The MIREX Grand Challenge: A Framework of Holistic User Experience Evaluation in Music Information Retrieval

Xiao Hu[1], Jin Ha Lee[2], David Bainbridge[3], Kahyun Choi[4], Peter Organisciak[4], J. Stephen Downie[4]

1. University of Hong Kong; 2. University of Washington; 3. University of Waikato; 4. University of Illinois

Abstract:
Music Information Retrieval (MIR) evaluation has traditionally focused upon system-centered approaches where components of MIR systems are evaluated against predefined datasets and golden answers (i.e., ground truth). There are two major limitations of such system-centered evaluation approaches: 1) the evaluation focuses on subtasks in music information retrieval but not entire systems; and 2) users and their interactions with MIR systems are largely excluded. This paper describes the first implementation of a holistic user experience evaluation in MIR, the MIREX Grand Challenge, where complete MIR systems are evaluated with user experience being the single overarching goal. It is the first time complete MIR systems have been evaluated with end-users in a realistic scenario. We present the design of the evaluation task, the evaluation criteria and a novel evaluation interface and data collection platform. This is followed by an analysis of the results, reflection of the experience and lessons learned, and plans for future directions.

Introduction

Evaluation has been critical to the field of Music Information Retrieval (MIR) since its beginning (Downie, 2003). The Music Information Retrieval Evaluation eXchange (MIREX) is the community-wide evaluation event held annually since 2005 (Downie et al., 2014). During the ten years of MIREX, a remarkable number of MIR tasks have been evaluated, covering nearly all aspects of MIR systems, from automated detection of low level music features such as keys, tempos, beats and onsets, to music classifications based on higher level semantics such as genre, mood and music similarity. As these tasks do not consider user interactions, they are often referred as *system-centered* tasks. These tasks were collectively defined by the MIR research community and are indispensable components of MIR systems. Over the last ten years, MIREX has evaluated 2,397 algorithm runs submitted by MIR researchers from over 30 countries, contributing tremendously to the development of the MIR field. However, as the field matures, the current state of the art in conventional system-centered tasks is sufficient to support an acceptable degree of efficiency and effectiveness in these tasks. There is thus a demand for complete MIR systems which can integrate multiple sub-tasks and support the actual activity that all MIR systems aim at: music discovery. These full-featured systems are important for further refinements and improvements of the field (Downie, Crawford & Byrd, 2009).

A complete MIR system should not only contain a back-end that can process, index and search for music pieces or segmentations, but also a user-facing front-end that can support user interactions. To date all MIR evaluation tasks have been focusing on one or more components in the back-end, with the front-end being largely ignored. Such evaluation paradigm employs methods and metrics that ignore user interaction and are therefore inadequate for evaluating complete MIR systems (Downie et al., 2014). The time is ripe for building fully developed MIR systems and consequently, a shift of evaluation paradigm to holistic evaluation of complete systems is needed.

In the history of Information Retrieval (IR), system-centered evaluation is indeed necessary for an IR domain to develop during its early stages and to form its own theoretical frameworks. This was seen in the traditional text-centric IR community, with the Text Retrieval Evaluation Conference (TREC) focusing on system-centered evaluation in its initial years (Voorhees & Harman, 2005). With the development of the field, the demand for complete, user-oriented IR systems calls for the shift of evaluation paradigm, methods and infrastructures to *user-centered* ones. TREC, for instance, started its first Interactive IR track in 2003, which explicitly involved users in the evaluation (Voorhees & Harman, 2005). In MIR, with the maturity of various techniques (e.g., genre classification, similarity search) in recent years, there have been articles underscoring the need for user involvement in the evaluation processes (e.g., Hu & Liu, 2010; Lee & Cunningham, 2013; Hu & Kando, 2012; Schedl, 2013). Here, we even go one step further: to not only put users in the center of evaluation but also take *user experience* as the single construct of evaluation. This is because user experience is the ultimate goal of systems that serve end-users, and all components of the IR process, from storage of materials, to indexing, and to the user interface, contribute to user experience. To date there has been no evaluation framework in MIR focusing on complete systems and user experience. To fill this gap, we designed, implemented and tested a holistic user experience evaluation platform, the MIREX Grand Challenge on User Experience 2014 (abbreviated as "GC14UX"), as the first step towards a more comprehensive evaluation framework for MIR. The word Grand Challenge is coined from Downie et al. (2009) where "grand" embodied the goal of developing, deploying and then evaluating *complete* MIR systems. This is indeed a grand step as can be seen from the text IR domain where full-featured systems such as SMART (Salton, 1971) and Managing Gigabytes (Witten, Moffat & Bell, 1999) had significantly pushed forward research in the field.

In summary, the term "holistic evaluation" in this study refers to a twofold view: 1) it evaluates complete MIR systems, as opposed to individual components such as audio indexing and music similarity calculation; 2) it emphasizes the evaluation of user experience as a whole instead of only evaluating the search results. The contributions reported here can be summarized as follows:

1. The article contributes to a theoretical framework in MIR evaluation. The holistic user experience evaluation approach advocated and implemented in this study is a radical revolution from existing evaluation frameworks including MIREX, MediaEval and CLEF (see below). These existing frameworks focus on subtasks related to components of MIR systems rather than complete ones. In addition, they largely follow the Cranfield paradigm (Voorhees, 2002) where retrieval systems are compared against common test collections with topics/queries, materials/collections and relevance/similarity judgments. Putting users and their interactions with complete MIR systems in the center of consideration changes the entire perspective. The evaluation infrastructure does no longer consist of pre-defined queries or ground truth; the evaluation context is shifted from simplified laboratory settings to real-world scenarios; and the evaluation criteria focus on users' interactions and perceptions. The proposed evaluation framework will hopefully inspire more studies on holistic user experience evaluation in MIR.
2. As this is, to our best knowledge, the first community-wide holistic user experience evaluation, there are numerous methodological decisions to make in designing and implementing the evaluation framework (a.k.a., testbed). In particular, the creation of a sharable music dataset of a large size shifts the existing "algorithm-to-data" paradigm that has long plagued MIR research due to copyright imposed on most music materials.

Furthermore, the evaluation methods, results and reflections on lessons learned are all highly valuable for further studies on user-centered and holistic evaluation in MIR and related domains such as multimedia IR.

3. The findings of this study, particularly the evaluation results will have important implications for designing MIR systems and applications that cater to the needs of real-world users. There have been MIR applications including commercial ones, but they are mostly not evaluated in a scientific and scalable manner. How they support or frustrate users largely remains an unanswered question (Lee & Price, 2015). This study strives to inspire MIR application developers to join the holistic user experience evaluation events to be held in the future.

The rest of the article will present related work that inspired this study, followed by a detailed description of the evaluation framework. The result of a trial evaluation will be presented and reflections will be discussed with regard to the lessons learned and future plans.

Related Work

MIR Evaluation: MIREX, MusiClef. MediaEval

Evaluation has been emphasized in MIR since the earlier days of the field (Downie, 2003). A direct precursor to MIREX, The Audio Description Contest (ADC) was the first community-wide evaluation event in MIR, held by the local committee of the 5th annual conference of the International Society of Music Information Retrieval (ISMIR) (Cano et al., 2006). In the ADC, a number of algorithms based upon music audio engineering/signal processing techniques were evaluated using a set of standardized tasks and datasets. Based on the insights gained in the event and previous discussions in the MIR community, MIREX was officially launched in 2005 and has been held annually since then. To date, MIREX has evaluated over 30 tasks, ranging from low-level music feature extraction (e.g., Audio Downbeat Estimation, Audio Chord Detection, Multiple Fundamental Frequency Estimation and Tracking) to application oriented techniques (e.g., Audio Genre Classification, Audio Mood Classification, Audio Music Similarity, Query-by-Singing/Humming). Notwithstanding the highly significant contribution of MIREX to the field, all tasks in MIREX have been following the system-centric, Cranfield evaluation paradigm where IR algorithms are run on a pre-built collection of music materials and evaluated against a ground truth which is usually hand annotated by a small panel of human judges. There is no user interaction in any MIREX tasks. The only user involvement is for relevance judgments in some tasks (e.g., similarity tasks) for constructing ground truth. In this case, the users are actually acting as human judges instead of real users. In fact they are not working with the MIR systems being evaluated at all, but rather another system dedicated for collecting human opinions for building ground truth.

MusiClef is a benchmarking activity that has been developed since 2011 from the Cross-Language Evaluation Forum (*CLEF*), focusing on evaluating music access and retrieval techniques that utilize both music content and contextual information (e.g., tags, comments, or reviews) (Orio, Rizo, Miotto, Schedl, Montecchio & Lartillot, 2011). One of the emphases of MusiClef is to develop evaluation tasks that are connected to real-life usage scenarios, such as categorizing soundtracks that could be used in TV shows and identifying different recordings of the same (historical) classical music hosted in music libraries and archives. Starting from 2013, MusiClef was merged into the MediaEval, a benchmarking initiative for various multimedia access and retrieval tasks. In 2013 and 2014, MediaEval included music emotion recognition tasks (Aljanaki, Yang & Soleymani, 2014) which supplement the music mood related tasks in MIREX. Although the tasks in MusiClef and MediaEval provide more evaluation scenarios and evaluation datasets, the evaluation approach still follows Cranfield paradigm. As with MIREX, no user interface was evaluated and no user interactions were involved.

User-centered Evaluation in MIR

The system-centered approach of MIR evaluation simplifies the entire MIR process by excluding the users, their information needs and behaviors to a large extent. In recent years, researchers started criticizing this approach and arguing that the goal of MIR systems is to help users meet their music information needs, and thus MIR evaluation must take users into account (e.g., Hu & Liu, 2010; Hu & Kando, 2012; Lee & Cunningham, 2013; Schedl & Flexer, 2013). Furthermore, in multiple studies, it has been found that the results of system-centered evaluation may not align with users' perceptions (Hu & Kando, 2012; Lee & Cunningham, 2013).

User-centered evaluation approaches measure user behaviors and perceptions during the IR process (Kelly, 2009). Over the years, efforts have been made to bridge the gap between system-centered and user-centered evaluation, yet there have been few studies on formal user evaluation of MIR systems and most of the studies targeted the general music listeners. Pauws and colleagues were pioneers in this regard. They conducted a series of controlled user experiments to compare novel playlists generation systems to baseline systems using user-centered measures such as users' ratings on playlist quality, time spent on the task, as well as user-reported usefulness, ease-of-use and preference.(Pauws & Eggen, 2002; Pauws & Wijdeven, 2005; Vignoli & Pauws, 2005). More recently, HCI-based evaluation has been adopted in several studies evaluating MIR interfaces (Hoashi, Hamawaki, Ishizaki, Takishima & Katto, 2009; Hu & Kando, 2012). Besides user effectiveness and satisfaction, Hoashi et al. (2009) also employed a set of user experience measures including perceived system accuracy, explicitness and enjoyability.

The lack of user-centered evaluation in MIR is coupled with the absence of complete MIR systems that could be released to the public for music searching and discovery (Downie et. al, 2009). To inspire the development of complete MIR systems, the evaluation needs to adopt a holistic view of systems that does ***not*** separate system components into retrieval algorithm, back-end database, front-end user interface, etc. A recent study by Lee & Price (2015) conducted a user study to evaluate popular commercial MIR services using Nielsen's usability heuristics (Nielsen, 1994). This exemplifies user-centered MIR evaluation which focuses on user experience rather than criteria only focusing on system performance. In this study, we advocate the notion of user experience as a first-class research objective in the MIR community, by designing and experimenting a holistic user experience evaluation framework that can be used to evaluate novel MIR systems being developed by MIR researchers around the globe.

Holistic Evaluation and User Experience in Digital Libraries

The literature on the evaluation of digital libraries (DL) has presented convincing efforts on holistic evaluation. Summarizing previously proposed DL evaluation models, Fuhr et al., (2007) presented a comprehensive evaluation framework and a set of recommendations including involvement of practitioners and real users, building on past experience of large evaluation initiatives, community building in evaluation research, and evaluation of user behavior in context. Through a comprehensive three-stage study, Zhang (2010) proposed a holistic DL evaluation framework which includes important criteria from heterogeneous stakeholder groups including administrator, developer, librarian, researcher and end user. For each of the stakeholder groups, evaluation criteria are organized into six levels: content, technology, interface, user, service, and context (Zhang, 2010).

These DL evaluation frameworks can be applied to complete MIR systems as they are domain-agnostic, and complete MIR systems can essentially be regarded as music digital libraries (MDL). Indeed, the GC14UX follows Furh et al.'s recommendations of being user-centered, and builds upon the decade's successful experiences of MIREX, involving the global MIR research community and striving to evaluate user experience in realistic context. As the first step towards holistic evaluation in MIR, this study focuses on the researcher and end-user stakeholder groups in Zhang's model (2010), particularly MIR researchers and the general public audience while leaving other stakeholder and user groups for future studies.

The notion of user experience has been examined and it has been recognized that when users interact with IR and DL systems, they are not only seeking for information, but are also enjoying the process (Toms, Dufour & Hesemeier, 2004). In other words, even if a system is usable and functions well, it may still lose users if it is not engaging them (O'Brien & Toms, 2008). In a review of 51 papers on user experience, Bargas-Avila and Hornbæk (2011) summarized that there were two definitions of user experience. The first refers to it as a synonym for usability (e.g., efficiency, effectiveness, satisfaction), whereas the other also uses the phrase to denote aspects related to the hedonic nature of users, such as emotion, enjoyment, engagement and aesthetics. With the second definition with a broader scope, the boundary of user experience is extended from perfunctory to pleasurable and memorable (O'Brien & Toms, 2008). In the MIR domain, it is also found that user satisfaction in music information seeking could depend on both hedonic (i.e., pleasure) and utilitarian (i.e., acquisition of information) outcomes (Laplante & Downie, 2011; Hu & Kando, 2014). Notwithstanding the importance of hedonic aspects of user experience, we set out to first focus on the usability-related aspects in this study as they are the most basic in user experience and have yet to be systematically evaluated in MIR.

The GC14UX Evaluation Framework

A holistic user experience evaluation requires careful designs on a practical music dataset, use cases, evaluation criteria, and methods to measure the criteria (Kelly, 2009). This section presents all important components in the GC14UX evaluation framework in detail.

Music Dataset

For MIR, the foremost common goal for most users is to look for music, and thus in MIR evaluation, a collection of music must be available. To be fair to all systems being evaluated, in the context of the Grand Challenge, it is necessary to standardize the music collection upon which the evaluation will be conducted. The norm for MIREX evaluation has been an "algorithm-to-data" model where datasets are centrally hosted by the MIREX team, and researchers submit MIR algorithms to the MIREX team who then runs the algorithms against the datasets. This is due to the intellectual property laws that govern most music materials. Under this constraint, it is extremely difficult for the MIR community to build realistic test collections that consist of large amount of music in great variety which can be frequently updated (Downie et al., 2014). Moreover, the fact that datasets cannot be distributed among researchers has hindered the development of new algorithms and systems as they cannot be easily benchmarked with other existing systems.

To overcome the limitations of using datasets of copyrighted music, we strive to construct a copyright-free dataset for GC14UX. In addition, for a non-trivial and authentic MIR evaluation task, the music collection should be of a substantial size. To fulfill these requirements, a subset of tracks in the Jamendo[1] collection that have the most open set of distribution terms was selected. Jamendo is one of the world largest digital services for free music. As of May 20, 2014, the Jamendo collection contained 14,742 tracks with the CC-BY license. CC-BY is a permission license, allowing others to distribute, modify, optimize and use the licensed work, even commercially, as long as credit is given for the original creation.

We sampled 10,000 tracks from the Jamendo CC-BY collection by ensuring each selected track contains metadata supplied by Jamendo and with at least two genre tags. The dataset contains the complete MP3 tracks and corresponding metadata represented in JSON format, retrieved using the site's API. The dataset was made available for participants (system developers) to download for building their systems. The size of 10K tracks is moderate for an authentic MIR system, but it balances the interestingness of the evaluation task and the management load of the participants. Even for this moderate size, the zipped version of the dataset was over 60 Gigabyte and we had to set up mirror sites in the U.S. and Hong Kong to facilitate downloading by potential participants.

The dataset covers a wide range of music, along with a set of bibliographic (e.g., title, album, artist) metadata and tags. Table 1 shows the most popular tags in three categories—genre, instrument and free tags—as well as the number of songs associated with each of the tags. All songs have at least two genre tags, while most of them (8,662 songs) also have one or more free tags. More than half of the tracks (5,820) also have instrument tags. All songs have duration metadata with about 80% of them under 5 minutes long.

Genre tags (134 unique tags)		**Free tags (2,450 unique tags)**		**Instrument tags (98 unique tags)**	
Tag	#. of songs	Tag	#. of songs	Tag	#. of songs
Electronic	4,445	Vocal	2,265	voice	2,103
Rock	2,633	Adventure	996	drum	1,481
instrumental	1,891	soundtrack	826	synthesizer	1,452
Ambient	1,559	energy/etic	783	bass	1,212
Acoustic	1,284	Chillout	613	electricguitar	928
Pop	1,278	entertainment	343	computer	921

Table 1. Most popular tags in the sampled Jamendo collection

Having a sizable copyright-free dataset is a major shift in MIR evaluation. It largely alleviates the problems caused by non-sharable and small datasets. Open datasets such as this Jamendo one allow researchers to replicate experiments and compare novel systems to existing ones. An additional advantage of using copyright-free music in user-centered evaluation is that it can be freely played to anyone, and thus a large and diversified population of listeners can be involved in the evaluation. This is particularly desirable for obtaining a significant sample of user-centered evaluation measures.

The Task and Scenario

The primary purpose of GC14UX is to forefront the goals of users when they interact with MIR systems. In other words, GC14UX emphasizes that MIR systems should support users fulfilling their goals. Every system designer must consider some kind of user scenarios and goals (i.e., when users come to my system, what will they do?) in developing a system, and the most successful commercial MIR systems (e.g., Pandora, YouTube, Spotify) all support particular user tasks extremely well (e.g., online radio, sharing music videos). How well the system supports these user tasks must be taken into consideration as we evaluate the system. For instance, it would be more reasonable to evaluate Pandora on how well it does in generating music recommendations rather than how well it supports known-item searches. On the other hand, for evaluating a system such as SoundCloud, a music storage and sharing service, it would make sense to consider how easy it is to upload and share users' music.

In GC14UX, in order to help users/evaluators situate their use of a system, and to take into account the free (and less popular) nature of the music collection, we defined the user task to be evaluated as "You are creating a short video about a memorable occasion that happened to you recently, and you need to find some (copyright-free) songs to use as background music." The task is designed such that it is both flexible (suitable for all kinds of music) and authentic (realistic for users) based on the user tasks commonly conducted in music services as identified by Lee and Waterman (2012).

The Evaluators

In user-centered evaluation, evaluators are ideally sampled from targeted users of the evaluated systems. As the task in GC14UX is designed for the general public, any adult aged 18 and above who is interested in searching for music is an eligible evaluator. As the GC14UX is the first trial of the holistic user experience evaluation framework, evaluators were initially solicited via listservs associated with the MIR community. To recruit a large sample of evaluators, however, it was encouraged in the listserv messages to spread the invitation to any friends/colleagues/students who might be interested in participating. Consequently, it is reasonable to assume many evaluators were drawn primarily from the researcher and students in the MIR and the Information Science fields. Human research ethics approval has been obtained before the evaluators were recruited. It was ensured that all participating systems get roughly equal number of evaluators.

The Criteria

As the focus of the evaluation is user experience, widely-used and well-accepted usability heuristics proposed by Nielsen were considered (Nielsen & Molich, 1990; Nielsen, 1994; Nielsen, 2005) and presented below:

1) Visibility of system status.
2) Match between system and the real world.
3) User control and freedom.
4) Consistency and standards.
5) Error prevention.
6) Recognition rather than recall.
7) Flexibility and efficiency of use.
8) Aesthetic and minimalist design.
9) Help users recognize, diagnose, and recover from errors.
10) Help and documentation.

Lee and Price (2015) also summarizes a list of different criteria that were used in previous literature on evaluating recommender systems (Pu et al., 2011; Zhang et al., 2012; Herlocker et al., 2004; Knijnenburg et al., 2012). As our task in GC14UX was about music discovery, we also considered the following criteria:

- Recommendation Accuracy: items recommended match user interests (Pu et al., 2011).
- Explanation: why items are recommended to a user (Knijnenburg et al., 2012; Pu et al., 2011).
- Interaction Adequacy: the system allows a user to state what they like/dislike (Pu et al., 2011).
- Perceived Ease of Use/Familiarity: a user's acquaintance with or knowledge of the system (Knijnenburg et al., 2012; Pu et al., 2011).
- User control/Control over the system: the users felt in control in their interaction with the system (Knijnenburg et al., 2012; Pu et al., 2011).
- Novelty/Serendipity: providing "non-obvious" recommendations (Herlocker et al., 2004; Zhang et al., 2012).
- Privacy: concerns about the amount and depth of information the user provides (Knijnenburg et al., 2012).
- Confidence/Trust: a user's belief that the system "works" (Knijnenburg et al., 2012; Pu et al., 2011).
- Overall Satisfaction: how well the system fulfills the overall needs of a user (Pu et al., 2011).

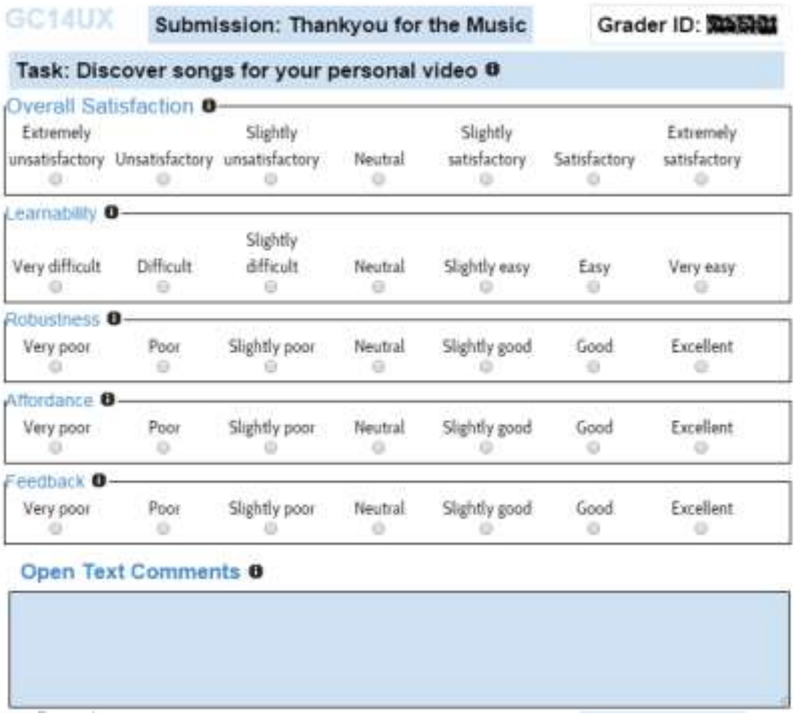
While it is tempting to solicit users' perceptions on all aspects of user experience, we also had to keep in mind that asking too many questions can potentially distract users from their music search processes. To minimize evaluators' cognitive load and ensure the evaluation process to be as close to an authentic situation as possible, we originally designed only one question on overall satisfaction (Pu et al, 2011), "How would you rate your overall satisfaction with the system?" However, with only this question, we would not know which aspect(s) the users are satisfied (or unsatisfied) with. In addition, as users' perceptions in MIR can be influenced by their music preferences (i.e., "feel" better when listening to music they like) (Hu & Kando, 2014), it is necessary to ensure users focus on their experience with the systems rather than their music preferences. Therefore we added a small number of questions regarding specific usability aspects: "learnability," "robustness," "affordance" and "feedback", each representing some aspects or combination of aspects from existing evaluation criteria in the literature. The explanations of the criteria and *approximate* mappings between them and the criteria from previous literature are presented in Table 2. In order to balance being comprehensive and minimizing the complexity of the evaluation assignments, some existing criteria are left out, considering they are less relevant to this study or overlapped with selected criteria. For example, the heuristic of "Flexibility and efficiency of use" is mainly concerned about providing different interfaces for expert and novice users, whereas in GC14UX the evaluated systems are all novel and experimental ones to which virtually all users are new to, and therefore novice. Due to the experimental nature of GC14UX, privacy is less important an issue as users will only interact with the systems for a limited time. In addition, "confidence/trust" is not selected in this study as it is about users' general attitude towards the system and to some extent represented by "overall satisfaction." (Knijnenburg et al., 2012; Pu et al., 2011).

Criteria	Explanation	Criteria in the literature
Learnability	How easy was it to figure out how to use the system?	Consistency and standards (Nielsen, 2005) Match between system and the real world (Nielsen, 2005) Recognition rather than recall (Nielsen, 2005) Help and documentation (Nielsen, 2005)

		Perceived ease of use (Knijnenburg et al., 2012; Pu et al., 2011)
Robustness	How good is the system's ability to warn you when you're about to make a mistake, allow you to recover, or retrace your step?	Error prevention (Nielsen,2005) Help users recognize, diagnose, and recover from errors (Nielsen,2005)
Affordance	How well does the system allow you to perform what you want to do?	User control and freedom (Nielsen,2005) User control/control over the system (Knijnenburg et al., 2012; Pu et al., 2011)
Feedback	How well does the system communicate what’s going on?	Visibility of system status (Nielsen, 2005) Explanation (Knijnenburg et al., 2012; Pu et al., 2011)

Table 2: Evaluation criteria and approximate mappings to those in the literature

Each of the questions is manifested in a closed format with a 7-point Likert scale ranging from “very unsatisfactory,””very difficult,” or “very poor” to “very satisfactory,””very easy,” or “excellent.” This was chosen over a higher granularity scale as those have been shown to increase cognitive load of the evaluators and possibly distract them from focusing on the MIR systems (van Schaik & Ling, 2003). In addition, 7-scale points have been shown to achieve optimal inter-rater reliability and more scale points did not increase reliability substantially (Cicchetti et al., 1985). The placement of anchor labels on the scales was also debated between simplicity and reliability. The evaluation interface would be much cleaner if anchor labels were only provided at either end of the scale; however, one of the problems with this form of scale is that humans’ interpretation on the relationship between points vary (e.g., Katter 1968, Fleiss 1971). After trying out different designs, we finally chose to present labels for all scale points, which is likely to yield higher test-retest reliability (Weng, 2004). The final evaluation form is shown in Figure 1.



GC14UX Submission: Thankyou for the Music Grader ID: [redacted]

Task: Discover songs for your personal video

Overall Satisfaction

Extremely unsatisfactory	Unsatisfactory	Slightly unsatisfactory	Neutral	Slightly satisfactory	Satisfactory	Extremely satisfactory
○	○	○	○	○	○	○

Learnability

Very difficult	Difficult	Slightly difficult	Neutral	Slightly easy	Easy	Very easy
○	○	○	○	○	○	○

Robustness

Very poor	Poor	Slightly poor	Neutral	Slightly good	Good	Excellent
○	○	○	○	○	○	○

Affordance

Very poor	Poor	Slightly poor	Neutral	Slightly good	Good	Excellent
○	○	○	○	○	○	○

Feedback

Very poor	Poor	Slightly poor	Neutral	Slightly good	Good	Excellent
○	○	○	○	○	○	○

Open Text Comments

Figure 1. GC14UX evaluation form

Besides the aforementioned closed questions, an open question is also provided to ask for evaluators' free-text comments. The instructions also encourage evaluators to input text comments in addition to ratings. The free-text input can facilitate qualitative user experience evaluations which can help understand which factors affect user experience and how to maximize the positive aspects of interface and interaction design (Schedl & Flexer, 2012).

The Evaluation Platform

To facilitate evaluators' interaction with the MIR systems in their authentic contexts, an online evaluation platform was built to allow remote access from anywhere a Web browser and an Internet connection were available. The platform was built upon the *Evalutron 6000* (*E6K)* system developed for remote access by community evaluators of the MIREX *Audio Music Similarity (AMS)* and *Symbolic Melodic Similarity (SMS)* tasks (Gruzd, Downie, Jones & Lee, 2007). The original E6K was designed to collect relevance judgments for query-candidate pairs generated by the submitted algorithms, so that ground truth datasets could be developed from the collected judgments. To support online access, E6K adopted a Client-Server architecture: of HTML, CSS and JavaScript on the client side; and PHP and MySQL on the server.

The evaluation platform for GC14UX adopted the main structure of the E6K including user account management and page organization, but with significant improvements. Specifically, the GC14UX platform presents the participating MIR systems by embedding each of them into an *iframe* (inline Frame), an HTML element designed to embed another document within the current HTML document. Figure 2 illustrates the overall structure of the GC14UX evaluation platform.

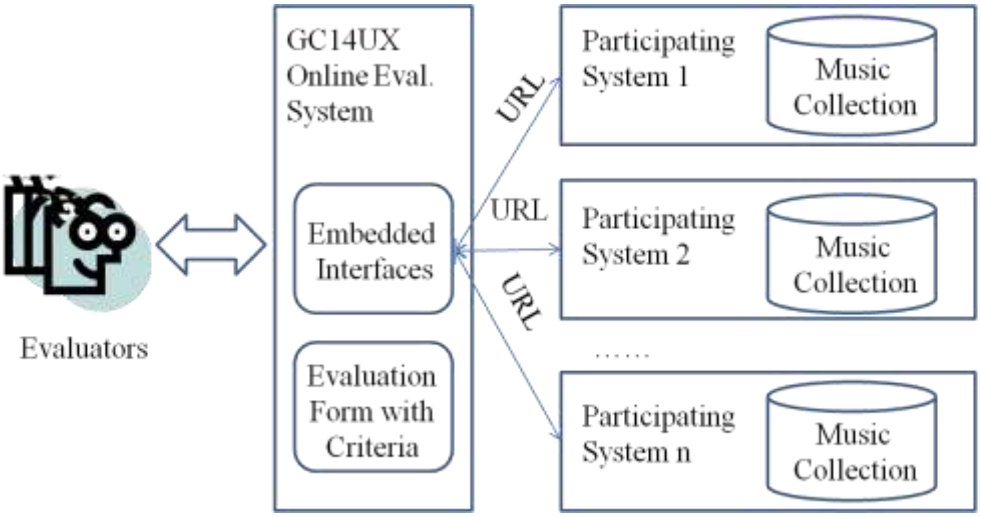
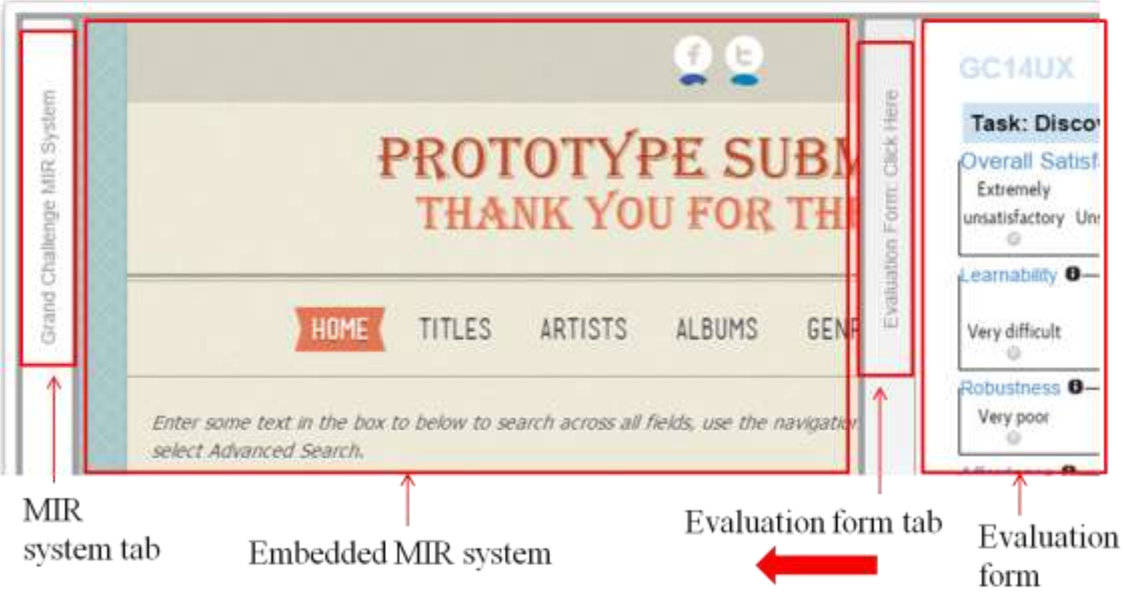


Figure 2. Overall structure of the GC14UX evaluation platform

Under this structure, the only requirement for participating MIR systems is to be implemented as websites accessible to users through a web browser. The systems are hosted by their developers so as to allow maximum design and implementation flexibility, though we provided a standard window size guideline so that all systems could comfortably fit in an *iframe*. At the time of system submission, instead of submitting the code to the data host as in the "algorithm-to-data" model, now participating teams only needed to submit the URLs to their systems through a submitter's web form (part of the GC14UX Online Evaluation platform). The resulting information was stored in a MySQL database table. The database information was then used to produce web pages for evaluators, where one of the submitted MIR systems is embedded in the page. Structuring things this way greatly improved the efficiency of managing the evaluation task (Downie et. al, 2014).

To maximize the screen space for the MIR systems, and to minimize the intrusion of the evaluation procedure to the user experience with the evaluated MIR systems, the evaluation form (Figure 1) was implemented as a sliding page. It is "folded in" (i.e., not visible) when a user is interacting with a MIR system, and is only presented when an evaluator clicks the "Evaluation form" tab. We liken the process to that of sliding a page in and out of an envelope. Figure 3 shows a screenshot of the evaluation interface *at the moment* while the evaluation form is sliding out. When the evaluation form page is completely slid out, the screen will show the complete evaluation form as in Figure 1, while the MIR system *iframe* is "folded in". A click on the "MIR system" tab will then make the evaluation form page slide back and be folded in at the right side of the screen and the MIR system *iframe* will take the screen and display complete interface of the MIR system under evaluation.



Enter some text in the box to below to search across all fields, use the navigatio select Advanced Search.

Figure 3. Evaluation platform interface in the middle of frame switch

Besides the input capturing front-end, a database is used as part of the back-end to save all answers from evaluators and manage assignments of each evaluator. As evaluators may want to modify their answers during the evaluation process, the database is designed such that later answers to questions replace old answers. To remind evaluators of their previous answers, the evaluation form shows the most recent answers of the evaluator when she/he last visited the form of an evaluated system.

Such a user-friendly and Web-accessible platform helps pave the way of attracting a larger number of evaluators from diversified locations and backgrounds, to conduct the evaluation in their natural environments. This is a remarkable advantage over traditional user experiment approach where users are recruited to interact with IR systems at specified time slots and locations (Andreasen et al., 2007). It also has the advantage of integrating (embedding) the evaluated system and the evaluation form into one place, without requiring the evaluation team to deploy the evaluated systems. This platform can be reused in other community-based user evaluation contexts where system developers are not conducting the evaluation and where multiple systems have to be evaluated and compared in the same settings. The source code of the platform has been uploaded to GitHub for free distribution under the University of Illinois/NCSA Open Source License[3].

The Procedure

After an evaluator logs in to the online evaluation platform, he/she is presented with an informed consent form which briefly introduces the research purposes and data collection procedure. After checking the consent, an evaluator will be assigned a number of assignments each of which involves interacting with one participating system and filling out an evaluation form with his/her perception of that system. The process is then repeated. The number of assignments will depend on the number of participating systems with the limit that no evaluator would evaluate a system more than once.

As the first run of the holistic evaluation, the GC14UX had three participating systems. Therefore each evaluator had three assignments each of which involved one participating system. This forms a within-subject evaluation. In future runs when there are a larger number of participating systems, each evaluator will be assigned to interact with a subset of the systems, such that the evaluators are not overloaded. The GC14UX evaluation platform is designed to ensure all participating systems will have roughly the same number of evaluators. To counterbalance possible effect of system sequence, the order of systems assigned is randomized for each evaluator.

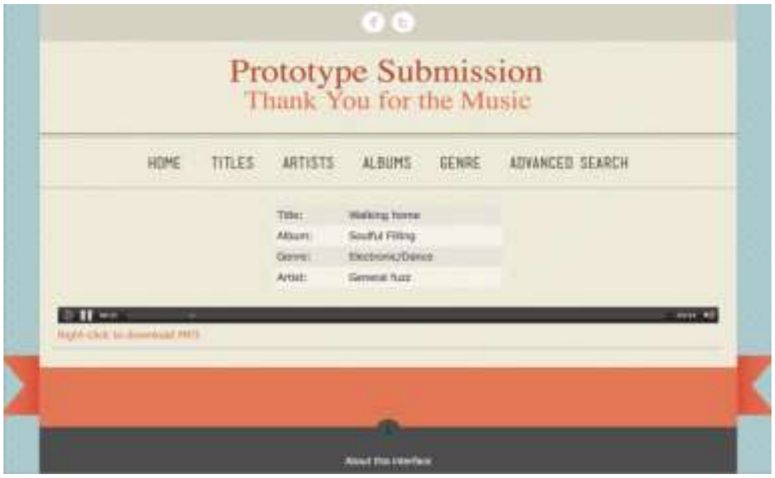
Before starting an assignment, the GC14UX evaluation platform shows a short instruction to the evaluator, asking him/her to 1) focus on evaluating the interaction and experience with the system as a whole, and not just the results; 2) be aware that the evaluated systems are using a collection of Creative Commons licensed, royalty-free music, and therefore the results may not include popular or familiar music; and 3) be aware that altering answers on an evaluated assignment is allowed and only the final set of answers will be analyzed. The evaluators can save partial answers and resume evaluation at a later time. All of them participated in this study voluntarily with no payment.

Results

Three systems participated in the GC14UX: "Tonic", "Moody" and "Thank you for the music." Implementing a complete interactive MIR system is a demanding task and often involves a group of researchers whose expertise cover a wide range of aspects from music signal processing to user interface design. Given that the GC14UX is the first iteration of holistic user experience evaluation in the MIR community, with a somewhat tight timeline[2], it was fortunate to have three participating systems to test the evaluation framework. The results and experience gained in this very first round will be highly informative for future improvements.

The Evaluated Systems

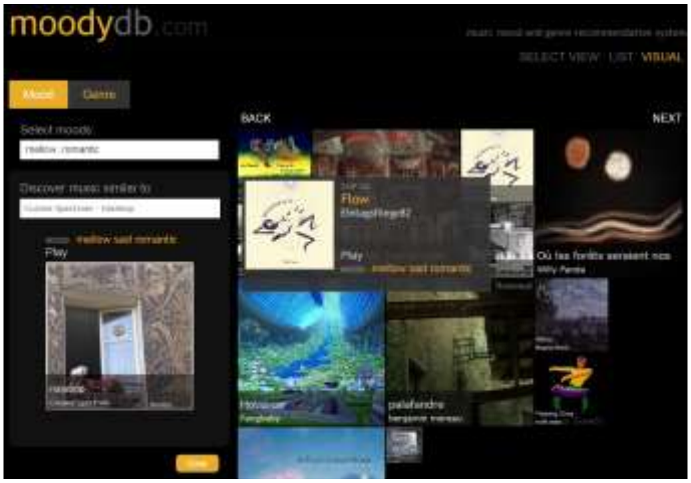
"Thank you for the music" (shortened name "Thank You") was included as a baseline interface to compare to, in that it is based on a traditional digital library framework where materials are indexed and searched by textual metadata, and makes no use of content-based analysis techniques. The system (Figure 4) was built using the open source Greenstone digital library framework (Witten, Bainbridge & Nichols, 2010) with a layout customized for the GC14UX. Figure 4 shows a page of one song in the collection, which consists of basic metadata of this song and an embedded audio player. The displayed song metadata is derived solely from the textual metadata provided in the Jamendo dataset. The navigation bar above the song lists the fields users can use to browse the music collection: title, artist, album and genre. The system also supports searching by the available metadata.



Title: Walking home
Album: Soulful Filling
Genre: Electronic/Dance
Artist: General fuzz

Figure 4. A screenshot of “Thank you for the music”

"Moody" is a music recommendation system that supports browsing by genre and mood. It allows users to choose from five mood labels and ten genre labels that are predefined based on recent MIR research. Informed by recent findings that a music piece may be of multiple moods and cross multiple genres (Hu & Downie, 2010; Lee, Hill & Work, 2012), Moody allows users to choose up to three genre/mood labels at the same time and retrieve songs satisfying all chosen genre/mood criteria. The genre labels of each song were obtained directly from the metadata in the Jamendo dataset, whereas the mood labels were based on a set of self-developed music mood classifiers. The Moody system also supports retrieving songs based on genre or mood similarity which is calculated based on genre/mood labels of the songs. Figure 5 shows the interface of Moody, which represents each retrieved song with its album image, giving it a colorful look. By clicking an image, users can play the song and view its basic metadata.



Select moods: mellow, romantic
Discover music similar to:

Figure 5. A screenshot of “Moody”

"Tonic" is a tag-based music retrieval system. It allows users to choose one or more predefined tags that can describe the music from many aspects including genre, mood, instrument, etc., and then retrieves songs relevant to the selected tags. The relevance is calculated based on the Jamendo metadata as well as self-built classifiers. At the same time, other tags associated with the retrieved songs are presented as suggestions to the users. The Tonic system has a unique, animation-based interface (Figure 6). It represents each selected tag as a bubble "floating" on the screen, and each recommended song as a smaller dot. Users can choose any of the dots to play and played songs are represented by their album images shown on the top of the screen (for previously played songs) or near the bottom (for the song currently played).

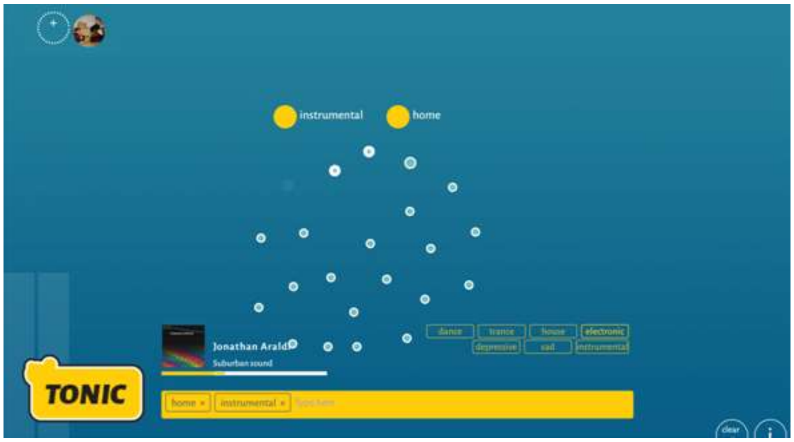


Figure 6. A screenshot of "Tonic"

Quantitative Scores

One benefit of implementing the online evaluation system is that it allows a relatively longer period of time to recruit more evaluators, which helps improve the validity of the results. From Oct. 13th to Nov. 14, 2014, 82 evaluators participated in the evaluation and 81 of them evaluated all three systems. In total, there were 244 sets of ratings. Table 3 summarizes the statistics of the answers on each of the closed evaluation questions, with bold numbers indicating the highest means among the systems. The Tonic system was rated the highest on "overall satisfaction," "affordance" and "feedback" while Thank You and Moody took the lead on Learnability and Robustness respectively. As described above, Tonic has more functions that other systems do not, such as listing songs that have been played in this session. This is probably why it was rated higher on "affordance." On the other hand, Thank You presents a simple and standard digital library interface that could be familiar to many evaluators, which may help explain its high rating on "learnability."

Criteria	**Statistics**	**Thank You** (N=81)	**Moody** (N=81)	**Tonic** (N=82)	**Kruscal-Wallis test**
Overall Satisfaction	Mean	4.15	4.63	**5.11**	*p* = 0.001**
	Std. Dev.	1.63	1.57	1.44	
	Median	4	5	5	
Learnability	Mean	**5.37**	5.33	5.29	*p* = 0.963
	Std. Dev.	1.23	1.28	1.56	
	Median	6	6	6	
Robustness	Mean	4.40	**4.53**	4.48	*p* = 0.769

	Std. Dev.	1.37	1.43	1.33	
	Median	4	5	4	
Affordance	Mean	4.49	4.65	**4.71**	*p* = 0.534
	Std. Dev.	1.61	1.50	1.46	
	Median	5	5	5	
Feedback	Mean	4.72	4.49	**4.79**	*p* = 0.756
	Std. Dev.	1.52	1.64	1.59	
	Median	5	5	5	

Table 3. Statistics and Kruskal–Wallis test results of ratings in the evaluation criteria.
** Difference is significant at *p* = 0.01 level (2-tailed).

To test whether the differences among the ratings of the three systems are significant, a non-parametric Kruskal–Wallis test was conducted on each criterion. Non-parametric tests are suitable as the ratings are ordinal data. The results (Table 3) show that only “overall satisfaction” had significant difference. Post hoc pair-wise tests discovered that the pair of significant difference was Thank you and Tonic ($p < 0.01$). The other pairs were not significantly different ($p = 0.17$ after Dunn-Sidak adjustment (Dunn, 1964). The fact that no criteria but overall satisfaction have significant difference may suggest there might be other important criteria contributing to overall satisfaction but not included in this evaluation. This will be discussed more in the *Reflection on MIR Evaluation* section.

To reveal possible relationships among the evaluation criteria, Spearman’s rank correlation coefficients were calculated for each pair of the criteria. Results are presented in Table 4. Overall satisfaction is significantly correlated with all other criteria, with “affordance” being most strongly correlated (rho = 0.70). This indicates that “affordance,” or the functionality of the systems has played a more important role than the other criteria when evaluators rated overall satisfaction. Table 4 also shows all criteria are significantly correlated but some correlations are weaker: robustness and affordance (rho = 0.30), learnability and robustness (0.31), for instance.

	Overall Satisfaction	**Learnability**	**Robustness**	**Affordance**	**Feedback**
Overall Satisfaction	1.000	.443**	.419**	.701**	.580**
Learnability		1.000	.312**	.317**	.438**
Robustness			1.000	.298**	.376**
Affordance				1.000	.539**
Feedback					1.000

Table 4. Spearman’s correlation between pairs of evaluation criteria (N = 244).
** Correlation is significant at *p* = 0.01 level (2-tailed).

Open-text comments

The open-text question asked evaluators to comment on any aspects of each system, which could reveal what matters most to them when they interact and evaluate complete, end-user oriented MIR systems. While a detailed analysis of the textual comments is reported separately in Lee et al. (2015), here we present a summary of the observations. A majority of comments, regardless of positive or negative, mentioned some aspects of interface designs. Phrases such as "visual appearance", "intuitive interface", "good looking UI [user interface]", “appealing and fun [application]” were frequently mentioned. Negative comments were often about confusions in user interface such as the bubble sizes in Tonic and two separate search boxes (genre and mood) in Moody. These indicate that presentation or interface design is one of the most important factors when users evaluate MIR systems.

There were relatively fewer comments on the functionality of the systems, and most of them were concerned about metadata-based versus content-based search. Some evaluators wondered whether the systems actually analyzed the music content, probably because all of the systems retrieve songs by text input, either common metadata, mood/genre labels or free tags. Although both Tonic and Moody built music audio classifiers at the back-end, it was not obvious to end-users. The relationships among elements (e.g., songs, tags) were mentioned as helpful information that systems should provide. Several evaluators complained about the quality of recommendations but at the same time they suspected that this might be attributed to the limitation of the test music collection rather than the performance of the systems.

Implications for MIR systems

The evaluation results have several implications for MIR system developers. First, the fact that users were significantly more satisfied with Tonic than the baseline Thank You indicates that searching and browsing by basic metadata were probably not sufficient. Users call for systems with at least some recommendation functions that can relate different songs or tags (Downie et al., 2009; Lee & Waterman, 2012). In addition, the fact that “affordance” had the strongest correlation to overall satisfaction emphasizes that the functionality of a system is still crucial for users. Second, the look of the system is highly important. It should be simple to operate, as the minimum requirement. Sleek, professional and fun interfaces would be highly preferred (Laplante & Downie, 2011). Third, the higher ratings on learnability compared to other criteria indicate that either all three systems did a good job in making intuitive interfaces, the evaluators were good self-learners, or both. The open text comments further suggested that all elements in the interface should be intuitive and easy to learn, or the users would not be satisfied. It is thus suggested that future system developers must conduct usability tests to identify and fix confusing interface elements (Lee & Price, 2015). Fourth, the music collections to which the systems provide access need to be sufficiently large and diverse for the target user groups (Hu & Kando, 2014).

The fact that the systems had significantly different scores only on the criterion of overall satisfaction reveals that user experience is affected by factors beyond usability. The open text comments also mentioned aesthetics and novelty (of the system and content) quite often (Lee et al., 2015). Therefore, it would be sensible for MIR system designers to pay attention to factors related to user engagement such as attractiveness, interactive visualization, and presentation of novel music (O’Brien & Toms, 2008; Laplante & Downie, 2011).

Admittedly, the evaluation is based on users’ one-time interaction with the systems, and thus what was measured is closer to evaluating users' first impressions of MIR systems, rather than long-term preferences. Nonetheless, positive first impressions are important or even necessary for subsequent adoption, usage, and even loyalty, especially when there is an abundance of available choices with intuitive, interactive and attractive interfaces. Although some evaluators mentioned that current MIR systems should provide content-based functions, we should note that many of the evaluators are likely to be familiar to MIR or IR in general. Future evaluations with general users without MIR or IR expertise will be necessary to confirm whether this concern is shared by general users.

Reflections on MIR Evaluation

As the first attempt for holistic user experience evaluation in MIR, this study gains valuable insights for future endeavors along this line of research. In the MIREX session of the International Conference on Music Information Retrieval (ISMIR) 2014, the GC14UX was openly discussed among the attendees. It was well acknowledged that holistic user experience evaluation has long been waited for in MIR. Furthermore, we hope GC14UX acts as a catalyst that inspires MIR researchers to pay more attention to developing complete MIR systems and evaluating user experience. The results from this first attempt demonstrate the type of roles future GCUX evaluations would provide: to give system developers feedback on more abstract facets of system design. Many MIR researchers expressed that they would like to see and/or participate in future rounds of holistic evaluation like GC14UX. Here we summarize and further analyze the raised opinions based on which suggestions for future rounds of holistic user experience evaluations are proposed.

More Use Cases

In GC14UX, there was only one use case, searching for free music as the background of a personal video. It is desirable to have more and diversified use cases as people look for music in a wide range of situations and for different purposes. In addition, there are multiple user groups for MIR systems besides the general audience: for instance, musicologists, musicians, music producers, video producers and music librarians. Different user groups may have different needs which could generate diversified use cases. Systems supporting different use cases should have equal status of being evaluated in the holistic user experience evaluation framework. In addition, in each round the use cases should be changed/updated so as to avoid "use case overfitting" where systems are optimized to cater certain specific use cases to the extent that they cannot be applied to other unseen use cases.

In future rounds of GCUX, multiple use cases should be designed, each of which could be taken as a subtask, and each system can participate in one or more subtasks. In MIR, use cases should be "closer to what would be useful for real users” (Lee & Cunningham, 2013, p.499). Based on the user requirements for music services identified by Lee and Waterman (2012), real-life music use scenarios that could be developed into evaluation tasks in the near future include: to listen to music recordings; to discover new music; to obtain music recordings; to identify/verify a particular song/artist/album; to learn more about artists/bands; to get recommendations; to watch performances/music videos; to create playlists/stations; to share music; and to curate/manage a themed music collection and its metadata.

More Diversified Music Collection

Royalty-free music has at least three advantages in MIR evaluation: 1) being freely sharable and distributable, 2) being able to compose large and diversified music datasets, and 3) being able to mitigate against the possible user experience bias induced by the differential presence (or absence) of popular or known music within the participating systems. However, it has the limitation of being unfamiliar to most users. Notwithstanding the demand for free and unfamiliar music, familiarity and popularity indeed play a role in many cases of music information seeking and use (Hu & Lee, 2012). One possible solution is to use short previews of commercial music that are made freely available via online music services such as 7digital (www.7digital.com). Admittedly the previews are short clips (usually 30 seconds long) instead of the full recordings, but, accompanied with rich metadata, it could be a viable approximation. An alternative is to negotiate with music service providers to make the music content or (a wide range of) extracted features available only to system developers. When end-users/evaluators would like to play the music, the systems would then point them back to the relevant service provider. This, for example, is the model of Google Scholars and Google Books where the full-text content is available to Google for building search indexes which then help drive user traffic to the original content providers (e.g., publishers, journals and databases).

In the long run, the MIR research community should collectively work with commercial labels on the issue that has been plaguing MIR research since the beginning: music availability for academic use. In the context of holistic user experience evaluation, this would put strict requirements on the evaluation platforms such that only evaluators can play and listen to the music included in the evaluation dataset via the evaluated MIR systems. This would help ensure the strictly limited use of the music is for academic evaluation only.

Evaluation Methods

System logs can record how users interact with a system and have a great potential in user experience evaluation. Log analysis has been used in a number of domains in finding out usage patterns and issues in system design (Srivastava, Cooley, Deshpande & Tan, 2000). It has the advantages of being unintrusive and objective, as well as being able to be batch processed using automated programs. In future rounds of holistic user experience evaluation, participants could also submit system logs that contain user system interactive behaviors such as query terms, link/button clicks, scrolling up and down, and mouse movements. The GC14UX evaluation system could also record, in an anonymized form, evaluation behaviors such as logging in and out, changing answers and switching between the evaluated systems and the evaluation forms. These can help reveal how evaluators actually work during the evaluation process.

Besides the online heuristic-based evaluation used in GC14UX, user-experiment based evaluation methods may also be helpful, such as usability testing via a think-aloud protocol and/or observations of user interactions. These methods can help uncover a variety of issues with a system's design in detail, but they involve labor-intensive analysis that is difficult to automate. It is therefore desirable to combine online evaluation and user experiment (Andreasen et al., 2007).

Evaluation Criteria

To keep the evaluation context as authentic as possible, questions asked during the evaluation should be as few as possible. On the other hand, to capture more feedback from evaluators, more questions would be desirable. It is difficult to strike a balance. For GC14UX, the closed questions covered most of Nielsen's heuristics of usability except for "aesthetic and minimalist design" which in fact could be important for user experience as reflected by the evaluators' open-text comments. Beyond usability measures, hedonic outcomes of music information seeking (Laplante & Downie, 2011) should be taken into consideration as well. Many of the attributes in the engagement model proposed by O'Brien and Toms (2008), such as affect, aesthetics and sensory appeal, can be particularly helpful in completing the holistic evaluation framework. Considering MIR systems are often for entertainment purposes, Hu and Kando (2014) proposed a set of evaluation criteria for casual-leisure MIR research, including novelty, aesthetics, enjoyableness and emotion status. These criteria can also be adopted in future rounds of holistic user experience evaluation.

System preference is another criterion in system evaluation (Kelly, 2009). Instead of asking for absolute scores, system preference asks evaluators to rank the systems based on their relative preferences. It is argued that preferences or rankings are more accurate than absolute scores as human evaluators may not always remember the scores they give to each system but they are more certain about which systems work better (Urbano, Morato, Marrero & Martín, 2010; Yang & Chen, 2011). In text IR evaluation, system preference is often adopted in conjunction with other user-centered criteria such as satisfaction, to collect a clear indication of users' attitude towards systems (Kelly, 2009). For future iterations of GC14UX, system preference could also be used.

If the MIR community feels the need to have a much more exhaustive and/or much more pre-defined list of evaluation criteria, one alternative is to recruit panels of experts and/or paid evaluators to conduct the evaluation. In this way, the systems are evaluated thoroughly, but not in authentic contexts. If time and resources permit, it is desirable to have two versions of evaluation: a short version along the lines used in GC14UX—or even a shorter one, for general users in their authentic contexts—and a detailed version with more comprehensive criteria developed by an assembled panel of experts.

Conclusions

In this article we have presented the very first holistic user experiment evaluation framework in MIR, the Grand Challenge on Usability Experience 2014 (GC14UX). A novel and copyright-free music collection, a use scenario, a set of evaluation criteria as well as a unique evaluation platform were constructed for this evaluation framework. Three diverse MIR systems participated in GC14UX and 82 evaluators interacted and rated the systems. The evaluation results reveal significant differences on overall satisfaction among the systems, as well as the relative strengths of the systems on four specific user experience aspects: learnability, robustness, affordance and feedback. In addition to the results, the carefully designed evaluation methods were thoroughly described and reflected. Through GC14UX, we gained invaluable experience in holistic user experience evaluation in MIR, based on which suggestions were made for future iterations and improvements.

Footnote: 1. http://www.jamendo.com/en/welcome

2. The first announcement was made to the MIR community in late February and system submission deadline was in late September.

3. The code is available at https://github.com/imirsel/grand-challenge

Acknowledgements

We sincerely thank all the anonymous evaluators participating in the GC14UX , the authors of the participating systems, and the colleagues in the MIR community who kindly shared their opinions and suggestions. We also thank the anonymous reviewers for their helpful comments. This study is partially supported by the Andrew W. Mellon Foundation. The University of Hong Kong also supported part of the study with a seed research fund.

References:

Aljanaki, A., Yang, Y. H., & Soleymani, M. (2014, October). Emotion in music task at mediaeval 2014. In *MediaEval 2014 Workshop, Barcelona, Spain*.

Andreasen, M. S., Nielsen, H. V., Schrøder, S. O., & Stage, J. (2007, April). What happened to remote usability testing?: an empirical study of three methods. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 1405-1414). ACM.

Bargas-avila, J. A, & Hornbæk, K. (2011). Old Wine in New Bottles or Novel Challenges ? A Critical Analysis of Empirical Studies of User Experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2689–2698. doi:10.1145/1978942.1979336

Cano, P., Gómez, E., Gouyon, F., Herrera, P., Koppenberger, M., Ong, B., Serra, X., Streich, S. & Wack, N. (2006). ISMIR 2004 audio description contest. *Music Technology Group of the Universitat Pompeu Fabra, Tech. Rep*.

Cicchetti, D. V., Shoinralter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement*, *9*(1), 31-36.

Downie, J. S. (2003). Music information retrieval. *Annual review of information science and technology*, *37*(1), 295-340.

Downie, J. S., Byrd, D., & Crawford, T. (2009, October). Ten years of ISMIR: Reflections on challenges and opportunities. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)* (pp. 13-18).

Downie, J. S., Ehmann, A. F., Bay, M., & Jones, M. C. (2010). The music information retrieval evaluation exchange: Some observations and insights. In *Advances in music information retrieval* (pp. 93-115). Springer Berlin Heidelberg.

Downie, J. S., Hu, X., Lee, J. H., Choi, K., Cunningham, S. J., Hao, Y. and Bainbridge, D. (2014). Ten years of MIREX (Music Information Retrieval Evaluation eXchange): Reflections, challenges and opportunities, In *Proceedings of the 15th International Conference on Music Information Retrieval (ISMIR)*,Oct. 2014, Taipei, Taiwan.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, *6*(3), 241-252.

Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., … Sølvberg, I. (2007). Evaluation of digital libraries. *International Journal on Digital Libraries*, 8(1), 21–38. doi:10.1007/s00799-007-0011-z

Gruzd, A. A., Downie, J. S., Jones, M. C., & Lee, J. H. (2007, June). Evalutron 6000: collecting music relevance judgments. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (pp. 507-507). ACM.

Harman, D. (1995). Overview of the second text retrieval conference (TREC-2). *Information Processing & Management*, 31(3), 271-289.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS),* 22(1), 5-53.

Hoashi, K., Hamawaki, S., Ishizaki, H., Takishima, Y., & Katto, J. (2009). Usability evaluation of visualization interfaces for content-based music retrieval systems. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)* (pp. 207-212).

Hu, X., & Downie, J. S. (2010, June). Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th annual joint conference on Digital libraries* (pp. 159-168). ACM.

Hu, X., & Lee, J. H. (2012). A cross-cultural study of music mood perception between American and Chinese listeners. In *Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR)* (pp. 535-540).

Hu, X., & Kando, N. (2012). User-centered measures vs. System effectiveness in finding similar songs. In *Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR)* (pp. 331-336).

Hu, X., & N. Kando (2014). Evaluation of music search in casual-leisure situations. In *Workshop on searching for fun 2014, Proceedings of the 5th Information Interaction in Context Symposium on-IIiX'14* (pp. 1-4). ACM Press.

Hu, X., & Liu, J. (2010, August). Evaluation of music information retrieval: Towards a user-centered approach. In *Proceedings of the 4th Workshop on Human-Computer Interaction and Information Retrieval (HCIR).*

Katter, R. V. (1968). The influence of scale form on relevance judgments. *Information Storage and Retrieval*, *4*(1), 1-11.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, *3*(1-2), 1-224.

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4-5), 441-504.

Laplante, A., & Downie, J. S. (2011). The utilitarian and hedonic outcomes of music information-seeking in everyday life. *Library and Information Science Research*, 33(3), 202–210. doi:10.1016/j.lisr.2010.11.002

Lee, J. H., & Cunningham, S. J. (2013). Toward an understanding of the history and impact of user studies in music information retrieval. *Journal of Intelligent Information Systems*, *41*(3), 499-521.

Lee, J. H., Hill, T., & Work, L. (2012, February). What does music mood mean for real users? In *Proceedings of the 2012 iConference* (pp. 112-119). ACM.

Lee, J. H., Hu, X., Choi, K. & Downie, J. S. (2015). MIREX Grand Challenge 2014 on User Experience: Qualitative Analysis of User Feedback. Accepted to *Proceedings of the 16th International Conference on Music Information Retrieval (ISMIR)*.

Lee, J. H., & Price, R. (2015). User experience with commercial music services: an empirical exploration. *Journal of the Association for Information Science and Technology*, doi: 10.1002/asi.23433.

Lee, J. H., & Waterman, N. M. (2012, October). Understanding user requirements for music information services. In *Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR)* (pp. 253-258).

Liem, C. C., Orio, N., Peeters, G., & Schedl, M. (2013). MusiClef 2013: Soundtrack selection for commercials. In *MediaEval*.

Nielsen, J. (1994). Heuristic evaluation. *Usability inspection methods*, *17*(1), 25-62.

Nielsen, J. (2005). Ten Usability Heuristics. http://www.useit.com/papers/heuristic/heuristic_list.html

Nielsen, J., & Molich, R. (1990, March). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 249-256). ACM.

Orio, N., Rizo, D., Miotto, R., Schedl, M., Montecchio, N., & Lartillot, O. (2011, October). MusiCLEF: A benchmark activity in multimodal music information retrieval. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)* (pp. 603-608).

Pauws, S., & Eggen, B. (2002, October). PATS: Realization and user evaluation of an automatic playlist generator. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR).*

Pauws, S., & van de Wijdeven, S. (2005, September). User evaluation of a new interactive playlist generation concept. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)* (pp. 638-643).

Pu, P., Chen, L., & Hu, R. (2011, October). A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems* (pp. 157-164). ACM.

Salton, G. (1971). *The SMART Retrieval System—Experiments In Automatic Document Processing*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA

Schedl, M. (2013). Ameliorating Music Recommendation: Integrating Music Content, Music Context, and User Context for Improved Music Retrieval and Recommendation. In *Proceedings of International Conference on Advances in Mobile Computing & Multimedia*, 3.

Schedl, M., Flexer, A., & Urbano, J. (2013). The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, *41*(3), 523-539.

van Schaik, P., & Ling, J. (2003). Using on-line surveys to measure three key constructs of the quality of human–computer interaction in web sites: psychometric properties and implications. *International Journal of Human-Computer Studies*, *59*(5), 545-567

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter*, *1*(2), 12-23.

Urbano, J., Morato, J., Marrero, M., & Martín, D. (2010). Crowdsourcing preference judgments for evaluation of music similarity tasks. In *ACM SIGIR workshop on crowdsourcing for search evaluation* (pp. 9-16).

Vignoli, F., & Pauws, S. (2005, September). A music retrieval system based on user driven similarity and its evaluation. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)* (pp. 272-279).

Voorhees, E. M. (2002, January). The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems* (pp. 355-370). Springer Berlin Heidelberg.

Voorhees, E. M., & Harman, D. K. (Eds.). (2005). *TREC: Experiment and evaluation in information retrieval* (Vol. 63). Cambridge: MIT press.

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, *64*(6), 956-972.

Witten, I.H., Bainbridge, D. & Nichols, D.M. (2010) *How to Build a Digital Library*, Second Edition. Burlington, MA: Morgan Kaufmann.

Witten, I. H., Moffat, A., & Bell, T. C. (1999). *Managing Gigabytes: Compressing and Indexing Documents And Images*. Morgan Kaufmann.

Yang, Y. H., & Chen, H. H. (2011). Ranking-based emotion recognition for music organization and retrieval. *Audio, Speech, and Language Processing, IEEE Transactions on*, *19*(4), 762-774.

Zhang, Y. (2010). Developing a holistic model for digital library evaluation. *Journal of the American Society for Information Science and Technology*, 61(1), 88–110. doi:10.1002/asi.21220

Zhang, Y. C., Séaghdha, D. Ó., Quercia, D., & Jambor, T. (2012, February). Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining* (pp. 13-22). ACM.