information Retrieval: Conceiving Melodic N-grams as Text*. Thesis, 1999, University of Western Ontario.
- [15] Meek, C. and Birmingham, W. Thematic Extractor. In *Proceedings of International Symposium on Music Information Retrieval* (Bloomington, IN, USA, 2001).