The Music Information Retrieval Evaluation eXchange (MIREX): Community-Led Formal Evaluations

J. Stephen Downie
jdownie@uiuc.edu
University of Illinois, USA

Andreas F. Ehmann
aehmann@uiuc.edu
University of Illinois, USA

Jin Ha Lee
jinlee l @uiuc.edu
University of Illinois, USA

Introduction

This paper provides a general overview of the infrastructure, challenges, evaluation results, and future goals of the Music Information Retrieval Evaluation eXchange (MIREX). MIREX [1] represents a community-based formal evaluation framework for the evaluation of algorithms and techniques related to music information retrieval (MIR), music digital libraries (MDL) and computational musicology (CM). MIREX is coordinated and managed by the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) at the University of Illinois at Urbana-Champaign. To date, since its inception in 2005, three annual MIREX evaluations have been performed covering a wide variety of MIR/MDL/CM tasks. The task definitions and evaluation methods for each annual MIREX are largely determined by community discussion through various communication channels with dedicated Wikis [4] playing a special role. Section 2 presents and defines the tasks associated with each of the past three MIREX evaluations.

In many respects, MIREX shares similarities to the Text Retrieval Conference (TREC) [5] in its overall approach to handling the evaluation of algorithms designed by the research community. Both MIREX and TREC are predicated upon the standardization of data collections; the standardization of the tasks and queries to be performed on the collections; and the standardization of evaluation methods used on the results generated by the tasks/queries [1]. However, associated with MIREX, there exist a unique set of challenges that cause MIREX to deviate from many of the methodologies associated with TREC. Section 3 covers some of these challenges, and the resultant solutions that comprise the overall framework and methodology of how MIREX evaluations are executed. Since MIREX is an ever-evolving entity, Section 4 will present key advances made between MIREX 2005 and MIREX 2006, as well as future goals for MIREX 2007 and beyond.

MIREX 2005, 2006, and 2007 tasks

The tasks associated with MIREX 2005, 2006 and 2007 are shown in Table 1.

Table 1. Task lists for MIREX 2005, 2006, and 2007 (with number of runs evaluated for each)

TASK	2005	2006	2007
Audio Artist Identification	7		7
Audio Beat Tracking		5	
Audio Classical Composer Identification			7
Audio Cover Song Identification		8	8
Audio Drum Detection	8		
Audio Genre Classification	15		7
Audio Key Finding	7		
Audio Melody Extraction	10	10 (2 subtasks)	
Audio Mood Classification			9
Audio Music Similarity and Retrieval		6	12
Audio Onset Detection	9	13	17
Audio Tempo Extraction	13	7	
Multiple F0 Estimation			16
Multiple F0 Note Detection			11
Query-by-Singing/Humming		23 (2 subtasks)	20 (2 subtasks)
Score Following	2		
Symbolic Genre Classification	5		
Symbolic Key Finding	5		
Symbolic Melodic Similarity	7	18 (3 subtasks)	3

A more detailed description of the tasks, as well as the formal evaluation results can be found on each year's associated Wiki pages [4]. The tasks cover a wide range of techniques associated with MIR/MDL/CM research, and also vary in scope. Tasks such as "Audio Onset Detection" (i.e., marking the exact time locations of all musical events in a piece of audio) and "Audio Melody Extraction" (i.e., tracking the pitch/fundamental frequency of the predominant melody in a piece of music) can be considered low-level tasks, in that they are primarily concerned with extracting musical descriptors from a single piece of musical audio. The motivation for evaluating such low-level tasks is that higher level MIR/MDL/CM systems such as "Audio Music Similarity and Retrieval" (i.e., retrieving similar pieces of music in a collection to a specified query song) or "Audio Cover Song Identification" (i.e., finding all variations of a given musical piece in a collection) can be built using many of the techniques involved in the low-level audio description tasks.

Table 2 provides a summary of participation in the past three MIREX evaluations. To date, 300 algorithm runs have been performed and evaluated.

Table 2. Summary data for MIREX 2005, 2006, and 2007

COUNTS	2005	2006	2007
Number of Task (and Subtask) "Sets"	10	13	12
Number of Teams	41	46	40
Number of Individuals	82	50	73
Number of Countries	19	14	15
Number of Runs	86	92	122

Challenges and Methodology

Although largely inspired by TREC, MIREX differs significantly from TREC in that the datasets for each task are not freely distributed to the participants. One primary reason for the lack of freely available datasets is the current state of copyright enforcement of musical intellectual property preventing the free distribution of many of the collections used in the MIREX evaluations. In addition, MIREX relies heavily on "donated" data and ground-truth. For tasks which require extremely labor-intensive hand-annotation to generate ground-truth — most notably low-level description tasks such as "Audio Onset Detection" — there is an overall reluctance of contributors to make their data and annotations freely available. As a result, it is nearly impossible to generate a representative dataset that encompasses all possible varieties, instrumentations, etc. of music. As such, there exists the potential of "tuning" or "overfitting" to a specific dataset at the expense of generalizablity of the algorithm to all varieties and types of music.

Due to the inability to freely distribute data, MIREX has adopted a model whereby all the evaluation data are housed in one central location (at IMIRSEL). Participants in MIREX then submit their algorithms to IMIRSEL to be run against the data collections. This model poses a unique set of challenges for the IMIRSEL team in managing and executing each annual MIREX. Firstly, data must be gathered and managed from various sources. For some tasks, differing formats for both the data and ground truth exist, as well as the potential for corrupted or incorrectly annotated ground-truth necessitating testing of the integrity of the data itself. The music collections used for MIREX tasks have already surpassed one terabyte and are continuously growing. In addition, many algorithms generate a large amount of intermediate data in their execution which must also be managed. In some cases, the intermediate data are larger in size than the actual music they describe and represent.

Moreover, IMIRSEL is responsible for supporting a wide variety of programming languages (e.g., MATLAB, Java, C/C++, PERL, Python, etc.) across different platforms (e.g., Windows, *NIX, MacOS, etc.). Despite guidelines dictating file input/output formats, coding conventions, linking methods, error handling schemes, etc., the largest amount of effort expended by IMIRSEL is in compiling, debugging, and verifying the output format and validity of submitted algorithms. Collectively, submissions to MIREX represent hundreds of hours of CPU computation time and person-hours in managing, setting up, and performing their execution.

Advances and Future Goals

One of the most significant advances made after MIREX 2005 was the incorporation of musical similarity evaluation tasks contingent upon subjective, human evaluation (MIREX 2006 and 2007). The addition of human evaluations of music similarity systems was born out of a community desire to reflect real-world applications and needs, and culminated in the "Audio Music Similarity and Retrieval" and "Symbolic Melodic Similarity" tasks. Both similarity tasks involve retrieving the top-N relevant or "similar" musical pieces in a collection using a specific musical piece as a query. A web interface with embedded audio players called the Evalutron 6000 was designed to allow evaluators to judge the similarity of a query "seed" with a retrieved "candidate" on both a broad scale (i.e., Not Similar, Somewhat Similar, Very Similar) and a fine, continuous, 10-point scale [3].

Another significant advance made manifest in MIREX 2006, and then repeated for MIREX 2007, was the application of formal statistical significance testing of returned results. These tests were applied in order to test whether performance differences between systems were truly significant. Because of its non-parametric nature, Friedman's ANOVA test was used on a variety of tasks to compare system performances. In general, these tests have shown that there are clusters of top performing techniques but these top-ranked techniques are not performing significantly better than their other top-ranked peers.

For future MIREX evaluations, IMIRSEL is presently developing a web service system that intends to resolve some of the key challenges associated with the execution of submitted algorithms by placing many of the responsibilities in the participant's hands. The web service, called MIREX DIY, represents a "black box" architecture, whereby a participant submits their algorithm/code through the web service, remotely begins its execution, and receives real-time feedback regarding the execution state of their algorithm. Execution failures can be monitored by the participant, and fixed if necessary. Upon successful execution and completion of the algorithm, performance results are returned to the participant. Eventually, such a system would allow submission and evaluation of algorithms year-round. Feel free to explore the MIREX DIY demo at http://cluster3.lis.uiuc.edu:8080/mirexdiydemo.

Acknowledgments

MIREX has received considerable financial support from both the Andrew W. Mellon Foundation and the National Science Foundation (NSF) under grant numbers NSF IIS-0340597 and NSF IIS- 0327371.

References

[1] Downie, J. S. 2006. The Music Information Retrieval Evaluation eXchange (MIREX). *D-Lib Magazine, 12,* 12 (December 2006: http://www.dlib.org/dlib/december06/downie/12downie.html.

[2] Downie, J. S., West, K., Ehmann, A., and Vincent, E. 2005. The 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005): Preliminary overview. In *Proceedings of the 6th International Conference on Music Information Retireval (ISMIR 2005)*, Queen Mary, UK, 2005, pp. 320-323.

[3] Gruzd, A. A., Downie, J. S., Jones, M. C., and Lee, J. H. 2007. Evalutron 6000: Collecting music relevance judgments. *ACM IEEE Joint Conference on Digital Libraries 2007*, p. 507.

[4] MIREX Wiki: http://music-ir.org/mirexwiki/.

[5] TREC: http://trec.nist.gov/.

Online Collaborative Research with REKn and PReE

Michael Elkink
melkink@uvic.ca
University of Victoria, Canada

Ray Siemens
siemens@uvic.ca
University of Victoria, Canada

Karin Armstrong
karindar@uvic.ca
University of Victoria, Canada

The advent of large-scale primary resources in the humanities such as EEBO and EEBO-TCP, and similarly large-scale availability of the full-texts secondary materials through electronic publication services and amalgamators, suggests new ways in which the scholar and student are able to interact with the materials that comprise the focus of their professional engagement. The Renaissance English Knowledgebase (REKn) explores one such prospect. REKn attempts to capture and represent essential materials contributing to an understanding of those aspects of early modern life which are of interest to the literary scholar - via a combination of digital representations of literary and artistic works of the Renaissance plus those of our own time reflecting our understanding of earlier works. REKn contains some 13,000 primary sources at present, plus secondary materials such as published scholarly articles and books chapters reflecting our understanding of these earlier works (some 100,000). These materials are accessed through a Professional Reading Environment (PReE), supported by a database system that facilitates their navigation and dynamic interaction, also providing access to inquiry-oriented analytical tools beyond simple search functions. The effect is that of providing an expert reading environment for those in our field, one that encourages close, comprehensive reading at the same time as it provides, conveniently, the building blocks of broad-based research inquiry. We are currently moving beyond the stage of proof-of-concept with these projects.

Our current research aim with these projects is to foster social networking functionality in our professional reading environment. For DH2008 we propose a poster/demo that details the current status of both the knowledgebase (REKn) and the reading environment (PREe) in relation to social networking, the direction each will take in the future, and a demonstration of the functional technologies we employ and our current implementation.

Rather than leveraging the power of an individual computer to perform complex computation on a personalized data set - which is the way most academics appear to work (see, for example, Siemens, et al., 2004), and is an approach exemplified