Task Complexity and Difficulty in Music Information Retrieval

Xiao Hu
xiaoxhu@hku.hk
Faculty of Education, University of Hong Kong, Hong Kong.

Noriko Kando
kando@nii.ac.jp
National Institute of Informatics, Tokyo, Japan

Abstract: There has been little research on task complexity and difficulty in music information retrieval (MIR), whereas many studies in the text retrieval domain have found that task complexity and difficulty have significant effects on user effectiveness. This study aimed to bridge the gap by exploring 1) the relationship between task complexity and difficulty; 2) factors affecting task difficulty; and 3) the relationship between task difficulty, task complexity, and user search behaviors in MIR. An empirical user experiment was conducted with 51 participants and a novel MIR system. The participants searched for six topics across three complexity levels. The results revealed that 1) perceived task difficulty in music search is influenced by task complexity, user background, system affordances and task uncertainty and enjoyability; and 2) perceived task difficulty in MIR is significantly correlated with effectiveness metrics such as the number of songs found, number of clicks and task completion time. The findings have implications for the design of music search tasks (in research) or use cases (in system development) as well as future MIR systems that can detect task difficulty based on user effectiveness metrics.

I. Introduction

Music information retrieval (MIR) as a research field has flourished in recent decades, with numerous approaches developed to leverage novel search technologies such as audio fingerprinting and pattern matching (Downie, 2003; Schedl, Gómez & Urbano, 2014; Downie et al, 2014). The impact of MIR systems on users' music searching behaviors is under-researched (Hu & Kando, 2012, Hu & Kando, 2014; Hu, et al., 2015; Lee & Cunningham, 2013). In particular, there has not been research which has focused on task complexity or difficulty in MIR, nor on their effects on user interactions in music search, although it has been found in the text information retrieval (IR) field that task complexity and difficulty can affect information seeking and use (Byström & Järvelin, 1995; Wildemuth & Freund, 2009; Liu, Liu, Gwizdka & Belkin, 2010; Kelly et al., 2015; Wildemuth, Freund, & Toms, 2014). Although the two concepts are sometimes used interchangeably in the literature, it is important to differentiate task complexity and task difficulty. In this study, we adopt the clarification in (Wildemuth et al., 2014) that task complexity is designed and agreed by researchers or experts based on existing frameworks of cognitive complexity and thus is regarded as objective; whereas task difficulty refers to users' perception of how difficult the tasks are.

This study aims to explore task complexity and task difficulty in the context of MIR and their effects on user effectiveness during music search. A task-driven user experiment was conducted in a
controlled setting. The tasks were designed according to a cognitive complexity framework (Arguello et al., 2012; Krathwohl, 2002). User behavioral data were collected from search sessions using a novel MIR system. User background and perceptions were collected with questionnaires and focus group interviews. Quantitative and qualitative data were analyzed to answer the following research questions:

RQ1: Is user perception of task difficulty correlated with task complexity in music search?
RQ2: What other factors of users, the MIR system and the tasks also influence perceived task difficulty?
RQ3: What are the relationships between task difficulty, task complexity, user effectiveness and satisfaction in music search?

The first two questions are intended to find out what constitutes task difficulty and whether findings in task complexity and difficulty in text retrieval also apply in the MIR domain. On one hand, music search, similar to searching for other information, does involve cognitive loads of the searchers. On the other hand, music search could be entertainment-oriented (Laplante & Downie, 2011; Hu & Kando, 2014), and thus searchers may have different interpretations of what is difficult. The third question is intended to investigate possible relationships between task complexity, task difficulty, user effectiveness and user satisfaction. Findings from this study may help further our understanding on why a task is difficult in music search, and provide methodological implications for determining task difficulty in future MIR studies.

II. Related Work

2.1 Task complexity and difficulty in text retrieval

In recent years, there has been active research on the assessment of task complexity and difficulty as well as their relationship with search results and user behaviors in the domain of text Information Retrieval (IR). A number of studies have examined how increasing task complexity affects the searching process, searchers’ performance, as well as search outcomes. Byström and Järvelin (1995) proposed a framework for task categories with different complexity levels, from automated information processing tasks to genuine decision making tasks. As a conceptual study, it revealed the relationships between task complexity, types of information, information channels and sources. Following this framework, a recent study by Bailey et al. (2015) re-classified the past topics in the Text Retrieval Conferences (TREC) into three classes of task complexity: remembering, understanding, and analysis.

Other studies have set out to measure user search behaviors with tasks of different complexity. A study by Aula, Khan, and Guan (2010) found that users’ search behaviors become less systematic when search tasks were more complex. Users changed their searching approaches, spent more time on the results page, or formulated more queries for difficult tasks. A more recent study (Arguello, Wu, Kelly & Edwards, 2012) revealed that more complex tasks required significantly more interaction between users and the system. Besides task complexity, perceived task difficulty has
also been shown to have significant implications for search behaviors. In Kim’s (2006) study, it was found that user’s perception of task difficulty served as a good indicator and predictor for search effectiveness in exploratory and factual tasks as reflected by the number of query reformulations, time spent and pages viewed. Haerem and Rau (2007) studied the relationships between task difficulty and user expertise. Novices and experts were found to exhibit varied performance in tasks with different objectives and perceived task difficulty. More recently, Liu et al. (2010) investigated user search behaviors and task difficulty. Users were found to not only spend more time on difficult tasks, but also issued more queries, viewed more retrieved items, and spent more time on retrieved items.

In an effort to summarize previous research on task complexity and difficulty, Wildemuth et al. (2014) reviewed 106 papers in Interactive Information Retrieval (IIR) and examined how the concepts of "task complexity" and "task difficulty" as well as task descriptions were defined in these studies. Despite the active research on this topic in the text IR domain, there has been little research on task complexity and difficulty in MIR. In this study, we aim to bridge this gap by exploring and assessing task complexity, task difficulty, their effects on search behaviors as well as implications for MIR task and system design.

2.2 Tasks in Music Information Retrieval

There have been a number of user studies involving various search tasks in MIR. An earlier study by Iwamura (2001) explored the possibility of searching for a song title when only the melody was known, and found it was not an easy task. A study by Salaba and Zhang (2012) investigated the extent to which a given MIR system supported two kinds of user tasks: prescribed tasks (e.g., identifying a work given its title) and an open task self-assigned by the users based on personal interests. Although the results showed that there were more failures in the open task than prescribed tasks, the study did not explicitly discuss the issue of task complexity or difficulty. Very recently, the Music Information Retrieval Evaluation eXchange (MIREX), a research community wide evaluation platform, hosted its first Grand Challenge on User Experience (GC14UX) where three novel MIR systems were evaluated with end-users’ interactions (Hu et al., 2015). This was the first time a user task was used in MIREX. The task was to “find background music for a self-created short video about a memorable occasion of the user” (p.5), which was intended to be flexible and authentic to users’ lives.

Other studies have explored different tasks in the context of user needs assessment (Lee & Cunningham, 2013). Aiming to design authentic music search tasks for system evaluation, a number of studies analyzed user queries raised in searching for music. Downie and Cunningham (2002) investigated features used to describe music information need by analyzing music-related requests posted to a popular music newsgroup. They found people often search for music by bibliographic description (e.g., title, artist), lyrics, genre, similar works and affect (i.e., mood, emotion). Lee (2010) followed this line of research by analyzing music queries posted on a Q&A website. Findings show that most queries raised were known-item searches, with a wide variety of descriptive information including the contexts of knowing or listening to the items. A subsequent study (Lee & Waterman,

2012) conducted a large scale survey among music users and identified user requirements for music services. Based on these findings, a recent study on MIR evaluation synthesized a list of music use scenarios based on which music seeking tasks can be developed, such as to listen to music recordings, to discover new music, to create playlists/stations, etc. (Hu et al, 2015).

However, none of these studies considered task complexity and difficulty as constructs in the MIR process nor their effects on users searching behaviors. As the implications of task complexity and difficulty may differ in text IR and MIR, this study will investigate the relationship between music search task complexity and difficulty, and draw inferences for their causes and effects.

### III. Research Design

To answer the research questions, a user experiment was conducted with tasks of different complexity levels and a novel MIR system. This section describes the system, tasks, measures and the experiment procedure.

#### 3.1 The System

Moodydb is a novel MIR system developed in-house by the research team (Hu et al., 2008b; Hu et al., 2011). It is a Web-based music mood classification and retrieval system. It is content-based in that it extracts salient spectral features (e.g., Mel-frequency Cepstral Coefficients, Casey et al., 2008) from music audio and automatically classifies music pieces into five mood categories: *passionate, cheerful, bittersweet, silly/quirky* and *aggressive*. The five categories have been used in the Audio Mood Classification (AMC) task in MIREX since 2007 (Hu et al., 2008a). For automated classification, Moodydb adopts the SMO (sequential minimal optimization) implementation of Support Vector Machines in the Weka toolkit (Witten & Frank, 1999). Moodydb also supports retrieving music by mood similarity which is calculated by the distance between the mood profiles of the songs. Given a seed song, Moodydb returns a set of songs with similar moods and displays the songs and their album covers on the screen. For technical details of the system, please refer to Hu et al. (2008b).

When a user starts using Moodydb, he or she types in part of a song title or artist name into the search box. Moodydb has automated completion function and displays a list of songs matching the textual query. After the user selects one of the songs as the seed song, Moodydb will retrieve a set of songs with similar moods to the seed song as recommendations. The user then can examine and play the songs. When the cursor hovers over a recommended song, a small panel will appear, displaying detailed information of the song, including basic metadata (e.g., title, artist), the mood label assigned by the content-based mood classifier, the album cover, and a “play” button. Due to intellectual property constraints, each audio clip played in Moodydb is limited to 30 second extracts from the middle of a song.

Moodydb provides two modes of presenting recommended songs, named *list* and *visual*, as shown in Figures 1 and 2. In the list mode, the recommended songs are arranged in two columns ranked from top to bottom of the page according to their mood similarity to the seed song. The higher a song is listed, the more similar it is determined to be to the seed song. Between the two songs in a
row, the left one is more similar than the right one. There are up to 10 recommended songs on each page. The visual mode uses the sizes of album images to indicate the similarity levels. The larger a song’s album image is, the more similar it is determined to be to the seed song (Hu et al., 2011). The experiment presented in this study assigned the two modes to an equal number of participants, and the perceived task difficulty of searches conducted using the two modes was compared.

![Figure 1. Moodydb interface with annotations: the list mode](image1)

![Figure 2. Moodydb interface with annotations: the visual mode](image2)

There are 750 songs in Moodydb, consisting mostly Western popular songs. It is acknowledged that the music collection is limited compared to commercial music services as it is constrained by the music audio files owned by the researchers. Despite this limitation, this system was used for several reasons: 1) mood is an important access point for users in searching for music (Lee & Downie, 2004, Vignoli 2004, Lee & Waterman, 2012), while there are few existing systems which support music search by mood; 2) with an in-house system it is possible to control the system features, such as different visualization modes; 3) Moodydb has been used in previous studies (Hu, Kando & Yuan,
2011; Hu & Kando, 2014), and thus using the same system allows for cross-study comparison; and 4) it is possible to obtain the server side logs for more detailed analysis. For example, the completion time and times of song plays on each task by each user were accurately recorded in the server logs.

3.2 The Tasks

As there is little research on task complexity and difficulty in MIR, Krathwohl’s cognitive complexity framework (2002) is adopted in this study. It is well recognized in the education field, and has been used in interactive IR (Arguello et al., 2012; Kelly et al., 2015). In this framework, the lowest level is “remember” which involves fact-finding tasks. The next level is “understand”, which is to “determine the meaning of information by interpreting, inferring, comparing and/or explaining” (Krathwohl 2002, p.215). A third level, “analyze” is also included in this study. Tasks in this level are more complex, which involves differentiating various parts of the material and detecting the relationships between the parts as well as those between the parts and the whole.

When designing the tasks, it is also necessary to consider the affordances of the Moodydb system: the properties of Moodydb that “determine how it could be used” (Norman, 1988, p.9). First, Moodydb is a recommender system, and thus tasks involving song recommendations would be the most suitable. Second, some characteristics of the songs are not shown in the interface, but could be a criterion of music seeking, such as voice gender or tempo/speed. Using those as selection criteria would involve a certain level of understanding of the songs and thus contribute to higher task complexity. Third, a song can have a number of aspects such as mood, genre, rhythm, etc. A task involving evaluation on these multiple aspects could further increase the complexity to the “analyze” level. The designed tasks in the three cognitive levels are listed in Table 1.

<table>
<thead>
<tr>
<th>Level</th>
<th>Complexity</th>
<th>Cognitive level</th>
<th>Task type</th>
<th>Task description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Easy</td>
<td>Remember</td>
<td>Find recommendations</td>
<td>Given an artist name, find a song of this artist in a given mood and other songs with the same mood label</td>
</tr>
<tr>
<td>2</td>
<td>Medium</td>
<td>Understand</td>
<td>Find recommendations with characteristics that are not displayed</td>
<td>Given a song, find other songs with the same mood label and in a voice of the opposite gender (voice gender cannot be seen from the interface)</td>
</tr>
<tr>
<td>3</td>
<td>Difficult</td>
<td>Analysis</td>
<td>Playlist generation</td>
<td>Given an artist, choose a song of this artist and find other songs for building a playlist, with consideration of multiple musical aspects</td>
</tr>
</tbody>
</table>
Level 1 task is intended to find songs using given criteria that can be evaluated from the metadata displayed on the system interface. It only requires the users to look for mood labels, without listening to or evaluate the songs. Level 2 goes one step further, requiring the users to listen to the songs and make judgments based on the music heard. This task approximates the “understand” level in the cognitive framework, although what constitutes “understanding” a piece of music is still an open question (Levitin, 2011). One may think “understanding” music would entail knowing the theme or subject of the music and/or being able to discern the music’s characteristics such as melody, speed and/or instrumentation. However, musical themes are often hard to judge within a short period of time of listening except for songs which are familiar to the users. In addition, some music characteristics (e.g., instrumentations, height of pitches) may be hard to judge for ordinary users. Therefore, the judgment in this Level 2 task is designed to be straightforward enough such that the judgment does not require deliberate thinking which might make the task too complex. The task in Level 3 involves more deliberate judgment as it asks users to generate a playlist which requires the retrieved songs to be coherent in some way. It also requires considerations of multiple musical aspects. Therefore, the task in Level 3 can be regarded as being at the “analyze” level.

For each task, two topics were designed to balance possible biases introduced by music familiarity and preferences. There were further considerations in instantializing the tasks. First, we tried to include famous artists to make the topics more interesting to the participants and thus closer to real-life experience (Hu et al., 2015). Second, the researchers tried to balance female and male artists. Third, as the dataset was limited, the researchers tried to include artists with multiple songs in the system, so that the participants could have more freedom of choice, making the topics closer to real-world situations. Finally, all topics in all levels asked for three (3) songs (answers), to make the metrics (e.g., task completion time) comparable across tasks. Table 2 shows the topics in each task.

Table 2. Topics in each complexity level (emphases are the same as those presented to the users)

<table>
<thead>
<tr>
<th>Level</th>
<th>Topic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>T1</td>
<td>Choose one song from &quot;The Beatles&quot; as the seed song, then find three (3) recommended songs with the same mood label as the seed song.</td>
</tr>
<tr>
<td>1</td>
<td>T2</td>
<td>Choose one song from the &quot;Black Eyed Peas&quot; as the seed song, then find three (3) recommended songs with the same mood label as the seed song.</td>
</tr>
<tr>
<td>2</td>
<td>T3</td>
<td>Use the song, &quot;Wannabe&quot; by the &quot;Spice Girls&quot; as the seed song, then find three (3) recommended songs with the same mood label and performed by a male voice.</td>
</tr>
<tr>
<td>2</td>
<td>T4</td>
<td>Use the song, &quot;Speed of Sound&quot; by &quot;Coldplay&quot; as the seed song, then find three (3) recommended songs with the same mood label and performed by a female voice.</td>
</tr>
<tr>
<td>3</td>
<td>T5</td>
<td>Choose a song by the artist &quot;Avril Lavigne&quot; as the seed song, and three (3) recommended songs to make a playlist with the seed song. Please consider as many music aspects as you can, such as mood, rhythm, vocal, artist, album image, etc. Please note the mood you feel about a song could be different from the mood</td>
</tr>
</tbody>
</table>
label given by the system, so you need to listen to the songs to make choices.

3. Choose a song by the artist "Eminem" as the seed song, and three (3) recommended songs to make a playlist with the seed song. Please consider as many music aspects as you can, such as mood, rhythm, vocal, artist, album image, etc. Please note the mood you feel about a song could be different from the mood label given by the system, so you need to listen to the songs to make choices.

3.3 Measures

To answer the research questions, we collected the measures of user perceived task difficulty, user effectiveness, satisfaction as well as users’ familiarity and preferences for the songs. The metrics, definitions and scales are presented in Table 3. Participant interactions were recorded by the Open Web Analytics (OWA) tracking tool, and user interaction metrics were calculated based on the logs.

Table 3. Measures of user behaviors and perceptions.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Metrics</th>
<th>Definition</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task difficulty</td>
<td>Perceived task difficulty</td>
<td>User self-reported ratings on difficulty of each search</td>
<td>7-point Likert scale; 1: very easy; 7: very difficult</td>
</tr>
<tr>
<td>User effectiveness</td>
<td>Task completion time</td>
<td>Time duration between the start and the end of a search</td>
<td>Non-negative integer numbers in seconds</td>
</tr>
<tr>
<td></td>
<td>Number of songs found</td>
<td>The number of songs found as answers to a search topic</td>
<td>Non-negative integer numbers</td>
</tr>
<tr>
<td></td>
<td>Number of clicks</td>
<td>Total number of clicks a participant made in a search</td>
<td>Non-negative integer numbers</td>
</tr>
<tr>
<td></td>
<td>Number of plays</td>
<td>Times a participants played music in a search</td>
<td>Non-negative integer numbers</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>Satisfaction with songs found</td>
<td>User self-reported satisfaction level</td>
<td>7-point Likert scale; 1: very unsatisfied; 7: very satisfied</td>
</tr>
<tr>
<td>Familiarity and preferences</td>
<td>Song or artist familiarity</td>
<td>User self-reported familiarity to the song or artist stated in the topic</td>
<td>5 options: “do not know it”, “I heard of it”, “have listened to”, “very familiar” “I am not sure” (not analyzed)</td>
</tr>
<tr>
<td></td>
<td>Preference of the seed song</td>
<td>User self-reported extent to which he/she likes the seed song</td>
<td>7-point Likert scale; 1: very much dislike the song; 7: very much like the song</td>
</tr>
<tr>
<td></td>
<td>Preference of the mood of the seed song</td>
<td>User self-reported extent to which he/she likes the mood of the seed song</td>
<td>7-point Likert scale; 1: very much dislike; 7: very much like</td>
</tr>
</tbody>
</table>

3.4 Experiment Procedure

The experiment was conducted in a batch manner, with six to nine participants in each batch performing the tasks at the same time. Participants filled a pre-experiment questionnaire at the
beginning and a post-experiment questionnaire at the end. The pre-experiment questionnaire gathered demographic information, computer and search expertise, as well as music background information. The post-experiment questionnaire asked about participants’ general experience regarding the search process.

Before conducting the tasks, there was a 10 minute training session on how to use the Moodydb system. Participants were given a chance to practice with the system and raise questions. Each participant was assigned all six topics presented in Table 2. The order of the topics and complexity levels were randomized using a Graeco-Latin square design such that 1) one third of the participants started with topics in level 1, 2, and 3 respectively; 2) the two topics in each level were arranged one after the other so as to reduce the cognitive load of switching back and forth between topics in different levels; and, 3) the two topics in each level alternated their positions across participants (e.g., half of the participants searched for Topic 1 before Topic 2, the other half searched Topic 2 before Topic 1). Half of the participants used each of the two presentation modes. After searching for each topic, participants completed a post-search questionnaire with questions on user perceptions. Topic switch was facilitated by an online system where participants submitted answers. After finishing all six topics, a group interview was conducted with the participants in each batch, to solicit detailed opinions on the tasks and searching processes. The entire procedure lasted about 1.5 hours and each participant was in control of the time spent on each topic. Participants signed consent forms before the experiment and were paid a nominal fee after the experiment.

IV. Results and Discussions

4.1 Participants

The participants were recruited through advertisements in universities in Hong Kong as well as Facebook accounts of the researchers’ social networks. To avoid possible bias introduced by different cultural backgrounds, it was a requirement that the participants be born and raised in Hong Kong and could speak and understand Cantonese. There were 51 participants (28 female) who joined the experiments in eight batches. Their average age was 20.9 years old, and their majors ranged from Social Sciences (13), Science (13), Engineering (7), Business (7), Arts (7) to Medicine (4). Statistics of participants’ background on music knowledge, music listening, searching, computer and English skills are shown in Table 4. Self-reported English abilities were collected as most of the songs had English lyrics.

<table>
<thead>
<tr>
<th>Background</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Music knowledge(^1)</td>
<td>3.14</td>
<td>0.75</td>
</tr>
<tr>
<td>Frequency of music search(^2)</td>
<td>3.65</td>
<td>1.06</td>
</tr>
<tr>
<td>Frequency of music listening(^2)</td>
<td>4.65</td>
<td>0.69</td>
</tr>
<tr>
<td>Expertise with computers(^1)</td>
<td>4.47</td>
<td>0.86</td>
</tr>
</tbody>
</table>
4.2 RQ1: Task Complexity vs. User Perceived Task Difficulty

As both measures are in ordinal scale, Spearman’s rank correlation coefficient was calculated and showed significant correlations between task complexity and perceived task difficulty: $\rho = 0.442$ ($p < 0.001$). The value indicates a moderate correlation (Corder & Foreman, 2009) and is comparable to that in IR studies (e.g., Kelly et al. (2015) reported a correlation of 0.413). Kruskal-Wallis tests were applied to see if there were any differences in users' perceived task difficulty across tasks at different complexity levels. The results showed that perceived task difficulty was significantly different across task complexity levels ($p < .001$). The follow-up pair-wise comparisons indicated that users’ perceived tasks difficulty was significantly lower for tasks at complexity Level 1 (median = 2) than that for tasks at complexity Level 2 and Level 3 (median = 3; $p < .001$). This result verified that task complexity designed based on the cognitive complexity framework, did match users' perceived difficulty to some extent; tasks with low complexity were perceived as being easier than other tasks. As reflected by a participant in the focus group interview:

“The more difficult ones are the ones with the more detailed instructions... and then you have to know their mood, beat and rhythm and stuff.” (Participant 8 commenting on Level 3 task)

However, there was no significant difference in perceived task difficulty between tasks in complexity Level 2 and Level 3 ($p = .483$). By task design described in Section 3.2, the Level 3 task involved comparing multiple songs in multiple aspects, whereas the Level 2 task focused on one aspect (i.e., voice gender) of the songs without considering the interrelationships of the songs. Therefore, according to the cognitive load framework, the two tasks are indeed at different complexity levels. The result that they were perceived as being of similar difficulty seems to indicate there are other factors affecting user perceived task difficulty. Similarly, in IR studies, it has also been found that users did not always regard more cognitively complex tasks as more difficult (Kelly et al. 2015). Other factors such as engagement, enjoyment and user expectations were mentioned as potentially related to task difficulty (Kelly et al. 2015), so were user characteristics such as domain knowledge as well as system factors such as coverage of corpus (Wildemuth et al., 2014; Liu, 2015). We will further analyze these factors in the following subsections.

4.3 RQ2: User and System Factors Influencing Task Difficulty

User Background

Participants’ backgrounds were examined with regard to their possible relationship with perceived task difficulty, including demographic information (e.g., gender, age), self-reported music knowledge, computer knowledge, and online search ability, etc. As each user conducted six
searches, the perceived difficulties of each of these searches were averaged. For binary variables, gender and capability of playing an instrument, Mann-Whitney U tests were conducted and the results showed no significant difference in perceived task difficulty between the two groups ($p = .917$ for gender, $p = .169$ for capability of playing an instrument). Participants’ majors of study were also compared. The results indicated insignificant differences, either with four major groups (Arts, Science, Social Sciences, Business, $p = .135$, Kruskal-Wallis test), or two major group (Hard Sciences, Others, $p = .233$, Mann-Whitney test).

For ordinal (e.g., frequency of music search) and numerical (i.e. age) variables, Spearman’s rank correlation coefficients were calculated between task difficulty and user self-reported background in various aspects. The results (Table 5) show that perceived task difficulty was moderately correlated ($r = .334$) with music knowledge and weakly correlated with users’ age ($r = -.280$) (Corder & Foreman, 2009). Interestingly, the correlation to music knowledge is positive, meaning that the more knowledgeable a user felt they were in music, the greater the difficulty she/he tended to report. Perhaps these users had more considerations based on their richer music knowledge during searches. Users’ age was negatively correlated with average perceive difficulty, indicating that younger users perceived more difficulty than older users. Given the fact that the correlation was rather weak, further research on the possible age effect on perceived task difficulty in MIR is warranted.

<table>
<thead>
<tr>
<th>Users’ background</th>
<th>$r_s$</th>
<th>$p$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Music knowledge</td>
<td>.334*</td>
<td>.017</td>
</tr>
<tr>
<td>Age</td>
<td>-.280*</td>
<td>.049</td>
</tr>
<tr>
<td>Frequency of music search</td>
<td>-.011</td>
<td>.940</td>
</tr>
<tr>
<td>Frequency of music listening</td>
<td>.056</td>
<td>.695</td>
</tr>
<tr>
<td>Computer knowledge</td>
<td>.094</td>
<td>.510</td>
</tr>
<tr>
<td>Online search ability</td>
<td>.071</td>
<td>.621</td>
</tr>
<tr>
<td>English reading ability</td>
<td>.063</td>
<td>.659</td>
</tr>
<tr>
<td>English listening ability</td>
<td>.027</td>
<td>.853</td>
</tr>
</tbody>
</table>

N= 51    *: significant at $p < .05$ level

Spearman’s correlations were also calculated between perceived task difficulty and user preference and familiarity. There was no significant correlation at the $p = .05$ level. In other words, song preference, mood preference, or familiarity to the seed songs/artists did not affect task difficulty perceived by users.

**System factors**
As there were two visualization modes in the Moodydb system, Mann-Whitney tests were used to compare user perceived task difficulty when using the two visualization modes. The result indicate no significant difference ($p = .183$). Other system related factors were identified by analyzing the focus group discussions where users were asked how difficult or easy they felt about the given search topics and why. Issues related to system affordances were mentioned quite often. The first quote below demonstrates a complaint that the system did not show information required by the task. The second was about the inaccuracy of mood labels.

“There are some tasks required us to find a song that is a female vocal but the original song is male vocal, I think it is difficult to find a specific vocal because the system cannot provide such details and I have to click play, play and play to look for the required song.” (Participant 3 commenting on Level 2 task)

“...but there were only 5 mood labels, then I found that a lot of songs had aggressive, bittersweet mood labels, but the songs may not belong to the aggressive category,...” (Participant 2 commenting on Level 3 task)

The size of the collection was also mentioned and some participants thought that the tasks might have been easier if the music collection was larger:

“I found ‘Cold Play, Fix U’ the hardest, find a female vocal with a similar mood, because it is hard to find and I had to click on every page, the results were limited, so I found it harder.” (Participant 30 commenting on Level 2 task)

In sum, task difficulty may have been affected by the adequacy and accuracy of information provided by Moodydb as well as the size of the music collection. Future studies are needed to quantify these possible factors.

Task-related Factors

It is interesting to note that tasks in complexity Level 2 (“understand”) and Level 3 (“analyze”) were perceived as having similar difficulty. In the focus groups, both tasks were referred to as the most difficult a roughly equal number of times (16 and 13 times for the Level 2 and 3 tasks respectively). For the Level 2 task, besides the lack of voice gender information (either textual or visual) provided by the system, participants also referred to the fact that distinguishing voice gender itself may not always be straightforward (Weninger, Wöllmer & Schuller, 2011).

“When I am just looking at the album, I am seeing where there is a female or a guy on it, then I found some of the albums does not have people on it, just like, images, so I think that’s the more difficult task.” (Participant 26 commenting on Level 2 task)

“I cannot identify through the name of the performer, I need to listen, sometimes I cannot verify the voice of the western (singers), because some female also have a lower voice, I’m not sure.” (Participant 28 commenting on Level 2 task)
Having no straightforward answer was recognized as one of the 19 reasons for task difficulty by Liu (2015). There is no exception in MIR. When no obvious or exact answers presented, users would have to listen more carefully to make judgments that comply with the task requirement.

Debriefs on the Level 3 task were slightly different. Besides the inherent cognitive complexity of involving multiple aspects of music, the uncertainty of having to rely on ones’ preference was mentioned by some participants.

“It is the hardest to follow own preference, as I did not know what to choose.” (Participant 31 commented on Level 3 task)

It is noteworthy that uncertainty is by no means a cause of difficulty for all users. For some users, uncertainty can implicate fun instead of difficulty or frustration (see below). Liu's framework (2015) also included uncertainty with tasks and classified it into the aspect of user-task interaction, indicating its user dependency.

Enjoyableness was also mentioned by participants. It was noted that a task was less difficult when they enjoyed it, even if the task was complex by design.

“The playlist task was easier ... I could listen to the songs. I kind of like enjoy it.” (Participant 10 commenting on Level 3 task)

This observation is in accordance to the hedonic nature identified in music information seeking (Laplante & Downie, 2011; Hu & Liu, 2010; Hu et al., 2015). In this study, while a number of participants commented that the Level 3 task was enjoyable, none of them considered the Level 2 task enjoyable. Some of them even reflected that the Level 2 task was the “least enjoyable”. The unenjoyable nature of the task might have made it appear more difficult even though its complexity level was medium by design. For further verification, we calculated Spearman’s correlation between task difficulty, user effectiveness and satisfaction on tasks in different complexity levels respectively. The significant results are shown in Table 6. For all tasks, more clicks and plays were involved when task difficulty increased. Task completion time had a moderate positive correlation with task difficulty for the Level 2 task, but no such correlation for the Level 3 task. In other words, users who took a longer time in the Level 2 task would feel the task was more difficult, but when they spent more time on the Level 3 task, they did not necessarily find it difficult. These observations suggest a new aspect of the metrics of "task completion time" in music retrieval. Similar to text IR, there could be a metric for negative user effectiveness and higher perceived task difficulty, but not for tasks that are enjoyable like the one in Level 3. When the task is enjoyable, task completion time may not necessarily indicate user effectiveness or be related to perceived difficulty.

Table 6. Significant correlations between task difficulty, user effectiveness and satisfaction in each complexity level

<table>
<thead>
<tr>
<th>Complexity</th>
<th>Metrics</th>
<th>rs</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>Number of clicks</td>
<td>.339**</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>Number of plays</td>
<td>.321**</td>
<td>.001</td>
</tr>
<tr>
<td></td>
<td>Task complete time</td>
<td>.246*</td>
<td>.013</td>
</tr>
</tbody>
</table>
It is noteworthy that satisfaction is negatively correlated with task difficulty for the Level 3 task only. When users felt more satisfied with their answers to the Level 3 task, they would report that the task was easier. However, for the Level 2 task, user satisfaction was not related to task difficulty. In other words, even if they were satisfied with the answers, they might have still felt the task was difficult. We conjecture this difference could be possibly attributed to the fact that the task was regarded as not enjoyable.

4.4 RQ3: Task Complexity, Perceived Task Difficulty, User Effectiveness and Satisfaction

Spearman’s rank correlation coefficient was calculated to investigate possible relationships between task complexity / difficulty and user effectiveness / satisfaction measures. Significant correlation coefficients are shown in Table 7, ranked by strength of correlation.

<table>
<thead>
<tr>
<th>Metric 1</th>
<th>Metric 2</th>
<th>$r$,</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task complexity</td>
<td>No. of plays</td>
<td>.585**</td>
<td>0.000</td>
</tr>
<tr>
<td>Task complexity</td>
<td>No. of clicks</td>
<td>.560**</td>
<td>0.000</td>
</tr>
<tr>
<td>Perceived task difficulty</td>
<td>No. of plays</td>
<td>.458**</td>
<td>0.000</td>
</tr>
<tr>
<td>Perceived task difficulty</td>
<td>No. of clicks</td>
<td>.448**</td>
<td>0.000</td>
</tr>
<tr>
<td>Task complexity</td>
<td>Task completion time</td>
<td>.412**</td>
<td>0.000</td>
</tr>
<tr>
<td>Perceived task difficulty</td>
<td>Task completion time</td>
<td>.389**</td>
<td>0.000</td>
</tr>
</tbody>
</table>

N= 306 *: significant at p < .05 level **: significant at p < .01 level

The coefficients in Table 7 show moderate ($|0.3| \leq r \leq |0.5|$) or strong ($|0.5| < r$) correlations (Corder & Foreman, 2009). Both task complexity and perceived task difficulty are strongly or moderately correlated with the number of clicks, number of plays and task completion time. It is noteworthy that the correlations with task complexity were stronger than those with perceived task difficulty. On the other hand, neither complexity nor perceived difficulty correlated with the number of songs retrieved or satisfaction. To further understand the relationship, Figures 3 and 4 show the box plots of user effectiveness and satisfaction measures across task complexity and perceived task difficulty respectively.
One-way ANOVA was applied to analyze whether user effectiveness metrics had any significant differences across tasks in different task complexity levels. The results showed significant differences on all these effectiveness metrics. For task completion time ($F = 31.11, p < .001$) there were significant differences between Level 1 and Level 3 ($p < .001$) as well as between Level 2 and Level 3 ($p < .001$). On average, users took significantly longer to complete tasks in Level 3 than Level 1 (180.15 more seconds) and Level 2 (140.03 more seconds). Although the difference were not significant between Level 1 and Level 2 ($p = .217$), on average Level 2 tasks took longer (40.57 more seconds) than those in Level 1. The results show that the designed task complexity level did have an effect on the users’ task completion time. The more complex the tasks were by design, the longer they took the users to complete.

For the number of songs found ($F = 15.88, p < .001$) although the tasks only requested for three songs, in many cases users wrote down more than three songs. Follow up pair comparisons indicated significant differences between Level 1 and Level 2 (difference of means was 0.41, $p < .001$), and between Level 3 and Level 2 (difference of means was 0.39, $p < .001$). Interestingly, tasks in Level 1 and Level 3 had similar numbers of songs found (mean = 3.41 and 3.39, $p = .969$), while the task in Level 2 got fewer answers on average (mean = 3.00). While it was expected that
more songs could be found for the least complex task, it was surprising that the Level 3 task (the most complex by design) did not get fewer answers.

For number of clicks ($F = 45.05, p < .001$) and number of plays ($F = 45.60, p < .001$), post hoc comparisons revealed significant differences between all pairs. It can be seen from Figure 3 (c) and (d), values of these two metrics increased with task complexity levels. In other words, more complex tasks did involve more user clicks and song plays.

A Kruskal-Wallis test was conducted to compare user satisfaction across task complexity levels. The result was insignificant ($p = .922$). Users were generally quite satisfied across tasks complexity levels (median = 5 in all levels, Figure 3 (e)).

**Perceived Task Difficulty**

Similar to task complexity, One-way ANOVA was conducted and the results indicate that all user effectiveness metrics were significantly different across perceived task difficulty ratings. Post hoc test on task completion time ($F = 7.56, p < .001$) disclosed that tasks in perceived task difficulty rating 1, i.e. the lowest difficulty, took significantly less time than those in perceived difficulty ratings 3 and above. Similarly, tasks in rating 2 took less time than those in ratings 5 and above. There were no other significant pairs.

Post hoc test on the number of songs ($F = 4.21, p < .001$) found that tasks in rating 5 got significantly more songs than those in rating 3. Also, rating 7 got significantly fewer songs than ratings 1, 2, 3, 4, and 6. These results show that the number of songs found did not have consistent patterns on the effects of perceived task difficulty (Figure 4 (b)), except for the fact that users found fewer songs for topics perceived as in rating 7 (the most difficult).
Figure 4. Boxplots of user effectiveness and satisfaction across different perceived task difficulty ratings

For number of clicks ($F = 10.90, p < .001$), there was a general trend that users had more clicks in topics with higher difficulty ratings, except for those in rating 7 (Figure 4 (c)). For number of plays ($F = 10.26, p < .001$), users played fewer songs for topics in ratings 1 and 2 than those in ratings 5, 6 and 7. The Kruskal-Wallis test on user perceived satisfaction shows significant difference across task difficulty ratings ($p = 0.03$), but post-hoc comparison did not find any pairs with significant difference at $p = 0.05$ level. This means the power of the test is low, which is likely due to the small sample size ($n$ ranged from 6 (rating 7) to 69 (rating 2)). However, from the boxplots (Figure 4 (e)), we can see that median satisfaction scores on higher difficulty ratings (5, 6, 7) were lower than those on lower difficulty ratings (1, 3, 4).

In general, these findings were in accordance with those in the text retrieval domain. First, the increases in user behaviors (i.e., number of clicks, number of plays) were related to higher task complexity levels and task difficulty ratings (Table 7). In text retrieval, it has been generally understood that more user search behaviors are related to the user’s perceived difficulty and/or the fact that users are struggling in the search (White & Dumais, 2009). Second, task completion time was related to task difficulty (Figure 4 (a)). In general, tasks perceived as easy (ratings 1 and 2) took less time than tasks perceived as difficult (levels 5, 6, 7). Liu et al. (2012) suggested that user search behaviors could be used to detect whether users encountered difficult tasks. As an attempt
to verify the extent to which such premise holds in MIR, we ran a linear regression analysis on predicting perceived task difficulty from the other measures (Table 8). All user effectiveness measures, except for number of plays, were statistically significant in predicting perceived task difficulty ratings.

Table 8. Regression analysis on task difficulty ratings

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Coefficient (B)</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2.535</td>
<td>4.768</td>
<td>.000**</td>
</tr>
<tr>
<td>Task complexity</td>
<td>.504</td>
<td>4.380</td>
<td>.000**</td>
</tr>
<tr>
<td>Task completion time</td>
<td>.001</td>
<td>2.637</td>
<td>.009**</td>
</tr>
<tr>
<td>Number of songs found</td>
<td>-.296</td>
<td>-2.217</td>
<td>.027*</td>
</tr>
<tr>
<td>Number of clicks</td>
<td>.017</td>
<td>3.257</td>
<td>.001**</td>
</tr>
<tr>
<td>Number of plays</td>
<td>.004</td>
<td>.240</td>
<td>.811</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>-.124</td>
<td>-2.105</td>
<td>.036*</td>
</tr>
</tbody>
</table>

N= 306*: significant at p < .05 level  **: significant at p < .01 level

To further examine the prediction power of the metrics, we conducted a classification experiment with similar setup in the study by Liu et al. (2012). Specifically, search sessions with task difficulty ratings 1 to 3 were combined into an “easy” class (202 sessions), whereas those with task difficulty ratings 5 to 7 were combined into a “difficult” class (69 sessions). A binary classification model was built using logistic regression to predict whether a session was easy or difficult based on task complexity and user effectiveness measures. The classification performance was evaluated with a five-fold cross-validation where 80% of the data were used as training and 20% as testing. As the largest class in this dataset was “easy”, we compared the performance of the predicted “easy” class to the baseline results of the majority vote (i.e., predicting every session as “easy”). As can be seen in Table 9, the prediction model outperformed the baseline and the F(0.5) measure was comparable to that of the model built in Liu et al. (2012) (0.87 for the model of whole-session level variables). Significant predictor variables included task complexity (B = -.567, p = .014), task completion time (B = -.002, p = .040), number of songs found (B = .740, p = .007), number of clicks (B = -.020, p = .032) and satisfaction (B = .227, p = .046). In other words, these measures contributed to predicting whether a task was perceived as easy or not.

Table 9. Binary classification accuracy, precision and F(0.5)

<table>
<thead>
<tr>
<th>Model</th>
<th>Overall accuracy</th>
<th>Precision</th>
<th>F(0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (majority vote)</td>
<td>.745</td>
<td>.745</td>
<td>.785</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>.775</td>
<td>.831</td>
<td>.840</td>
</tr>
</tbody>
</table>

VI. Conclusions and Future Study

Through a user experiment, this study revealed that task difficulty in music search was affected by factors of task, user and system. The task in the lowest complexity level ("remember") was perceived as being easier than those in higher levels of complexity. Users’ self-reported music
knowledge was positively correlated with tasks difficulty whereas users’ age had weak negative correlation. Although the visualization mode of the Moodydb system did not affect perceived difficulty, adequacy and accuracy of system provided information, as well as size of the music collection were mentioned as reasons that made a task difficult. Furthermore, related to task-user interaction, no straightforward answer, uncertainty and enjoyability of the tasks, may have also affected task difficulty. In particular, enjoyability of tasks seemed to be a mediating factor between task completion time and perceived difficulty (Table 6). If a task is enjoyable (i.e. the task in complexity Level 3), it may not be difficult even if users spent a significant amount of time on it. These findings have implications on designing tasks for MIR research. Besides the cognitive complexity framework, the uncertainty and enjoyability of the tasks, user backgrounds and system affordances also need to be considered in designing and/or controlling task difficulty.

This study also revealed several user effectiveness metrics that were correlated with task difficulty: number of clicks, number of plays, task completion time (Table 7). In addition, these metrics, together with task complexity, can predict to some extent whether a task was difficult (Tables 8 and 9). This is one step further towards MIR systems that can detect user perceived difficulty based on user behavior metrics. Such systems can then further provide further user support when users encounter difficulties.

Future studies should include quantifying more system factors that may affect task difficulty in music search, such as the amount of information provided and collection size. Users’ cultural background may also be relevant as music which is culturally unfamiliar to users might be more difficult to seek. Moreover, real-time prediction of task difficulty based on user behavior metrics could be explored, toward detecting difficult tasks during music search.

Footnote:

Acknowledgment:
This study is partially supported by the Joint Research Grant No.2013-3-3 of the National Institute of Informatics in Japan, JSPS Grants-in-Aid for Scientific Research No.22300048, and the Seed Fund for Basic Research scheme in the University of Hong Kong. We thank the anonymous reviewers for their helpful comments.

References:
This is a pre-print version of the following article: Hu, X. & Kando, N. (In press). Task Complexity and Difficulty in Music Information Retrieval. Journal of the Association for Information Science and Technology.


This is a pre-print version of the following article: Hu, X. & Kando, N. (In press). Task Complexity and Difficulty in Music Information Retrieval. *Journal of the Association for Information Science and Technology*.


Levitin, D. J. (2011). *This is your brain on music: Understanding a human obsession*. Atlantic Books Ltd.


This is a pre-print version of the following article: Hu, X. & Kando, N. (In press). Task Complexity and Difficulty in Music Information Retrieval. *Journal of the Association for Information Science and Technology*.


