The Music Information Retrieval Evaluation eXchange: Some Observations and Insights

J. Stephen Downie, Andreas F. Ehmann, Mert Bay, and M. Cameron Jones

Abstract. Advances in the science and technology of Music Information Retrieval (MIR) systems and algorithms are dependent on the development of rigorous measures of accuracy and performance such that meaningful comparisons among current and novel approaches can be made. This is the motivating principle driving the efforts of the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) and the annual Music Information Retrieval Evaluation eXchange (MIREX). Since it started in 2005, MIREX has fostered great advancements not only in many specific areas of MIR, but also in our general understanding of how MIR systems and algorithms are to be evaluated. This chapter outlines some of the major highlights of the past four years of MIREX evaluations, including its organizing principles, the selection of evaluation metrics, and the evolution of evaluation tasks. The chapter concludes with a brief introduction of how MIREX plans to expand into the future using a suite of Web 2.0 technologies to automated MIREX evaluations.

Keywords: MIREX, Music Information Retrieval, Evaluation.

1 Introduction

Since 2005, a special set of sessions has convened at the annual International Conference on Music Information Retrieval (ISMIR). At these special sessions, which include a poster exhibition and a plenary meeting, Music Information Retrieval (MIR) researchers from around the world come together to compare, contrast and discuss the latest results data from the Music Information Retrieval Evaluation eXchange (MIREX). MIREX is to the MIR community what the Text Retrieval Conference (TREC) is to the text information retrieval community: A set of

J. Stephen Downie · Andreas F. Ehmann · Mert Bay · M. Cameron Jones
International Music Information Retrieval Systems Evaluation Laboratory
Graduate School of Library and Information Science,
University of Illinois at Urbana-Champaign,
501 East Daniel Street, Champaign, Illinois, USA 61820
jdownie@illinois.edu

Z.W. Raś and A.A. Wieczorkowska (Eds.): Adv. in Music Inform. Retrieval, SCI 274, pp. 93–115.
springerlink.com

community-defined formal evaluations through which a wide variety of state-of-the-art systems, algorithms and techniques are evaluated under controlled conditions. MIREX is managed by the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) at the University of Illinois at Urbana-Champaign (UIUC).

This chapter builds upon and extends Downie [5]. While some overlap is unavoidable, the reader is strongly encouraged to consult the earlier work for it covers important issues that will be missing detailed discussion in the current chapter. Unlike Downie [5], the present chapter will focus on the evolution of MIREX throughout the years. It will highlight some specific issues, problems and challenges that have emerged during the running of MIREX.

Section 2 reviews the history, special characteristics and general operations of MIREX. Section 3 examines the important primary metrics used to evaluate the algorithms. Section 4 takes an in-depth look at two closely related tasks, Audio Music Similarity (AMS) and Symbolic Melodic Similarity (SMS), as a kind of case study in the evolution of MIREX. Section 5 summarizes the chapter and briefly introduces the NEMA (Networked Environment for Music Analysis) project which is designed to support and strengthen MIREX and the ongoing formal evaluation of MIR systems.

2 History and Infrastructure

While MIREX officially began in 2005, it took a considerable amount of time and collective effort to make it a reality. In 1999 Downie led the *Exploratory Workshop on Music Information Retrieval* as part of the 1999 ACM Special Interest Group Information Retrieval (SIGIR) Conference at Berkeley, CA. One of the primary goals of this workshop was the exploration of "(…) consensus opinion on the establishment of research priorities, inter-disciplinary collaborations, evaluation standards, test collections (…) and TREC-like trials" [2]. In 2001, the attendees of ISMIR 2001 at Bloomington, IN passed a resolution calling for the establishment of formal evaluation opportunities for MIR researchers. This resolution helped garner modest feasibility study grants from the Andrew W. Mellon Foundation and the National Science Foundation. With the funds provided by the grants, a sequence of workshops was convened and a collection of evaluation white papers was compiled [3]. The recommendations based upon these workshops and white papers were subsequently published as [4] and worked into several grant applications. In late 2003, both the Andrew W. Mellon Foundation and the NSF awarded the substantial grants that were to make MIREX possible. ISMIR 2004 was held in Barcelona, ES and was hosted by The Music Technology Group (MTG) of the University Pompeu Fabra. At this meeting the MTG convened an evaluation session called the Audio Description Contest (ADC) [1]. While more limited in scope than MIREX, it is from the ADC that MIREX learned many valuable lessons. MIREX made its debut as part of ISMIR 2005, held at Queen Mary College, University of London, in September, 2005.

TREC, ADC and MIREX share a common intellectual foundation. All three are predicated upon the standardization of:

1. Test collections of considerable size;
2. Tasks and/or queries to be performed against the collections; and,
3. Evaluation procedures to compare performances among systems.

It is important to note that MIREX differs from TREC in one significant way. Unlike TREC, where the evaluation datasets are sent out to the participant labs, MIREX operates under an "algorithm-to-data" model. This means that algorithms are sent to IMIRSEL to be run on IMIRSEL equipment by IMIRSEL personnel and volunteers. While the algorithm-to-data model puts a considerable burden on IMIRSEL resources in terms of workload, data management and equipment, it is currently the only feasible solution to working within the boundaries of the highly restrictive and litigious legal environment surrounding music intellectual property law.

Beyond computational infrastructure, IMIRSEL also hosts the basic communications infrastructure for MIREX which includes the MIREX wikis[1] and the MIREX mailing lists. The MIREX wikis serve two purposes. First, during the spring and summer each year, they are used by the community to define the task sets, evaluation metrics and general rules for the year's upcoming MIREX. Second, the wikis are used to publish and archive the raw and summarized results data for each task and associated algorithms just prior to ISMIR convening each autumn. These results data are used by participants to help them put together their mandatory poster presentations and for further use in follow up publications. The MIREX "EvalFest"[2] mailing list is the general purpose mailing list that is used to solicit task ideas and collections and to foster broad discussions about evaluation issues. On a case-by-case basis, IMIRSEL also creates task-specific mailings lists through which finely detailed discussions and debates about metrics, collections and input/output formats, etc. are undertaken.

Table 1 MIREX Descriptive Statistics 2005-2008

	2005	2006	2007	2008
Number of Task (and Subtask) "Sets"	10	13	12	18
Number of Individuals	82	50	73	84
Number of Countries	19	14	15	19
Number of Runs	86	92	122	169

MIREX has enjoyed considerable growth over its four year history. As Table 1 indicates, the number of task sets (including subtasks) has grown 80% from 10 (2005) to 18 (2008). The number of individual algorithms evaluated has similarly grown 95% from 86 (2005) to 169 (2008). In total, MIREX has evaluated 469 algorithm runs.

The range of MIREX tasks broadly reflects the varied interests of the MIR research community. Many tasks, such as Audio Artist Identification (AAI), Symbolic

[1] See http://music-ir.org/mirexwiki
[2] Subscription information at https://mail.isrl.illinois.edu/mailman/listinfo/evalfest

Genre Classification (SGC), Audio Genre Classification (AGC), Audio Mood Classification (AMC), and Audio Tag Classification (ATC), represent classic machine learning train-test classification evaluations. These classification tasks accounted for 28% (129) of MIREX's 469 evaluation runs. Other tasks, such as Audio Beat Tracking (ABT), Audio Chord Detection (ACD), Audio Melody Extraction (AME), Audio Onset Detection (AOD) and Multiple F_0 Estimation (MFE), etc. are "low-level" tasks that represent tools and techniques upon which many MIR systems depend. For example, many systems use melody extractors as a first step toward building searchable indexes. The development and evaluation of low-level MIR subsystems is important to the MIR community as this category of evaluation comprised 201 (43%) of the MIREX evaluation runs. Audio Cover Song Identification (ACS), Audio Music Similarity (AMS), Query-by-Singing/Humming (QBSH), Query-by-Tapping (QBT) and Symbolic Melodic Similarity (SMS) are those tasks related to what most people would consider to be MIR, that is, the idea of searching for music given some type of music query. QBSH has been the single most evaluated task with over 59 individual runs

Table 2 The MIREX Tasks and Number of Runs per Task 2005-2008

KEY	TASK NAME	2005	2006	2007	2008
AAI	Audio Artist Identification	7		7	11
ABT	Audio Beat Tracking		5		
ACD	Audio Chord Detection				15*
ACC	Audio Classical Composer ID			7	11
ACS	Audio Cover Song Identification		8	8	8
ADD	Audio Drum Detection	8			
AGC	Audio Genre Classification	15		7	26*
AKF	Audio Key Finding	7			
AME	Audio Melody Extraction	10	10*		21**
AMC	Audio Mood Classification			9	13
AMS	Audio Music Similarity		6	12	
AOD	Audio Onset Detection	9	13	17	
ATC	Audio Tag Classification				11
ATE	Audio Tempo Extraction	13	7		
MFE	Multiple F_0 Estimation (Frame Level)			16	15
MFN	Multiple F_0 Note Detection			11	13
QBSH	Query-by-Singing/Humming		23*	20*	16*
QBT	Query-by-Tapping				5
SF	Score Following		2		4
SGC	Symbolic Genre Classification	5			
SKF	Symbolic Key Finding	5			
SMS	Symbolic Melodic Similarity	7	18 **	8	

* task comprised two subtasks ** task comprised three subtasks

evaluated over 2006-2008 (or 13% of runs evaluated). As a category, these searching tasks represented 139 or 30% of all MIREX runs.

3 Overview of MIREX Primary Evaluation Metrics

Table 3 Top MIREX Scores and Primary Evaluation Metrics 2005-2008 (normalized 0-1)

KEY	PRIMARY METRIC	2005	2006	2007	2008
AAI	Average Accuracy	0.72		0.48	0.48
ABT	Average Beat P-Score		0.41		
ACD1	Overlap Score				0.66
ACD2	Overlap Score				0.72
ACC	Average Accuracy			0.54	0.53
ACS	Average Precision		0.23	0.52	0.75
ADD	Average F-Measure	0.67			
AGC1	Average Hierarchical Accuracy	0.83		0.68	0.66
AGC2	Average Accuracy				0.65
AKD	Average Hierarchical Accuracy	0.90			
AMC	Average Accuracy			0.62	0.64
AME1	Average Accuracy	0.71	0.73		0.70
AME2	Average Accuracy		0.83		0.85
AME3	Average Accuracy				0.76
AMS	Average Fine Score		0.43	0.57	
AOD	Average F-Measure	0.80	0.79	0.81	
ATE	Average F-Measure				0.28
ATE	Average Tempo P-Score	0.69	0.81		
MFE	Average Accuracy			0.62	0.67
MFN	Average F-Measure			0.61	0.61
QBSH1	Average Precision (Mean Reciprocal Rank)		0.93	0.93	0.93
QBSH2	Average Precision		0.93	0.94	0.94
QBT	Average Precision (Mean Reciprocal Rank)				0.52
SF	Average Precision		0.83		0.67
SGC	Average Hierarchical Accuracy	0.77			
SKD	Average Hierarchical Accuracy	0.91			
SMS1	Average Dynamic Recall [Binary Score][3]	0.66	0.72 [0.73]		
SMS2	Average Dynamic Recall [Binary Score]		0.82 [0.44]		
SMS3	Average Dynamic Recall [Binary Score]		0.78 [0.83]		
SMS4	Average Dynamic Recall [Fine Score]			0.72 [0.56]	

[3] The SMS tasks used two different metrics as “primary” scores. Each year the Average Dynamic Recall (ADR) score was reported along with either the Binary Score or the Fine Score (presented in square brackets). More information about these metrics found in Sections 3.6 and 4.2.

The selection or creation of appropriate evaluation metrics is crucial to the proper scientific evaluation of MIR system performance. The selection of evaluation metrics also has a strong emotional component as participants strive to show off the success of their algorithms and systems in the best possible terms. Thus, it is not surprising that the selection/creation of MIREX evaluation metrics undergoes a great deal of sometimes heated discussion while the tasks are being developed by the community. Table 3 summarizes the primary evaluation metrics used in each task along with the top-ranked score using that metric for each year. These are the primary, or "official," metrics used to rank order the MIREX results each on the master MIREX results poster published each year at ISMIR. It is important to note, however, that most tasks are actually evaluated using a wide variety of metrics. For each task, the results using the other metrics are summarized and posted on each respective results wiki page. Notwithstanding all the debate over which metric should become the "official" metric, we are discovering a general trend among the tasks that the rank order of system performances within a task appears remarkably stable regardless of the metric chosen [14]. As one can see in Table 3, MIREX uses many different primary metrics; some are used over a range of related tasks, others are tailored specifically to one. We will now discuss the primary evaluation metrics used by MIREX in evaluating MIR performance along with some of the justifications for using the metrics.

3.1 Average Accuracy and Hierarchical Accuracy

In classification tasks such as Audio Artist Identification (AAI), performance can be measured using classification accuracy. Given N_{total} pieces to be classified, the average accuracy of a classifier, *Acc*, can be defined as

$$Acc = \frac{N_{correct}}{N_{total}} \tag{1}$$

where $N_{correct}$ is the number of correctly classified instances. Average accuracy is also applicable to such tasks as Audio Melody Extraction (AME) where accuracy measures the number of analysis frames where fundamental frequencies (F_0s) are correctly estimated versus the total number of frames in a piece.

For Multi-F_0 Estimation (MFE) accuracy is calculated as

$$Acc = \frac{TP}{TP + FP + FN} \tag{2}$$

where *TP* is the count of true positives, *FP* is the count of true negatives and *FN* is the count of false negatives. In the MFE task, where the number of active F_0s in the ground-truth changes per frame, *TP* is the number of correctly detected F_0s per frame summed over all frames. *FP* is the number of detected F_0s that are not in the ground-truth list for that frame summed over all frames and *FN* is the number of F_0s in the ground-truth list minus the number of detected F_0s for that frame summed over all frames.

In some cases, misclassifications can occur that are not as “erroneous” or “offensive” as others. For instance, it is generally consider “better” to misclassify a Baroque work as Classical than Heavy Metal. Similarly, having a system misclassify a hard-driving Blues song as a Rock & Roll song is quite understandable as it is the same kind of “error” that many humans might make. In the cases of the Symbolic and Audio Key Detection (SKD and AKD) tasks, misclassifying a key with its perfect fifth is also considered a somewhat acceptable mistake (i.e., many humans make the same error). For these reasons, some tasks are also evaluated using hierarchical accuracies, which discount certain acceptable errors. For example, in the two key finding tasks, correct keys were awarded 1.0 point, perfect fifth errors were given 0.5 points, relative major/minor errors 0.3 points, and parallel major/minor errors 0.2 points. Therefore, the hierarchical accuracy,
Acc_H can be expressed as

$$Acc_H = \frac{N_{correct} + 0.5E_{fifth} + 0.3E_{relative} + 0.2E_{parallel}}{N_{total}} \tag{3}$$

where E_{fifth} represents the number of perfect-fifth errors, $E_{relative}$ the number of relative major/minor errors, and $E_{parallel}$ the number of parallel major minor errors. Similarly, for the Audio Genre Classification (AGC) and Symbolic Genre Classification (SGC) tasks, a genre hierarchy was employed such that errors between similar genres were only discounted a half point (e.g., Jazz and Blues, Classical and Romantic, etc.).

3.2 Precision, Recall, and F-Measure

Consider a system that when given a query, returns a list of documents/items that it “believes” are the proper responses to the query. If an item is returned, and it is relevant, it can be considered a true positive, *TP*. If a document is returned and it is not relevant it is a false positive, *FP*. If a document is not returned, but is relevant it is a false negative, *FN*. Finally if a document is not returned and is not relevant, it is a true negative, *TN*. Using this system of *TP*, *FP*, *TN* and *FN* documents we can now define the two “classic” information retrieval evaluation metrics whose use predates the use of computers: *precision* and *recall.*

We can define the precision, *P*, as

$$P = \frac{TP}{TP + FP} \tag{4}$$

Put simply, precision is the ratio of relevant returned documents to the total number of returned documents. Recall, *R*, on the other hand is the ratio of relevant returned documents to the total number of relevant documents available in a system, and is expressed as

$$R = \frac{TP}{TP + FN} \tag{5}$$

These two measures can be combined into a single measure called the F-Measure, F, which is the harmonic mean of precision and recall:

$$F = \frac{2PR}{P+R} \tag{6}$$

As an example of how *precision*, *recall*, and F-Measure are appropriate to a MIR task, consider the Audio Onset Detection (AOD) task. AOD concerns itself with finding the start times of all sonic events in a piece of audio. In the AOD task, each system outputs its predicted onset times for a piece of audio, and these predictions are compared to a ground-truth of manually annotated onsets. Assume, for example, the ground-truth annotation of an audio snippet contains 100 onsets of audio events. Furthermore, let us assume an algorithm returns 80 onsets, 60 of which are correct. In this case, the precision would be 60/80 (0.75) and the recall 60/100 (0.60).

To better understand why we are interested in the F-Measure, let us now consider some extreme situations. Assume an algorithm predicts that there are one million onsets and, because these predicted onsets are so densely spaced, it has subsequently managed to predict the locations of the 100 true onsets in the piece. In this case, because all of the onsets that were in the ground-truth were correctly recalled, we have a case of perfect *recall*, i.e., 100/100 (1.0). However, only 100 of the one million returned onsets were correct, causing *precision* to drop to 100/1,000,000 (0.0001). Conversely, let us assume that an algorithm only returns one single onset, which is correct. In this case the *precision* is perfect 1/1 (1.0), but the *recall* is quite small, 1/100 (0.01). In general, *recall* and *precision* are traded at the expense of the other. The two measures are combined in the F-Measure. If either *recall* or *precision* are very low, the F-measure will be as well. Therefore, F-Measure rewards those systems that have the best balance of simultaneously high *recall* and *precision* scores.

3.3 Mean Reciprocal Rank

In query-based tasks such as Query-by-Singing/Humming (QBSH), system performance can be measured with mean reciprocal rank. Consider a system that returns a ranked list of results given a query. For example, QBSH systems are designed to return a ranked list of songs in response to a user singing a melody into a microphone. If the desired response to a query such a "Twinkle Twinkle Little Star" is third in the returned list (i.e., rank 3), the reciprocal rank is 1/3. If we take the mean of the reciprocal ranks over all queries, we arrive at the mean reciprocal rank (MRR) which can be formally expressed as

$$MRR = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{rank_n} \tag{7}$$

where N is the number of queries, and $rank_n$ is the rank of the correct response of query n in the returned response list. This metric rewards systems for placing the desired items near the top of the ranked lists and quickly penalizes those systems that returned the desired items lower in the list.

3.4 Audio Tempo and Audio Beat P-Score

In some MIREX tasks, more heuristic measures, derived from the outcomes of real-world user experiments, are used for evaluation. For example, in the Audio Tempo Extraction (ATE) task, two dominant tempi were required to be returned by each algorithm. This two tempi approach was adopted based on the work of McKinney and Moelants [13] which showed that perceived tempi vary among real listeners in a relatively discrete and integer-based manner (i.e., some persons hear a given song at 60 beats per minute (bpm) while others hear the same song at 120 bpm). The ATE ground-truth contained the two true dominant tempi, as well as a salience of the first tempo. Denoting this salience as α, we compute the P_{tempo} score as

$$P_{tempo} = \alpha T_1 + (1-\alpha)T_2 \tag{8}$$

where T_1 is 1 if the first tempo is identified within ± 8% of the ground-truth tempo value and 0 otherwise. Likewise, T_2 takes the value 1 if the second tempo was identified and 0 otherwise.

In the Audio Beat Tracking (ABT) task, the work of Moelants and McKinney also was used as a basis for the choice of metric. For the ABT task each evaluated system provided beat times extracted from a piece of audio, akin to the tapping of a foot along with the music. These extracted beat times are compared to 40 ground-truth beat tracks for each musical piece collected from humans tapping to the beat of the piece. The beat times for the algorithms and each ground-truth are converted to an impulse train (at a 100 Hz sampling rate), and a cross correlation of the algorithm output impulse train, $y[n]$, and the ground-truth, $a_s[n]$, is measured. The cross correlation is calculated across a small window of possible delays between $-W$ and $+W$ (where W is 1/5th of the beat length). The correlations are then averaged across the S (S=40) ground-truths. Therefore, the beat P-Score, P_{beat}, can be expressed as

$$P_{beat} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{N_p}\sum_{m=-W}^{+W}\sum_{n=1}^{N} y[n]a_s[n-m] \tag{9}$$

where N_p is a normalization factor for the number of beats and is calculated as

$$N_p = \max\left(\sum y[n], \sum a_s[n]\right) \tag{10}$$

3.5 Chord Detection Overlap Score

The chord detection overlap score was specifically designed for the MIREX Audio Chord Detection (ACD) task. For the ACD task, each system had to return the chord names along with their associated onset and offset times within a piece of music. A good performance criterion in such a situation is to measure the amount of overlap in time between the detected chords and the ground-truth. A system's returned chords can be represented as

$$C_{dt}(c,t) = \begin{cases} 1, & \text{if chord } c \text{ is active at time } t \\ 0, & \text{else} \end{cases} \tag{11}$$

Then the overlap score can be calculated as

$$\text{Overlap Score} = \frac{\int_t C_{dt}(c,t) \cdot C_{gt}(c,t)dt}{\int_t C_{gt}(c,t)dt} \tag{12}$$

where $C_{dt}(c,t)$ and $C_{gt}(c,t)$ are the detected and the ground-truth chords.

In the 2008 MIREX ACD task, the systems were restricted to return one of 25 different chord types rooted on the 12 pitches of the chromatic scale (i.e., 12 major chords, 12 minor chords, and no chord). Therefore, c is a discrete variable. Moreover, systems were only allowed to return one active chord at any given time.

3.6 Average Dynamic Recall

For such similarity tasks as Audio Music Similarity (AMS) and Symbolic Melodic Similarity (SMS) where systems return relevant items as a list ranked according to "similarity" to a given query, the underlying relevance measure is subjective because it is based on the biases, tastes, levels of expertise and foibles of the human assessors providing the assessments. For example, in the MIREX 2005 Symbolic Melodic Similarity (SMS05) task, systems returned a ranked list of songs whose melodies were "similar" to the query song. The ground-truth for SMS05 task was generated by humans manually scoring every query against every returned result in a set of pre-MIREX experiments conducted by researchers at Utrecht University [16]. It was Rainer Typke, then a graduate student at Utrecht University, who took the lead on proposing and organizing the SMS05 task. Using the *Répertoire International des Sources Musicales (RISM). Serie A/II, manuscrits musicaux après 1600* collection [15] of incipits, Typke and his colleagues at Utrecht created a ground-truth set of similarity judgments. For each of 11 queries, the Utrecht team had 35 music experts rank order the pre-filtered individual results (about 50 per query) based on similarity to the original query. The median ranks assigned to the retrieved incipits were subjected to the Wilcoxon rank sum test. This statistical testing procedure allowed the Utrecht team to create groups of results that contained items of comparable similarity while at the same time being able to order the groups themselves with regard to similarity to the query. It was these 11 lists of group-ordered incipits that formed the SMS05 ground-truth set.

Because the ground-truth generation involved highly subjective human judgments, it is quite reasonable to treat the ground-truth rank list not in an absolute item-by-item sense but rather in a more relativistic sense of ranked groups of equivalently similar items. For example, the first and the second items in a list might have slightly different scores but for all intents and purposes they are equivalently similar to a given query. To reflect this state of affairs Typke and his Utrecht colleagues developed the Average Dynamic Recall (ADR) metric [17].

ADR measures the average recall over the first *n* documents with a dynamic set of relevant documents where *n* is the number of documents in the ground-truth list. The set of relevant documents grows with the position in the list but not just by one item. For example if the ground-truth has two groups of equally relevant items such as <(1, 2), (3, 4, 5)>, then at position number 2 there are 2 relevant documents whereas, at position number 3 there are a total of 5 relevant documents. However, at each position in the results list the recall is calculated by dividing the number of relevant items with the position number, not the total number of relevant documents. Thus, the above definition can be written formally as

$$r_i = \#\{R_w \mid w \le i \wedge \exists j,k : j \le c \wedge R_w = G_k^j\} / i \qquad (13)$$

$$ADR = \frac{1}{n}\sum_{i=1}^{n} r_i \qquad (14)$$

where $< R_1, R_2, \ldots >$ is the returned results list. The ground-truth list has *g* groups such as $< (G_1^1, G_2^1, \ldots, G_{m1}^1), (G_1^2, G_2^2, \ldots, G_{m2}^2), \ldots . (G_1^g, G_2^g, \ldots, G_{mg}^g) >$. The ranking does not matter within each group. For example we do not know if $rank(G_i^j)$ is less than $rank(G_k^j)$. However, we do know that $rank(G_i^j)$ is less than $rank(G_i^k)$ given that $j < k$. Also *c* in the above equation is the group number that contains the i^{th} item in its group. The key point to remember about ADR is that it represents an attempt to mitigate the distorting effects of relying solely on absolute (mostly minute) differences in human-generated similarity scores by allowing for grouping of functionally similar items.

4 Evolution of Similarity Evaluation Tasks

The set of similarity tasks, Audio Music Similarity (AMS) and Symbolic Melodic Similarity (SMS), comprises those tasks that most closely resemble a classic information retrieval scenario. That is, for a given piece of music submitted as a query, the systems under evaluation are expected to return a ranked list of music pieces that are deemed to be similar to the query. In many ways, it is this scenario that most people think of when they think of MIR systems in real-world deployments. In this section we explore this issues raised in running the set of similarity tasks run by MIREX between 2005 and 2007. We will also examine how responding to these issues caused the structure of the similarity tasks to evolve over time.

4.1 2005 Symbolic Melodic Similarity

In 2005, there was no running of an AMS task as the community could not decide how it would set up the ground-truth for such a task. However, as mentioned in Section 3.6, it was Rainer Typke who was the intellectual leader of SMS05 because he had the ground-truth set in hand that he and his Utrecht colleagues had already created (SMS1 in Table 3). SMS05 had 7 algorithms evaluated. The best algorithm had an ADR score of 66% while the worst had a 52% ADR score. Overall, participants were pleased with the 2005 running of SMS. However, several important issues arose that would influence future similarity task runs at MIREX. First, it was obvious that the 11 queries contained in the Utrecht ground-truth set was not a big enough set upon which to make broad generalizations about system behaviors. Second, there was some concern that Utrecht's pre-filtering step might have removed potentially relevant items from the ground-truth. Third, generating new ground-truth data for MIREX 2006 using the Utrecht method would not be possible given the time and manpower constraints of MIREX. Fourth, the RISM collection does not encompass a wide enough range of music to represent all the musical tastes, genres and styles of interest to MIR developers and users. Fifth, no one in the MIREX community could come up with a feasible way to replicate the Utrecht method to generate ground-truth for a meaningful collection of audio-based music files.

4.2 2006 Symbolic Melodic Similarity and Audio Music Similarity

Given the issues raised after the running of SMS05, the IMIRSEL team worked with the MIREX community to define a general framework for MIREX 2006 that could be used to construct both the SMS06 and Audio Music Similarity (AMS06) tasks. The principal difference that would set the 2006 tasks apart from the 2005 task would be the adoption of a more TREC-like evaluation scenario. That is, rather than creating and using a pre-compiled ground-truth set, the evaluation of retrieved results would be conducted *post-hoc* using human judges (or "graders") to score the similarity between queries and their respective returned items.

Acceptance of the *post-hoc* set up was not controversial. However, as with many things community-based, three issues became hotly debated. The first point of contention was the choice of evaluation metric. The second was the number of graders that would be used to evaluate the results. The third issue was basic feasibility.

Many community members, including the authors of this chapter, favoured a simple binary measure of similarity/relevance. That is to say, a returned piece (also known as a "candidate") was, or was not, similar/relevant to the query piece. The binary approach would make the use of classic precision (see Section 3.2) easy to calculate. Others fought vigorously for some type of graduated "broad" judgment (e.g., Not Similar (NS), Somewhat Similar (SS), Very Similar (VS)) to better reflect the nuances in perceptions of similarity. Deciding upon methods for weighting the relative importance of NS, SS and VS opened up another debate thread. A continuous fine-scaled approach was also suggested to more "accurately" capture subtle differences in similarity. This fine-scale could be represented using some kind of slider that a grader could position between 0.0 (NS) and 10.0 (VS). The sum or averages of the slider-generated values could then be used to evaluate the success of each system for each query/candidate pair.

The number of graders per query/candidate pair was another contentious issue. One group, including this set of authors, wanted to mimic the traditional TREC approach of one grader per query. Others argued strongly that music relevance and text relevance were not equivalent and that music similarity most likely required a number of judges to overcome personal biases with regard to expertise and taste.

Issues of feasibility became intertwined with the number-of-graders debate. As stated before, MIREX has an algorithm-to-data model that means the bulk of the evaluation work has to performed and managed by the IMIRSEL team. Under ethics guidelines as advised by the University of Illinois' Institutional Review Board, volunteers cannot be expected to perform more than 3-4 hours of work without compensation (and added administrative rights and protections). This range of 180-240 minutes set an upper bound on the scope of the similarity tasks. Thus, the IMIRSEL team had to juggle a set of factors that would influence the scope (and hence the workload) of the similarity evaluations. These factors included:

1. Number of *A*lgorithms submitted
2. Number of *Q*ueries
3. Number of *C*andidates returned per query
4. Number of *M*inutes spent evaluating each query/candidate pair
5. Number of graders per query/candidate *P*air
6. Number of *G*raders available

These combined to form the basic feasibility equation outlined below.

$$(A * Q * C * M * P) / G \leq 240\text{minutes (per Grader)} \tag{15}$$

As one can see, there are many trade-offs involved. When developing the guidelines for the 2006 similarity tasks under these conditions it was also problematic that one does not know in advance with any certainty such things as how many algorithms will actually be submitted (*A*), how many people will volunteer to be graders (*G*) and/or how long it will actually take to evaluate each query/candidate pair (*M*).

So, given all the debate over the issues of metrics, graders and feasibility, what decisions were actually made? On the metric issue, Table 4 shows the breakdown of the set of new *post-hoc* similarity metrics tabulated for both AMS06 and SMS06. Fine-scaled scoring (FINE in Table 4) and graduate BROAD scores (NS, SS, and VS) were both included. Binary scoring (Greater0 and Greater1) was accomplished by treating Somewhat Similar (SS) scores as either Not Similar (NS) or Very Similar (VS). A trio of different broad score weighting schemes was created (PSum, WCsum and SDsum) to emphasize the relative importance of systems returning Very Similar (VS) results (see Table 4). As Table 5 shows, both SMS06 and AMS06 tasks ended up having 3 graders per query/candidate pair. Most importantly, the combination of factors outlined in the feasibility equation yielded an average number of query/candidate pairs per grader of 205 for SMS06 and 225 for AMS07. Under the relatively realistic assumption of 1 minute per evaluation, this brought the time commitments of the graders within our upper bound of 240 minutes.

Table 4 Basic Metrics used in the 2006 Similarity Tasks

METRIC	COMMENT	RANGE or VALUE
FINE	Sum of fine-grained human similarity decisions	0-10
PSum	Sum of human broad similarity decisions	NS=0, SS=1, VS=2
WCsum	'World Cup' scoring (rewards Very Similar)	NS=0, SS=1, VS=3
SDsum	'Stephen Downie' scoring (strongly rewards Very Similar)	NS=0, SS=1, VS=4
Greater0	Binary relevance judgment	NS=0, SS=1, VS=1
Greater1	Binary relevance judgment using only Very Similar	NS=0, SS=0, VS=1

Table 5 Summary Statistics for the AMS and SMS Tasks 2006-2007

	SMS06	AMS06	SMS07	AMS07
Number of algorithms	7	6	6	12
Number of queries	17	60	30	100
Total number of candidates returned	1360	1800	2400	6000
Total number of unique query/candidate pairs graded	905	1629	799	4832
Number of graders available	20	24	6	20
Number of evaluations per query/candidate pair	3	3	1	1
Number of queries per grader	15	7~8	1	5
Number of candidates returned per query	10	5	10	5
Average number of query/candidate pairs per grader	225	205	133	242
Number of grading events logged	23491	46254	3948	21912

For SMS06, each system was given a query and asked to return the 10 most melodically similar songs from a given collection. The collections were *RISM* (monophonic; 10,000 pieces; SMS1 in Table 3), *Karaoke* (polyphonic; 1,000 pieces; SMS2), *Mixed* (polyphonic; 15,741 pieces; SMS3). There were 6 *RISM* queries, 5 *Karaoke* queries and 6 *Mixed* queries for a total of 17 queries. Then, for each query, the returned results from all participants were grouped and anonymized into collections of query/candidate pairs. These pairs were evaluated by human graders, with each query/candidate pair being evaluated by the 3 different graders. Each grader was asked to provide a categorical BROAD score with 3 categories: NS, SS, VS as explained previously, and one FINE score (in the range from 0 to 10). Along with the basic FINE and BROAD scores, Utrecht's Average Dynamic Recall (ADR) was also calculated to provide some continuity with SMS05.

For AMS06, each system was given 5000 songs chosen from IMIRSEL's *USPOP*, *USCRAP* and *CoverSong* collections. The *USPOP* collection was donated to IMIRSEL by Dan Ellis's Lab Rosa at Columbia University and represents hit pop songs from 2002. The *USCRAP* collection was an eclectic mix of tracks bought by IMIRSEL from a music wholesaler specializing in remaindered CDs (thus the music quality was quite varied). The 330 member *CoverSong* collection is described in [6] and contains a set of 30 titles each of which is represented by 11 variants. The *CoverSong* collection was included in the AMS06 data set to test an ancillary hypothesis that explored whether standard spectral-based similarity techniques were suitable to detect the musically similar but acoustically disimilar cover songs.[4] Each system returned a 5000x5000 distance matrix that recorded the similarities between each song in the collection. After all the matrices were submitted, IMIRSEL randomly selected 60 songs to act as "queries." The first 5 most highly ranked songs out of the 5000 were extracted for each query from each system's matrix (after filtering out the query itself, returned results from the same artist and members of the *CoverSong* collection). Then, for each query, the returned results from all participants were grouped and anonymized into collections of query/candidate pairs. Like SMS, each query/candidate pair was evaluated by 3 different graders using the same set of BROAD and FINE scoring metrics.

It is interesting to note here that the AMS community did not adopt the ADR metric as its primary evaluation metric. The differences in primary metrics between SMS06 and AMS06 reflects the fact that the symbolic and audio research communities are quite independent of each other and hold different worldviews on the notion of what the similarity task is meant to achieve. As the respective task names suggest, the SMS community is interested in notions of "melodic" similarity regardless of such externalities as timbre or orchestration while the AMS community is interested in notions of "musical" similarity which does include timbre and orchestration as contributory factors.

In order to collect grader scores and manage the whole grading process for both SMS06 and AMS06, the IMIRSEL team developed the Evalutron 6000 (E6K) [7, 9, 10]. The E6K (Figure 1) was coded by IMIRSEL team member Anatoliy Gruzd using the "CMS Made Simple" open-source content management system[5] which both reduced the development time and simplified system management. As a web-based application, E6K adheres to a Client-Server model: the client consists of HTML, CSS and JavaScript; and, the server – PHP and MySQL. E6K's web-based approach has the benefit of allowing graders to use the system from anywhere they have a browser and an Internet connection. This was particularly important for MIREX given the international scope of its participants. E6K employs the popular Web 2.0 programming technique referred to as AJAX (Asynchronous JavaScript and XML)[6] to save similarity/relevance judgments and other interaction events in real time, allowing graders to leave the

[4] The answer to this hypothesis was a resounding no [5].

[5] See http://www.cmsmadesimple.org

[6] See http://en.wikipedia.org/wiki/Ajax_(programming).

system and come back as they wish. The ability of graders to come and go as they saw fit was a formal requirement mandated by UIUC's research ethics board. As a side benefit, however, adoption the AJAX approach helped to prevent data loss during unexpected service interruptions or system failures.

The E6K gives graders a choice of three audio players: Flash, Windows Media Player, and Quicktime to ensure cross-platform usability. All players draw from a common set of query/candidate MP3 files. The E6K tracks and records all user-interactions with the system. For MIREX 2006, this consisted of 69,745 logged events across both SMS06 and AMS06.

To better understand the effects that the various evaluation design decisions had on running both the AMS06 and SMS06 tasks, the IMIRSEL team performed a range of analyses on the results data [10]. One such analysis examined the inter-grader reliability (or inter-subjectivity) of the judgments made by trio of graders assigned to each query/candidate pair. The IMIRSEL team chose Fleiss's Kappa to evaluate the inter-grader reliability as it was a metric that it had worked with before. It is a measure of inter-grader reliability for nominal data and is based upon Cohen's two-grader reliability Kappa but is intended for use with an arbitrary number of graders [8]. Fleiss's Kappa is defined as

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \tag{16}$$

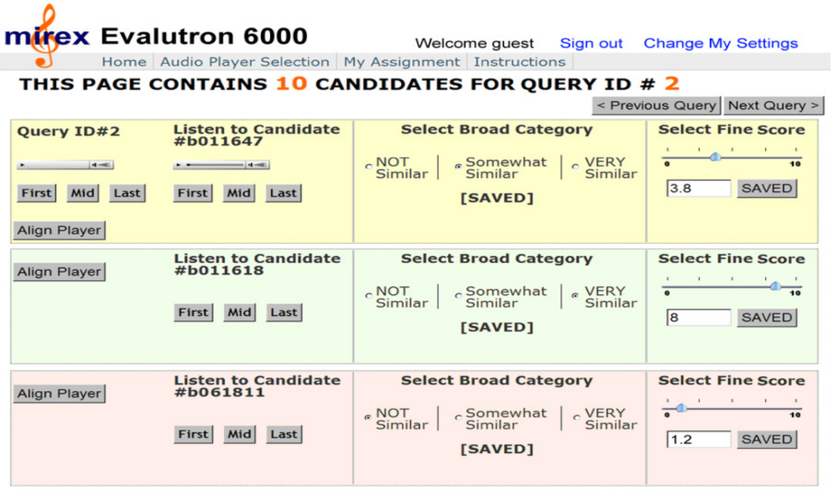
where:

$$\overline{P} = \frac{1}{Nn(n-1)} \sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \tag{17}$$

$$\overline{P}_e = \sum_{j=1}^{k} \left(\frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \right)^2 \tag{18}$$

In the context of the AMS06 and SMS06 tasks, N is the total number of query/candidate pairs that need grading; n is the number judgments per query/candidate pair; k is the number of BROAD response categories (here equal to 3); and, n_{ij} is the number of graders who assigned the i-th query/candidate pair to the j-th BROAD category. Fleiss's Kappa scores range from 0.0 (no agreement) to 1.0 (complete agreement).

Two sets of BROAD category configurations were evaluated. First, we computed the Kappa score using all three BROAD categories (VS, SS, NS). Second, we computed the Kappa score after collapsing the VS and SS categories into a single "similar" (S) category to create a classic binary classification scheme of Similar (S) and Not Similar (NS). Table 6 below presents the Kappa scores for AMS06 and SMS06 under the 3-level and 2-level schemes.



THIS PAGE CONTAINS 10 CANDIDATES FOR QUERY ID # 2

Query ID#2	Listen to Candidate	Select Broad Category	Select Fine Score
Query ID#2	#b011647	○ NOT Similar ● Somewhat Similar ○ VERY Similar [SAVED]	3.8 SAVED
	#b011618	○ NOT Similar ○ Somewhat Similar ● VERY Similar [SAVED]	8 SAVED
	#b061811	● NOT Similar ○ Somewhat Similar ○ VERY Similar [SAVED]	1.2 SAVED

Fig. 1 Screenshot of the Evalutron 6000 (E6K) interface

Table 6 Fleiss's Kappa Scores for SMS06 and AMS06

	3-level (VS, SS, NS)	2-level (S, NS)
SMS06	0.3664	0.3201
AMS06	0.2141	0.2989

According to Landis and Koch [11] who studied the consistency of physician diagnoses of patient illnesses and then derived a scale for interpreting the strength of agreement indicated by Fleiss's Kappa, the scores reported in Table 6 show a "fair" level of agreement. While not ideal, the fair level of agreement is remarkable given the diversity in levels of music skill, tastes and cultural backgrounds of the MIREX 2006 graders. Also noteworthy is the difference in score agreements between the SMS06 and AMS06 tasks with the SMS06 graders being in stronger agreement with each other than the AMS06 graders. For example, when compared under the 3-level scheme, 7.1% of the query/candidate pairs had no agreement among the graders while only 2.2% percent of the SMS06 query/candidate pairs had no agreement. Partial disagreements (where two graders agreed and the third did not) occurred in 51.9% of the SMS06 cases and 62.8% of the AMS06 cases. 30.1% of the query candidate/pairs for AMS06 had perfect agreement among the three graders while the SMS06 grader reached perfect agreement on 45.9% of their query candidate pairs. Under the 2-level scheme, things even out between the two tasks as the AMS06 graders reached perfect agreement on 48.3% of the query/candidate pairs and the SMS06 graders reached perfect agreement on 49.8% of the query/candidate pairs. We believe this overall disparity in agreement levels between the two tasks is attributable to the different notions of similarity held by the two task communities.

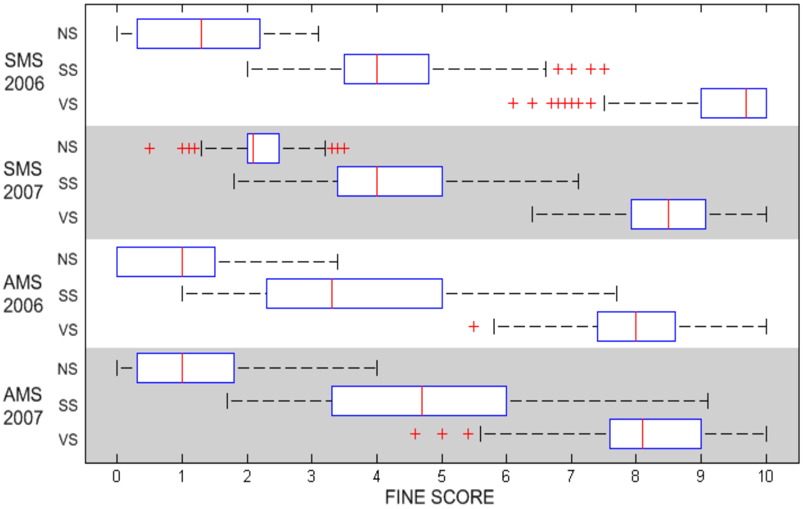


Fig. 2 Distribution of FINE scores within BROAD categories for the SMS06, SMS07, AMS06 and AMS07 tasks. The Box-and-Whisker plots show the median FINE score bounded by the first and third quartiles in the box, with whiskers extending to one-and-half times the inter-quartile range (i.e., the distance between the first and third quartiles) and outliers denoted with +'s.

To measure the consistency between the BROAD category and FINE scores, we calculated the distribution of FINE scores within each BROAD category. Figure 2 shows box-and-whisker plots for both SMS and AMS tasks from 2006 and 2007. The boxes have vertical lines at the 1st, 2nd, and 3rd quartiles. The whiskers bound the minimum and maximum values which fall within 1.5 times the inter-quartile range (IQR), outliers are denoted by + symbols. The box-and-whisker plots illustrate the relative differences between tasks in terms of the assignment of FINE and BROAD scores to musical works. Not only do they reveal the distribution of FINE score responses for a given BROAD score for a given task, they also speak to the variation among graders of what constitutes two pieces being not similar (NS), somewhat similar (SS), or very similar (VS).

In all four of the AMS06, SMS06, AMS07 and SMS07 tasks, the NS and VS categories have the most compact distributions of FINE scores. Similarly, the SS category has the largest IQR in both sets of SMS and AMS tasks. SS scores overlap with the NS values from both task sets. In AMS, however, note how SS greatly overlaps ***both*** the NS and VS values. The data presented in Figures 2 lead us to two important observations. First, and again, there appears to be a fundamental difference in the interpretations of "similarity" between the SMS and AMS task communities. Second, the SS category, regardless of task, appears problematic. The overly broad term "somewhat similar" (SS) is open to many interpretations and meanings, allowing graders to judge two pieces "similar" at any number of ranks. This is not only what would be expected given our natural intuitions about labels like “somewhat similar”, but is also evidenced in the data which clearly illustrate a wide distribution of FINE scores corresponding to the SS category for all tasks.

4.3 2007 Symbolic Melodic Similarity and Audio Music Similarity

The single biggest difference between the running of the 2006 similarity task set and its 2007 running was the adoption of a single grader per query/candidate pair model for both SMS07 and AMS07. This single grader model was adopted for two reasons. First, it greatly reduced the administrative overhead of finding and managing a large number of graders needed. This lessening of load allowed the AMS07 community to significantly increase the number of queries that could be evaluated from 60 in AMS06 to 100 in AMS07. Similarly, the SMS07 query set increased to 30 from its 2006 17 queries. Second, the two communities were convinced that general state of agreement among graders shown in the analyses of the 2006 data indicated that there was less of a need to control for inter-grader variance by having multiple graders than they had originally thought.

Both the AMS07 and SMS07 tasks kept the FINE and the 3-level BROAD categories scoring systems available on the E6K system. Notwithstanding the problematic nature of the SS category, the 3-level system was kept in part to have consistency between years, in part to see if any differences could be noted across years, and in part because it really cost very little to collapse the VS and SS categories into a single S category to create a binary relevance score. The SMS07 community also kept ADR as its primary evaluation metric while AMS07 community continued to ignore it in favour of average FINE score.

SMS07 differed from SMS06 in several significant ways. First, the underlying dataset chosen (SMS4 in Table 3) was changed to 5274 pieces drawn from the Essen Associative Code and Folksong Database.[7] Second, for each query, four classes of error-mutations were created to test the fault-tolerance of the systems. Thus the query set comprised the following 5 query classes:

1. No errors
2. One note deleted
3. One note inserted
4. One interval enlarged
5. One interval compressed

For each query (and its 4 mutations), the returned candidates from all systems were then grouped together (query set) for evaluation by the human graders. The graders were provided, via the E6K system, with the perfect version to represent the query. It was against this perfect version that the graders evaluated the candidates. Graders did not know whether the candidates came from a perfect or mutated query. Each query/candidate pair was evaluated by 1 individual grader. Furthermore, each query was the sole responsibility of only 1 grader to ensure uniformity of results within that query.

[7] More information about the Essen Collection available at http://www.esac-data.org

While the basic premise of AMS07 was kept, there were several subtle changes from AMS07. First, the size of the returned matrices increased to 7000 X 7000 as the dataset also included additional new music drawn from IMIRSEL's *Classical* and *Sundry* collections. Second, 30 second clips were used for AMS07 rather than the full songs used in AMS06 to speed up processing and to ensure that the graders would listen to the exact same music as the systems. This change came about after analyzing the 2006 AMS E6K time-on-task data where it became apparent that many graders were making their judgments without listening to the entire pieces of music provided. Second, 100 songs were randomly selected from the returned matrices with the constraint that the 100 songs equally represent each of the 10 genres found in the database (i.e., 10 queries per genre). The 5 most highly ranked songs were then returned per query (after filtering for the query song and songs from the same artist). As with SMS07, a query became the sole responsibility of 1 grader to ensure uniformity of scoring within each query.

The actual running of the 2007 similarity tasks was much less stressful on the IMIRSEL team primarily because the need to manage multiple graders per query/candidate pair was eliminated. The SMS07 ADR highest score of 0.72 along with the highest FINE score of 0.56 indicates that the best performing SMS systems are quite tolerant of query input errors. These scores compare favourably with the SMS1 scores (found in Table 3) from 2005 (0.66 ADR) and 2006 (0.72 ADR). The AMS07 highest average FINE score increased to 0.57 from 0.43 for AMS06. We believe this increase in score is jointly attributable to improvements in the algorithms (primarily) and the shortening of the query/candidate pair lengths to 30 seconds to ensure the synchronization of "listening" between the systems and the graders (secondarily).

There was no running of either an SMS or AMS task during MIREX 2008. The first reason for this was a general consensus that developers needed more time to make non-trivial improvements to their systems. The second and perhaps more daunting reason is the data shortage issue. For SMS in particular, there is an acute shortage of available trustworthy symbolic music from which to build meaningful, large-scale test collections. Even with AMS, where more data is available, it still takes considerable effort and expense to acquire and then prepare the audio files (i.e., cutting to length, normalizing data formats, etc.) for use in a proper test collection. Of course, it is possible to re-use the datasets from previous years. However, constant reuse of data will most likely lead to the "overfitting" of algorithms to the data leading to meaningless apparent improvements in results.

If the two communities can decide on how to deal with the data issue, the IMIRSEL team looks forward to running future iterations of both the AMS and SMS similarity tasks. Based on the smoothness with which both SMS07 and AMS07 ran, we recommend keeping the basic 2007 format for each task with one caveat. As Figure 2 shows, the SS BROAD category continues to be problematic with its wide variance in both the AMS and SMS tasks. We need to encourage these communities to think hard about either how to reduce this variability or to consider eliminating the SS BROAD category altogether, particularly with regard to the AMS task.

5 Summary and Future Directions

In this chapter we have examined the history and infrastructure of MIREX. MIREX is a community-led endeavour that reflects the wide range of research streams being undertaken by MIR researchers from all around the world. The growth of MIREX over the period from 2005 to 2008 has been remarkable with the number of algorithms evaluated increasing steadily each year. We have discussed the numerous primary evaluation metrics used to "officially" report the performance results of each tasks. We have highlighted that the choice and/or creation of particular primary metric is influenced by a mixture of scientific imperatives, previous empirical research, traditions of practice, participant desires to succeed and the pragmatics of actually getting the evaluations completed within constraints of acceptable time and effort expenditures. By looking at the two similarity tasks, AMS and SMS, we have illustrated how MIREX tasks have evolved through time as the MIREX research community builds upon its successes and addresses the problems it encounters.

We have also suggested that one major problem facing the MIREX community is the lack of useable data upon which to build realistic test collections. In an effort to solve this problem, along with the potential problem of having IMIRSEL overcome by the increase in the number of algorithm submissions, a new international research collaboration called the Networked Environment for Music Analysis (NEMA) has been formed. NEMA comprises research labs from the Universities of Waikato, Illinois at Urbana-Champaign, Southampton, London (both Goldsmiths and Queen Mary Colleges), and McGill. One important goal of NEMA is the construction of a web-service framework that would make MIREX evaluation tasks, test collections and automated evaluation scripts available to the

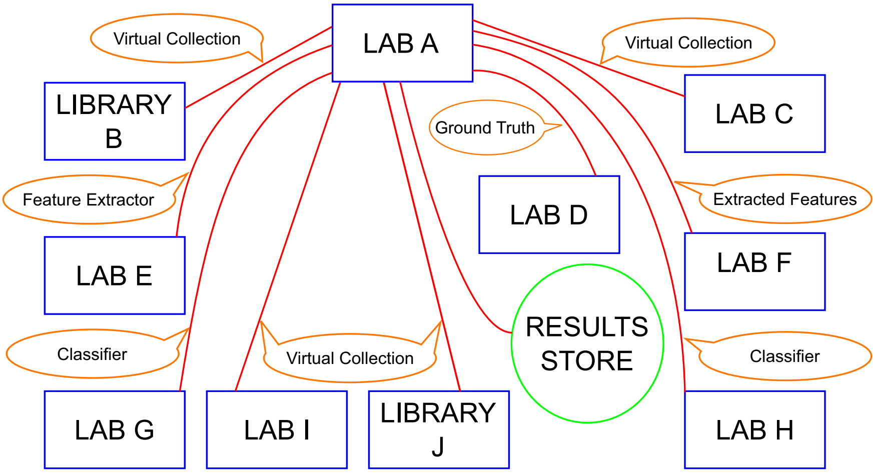


Fig. 3 An illustration of how NEMA will be used to gather remote resources for evaluation

community on a year-round basis. It would also expand the availability of data by developing the On-demand Metadata Extraction Network (OMEN) [12]. OMEN is designed to acquire metadata and features from remote music repositories while at the same time respecting copyright laws. NEMA hopes to incorporate OMEN within a system that would allow researchers to locate and use bits and pieces of algorithms from other researchers. Thus, as Figure 3 illustrates, a team at Lab A could build a hybrid experimental system quickly from the classifiers and feature extractors from other participating labs. It could run a MIREX-based evaluation from the comfort of its lab any time it chose using its new system. It could then report its findings and, if successful, redeposit its new hybrid algorithm into the NEMA repository for others to build upon.

Acknowledgments. The IMIRSEL team wishes to thank the Andrew W. Mellon Foundation for its support. This work has also been support by the National Science Foundation Grant IIS- 0328471.

References

1. Cano, P., Gomez, E., Gouyon, F., Herrera, P., Koppenberger, M., Ong, B., Serra, X., Streich, S., Wack, N.: ISMIR 2004 audio description contest. MTG Technical Report, MTG-TR-2006-02 (Music Technology Group, Barcelona, Spain) (2004)
2. Downie, J.S. (ed.): The exploratory workshop on music information retrieval. Graduate School of Library and Information Science, Champaign (1999)
3. Downie, J.S. (ed.): The MIR/MDL evaluation project white paper collection, edition #3: Establishing music information retrieval (MIR) and music digital library (MDL) evaluation frameworks: Preliminary foundations and infrastructures. Graduate School of Library and Information Science, Champaign (2003)
4. Downie, J.S.: The scientific evaluation of music information retrieval systems: Foundations and future. Computer Music Journal 28(3), 12–23 (2004)
5. Downie, J.S.: The Music Information Retrieval Evaluation Exchange (2005-2007): A window into music information retrieval research. Acoustical Science and Technology 29(4), 247–255 (2008)
6. Downie, J.S., Bay, M., Ehmann, A.F., Jones, M.C.: Audio cover song identification: MIREX 2006-2007 results and analyses. In: Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008), pp. 468–473 (2008)
7. Downie, J.S., Lee, J.H., Gruzd, A.A., Jones, M.C.: Toward an understanding of similarity judgments for music digital library evaluation. In: Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2007), pp. 307–308 (2007)
8. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin 76(5), 378–382 (1971)
9. Gruzd, A.A., Downie, J.S., Jones, M.C., Lee, J.H.: Evalutron 6000: Collecting music relevance judgments. In: Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries (JCDL 2007), p. 507 (2007)
10. Jones, M.C., Downie, J.S., Ehmann, A.F.: Understanding human judgments of similarity in music information retrieval. In: Proceedings of the Eighth International Conference on Music Information Retrieval (ISMIR 2007), pp. 539–542 (2007)

11. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics 33, 159–174 (1977)
12. McEnnis, D., McKay, C., Fujinaga, I.: Overview of OMEN. In: Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR 2006), pp. 7–12 (2006)
13. McKinney, M.F., Moelants, D.: Tempo perception and musical content: What makes a piece fast, slow or temporally ambiguous? In: Proceedings of the International Conference on Music Perception and Cognition, Evanston, IL, USA (2004)
14. Pampalk, E.: Audio-based music similarity and retrieval: Combining a spectral similarity model with information extracted from fluctuation patters. In: MIREX 2006, Submission Abstracts (2006), `http://www.music-ir.org/evaluation/MIREX/2006_abstracts/AS_pampalk.pdf`
15. Répertoire International des Sources Musicales (RISM). Serie A/II, manuscrits musicaux après 1600. München. K. G. Saur Verlag, Germany (2002)
16. Typke, R., den Hoed, M., de Nooijer, J., Wiering, F., Veltkamp, R.C.: A ground truth for half a million musical incipits. Journal of Digital Information Management 3(1), 34–39 (2005)
17. Typke, R., Veltkamp, R., Wiering, F.: A measure for evaluating retrieval techniques based on partially ordered ground truth lists. In: IEEE International Conference on Multimedia and Expo (ICME 2006), pp. 1793–1796 (2006)