Acoust. Sci. & Tech. **29**, 4 (2008)

INVITED REVIEW

The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research

J. Stephen Downie

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Abstract: The Music Information Retrieval Evaluation eXchange (MIREX) is the community-based framework for the formal evaluation of Music Information Retrieval (MIR) systems and algorithms. By looking at the background, structure, challenges, and contributions of MIREX this paper provides some insights into the world of MIR research. Because MIREX tasks are defined by the community they reflect the interests, techniques, and research paradigms of the community as a whole. Both MIREX and MIR have a strong bias toward audio-based approaches as most MIR researchers have strengths in signal processing. Spectral-based approaches to MIR tasks have led to advancements in the MIR field but they now appear to be reaching their limits of effectiveness. This limitation is called the "glass ceiling" problem and the MIREX results data support its existence. The *post-hoc* analyses of MIREX results data indicate that there are groups of systems that perform equally well within various MIR tasks. There are many challenges facing MIREX and MIR research most of which have their root causes in the intellectual property issues surrounding music. The current inability of researchers to test their approaches against the MIREX test collections outside the annual MIREX cycle is hindering the rapid development of improved MIR systems.

Keywords: Music information retrieval, Evaluation, MIREX

PACS number: 43.75.Xz [doi:10.1250/ast.29.247]

1. INTRODUCTION

Music Information Retrieval (MIR) is a multidisciplinary research field that draws upon the traditions, methodologies, and techniques of a remarkably wide range of disciplines [1]. An incomplete listing of these disciplines includes acoustics, psychoacoustics, signal processing, computer science, musicology, library science, informatics, and machine learning, etc. The primary goal of MIR research, regardless of the disciplinary paradigm under which it is undertaken, is the provision of a level of access to the world's vast store of music on a level equal to, or exceeding, that currently being afforded by text-based search engines. Because music is a complex amalgam of acoustic, rhythmic, harmonic, structural, and cultural phenomena, the grand challenge facing MIR research is the development of retrieval systems that deal with music on its own terms. That is, MIR researchers strive to build retrieval systems wherein the music itself, whether it be in represented in auditory (e.g., MP3s, WAV, etc.) or symbolic (e.g., MIDI, score, etc.) formats (or both), is the principal mechanism by which users interact with the systems. Simply put, MIR research wants to develop systems that allow users to search music content using queries that are framed musically. Musically framed queries can include such techniques as query-by-singing, query-by-example (e.g., submitting a known MP3 to find similar pieces) and/or query-by-notation (e.g., placing notes on a musical staff to form the query), etc. For those new to the MIR field, [2] and [3] are recommended introductory overviews. Perusal of the collected proceedings of the International Conference on Music Information Retrieval (ISMIR) is also recommended [4].

If one wishes to understand the important issues, trends, and advancements in MIR research, one should begin with an examination of the infrastructure, challenges, evaluation results, and future goals of the Music Information Retrieval Evaluation eXchange (MIREX). MIREX represents a community-based framework for the formal evaluation of algorithms and techniques related to MIR. MIREX is coordinated and managed by the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) [5] at the University of Illinois at Urbana-Champaign. Since its inauguration in 2005 [6], three annual MIREX evaluations have been performed covering a range of tasks that closely models many of the major MIR research areas. 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
AA	Audio Artist Identification	7		7
AB	Audio Beat Tracking		5	
ACC	Audio Classical Composer Identification			7
ACS	Audio Cover Song Identification		8	8
AD	Audio Drum Detection	8		
AG	Audio Genre Classification	15		7
AK	Audio Key Finding	7		
AME	Audio Melody Extraction	10	10*	
AMC	Audio Mood Classification			9
AMS	Audio Music Similarity and Retrieval		6	12
AO	Audio Onset Detection	9	13	17
AT	Audio Tempo Extraction	13	7	
MFE	Multiple F0 Estimation			16
MFT	Multiple F0 Note Tracking			11
QBSH	Query-by-Singing/Humming		23*	20*
SF	Score Following		2	
SG	Symbolic Genre Classification	5		
SK	Symbolic Key Finding	5		
SMS	Symbolic Melodic Similarity	7	18#	8

*task comprised two subtasks #task comprised three subtasks

By looking at the background, structure, challenges, key contributions, and future of MIREX this paper will provide a window into the larger world of MIR research. Section 2 outlines the basic background and infrastructure of MIREX along with an explication of how MIREX reflects the interests of the MIR community. In Section 3 are highlighted the key challenges that face both MIREX and MIR research in general. Section 4 discusses in some depth two of the key contributions that MIREX has made to MIR research. Section 5 summarizes our exploration of MIREX and MIR issues and introduces a new research consortium that has been put together to increase the viability and productivity of both MIREX and MIR research into the future. The paper concludes with information on how get involved with the MIREX and MIR communities.

2. MIREX 2005, 2006, AND 2007 TASKS

2.1. Background and Infrastructure

In 2004, the Music Technology Group (MTG) of the Universitat Pompeu Fabra, as hosts of ISMIR 2004, convened the "Audio Description Contest" (ADC) [7]. ADC was a direct antecedent to MIREX from which the more ambitious MIREX learned many lessons. Both ADC and MIREX were inspired by, and therefore share, many similarities to the Text Retrieval Conference (TREC) framework [8,9] with regard to their overall evaluation methodologies. ADC, MIREX, and TREC are predicated upon the standardization of the:

1. test collections of significant size;
2. tasks and/or queries to be performed on the test collections; and,
3. evaluation methods to be used to evaluate the results generated by the tasks/queries.

Like TREC, the task definitions and evaluation methods for each annual MIREX are largely determined by community discussion through various communication channels. For MIREX, the community makes use of a special mailing list for community-wide input (>300 subscribers) and a set of dedicated Wiki spaces[1] for task-specific debates and definitions. The community traditionally begins task discussions in early spring when eager researchers call out for expressions of interest in a particular task. If interest is shown, the self-appointed task leader(s) set up a Wiki space where the interested parties refine their ideas concerning test collections, tasks to be evaluated, the metrics to be used, and the input/output formats to which the algorithms will be expected to adhere. If all these pieces come together, and there are at least three researchers willing to participate, MIREX will incorporate the task into its official set of evaluations. The evaluations themselves are run in July and August each year. The final results, along with some raw performance data, are then posted to the MIREX Wiki at a time prior to the annual ISMIR conference (September or October). At the annual ISMIR conference, MIREX holds its poster session (to which all participants are expected to contribute) and convenes a half-day MIREX plenary meeting to discuss successes and failures and to plan for future MIREX evaluations. Readers are especially encouraged to visit the MIREX Wiki to see both task debates and results reports. Participants are expected to submit extended abstracts describing the techniques being evaluated and these are also available on the MIREX Wiki.

2.2. MIREX as a Reflection of MIR Research

Table 1 highlights some interesting facts about MIREX specifically, and MIR, in general. Nineteen different tasks have been defined and run over the 2005–2007 period. Because some tasks have been run over multiple years and some of these comprise several subtasks, a total of 35 task

[1]See http://music-ir.org/mirexwiki. For those unfamiliar with Wiki systems, they are websites that allow participants to collaboratively create, edit, interlink, and manage web pages via their web browsers using a simple mark-up language. An excellent introduction to Wiki systems can be found at http://en.wikipedia.org/wiki/Wiki.

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

	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

"sets" have been run. Of the 19 distinct tasks defined, only three (SG, SMS, and SK) were tasks in the symbolic domain. The remaining 16 tasks (84%) are those which audio processing techniques are employed. Two tasks, SF and QBSH, while included in the audio-based task class, are actually hybrid tasks in that they both take audio input and try to match it against an underlying symbolic representation. As Table 2 shows, there have been a steadily increasing number of algorithms evaluated over MIREX's three year history (300 algorithms in total). Again, we see that 257 (86%) of these 300 runs are evaluating audio-based techniques. There are several reasons why audio-based research predominates both MIREX specifically, and MIR, in general. First, many MIR researchers come from a signal processing research discipline (e.g., electrical engineering, acoustics, speech processing, etc.). It is thus understandable that these researchers would apply their audio skills to MIR problems. Second, it is relatively easy to collect music in its audio form from such sources as personal CD collections, MP3 collections, iTunes, etc. Large-scale collections of digitally encoded symbolic/notation-based music that are readily available and not restricted to the "Classical" music styles, on the other hand, are quite rare. This rarity represents a substantial impediment to the advancement of symbolic MIR techniques. Third, and possibly more troublesome for those with expertise in the music domain, it can be suggested that dealing with music in its audio forms requires less music-specific knowledge than dealing with its symbolic forms (i.e., one needs to be able to read and understand music to work with symbolic music representations in a non-trivial manner).

Note that in a field called Music Information Retrieval, only three (16%) of the nineteen tasks (AMS, SMS, and QBSH) represent information retrieval (IR) tasks in the standard sense of a queries being presented and items being returned. Five (26%) of the MIREX tasks (AA, ACC, AMC, AG, and SG) are structured within the traditional train-test cross-validation paradigms of the machine learning community. Nine (47%) of the tasks can be described best as evaluating "low-level" MIR subcomponents (AB, AD, AK, AME, AO, AT, MFT, MFE, and SK). This low-level classification signifies that the techniques being evaluated are those that will necessarily be incorporated into future MIR systems if these systems are to truly deal with music on its own terms and with its manifold complexities. For example, at the lowest-level, is the set of audio onset (AO) techniques which are designed to denote the exact locations of musically significant events in audio streams. If the AO problem is solved satisfactorily, it will help the performance of other low-level tasks such as AB, AD, AME, AT, MFT, and MFE, etc. These low-level tasks will in turn be used to extract/construct such musically necessary features as melodic shape, harmonic progressions, and rhythmic patterns, etc. upon which useful MIR systems will rely. Thus, the three-fold disparity between the number of "low-level" and IR tasks MIREX has run is, in my opinion, an accurate reflection of the general consensus of the MIR community that MIR research needs first to devote its attention to the low-level research questions upon which the success of future MIR systems are dependent. Before moving on to the challenges posed by running MIREX, I would like to draw the reader's attention to the lack of MIREX tasks dealing with usability and interface issues. Both of these domains are important and vibrant MIR research areas. Because the evaluation of usability and interfaces involves a certain amount of qualitative judgment on the part of the evaluators, it has proven difficult to properly design formal MIREX usability and interface task definitions. Discussions are ongoing concerning how to overcome this serious MIREX shortcoming.

3. THE CHALLENGES OF MANAGING MIREX

Although largely inspired by TREC, MIREX differs significantly from TREC in that the datasets for each task are not freely distributed to the participants. The primary reason for the lack of freely available datasets is the current state of musical intellectual property copyright enforcement. The constant stream of news stories about the Recording Industry Association of America (RIAA) bringing lawsuits against those accused of sharing music on peer-to-peer networks has had a profoundly chilling effect on MIR research and data sharing. Notwithstanding a potential defense under "fair use" or "fair dealing" copyright doctrines, no senior researcher or lab administrator wants to be named in such a lawsuit nor incur the expense of mounting what might prove to be an unsuccessful defense. Thus, due to this inability to freely distribute test collection 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 centralized algorithm-to-data model poses a unique set of challenges for the IMIRSEL team and the community at large in managing and executing each annual MIREX. In the following discussion are highlighted the seven salient challenges that MIREX continues to face.

1. The simple acquisition of test collection data is fraught with time-consuming perils. Sometimes the data are donated by (or purchased at cost from) labs interested in a particular task. Sometimes (rarely) they are donated by recording companies. Sometimes (most often) they are purchased outright from commercial sources (which can be a strain on IMIRSEL's research budget). Thus, hundreds of hours each year are consumed by the locating, gathering, and managing of a wide variety of evaluation content. However, this is nothing compared to the negotiation of formal terms-of-use agreements among institutional stakeholders. Negotiating these terms-of-use agreements can consume shocking amounts of time as institutional legal and administrative teams get involved.
2. The acquisition of ground-truth data poses its own set of challenges. Quality ground-truth data is very expensive to produce. Even if the data are created by volunteers, considerable resources must be allocated in their creation. This can make even the most generous of labs hesitant to share ground-truth data sets with MIREX. The shortage of ground-truth data has led to the re-running of such tasks as AT, AME, AO, and ACS using the same ground-truth as previous years. In fact, AO has been run in each of MIREX's 2005, 2006, and 2007 sessions using the same ground-truth. This situation puts these evaluation tasks in jeopardy of being overfitted by the algorithms being evaluated which could severely diminish the future utility of the results data.
3. Some evaluation tasks fall under the rubric of human subjects research which can magnify the administrative overhead enormously. For example, the AMS and SMS tasks, where systems are evaluated *post hoc* by volunteers who compare queries and returned results for similarity, were both deemed to be human subjects research by the Institutional Review Board of the University of Illinois. This determination set into motion a whole suite of legal requirements and safeguards that must be followed under United States federal law. These safeguards include the construction, and external approval, of formal research protocols, the creation of informed consent mechanisms, the screening of underage evaluators, and the special treatment of results data to ensure evaluator confidentiality.
4. Experience has shown the MIREX team that there is a high potential for corrupted or incorrectly annotated test collection and/or ground-truth data. This has forced MIREX to adopt a continuous regime of data integrity testing. This issue arose, for example, in the AO task where some of the ground-truth annotations were found to be mislabeled. In the context of the AG and AA tasks, metadata information taken from an online source was found to be incorrect and had to be correct by hand. For AK and SK, we discovered that some of the key signature labels on the evaluation data were also incorrect. Since these key signature labels were intended to form the ground-truth, IMIRSEL had to call in an undergraduate music major with perfect pitch to verify the key of each of 1252 test collection files.
5. There are infrastructure capacity issues brought about by MIREX's present algorithm-to-data model. For example, the MIREX music collections currently comprise more than two terabytes of audio data representing some 30,000 tracks divided among popular, classical, and Americana sub-collections. Furthermore, many algorithms generate large amounts 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. Algorithms using Short-Time Fourier Transform (STFT) techniques are especially prone to this interim data explosion problem. Because of space limitations, MIREX has been discarding the features sets generated by the various algorithms. We see this as a significant loss to the MIR community as these feature sets could in turn be re-used by researchers in novel experiments. Even though the raw evaluation outputs generated by the algorithms can also be quite large and diverse, MIREX does have a formal policy of keeping these raw outputs. The policy is designed to encourage the re-use of the raw output data by the community in secondary analyses of the evaluation tasks. Notwithstanding the archiving burden this policy incurs, providing access to these outputs sets has helped the community to uncover, and then correct, evaluation errors made with regard to, for example, the AO and AB tasks.
6. The management of submitted algorithms is the largest consumer of human resources at IMIRSEL. In an effort to encourage the maximum number of participants, MIREX places almost no restrictions on the computing language used to build the systems under evaluation. Because these algorithms are run by IMIRSEL, it makes the IMIRSEL team responsible for supporting a wide variety of programming languages (e.g., MATLAB, Java, C/C++, PERL, Python, MAX, etc.) across different platforms (Windows, *NIX, MacOS). Despite guidelines dictating file input/output formats, coding conventions, linking methods, error handling schemes, etc., the largest amount of effort expended by IMIRSEL on behalf of MIREX is in compiling, debugging, and verifying the output format and validity of submitted algorithms. Collectively, managing and monitoring

the algorithms submitted to MIREX consumes nearly a 1000 person-hours each year. Similarly, MIREX algorithms can be very computationally expensive, especially those computing exhaustive distance matrices. For example, AMS 2007 had participants building 7000 × 7000 matrices which were then used to provide the ranked lists of results output for each randomly selected query input. Submissions performing iterative machine learning parameter optimizations (e.g., some AG and AA submissions) have been notoriously expensive. Runs lasting 24–48 hours are not rare. MIREX currently has a 72 hour runtime limitation rule. The recent acquisition of several quad-core multi-processor computers at IMIRSEL has gone a long way to mitigating the effects of dedicating a CPU to one algorithm for 72 hours. While it is tedious to monitor an algorithm run for 72 hours (and even more tedious to have that run fail to write out its results properly), MIREX has not been pushing the community to write more efficient code. MIREX is more concerned with output results than computational efficiency and wants to encourage the submission of non-optimized proof-of-concept techniques.

7. The time constraints imposed by constant debugging of code, along with our inability to share test collections, have made the off-cycle, on-demand re-running of evaluation tasks next to impossible. Furthermore, participants only see the "final" results sets which are made available shortly before the MIREX plenary meeting. This makes the *de facto* research cycle at least a year long. Thus, MIR researchers who have novel MIR techniques cannot determine if their techniques are reaching state-of-the-art effectiveness in a timely manner.

4. TWO KEY CONTRIBUTIONS

4.1. Toward Shattering the "Glass Ceiling"

In 2004, Aucouturier and Pachet [10] published a paper wherein they noted that use of naïve[2] audio-based timbral feature sets to perform MIR similarity tasks has real limitations in terms of successfully identifying musically similar pieces of audio. Despite running over a hundred combinations of machine learning algorithms and Mel Frequency Cepstral Coefficients (MFCC) input parameter variations, they could only see a 15% improvement from baseline (using the *R*-precision metric). *R*-precision is the precision after *R* items have been retrieved, where *R* is the number of relevant items for a given search query. Precision, in turn, measures the ability of a system to retrieve only relevant items. Precision is defined as:

$$precision = \frac{|\{relevant\ items\} \cap \{retrieved\ items\}|}{|\{retrieved\ items\}|}$$

Aucouturier and Pachet's best run had an R-precision of only 65%. They coined the phrase "glass ceiling" to represent this limitation. Similarly, in 2001, Whitman, Flake, and Lawrence [11] coined the phrase "album effect" to represent the situation where experimental AA results were being inflated by learning algorithms naïvely picking up on the trivial production qualities of albums (which tend to be consistent across tracks within a given CD) rather than the music qualities of the artists themselves. Thus, when experiments are done where individual albums are used exclusively in only the training or test sets, AA performance degrades substantially. The album effect has been explored quite cogently by Kim, Williamson, and Pilli [12]. Pampalk, Flexer, and Widmer [13] noted similar effects when they conducted a set of AG performance experiments with, and without filtering, for individual performing artists across train and test sets (i.e., they tested with, and without, the application of "artist filtering"). They report stunning changes in performance: 79% accuracy without artist filtering and 27% with artist filtering! Flexer [14] extended this line of research and concluded that all AG work needs to be done with artist filtering. One could argue, given that each of these phenomena are predicated on the application spectral/timbral analyses, that they in fact represent aspects of the same underlying problem (i.e., that spectral-based approaches are not capturing "music information" in a truly meaningful way). Simply put, timbral similarity is not equating to music similarity [15].

Evidence of the "glass ceiling," "album effect," and "artist filtering" issues can be found throughout the MIREX results. In fact, I contend that this evidence has been one of most important contributions of MIREX to MIR research. In 2005, the best accuracy result for the AA task was 72.45% (Mandell and Ellis). In 2007, the best AA result was 48.14% (IMIRSEL_SVM (Support Vector Machine)). In 2005, the best accuracy result for AG was 82.34% (Bergstra, Casandre, and Eck). In 2007, it was 68.29% (IMIRSEL_SVM). Does this indicate that AA performance has dropped by 24.31% and AG performance has dropped by 14.05% over the course of two years? No, but it does present further empirical evidence that the glass ceiling, album effect, and artist filtering issues are real. With regard to the AA task, the 2005 task definition did not include any sort of album filtering. In 2007, the test collection data were partitioned such that no track from the same single album would appear simultaneously in both

[2]The term naïve is used here to denote the fact that the timbral extraction signal processing techniques being widely used in MIR research and discussed by Aucouturier and Pachet are not based upon music theory. They come to MIR from a long tradition of use in the speech community. This long tradition of use means that they are many pre-existing implementations that MIR researchers have simply and atheoretically applied to the music domain.

the training and test sets in any cross-validation fold (i.e., album filtering). In 2005, no filtering of any sort was performed on the AG test collection. However, in 2007, the AG test collection data were partitioned whereby no single artist could appear simultaneously in both the train and test sets in any individual fold of the cross-validation runs (i.e., artist filtering).

Intrigued by the glass ceiling issues, I designed the ACS task in 2006 to specifically address the distinction between "timbral similarity" and "music similarity." At the heart of the ACS 1000 song test collection are 330 tracks consisting of music drawn from a wide range of genres and styles. These 330 tracks comprise 30 subcollections of 11 tracks each that are "cover versions" of each other. Thus, each set of 11 tracks are musically similar in the sense that they are variations on the same piece of music. The remaining 670 tracks are unrelated "noise" tracks intended to increase the search space. The use of the 30 subcollections present a glass ceiling challenge to researchers in that no subcollection contains tracks from the same artist or album. The state-of-the-art in spectral similarity techniques is further challenged by the extraordinarily different instrumentations used in each of the cover versions. For both the 2006 and 2007 ACS sessions, each of the 330 cover songs were used as queries and the systems were required to return 10 results for each query. Systems were evaluated on the number of the songs from the same class/set as the query that were returned in the list of 10 results for each query (i.e., precision measured using the top-ten results, also known as precision@10). In 2006, the best performing system (Ellis) achieved a 23.1% precision average. In 2007, the best performing system (Serra and Gomez) scored a 50.09% precision average. Three 2007 submissions, Ellis and Cotton (36.58%), Bello (26.33%), and Jensen, Ellis, Christiansen, and Holt (23.09%), did as well as, or better than, the best 2006 score.

While it is encouraging to note the remarkable increase in performance over one year, it is more telling to report the dismal performance of the IMIRSEL submission with its 10.03% precision average. The IMIRSEL performance was the worst of all eight 2007 ACS submissions. The IMIRSEL result is noteworthy because the IMIRSEL ACS submission was based upon the same naive spectral feature set (i.e., MFCCs, zero crossing rates, spectral flux, and spectral centroid data) as the IMIRSEL_SVM submission that ranked amongst the top submissions in both the 2007 AA (48.14%) and AG (68.29%) tasks. The unlike the IMIRSEL_SVM, the top performing ACS 2007 submissions were specifically designed to move beyond simple spectral-based similarity approaches (i.e., designed to capture such higher-order musical features as tonality, rhythm, and harmonic progressions, etc.). These new, more advanced, ACS systems are definitely leading the way for MIR research to move away from "timbral similarity" toward a more robust and meaningful conception of "musical similarity."

4.2. Introduction of Friedman's ANOVA and the Tukey-Kramer HSD

After completing MIREX 2005, it became apparent that MIREX needed to do more for the community than present task results as a set of rank ordered lists[3]. The community needed to know whether significant differences in system performances truly exist. For example, they needed to know whether System A with a hypothetical score of "72%" was really performing better than System B ("68%") and/or System C ("65%"), etc. Taking inspiration from the TREC analysis work of Tague-Sutcliffe and Blustein [16], the IMIRSEL team began exploring the use of Friedman's ANOVA. Friedman's is a non-parametric test (i.e., does not assume the normal distribution of the underlying data) [17]. Since many retrieval result sets have non-normal distributions [18], the Friedman test has been used in the TREC community for a number of years.

Friedman's ANOVA is a global test of significance. As a global test, it cannot tell us between which specific systems there exist significant differences. Thus, if a Friedman test does indicate the presence of a significant difference, we must then turn to a set of *post-hoc* pair-wise comparisons of each of the system results to locate the presence (or absence) of performance differences among the individual systems. To conduct the MIREX *post-hoc* comparisons, IMIRSEL chose the Tukey-Kramer "Honestly Significant Difference" (HSD) technique.

The Tukey-Kramer HSD is much superior to the commonly misused multiple Student's *t*-tests [18]. The problem with the typical naïve application of multiple *t*-tests to do *post-hoc* pair-wise comparisons lies in the fact that the probability of incorrectly rejecting the null hypothesis of no difference (i.e., H_0: $\mu_{(x)} = \mu_{(y)}$ at some confidence α (e.g., $\alpha = 0.05$)) increases in direct proportion with the number of pair-wise comparison conducted. For a complete systematic pair-wise comparison analyses there will be $x = c(c-1)/2$ comparisons made where c is the number of items in the set of interest. In our case, the items of interest are the final scores of each submitted system within a given task. When doing such multiple comparisons, the experiment-wide a level is defined as $\alpha = 1 - (1 - \alpha_{(\text{per comparison})})^x$ where x is the number of comparisons made. Thus, adapting Tague-Sutcliffe and Blustein's example to our own MIREX situation, in the case of the

[3]The ranking metric for each task is determined by participant consensus. Overall, the choice of a specific ranking criterion does appear to significantly effect the ordering. For example, the AMS task captures performance using six different metrics derived from evaluator judgments of the similarity between a query "seed" and a retrieved "candidate" based upon both a broad scale.

Table 3 Tukey-Kramer HSD pair-wise comparisons of top-ranked systems ($\alpha = 0.05$).

Comparison		Task						
Rank	**Rank**	**ACS06**	**AMS06**	**QBSH06**	**ACS07**	**AMS07**	**QBSH07**	**SMS07**
1	2	**TRUE**	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
1	3	**TRUE**	FALSE	FALSE	**TRUE**	FALSE	FALSE	FALSE
1	4	**TRUE**	FALSE	FALSE	**TRUE**	FALSE	**TRUE**	FALSE
2	3	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
2	4	FALSE	FALSE	FALSE	FALSE	FALSE	**TRUE**	FALSE
3	4	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE

AMS 2007 data, where there are 12 systems to be compared, there are $12(11)/2 = 66$ possible pair-wise comparisons. Thus, if each of these comparisons were tested at the $\alpha = 0.05$ level, the probability of incorrectly rejecting H_0 at least once would be $1 - (0.95)^{66} = 0.97$, i.e., almost a certainty. Analysts have many options for compensating for the multicomparison problem including the Bonferroni, Scheffé, and Tukey-Kramer techniques. For MIREX, IMIRSEL chose the Tukey-Kramer HSD method because it is not as conservative as the Bonferonni or Scheffé techniques (i.e., it is less prone to miss a true significant difference).

For MIREX 2006, IMIRSEL selected a set of tasks that have a strong resemblance to traditional IR tasks to undergo Friedman's ANOVA and Tukey-Kramer HSD testing. The set included ACS, AMS, and QBSH. In 2007, we expanded the set to include SMS. IMIRSEL used MATLAB to conduct the Tukey-Kramer HSD analyses using the following Statistics Toolbox command for each analysis:

```
[c,m,h,gnames] = multcompare(stats, 'ctype',
                             'tukey-kramer', 'estimate',
                             'friedman', 'alpha', 0.05)
```

Similar to the results found in past TREC evaluations, the Tukey-Kramer HSD data for MIREX 2006 and 2007 indicate that most MIR systems generally tend to perform on par with those closely ranked with them (with a few outlying exceptions) [15]. Therefore, it is more precise, in most cases, to characterize the relative performance success across submissions in terms of groups (i.e., top, middle, bottom, etc.) rather than baldly stating that System X is "best" at, or the "winner" of, a given MIREX task. Table 3 illustrates this point. Table 3 presents the pair-wise comparison results from the Tukey-Kramer HSD analyses for the four top-ranked submissions in each of our analyzed tasks. The null hypothesis for each pair-wise comparison is, at $\alpha = 0.05$, H_0: $\mu_{(\text{rank } x)} = \mu_{(\text{rank } y)}$, and the alternative hypothesis is H_a: $\mu_{(\text{rank } x)} \neq \mu_{(\text{rank } y)}$. In Table 3, rejection of H_0 is signified by TRUE (i.e., there is a significant difference between systems).

There are 72 pair-wise comparisons represented in Table 3. Of these, 65 (~90%) show no significant differences and only 7 (~10%) show significant differences between systems. Four of the seven tasks (AMS06, QBSH06, AMS07, and SM07) have no differences among the top-ranked systems. It is noteworthy that only in the ACS07 case do we see the 1st-ranked system consistently different from the lower-ranked systems. This system was the Serra and Gomez submission that was specifically designed to move beyond simple timbral similarity techniques and engaged the task through the use of the more musically meaningful idea of tonal sequencing.

Finally, like TREC, our Friedman analyses are indicating that the variance across queries is also statistically significant [15]. This significant variance across queries tells us that systems are individually performing better on ***different*** subsets of the query pools. This finding is important because understanding the interaction between queries and systems can lead researchers to develop better hybrid systems that account for differences in query characteristics. For MIREX 2008, we will attempt to convey a rational breakdown of the query variance information to the participants to see if this does indeed help the community build better systems. It is our hope that, for example, developers could combine System A's techniques, which did best on query subset X, with System B's techniques, which did best on query subset Y, etc.

5. CONCLUSIONS AND FUTURE WORK

5.1. Summary

By looking at the background, structure, challenges, and contributions of MIREX, this paper has provided some insights into the world of MIR research. Because the MIREX tasks are defined by the community they reflect the interests, techniques, and research paradigms of the community as a whole. Currently, both MIREX and MIR have strong biases toward audio-based and machine learning approaches as most MIR researchers have strengths in these areas. This has led to advancements in the MIR field but some of their simpler spectral-based techniques do appear to be reaching their limits of effectiveness. This limitation is called the "glass ceiling"

(i.e, Not Similar, Somewhat Similar, Very Similar) and a fine, continuous, 10-point scale [19]. Pampalk's [20] in-depth analysis of the 2006 AMS results data shows that the choice of metric did not alter the rank ordering of the 2006 AMS results.

problem and the MIREX result data support its existence. Recent results from MIREX 2007 suggest that community members are becoming more aware of the glass ceiling issue and have begun to explore techniques which strive to capture and exploit features that are more musically meaningful. The *post-hoc* analyses of results data indicate that there are groups of systems that perform equally well within the various tasks. However, within these groups, the individual systems appear to have unique abilities to deal with specific subsets of queries suggesting that MIR researchers should start developing hybrid approaches that combine the best aspects of the individual systems. There are many challenges facing MIREX with most of these having their root causes in the intellectual property issues surrounding music. The acquisition, validation, and storage of test collection and ground-truth data are ongoing challenges. Because MIREX cannot legally distribute its test collections, it has adopted a centralized algorithm-to-data model wherein all the submissions are run by IMIRSEL on behalf of MIREX. The current algorithm-to-data model is working for now but it is starting to be a limiting factor in the future growth of MIREX. Some challenges, like that of the inability of researchers to test approaches against the MIREX test collections outside the annual MIREX cycle, are hindering the rapid development of improved MIR systems.

5.2. Toward an Expanded Vision for MIREX

To ensure the future viability of MIREX, IMIRSEL has put together a consortium of six research labs to collectively work upon overcoming the current challenges facing MIREX. This consortium, called the Networked Environment for Music Analysis (NEMA), includes MIR labs from UIUC, USA (Downie, PI); McGill University, CA (Ich Fujinaga, Co-PI); Queen Mary College, University of London, UK (Mark Sandler); Goldsmiths College, University of London, UK (Tim Crawford); University of Southampon, UK (David De Roure); and, University of Waikato, NZ (David Bainbridge). In January of 2008, the NEMA consortium received $1,200,000 USD in funding from the Andrew W. Mellon Foundation. This funding will support MIREX though its next three cycles (2008, 2009, and 2010). At the same time, the NEMA group will be developing an open and extensible webservice-based resource framework that facilitates the integration of music data and analytic/evaluative tools that can be used by the global MIR research community on a basis independent of time or location. The NEMA team hopes to establish ongoing connections with other large-scale MIR research groups such as CrestMuse[4] (Japan), OMRAS II[5] (United Kingdom), and MTG[6] (Spain), etc. to maximize the benefits of international research interchanges. The key problems that will be addressed by the NEMA consortium are best summarized as:

1. ***Resource accessibility.*** For example, new means to provide access to good ground-truth sets, to broad-based music collections, to feature sets, and to pre-built models, etc. must be found. Also, in the case of music collections where items from the music collections will not be able to move about, new ways of bringing researchers and their tools to the data need to be constructed. It is important to envision a future where many different collections of music materials are independently made available in such a way as to create a much larger and diverse "super-collection." Such super-collections are needed to address over-fitting issues. They are also needed to allow for better scalability/stress testing of approaches. Finally, new methods of creating and providing on-demand computational and storage resources to the MIR community need to be explored.
2. ***Resource discovery.*** For example, even if the aforementioned resources were readily available it is still necessary to create appropriate music-specific location and discovery tools so that individual items or resource subsets might be put to use.
3. ***Resource sharing/re-use.*** For example, new standards for ground-truth and feature sets must be developed to facilitate their re-use. Mechanisms need to put into place to make it easy for researchers to store and then make their sets available to others. In the same manner, mechanisms must be put in place to overcome the interoperability problems that limit the re-use of research code, including feature extractors, classifiers, and pre-built classification models, etc.
4. ***Resource customization.*** For example, new ways need to be developed to help researchers amalgamate aspects of independently produced feature sets to create novel feature sets. New techniques must be found to easily create on-demand "virtual" collections that span across several real-world collections regardless of their physical location. Again, interoperability problems among research code sets must be overcome so that researchers can create customized hybrid systems that integrate tools from many different research labs.

By making progress toward overcoming these problems, the NEMA group offers the promise of a new and expanded MIREX and an improved research paradigm for MIR. Under this new paradigm, it should become possible for MIR researchers to overcome limitations of time-

[4]See http://www.crestmuse.jp.
[5]See http://www.omras2.com.
[6]See http://mtg.upf.edu.

specific and location-specific resources. In the new NEMA reality, for example, it should become commonplace for researchers at **Lab A** to easily build a virtual collection from **Library B** and **Lab C**, acquire the necessary ground-truth from **Lab D**, incorporate a feature extractor from **Lab E**, amalgamate the extracted features with those provided by **Lab F**, build a set of models based on pair of classifiers from **Labs G** and **H** and then validate the results against another virtual collection taken from **Lab I** and **Library J**. Once completed, the results and newly created features sets would be, in turn, made available for others to build upon.

5.3. Concluding Remarks: Getting Involved

If you are interested in general MIR research developments I suggest that you subscribe to the music-ir@ircam.fr mailing list. Subscription instructions can be found at http://www.ismir.net. This active list discusses the wide range of MIR issues. If you are interested in MIREX participation, you need to subscribe to the evalfest@mail.lis.uiuc.edu list. EvalFest subscription instructions, and links to its archive, are found at https://mail.isrl.uiuc.edu/mailman/listinfo/evalfest. I also recommend you consider submitting to, and attending, the ever growing ISMIR series of conferences. ISMIR began in 2000 at Plymouth, Massachusetts, USA with 95 participants and yielded 35 published items. ISMIR 2007, held in Vienna, Austria, had over 250 registrants and published 131 peer-reviewed items across the paper, poster, and demo categories. Future ISMIR conferences are scheduled for: 2008 in Philadelphia, USA; 2009 in Kobe, Japan; and, 2010 in Utrecht, Netherlands.

ACKNOWLEDGEMENTS

MIREX has received considerable financial support from both the Andrew W. Mellon Foundation and the National Science Foundation (NSF) under grants NSF IIS-0340597 and NSF IIS-0327371. I would like to thank the editor and the reviewers for their excellent revision suggestions.

REFERENCES

[1] J. Futrelle and J. S. Downie, "Interdisciplinary research issues in music information retrieval: ISMIR 2000–2002," *J. New Music Res.*, **32**, 121–131 (2003).

[2] J. S. Downie, "Music information retrieval," *Annu. Rev. Inf. Sci. Technol.*, **37**, 295–340 (2003).

[3] N. Orio, "Music information retrieval: A tutorial and review," *Found. Trends Inf. Retr.*, **1**, 1–90 (2006).

[4] M. Fingerhut, Ed., *Cumulative ISMIR Proceedings*, Available: http://www.ismir.net/proceedings.

[5] J. S. Downie, J. Futrelle and D. Tcheng, "The International Music Information Retrieval Systems Evaluation Laboratory: Governance, access, and security," *Proc. ISMIR 2004*, pp. 9–14 (2004).

[6] J. S. Downie, K. West, K. A. Ehmann and E. Vincent, "The 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005): Preliminary overview," *Proc. ISMIR 2005*, pp. 320–323 (2005).

[7] P. Cano, E. Gomez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich and N. Wack, "ISMIR 2004 Audio Description Contest," *MTG Tech. Rep.*, MTG-TR-2006-02 (Music Technology Group, Barcelona, Spain, 2004).

[8] J. S. Downie, "The scientific evaluation of music information retrieval systems: Foundations and future," *Comput. Music J.*, **28**, 12–23 (2004).

[9] E. Voorhees and D. Harmon, "The Text Retrieval Conference," in *TREC Experiment and Evaluation in Information Retrieval*, E. Voorhees and D. Harmon, Eds. (MIT Press, Cambridge, Mass., 2005), pp. 3–20.

[10] J.-J. Aucouturier and F. Pachet, "Improving timbre similarity: How high is the sky?" *J. Negative Results Speech Audio Sci.*, **1**, http://www.csl.sony.fr/downloads/papers/uploads/aucouturier-04b.pdf, in press (2004).

[11] B. Whitman, G. Flake and S. Lawrence, "Artist detection in music with Minnowmatch," *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pp. 559–568 (2001).

[12] Y. E. Kim, D. S. Williamson and S. Pilli, "Towards quantifying the 'album effect' in artist identification," *Proc. ISMIR 2006*, pp. 393–394 (2006).

[13] E. Pampalk, A. Flexer and G. Widmer, "Improvements of audio-based music similarity and genre classification," *Proc. ISMIR 2005*, pp. 260–263 (2005).

[14] A. Flexer, "A closer look on artist filters for musical genre classification," *Proc. ISMIR 2007*, pp. 341–344 (2007).

[15] C. McKay and I. Fujinaga, "Musical genre classification: Is it worth pursuing and how can it be improved?" *Proc. ISMIR 2006*, pp. 101–106 (2006).

[16] J. Tague-Sutcliffe and J. Blustein, "The statistical analysis of the TREC-3 data," in *Overview of the Third Text Retrieval Conference*, D. Harmon, Ed. (NIST, Gaithersburg, Md., 1995), pp. 385–398.

[17] M. L. Berenson, D. M. Levine and M. Goldstein, *Intermediate Statistical Methods and Applications: A Computer Package Approach* (Prentice-Hall, Englewood Cliffs, N.J., 1983).

[18] W. J. Conover, *Practical Non-Parametric Statistics* (Wiley, New York, 1980).

[19] M. C. Jones, J. S. Downie and A. F. Ehmann, "Understanding human judgments of similarity in music information retrieval," *Proc. ISMIR 2007*, pp. 101–106 (2007).

[20] E. Pampalk, "Audio-based music similarity and retrieval: Combining a spectral similarity model with information extracted from fluctuation patterns," *MIREX 2006*, Submission Abstracts (2006) Available: http://www.music-ir.org/evaluation/MIREX/2006_abstracts/AS_pampalk.pdf.

J. Stephen Downie is an Associate Professor at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. He holds a BA in Music Theory and Composition (1988), MLIS (1993) and PhD (1999) in Library and Information Science, all from the University of Western Ontario. Professor Downie is Director of the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL). He has been very active in the establishment of the Music Information Retrieval (MIR) and Music Digital Library (MDL) communities through his ongoing work with the ISMIR series of MIR conferences as a member of the ISMIR steering committee.