THE 2005 MUSIC INFORMATION RETRIEVAL EVALUATION EXCHANGE (MIREX 2005): PRELIMINARY OVERVIEW

J. Stephen Downie
GSLIS
University of Illinois at Urbana-Champaign
jdownie@uiuc.edu

Kris West
School of Computing Sciences, University of East Anglia
kw@cmp.uea.ac.uk

Andreas Ehmann
Electrical Engineering
University of Illinois at Urbana-Champaign
aehmann@uiuc.edu

Emmanuel Vincent
Electronic Engineering
Queen Mary
University of London
emmanuel.vincent@elec.qmul.ac.uk

ABSTRACT

This paper is an extended abstract which provides a brief preliminary overview of the 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005). The MIREX organizational framework and infrastructure are outlined. Summary data concerning the 10 evaluation contests is provided. Key issues affecting future MIR evaluations are identified and discussed. The paper concludes with a listing of targets items to be undertaken before MIREX 2006 to ensure the ongoing success of the MIREX framework.

Keywords: MIREX 2005, evaluation.

1 INTRODUCTION

This extended abstract provides a brief overview of the 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005) contest run in conjunction with the 6th International Conference on Music Information Retrieval (ISMIR 2005) held in London, UK, 11 September to 15 September, 2005. Although 2005 is the inaugural year for MIREX, MIREX 2005 should be considered a direct descendant of the very successful Audio Description Contest (ADC 2004) organized by the Music Technology Group (MTG), Universitat Pompeu Fabra (UPF), in Barcelona, Spain, as part of the ISMIR 2004 conference (see http://ismir2004.ismir.net/ISMIR_Contest.html).

Both ADC 2004 and MIREX 2005 were convened in response to the long-held desire of the MIR community to establish formal evaluation frameworks and metrics with which researchers could scientifically compare and contrast their wide variety of approaches to solving MIR tasks. Downie [1] provides an introduction to the issues involved in the establishment of such frameworks and metrics.

In Section 2 we outline the basic organizational scheme and infrastructure for MIREX 2005. In Section 3 we discuss some of the important issues brought to the fore while organizing MIREX 2005. Section 4 is devoted to listing the set of target items that are designed to build upon the successes of MIREX 2005 so that future iterations of MIREX can be more useful to the MIR community.

2 MIREX 2005 ORGANIZATION

2.1 Defining the MIREX 2005 contests

MIREX 2005 was co-chaired by Downie of the Graduate School of Library and Information Science (GSLIS), University of Illinois at Urbana-Champaign (UIUC) and Vincent of Electronic Engineering, Queen Mary, University of London. Like ADC 2004, the choice of evaluation scenarios and metrics for MIREX 2005 was based on the expressed interests of the MIR community itself. The two primary media for community decision making were the MIREX mailing list (157 subscribers) (https://mail.isrl.uiuc.edu/mailman/listinfo/evalfest) and the MIREX Wiki (http://www.music-ir.org/mirexwiki/). The mailing list and Wiki were established to help organize MIREX 2005 and through these interfaces proposals for evaluation tasks to be performed at MIREX 2005 were received. Each proposal included an approach to evaluating a MIR task, one or more evaluation metrics to be used in scoring performance on that task, and the nomination of potential databases that could be used to evaluate performance. Each proposal underwent significant refinement through this community dialogue process. Refinements included the addition of new data sets and major/minor modifications to the evaluation metrics.

After lively community debate on both the Wiki and the mailing list, a roster of 10 evaluation contests was settled upon for MIREX 2005. Special mention must be made of the “contest leaders” for they played pivotal roles in the moderation of the community dialogue and the finalizing of the contest scenarios. The contest names and summary data about each contest can be found in Table 1.

2.2 MIREX 2005 Infrastructure

The locus of the MIREX 2005 evaluation work was the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) housed in GSLIS, UIUC [2]. IMIRSEL provided MIREX 2005 with three key components:

1. the aforementioned communications mechanisms;
2. the computational infrastructure; and,
3. the M2K evaluation frameworks for each contest (see Section 2.2.2).

Table 1. MIREX 2005 summary data.

Contest Name	Submissions	Countries	Individuals	Contest Leaders
Audio Artist Identification	8	5	13	K. West
Audio Drum Detection	7	7	10	K. Tanghe
Audio Genre Classification	13	11	21	K. West
Audio Key Detection	5	3	6	C.-H. Chuan & E. Chew
Audio Melody Extraction	8	7	12	G. Poliner & D. Ellis
Audio Onset Detection	7	5	11	P. Leveau, P. Brossier & E. Vincent
Audio Tempo Detection	8	6	12	M. McKinney & D. Moelants
Symbolic Genre	5	4	9	C. McKay
Symbolic Key Detection	5	3	6	A. Mardirossian & E. Chew
Symbolic Melodic Similarity	6	6	15	R. Typke

Table 2. Computational infrastructure for MIREX 2005.

Machine Names	OS	Processor	RAM	Disk(s)
FAST	WIN XP	AMD Athlon XP 2600+ 1.9 GHz	2GB	80 GB
RED, YELLOW, GREEN	WIN XP	Intel Pentium 4 3.0 GHz	3GB	80 GB + 80 GB
BIBE	OS X	PowerPC G4 450 MHz	768 MB	20 GB + 20 GB
LINUX	RedHat 9	AMD Athlon XP 2600+ 1.9 GHz	1GB	80 GB
BeerClusterHead	CentOS	Dual AMD Opteron 64 1.6 GHz	4GB	1.8 TB NFS RAID
BeerClusterSlaves (x4)	CentOS	Dual AMD Opteron 64 1.6 GHz	4GB	160 GB Local disks

2.2.1 Computational infrastructure

The submissions to MIREX 2005 were designed to run on one or more of the Windows, Linux/Unix and Mac OS X architectures. Table 2 summarizes the hardware setup at IMIRSEL that was used to run the evaluation experiments. Post-evaluation standardized benchmarking of the individual computer processing speeds is planned so contestants can better ascertain the relative speed performances for those contests that spanned different computer architectures.

2.2.2 Evaluation frameworks infrastructure using M2K

In order to enable, coordinate and evaluate submissions to MIREX, a software framework was developed by the IMIRSEL team. This software framework had to be able to support submissions in a variety of formats including (but not limited to): C, C++, Java, Python and Matlab. The final solution is based in the Data-to-Knowledge (D2K) Toolkit and is included as part of the Music-to-Knowledge (M2K) Toolkit [2]. Both D2K and M2K are implemented in Java, giving them near total platform independence. M2K is an open-source initiative, meaning that any individual or group may leverage or modify this software and it can be evolved to support future evaluations. M2K is freely available from http://music-ir.org/evaluation/m2k.

The MIREX evaluation frameworks are implemented in M2K's modular format. Modules are connected by an XML-based itinerary which describes the particular process flow for each evaluation task. Figure 1 is a sample M2K MIREX evaluation itinerary. These frameworks are extremely flexible and can be customized by participants to suit the specific topologies of their submissions. This represents a significant advance over traditional evaluation frameworks and supports the central evaluation paradigm necessitated by the unique challenges posed by MIR evaluation.

3 DISCUSSION

3.1 Importance of MIREX

To get a sense of the importance that the MIR community has attached to MIREX 2005—in particular—and to the need for scientific evaluation—in general—it is worthwhile to note the strong evidence of growth between ADC 2004 and MIREX 2005. For example, ADC 2004 attracted 20 individual participants from 12 research labs; whereas, MIREX 2005 has 82 individual participants representing 41 different labs. ADC 2004 comprised 5 audio-based contests: Melody Extraction (4 submissions), Artist Identification (2 submissions), Rhythm Classification (1 submission), Music Genre Classification (5 submissions) and Tempo Induction (6 submissions) for a total of 18 primary submissions[1] [3]. A comparison of these data with the 72 primary submissions, distributed across 10 contests, for MIREX 2005 reveals an encouraging broadening of community interest in formal evaluation tasks. Furthermore, the number of primary submissions per contest for MIREX 2005 ranges from 5 (Symbolic Key Detection) to 13 (Audio Genre Classification) and represents a healthy deepening of researcher participation.

[1] The submission counts given above reflect only the "primary" submissions as some teams submitted for evaluation several algorithmic variants of their techniques to each contest.

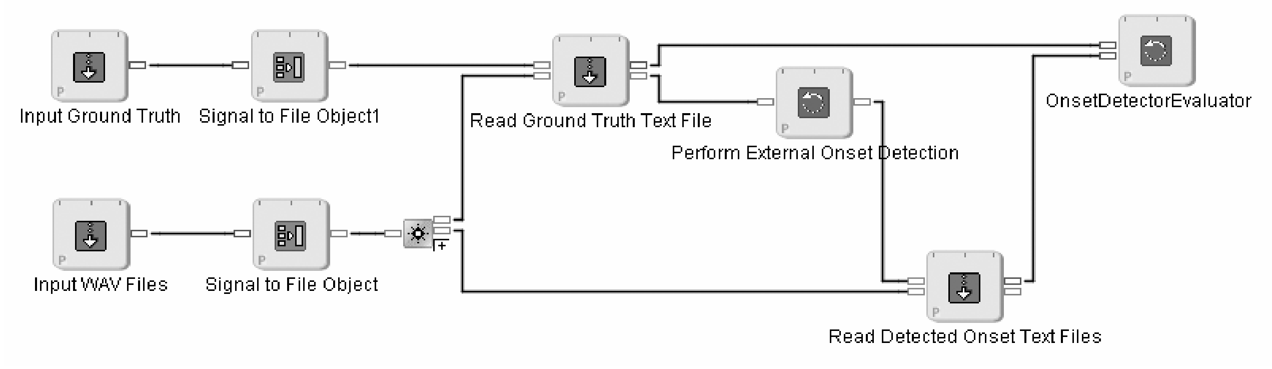

Figure 1. A sample MIREX evaluation framework implemented in M2K.

3.2 Continuing Challenges

Notwithstanding the advancements made by ADC 2004 and MIREX 2005 on deepening and broadening the scope of formal MIR evaluations, there remain several serious challenges facing the MIR community that must be overcome in order to conduct future MIREX contests that consistently provide meaningful and fair scientific evaluations. These challenges include:

1. the continued near impossibility of establishing a common set of evaluation databases or the sharing of databases among researchers due primarily to intellectual property restrictions and the financial implications of those restrictions;
2. the ongoing lack of established evaluation metrics for the plethora of tasks in MIR; and,
3. the general tendency in the field to use small databases for evaluation with the difficulties associated with the creation of ground-truth data being a primary cause of this state of affairs.

The MIREX 2005 organizers recognize that these aforementioned hurdles will not be overcome in the near future. We do, however, want to briefly highlight some of the implications of these ongoing challenges.

3.2.1 The constant ad hoc evolution of metrics

More than half of the evaluation tasks proposed for MIREX 2005 had evaluation metrics established or significantly refined through the MIREX communication process. While many of the metrics decided upon are based on principled evaluation procedures used in other fields, a close reading of the ongoing metrics discussions reveals a decidedly *ad hoc* approach to MIR evaluation metrics: in almost every case the initially proposed metrics were significantly refined through participant discussion and in several cases completely new metrics were established. Furthermore, tasks that enjoyed relatively well-established evaluation procedures, such as Audio and Symbolic Genre Classification and Artist Identification, had interesting evaluation metrics added to them, such as the discounting of confusion in the classifications through the use of hierarchical taxonomies.

The MIREX 2005 team fully appreciates the delicate balancing act between the necessity of community input on metric decisions (which tend to generate ever changing evaluation metrics) and the need to establish—*perhaps even impose*—universal, standardized metrics so that multi-year comparisons can be made. At this point, we have no simple solution to offer. We are, however, flagging the "*ad hoc* evolution of metrics" issue as a high-priority "target item" (Section 4).

3.2.2 The need for tests of statistical significance

Due to the financial implications of collecting large audio databases for evaluation and the significant burden of annotating them, there is a general tendency to use relatively small databases for evaluation. The establishment of central evaluation paradigms like ADC 2004 and MIREX 2005, has helped to alleviate, but not eliminate, this problem. Audio databases and annotation sets are valuable resources and a reluctance to surrender that data to a wider community is understandable. The MIR community has been remarkably open with their resources and has allowed the establishment of databases of much greater magnitude and coverage than existed prior to MIREX 2005. Despite this show of community goodwill, however, these databases are still relatively limited when compared to industrial-scale real-world problems. Because of these limitations we need to interpret the contest results achieved with care.

To mitigate this database-size limitation problem the IMIRSEL team has established tools in M2K that allow multiple, principled statistical significance tests to be applied to the comparison of results in every evaluation task performed at MIREX 2005. These techniques include the:

1. Student's t-Test;
2. Sign Test; and,
3. McNemar's Test (see [4]).

The Student's t and the sign tests are methods of assessing the significance of differences in ***overall*** performance between two systems. McNemar's test, however, takes into account the use of the same dataset in the comparative evaluation of two algorithms and assesses the significance of differences in performance on an ***item-by-item*** basis. Thoughtful application of these tests in combination can yield important insights into true system performance. For example, if a t-test yields a non-significant difference between two algorithms, the results from a McNemar's test on these algorithms can help determine whether the examined systems are exhibiting similar error functions. Opening dialogue concerning the application of tests of statistical significance is being flagged as another "target item" (Section 4).

3.2.3 Need for collaborative annotations

Due to the large number of tasks to be evaluated, the MIREX organizers could not possibly annotate all the evaluation data themselves. As a consequence, participants were encouraged to contribute their annotated data and to conduct new annotations. Some of these participants could be suspected of using the annotated data they contributed to fine tune their algorithms. However, if they were not trusted by the majority of other participants, only a small subset of evaluations could have been run. This issue will become even more stringent when new tasks are evaluated. In the future, collaborative annotation of the testing data by a large number of participants (all if possible) will be needed. The creation of an online collaborative annotation tool is another of our "target items" (Section 4).

4 FUTURE WORK: KEY TARGET ITEMS

The MIREX 2005 organizers and the IMIRSEL team have set up a list of 8 priority target items designed to improve upon the successes of ADC 2004 and MIREX 2005. These are items that we believe should be implemented prior to MIREX 2006. We have tasked ourselves to:

1. Establish a communications mechanism specifically devoted to the establishment of standardized and stable evaluation metrics to replace the undesirable *ad hoc* procedures currently being used.
2. Open discussions on the selection of more statistical significance testing procedures. The current three implemented are only a beginning and are not universally applicable because evaluation result data do not always conform to their underlying assumptions.
3. Work with the MIR community to establish new annotation tools and procedures to overcome the shortage of available ground-truth data. Ideally, these tools should be open-sourced or perhaps made available via M2K in conjunction with the proposed webservices system mentioned in Item #8.
4. Establish a more formal organizational structure for future MIREX contests. This year, the contest leaders became so by "default". We need to have contest leaders who have formally accepted the various administrative tasks associated with setting up the individual contests including acting as liaisons between participants and the MIREX organizers.
5. Convene an online forum to produce a high-level development plan for the future of the M2K Toolkit to solicit advice and opinions from the members of the MIR community on the services and formats that would be desirable in later versions of M2K.
6. Continue to develop the evaluator software and establish an open-source evaluation API. This will involve the redevelopment of the existing evaluation modules, adding a greater degree of abstraction to allow for optimal reuse of code and aid in the development of evaluators for new MIR tasks. This may include the provision of 'command line' versions of the evaluation systems.
7. Make useful evaluation data publicly available year round. Care will have to be taken in doing this as making all of the data used to evaluate a task available will preclude its use in future, fair evaluations as fine-tuning or over-fitting could be performed on this data. Therefore, distributable "development" data sets must also be established. Again, Item #8 should play a major role in making this a reality.
8. Establish a webservices-based IMIRSEL/M2K online system prototype which would allow MIR researchers to run evaluations on centrally held datasets and to compare their results against the earlier results achieved by others on those data and query sets. Mounting community-accessible annotation tools should be seen as part of these webservices.

ACKNOWLEGMENTS

The IMIRSEL team is supported by the Andrew W. Mellon Foundation and the National Science Foundation (NSF) Grant Nos. NSF IIS-0340597 and NSF IIS-0327371. E. Vincent is funded by the EPSRC grant GR/S75802/01.We thank those who provided content and ground-truth data. We thank the Automated Learning Group (ALG) at the National Center for Supercomputing Applications (NCSA) at UIUC and Paul Lamere of Sun Labs for their kind assistance. IMIRSEL's X. Hu, J. Futrelle, M. Callahan, C. Jones, D. Tcheng, M. McCrory, S. Kim and J-H. Lee are also thanked. Special thanks to the GSLIS technology services team. Finally, we thank the ISMIR 2005 organizing committee for making both ISMIR and MIREX 2005 a success.

REFERENCES

[1] Downie, J. The scientific evaluation of music information retrieval systems: Foundations and future. Computer Music Journal, 28, 2, (2004), 12-33.

[2] Downie, J., Futrelle, J., and Tcheng., D. "The International Music Information Retrieval Systems Evaluation Laboratory: Governance, access and security", Proceeding of the Fifth International Conference on Music Information Retrieval (ISMIR), Barcelona, Spain, 2004.

[3] Cano, P., Gómez, E., Gouyon, F., Herrera, P., Koppenberger, M., Ong, B., Serra, X., Streich, S., and Wack, N. ISMIR 2004 Audio Description Contest. Under review for journal publication.

[4] Gillick, L., and Cox, S. "Some statistical issues in the comparison of speech recognition algorithms", Proceedings of IEEE Conference on Acoustics, Speech and Signal Processing, Glasgow, UK, 1989.