HOW SIMILAR IS TOO SIMILAR?: EXPLORING USERS' PERCEPTIONS OF SIMILARITY IN PLAYLIST EVALUATION

Jin Ha Lee
University of Washington
jinhalee@uw.edu

ABSTRACT

The Audio Music Similarity and Retrieval (AMS) task in the annual Music Information Retrieval eXchange relies on human-evaluation. One limitation of the current design of AMS is that evaluators are provided with scarce contextual information as to why they are evaluating the similarity of the songs and how this information will be used. This study explores the potential use of AMS results for generating playlists based on similarity. We asked participants to listen to a subset of results from the 2010 AMS task and evaluate the set of candidates generated by the algorithms as a playlist generated from a seed song (the query). We found that while similarity does affect how people feel about the candidate set as a playlist, other factors such as variety, metadata, personal preference, familiarity, mix of familiar and new music, etc. also strongly affect users' perceptions of playlist quality as well. We discuss six user behaviors in detail and the implications for the AMS evaluation task.

1. INTRODUCTION

Audio Music Similarity and Retrieval (AMS) is one of the evaluation tasks conducted in Music Information Retrieval Evaluation eXchange (MIREX). AMS task relies on human evaluation for ground truth. Evaluators are asked to listen to a set of query-candidate pairs and indicate how similar they think the songs are on a broad scale (i.e., very similar, somewhat similar, not similar) as well as a fine scale (i.e., a score between 0-100). In 2010, the number of test queries was 120 and each participating algorithm returned 5 results per query [7]. Based on the human evaluation, average precision scores are calculated for each algorithm and the ranking of algorithms is determined.

One limitation with the design of the AMS task is that evaluators are rating only the similarity between each of the query-candidate pairs, not the candidate set as a whole. Moreover, the evaluators are not given any background information on a use scenario; why they are evaluating the similarity of the songs and how those data will be used.

The objective of this study is to explore one of the potential uses of the AMS evaluation task. One way of using music similarity data is to generate playlists or recommendations for users. How would users respond to the AMS results if they were presented as playlists generated for users to listen to? From the previous studies on playlists, we already know that users value both variety and coherence in their playlists [4, 11], in other words, they want playlists with songs that are similar to each other, but not too similar. How does this Goldilocks-style similarity translate to AMS similarity metrics? When you compare the fine score given for AMS results and the users' evaluation of the results as playlists, how similar or different are they? Also can we learn anything new about what users expect from playlists in addition to what we already know? This paper presents the findings from eight interviews conducted in order to answer these questions.

2. DESIGN OF THE STUDY

2.1 Test Collection

We conducted in-depth interviews asking participants to listen to a subset of the results of the MIREX 2010 AMS task and evaluate the candidate set as playlists. 7 queries, each from different genres (i.e., blues, classical, country, electronica, hip-hop, jazz, rock) were selected as test queries. Of these 7 queries, each participant was asked to choose at least 3 queries to evaluate. This was to ensure that participants have some freedom to choose the genre that they are familiar with and most likely to listen to in real life. Table 1 shows the list of queries tested in our study.

A total of 8 algorithms participated in 2010 AMS task, however, in order to reduce user fatigue we tested candidate sets of only 3 algorithms for each query. The candidate sets for blues, rock, and hip-hop were selected based on their average fine scores such that they would have similar scores within genre and represent a spectrum of scores between genres. The classical, electronica, jazz, and country candidate sets were chosen to represent a variety of average fine scores for the same query [See Table 2].

Genre	Query Title	Query Artist
Blues	Somebody's Been Talkin'	Blind Boy Fuller
Classical	Concerto No. 1 in C major Part 2	Monica Huggett
Country	Sylvia's Mother	Bobby Bare
Electronica	Q-Works	Q-Factor
Hip-Hop	Paper Chase	The Mountain Brothers
Jazz	Time's Lie	Gary Meek
Rock	Spank Thru	Nirvana

Table 1. List of test queries

Genre	No. of users	Algorithm	Fine Score	Average User Rating	Standard Devia-tion
Blues	3	BWL1	100	3.67	0.47
		PSS1	100	4.50	0.40
		SSPK2	100	4.17	0.85
Rock	6	SSPK2	85.8	3.50	0.96
		PS1	83	3.83	0.62
		TLN1	81.6	2.75	1.31
Hip-Hop	3	PSS1	69.2	2.67	1.03
		SSPK2	66.6	3.77	0.56
		TLN1	66.6	3.67	0.24
Classical	3	SSPK2	84	3.33	0.47
		PS1	78	3.50	1.78
		BWL1	66	3.00	0.41
Electroni-ca	3	PSS1	81.8	3.50	0.71
		PS1	63.6	3.67	1.25
		TLN1	41	2.33	1.70
Jazz	4	TLN1	73	3.50	0.79
		SSPK2	57	2.70	1.63
		PSS1	41	2.38	1.29
Country	3	SSPK2	77.2	3.83	0.24
		PS1	64.8	3.67	1.25
		BWL1	50	3.67	0.47

Table 2. List of algorithms, the number of participants, similarity scores, and average ratings from participants

2.2 Task Design

Each candidate set was presented to participants as a playlist consisting of 5 songs. The participants listened to the 30 second clips of these songs, multiple times if desired. We used the 30 second clips rather than the whole songs to be consistent with the AMS task and evaluation. The participants were asked to imagine that these playlists were generated by 3 different systems that used the query as the seed song. After listening to the 3 candidate sets per query, they were asked to rate each playlist on a 5 point scale and also rank them. We asked the participants the reasons for liking or disliking the playlists, and also to imagine an ideal playlist and what kinds of characteristics that playlist would have or not have. The interview data were analyzed using a grounded theory approach which allows us to generate a theory from empirical data [6].

2.3 Participants

Participants were recruited by using a snowball method starting with the colleagues of the lead researcher who are interested in music. They were selected so that they reflect some variance in their preferred music genre and style. 3 of the participants were in 20s, 2 were in 30s, and 3 were in 40s. 6 participants were male and 2 were female. Most participants listen to music at least occasionally, although the degree of their interests did vary. In the discussion below, responses from different participants are identified by their assigned number (i.e., P1–P8).

3. DATA AND DISCUSSION

3.1 Overview

Table 2 shows the list of algorithms that were tested for each query. The 4th column shows the average fine scores assigned to the candidate lists by the human evaluators in AMS task [7]. The 5th column shows the average rating from our participants.

3.2 User Behaviors

In the following, we provide a list of patterns that emerged from the user behaviors we observed.

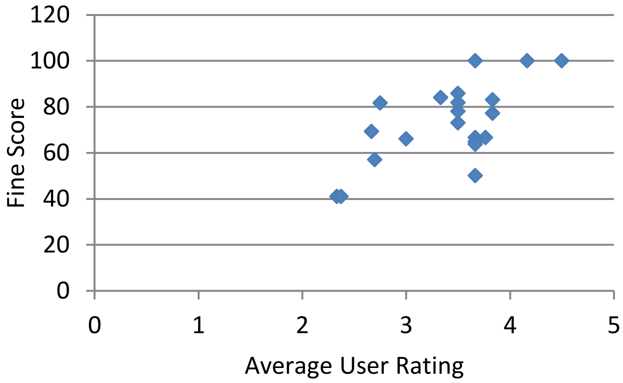

Figure 1. Fine scores and average user ratings

3.2.1 Similarity does matter, but variety is also important

When we compared the fine scores with the average ratings from our participants for the 21 algorithms, we did observe some correlation (*Pearson's r* = 0.66). Figure 1 shows the scatterplot of the fine scores and the average user ratings. There seems to be a stronger correlation between the fine scores and the user ratings for the playlists with either very high similarity score or very low similarity score. The playlists with medium similarity scores (around 60-80), on the other hand, do show greater variation in user rating. This suggests that similarity does have some effect on how people feel about the playlists, but other factors are having an impact as well.

One common theme that emerged from all the responses was the importance of variety as all eight participants said that they want some variety in the playlist. In fact, a number of participants (P1, P2, P4, P5, P6) reacted positively to the playlists that had coherent set of songs, but also said some lists were too similar and needed more variation (P1, P3, P4, P6, P7, P8).

P8: *It's kind of* ***monotonous****.* ***The songs all had pretty much the same tempo****…it was* ***all on 10 all the time, and there's no variety****. And again* ***all the tempos were the same, they all had big drums, and consistently all big sound****.*

From this quote, we can see that for P8, the same tempo, instrumentation, and style were the reasons why he felt that the songs on the playlist were *too* similar. This participant also said it was okay to have multiple songs from the same artist, although other participants had different opinions.

P8: (Songs from the same artists are) ***usually okay****, because* ***if I like the band I want to hear more of it****, and if I don't like them I already changed the channel.*

P1: *There's no lyrics in these so they kind of fit, but* ***three songs by the same artist*** *on the same playlist is kind of, out of five,* ***is a little bit, I think, extreme****.*

As you can see, variety and coherence meant different things for different participants. To name a few examples, P7 preferred variation across genres, P1 wanted various artists, and P4 liked diverse musical style. Participants also focused on different aspects when they said the playlist was coherent. For example, P5 focused on lyrical content, whereas P2 focused on tempo, P1 and P3 focused on mood, etc. When users perceive the variety or coherence of a playlist, they will react differently to different features, thus the transparency of the system seems important as discussed in [2, 13]. P4 expressed frustration that many of the current systems provide playlists or recommendations without telling the user how exactly the songs were selected.

P4*: There should be* ***why did you get this song*** *maybe because I don't understand most of the times, like* ***I would put in the seed song and get these and I'm like, what is going on here?***

In fact, in Åman and Liikkanen's survey of music recommendation systems [1], most of the systems received very low scores for transparency. Some music recommender systems do provide at least limited information about the songs to users. For instance, Pandora provides information about the features of the selected songs that are taken into account when they generate their playlists (e.g., it features pop/rock qualities, a subtle use of paired vocal harmony, mixed acoustic and electric instrumentation). The users, however, are not able to specify which of these features they want the system to focus on in generating playlists. For instance, some user may use *Everlong* by Foo Fighters as his seed song because he wants songs with similar instrumentation whereas another user may use the same seed song wanting songs with romantic lyrics.

3.2.2 Metadata affects how users feel about the music

In addition to relying on the musical features, various types of metadata can be used to further improve the selection of songs in a playlist. In previous works, use of song/album title, artist, genre, user rating, etc. have been discussed [10, 11, 12, 14]. In addition to these, we believe metadata such as lyrics can play a significant role. P6, for instance, gave the highest rating for one of the blues list because of the similar lyrical content. P2, P4 and P5 said they also create their own lists based on lyrics. Moreover, P5 and P7 said they want to be able to filter songs that have graphic lyrics.

Other potentially useful metadata mentioned by our participants was the theme. P4, P5, and P8 said that they create their own lists based on a theme/story (e.g., trains, 4th of July). The theme of the song may be automatically extracted, inferred from the title or lyrics, or assigned by users though social tagging. Time period was also important for our participants. P1, P2, and P3 mentioned that the songs in the given playlist were from different era and that negatively affected how they felt about it (e.g., mix of different classical music periods, rock music from 80s and 90s). P2 said that the song from different time periods "disrupted the flow" of the playlist. P1 also mentioned the difference in the quality of sound recordings from different eras.

P3: *The slightly jarring thing is…that first* ***seed sounded to me more like sort of baroque music so like in the 16th or 17th century*** *and then from what I remember of* ***the ones in the playlist, they were sort of like 18th century and some 19th century*** *stuff at the end…****not the same time period****.*

P1: *A lot of this older country music that was recorded before modern recording equipment, has* ***this kind of echoy, tin canny sound*** *to it, right? And* ***I don't like that.*** *That* ***really muffled poor recording****…****I just can't get over that.***

3.2.3 Having a song users love or hate can significantly affect how they feel about the whole playlist

Personal preference of music highly affected how people felt about the playlist as a whole. Participants seemed delighted when they heard the song or artist they liked. Some specifically stated that they were rating the playlist higher because they really liked one of the songs.

P8: *Actually* ***the other one probably should have been 2*** *(which he rated 3),* ***but I really love that Foo Fighter song*** *so… How can you dislike that melody line? La la la la la…*

P1: *Learn to Fly was very poppy so* ***it didn't quite fit but*** *you know,* ***I like that song so that doesn't really bother me too much****. And* ***I like Soundgarden so that's good****.*

The overall preference of the genre or musical style also mattered. For instance, P3 said "*I just prefer more upbeat songs in general*", or some participants preferred a particular sub-genre (e.g., P1 likes country rock, P4 likes swing jazz, P6 likes alt rock) and seeing songs that fit those criteria on the playlist made them respond more positively. Sometimes user's preference overrode the similarity.

P4: ***I really like this playlist cause I like the music on it*** *because I'm more,* ***I'm way more into the swing side of jazz*** *but* ***I don't really think any of these songs go with the seed song very much.***

While explaining the characteristics of their ideal playlist, a number of participants (P1, P4, P8) specifically said that they want to see more songs by the seed artist. Currently in AMS task, the seed artist is filtered from the results.

P4: *I* ***never seem to get in your playlist the same artist as your seed song, what is up with that? Obviously I like this artist, why do you not intersperse more of that artist?***

Participants also had strong reactions to the music that they hated as discussed in [3]. Having even a single song that they dislike significantly affected how they rated the playlist as a whole. For instance, P3 and P4 who listened to the jazz set commented that one of the songs sounded like elevator music which made them absolutely hate the list.

P3: *Oh, god, elevator music.* ***I loathe elevator music.***

P4: *This is* ***like elevator music****. It's too early in the morning for this...and* ***I'm offended that like elevator music is associated with Jazz.*** *There is kind of a bit of* ***an offensive thing*** *going on there. You put in Jazz and you get like,* ***is that what people think?*** *This gets a* ***zero****.*

This suggests that perhaps providing a way to permanently ban a song, like in systems like Musicovery, is important to users. 4 participants (P3, P4, P7, P8) specifically said that they want to be able to remove songs that they hate so that it will never appear in any of their playlists. P3 said "*I'm not sure what I like, but I know for sure what I DON'T like*." P8 said "*all the songs have to be able to stand by itself*" without "*the second rate songs.*"

P8: *(answering the question "Anything that this ideal playlist should not have?") Van Halen.* ***It shouldn't have music that sucks.***

3.2.4 Users like learning new things, but they still want them contextualized in familiar territory

Several participants stated that they like learning something new and being exposed to new songs by listening to these playlists.

P3: *I kind of want the system that* ***educates me****. You know, that* ***picks things that I don't really know about****.*

P7: *The Van Halen cover of the Who. I had no idea that song existed so that's (good),* ***I love learning things, that's why I go to things like Pandora.***

Finding new songs by a known artist was also a positive thing, like P8 who was happy to learn one of the Foo Fighters songs. P7 said re-discovering songs that were once familiar but forgotten was also a positive experience. Another notable pattern was that all participants wanted a mix of familiar and new songs, although there were disagreements on the ideal proportion of familiar and new songs. Fields [5] also advocates a playlist (familiar songs) with recommendation (new songs), although playlists are typically distinguished from recommendations [4]. Having familiar and new songs together on the list can perhaps help users by enabling them to establish the context, understand the connections between the songs better and remember the new songs better.

P6: *A mix of things is good, cause I would like to* ***discover new artists****, it's always a good way to (be)* ***introduced through somewhat similar artists you already like****…I'll* ***feel more comfortable getting into the genre*** *if* ***there's a few (songs) I kind of knew, and then I could kinda determine my likes from there.***

3.2.5 Users tend to be more generous for unfamiliar music

Participants also reacted differently to the genre based on their familiarity. Better familiarity with the genre seemed to lead to stronger criticisms and disappointment, higher expectation, and more intense reaction. It also led to lower ratings overall, compared to the playlists in a genre that participants were less familiar with. For instance, when evaluating the same electronic playlist, P8 thought they all sounded okay and similar enough to the seed whereas P1 pointed out the mix of different sub-genres and gave lower scores. The participants, in fact, were aware of this behavior themselves.

P8: *I feel like* ***I can learn about the genres that I don't know much about****...so I'm* ***way more likely to just sort of***

go along, go with the flow and see what I learn** whereas usually if I'm listening to something **I'm familiar with…I want the songs to have like consistent values, production values, song writing values,** um, **I'm just way pickier.

P2: *Because I know this genre so well, I feel like **I become pickier**, yeah, **my expectations are higher.***

3.2.6 Users know and will tell you about their boundaries

Overall, participants tended to be very aware and assertive about their boundaries: what they like and dislike, how much variation they can tolerate, and other little quirks. For instance, both P5 and P8 said they did not like songs that were anti-social. P4 said the mix of vocal and instrumental music was intolerable.

P4: *That third choral thing has to go. **That was wrong.** I liked the rest of the playlist but that one just…**I couldn't see why it came up, I really didn't like it**…To me they would be different channels, **different categories**.*

Participants also had different reactions to the randomness of the playlists. P5 said "*I didn't like the song in foreign, Northern European? Language because I couldn't figure out what it was about*" although P3 liked the randomness of the German jazz song.

P3: *That was the **most bizarre combination… it would be kind of fun** to be sort of given this and just be playing it in the car thinking **oh I wonder what's gonna happen next**... this is like a **weird mystery gift**, you know, like the Christmas present from your mad auntie.*

4. IMPLICATIONS ON SIMILARITY EVALUATION

Based on our interview data, we have four recommendations for possibly improving the AMS task in MIREX. The former two recommendations aim to facilitate obtaining more objective results, and the latter two are for making the AMS task more user-centric.

4.1 Specification of Features

Our participants considered a variety of features when they evaluated the playlists. Examples of commented features include mood, genre, lyrical content, tempo, instrumentation, delivery, time period, style, and so on. They assumed that the songs on the playlists were selected because of some combination of these different features, although uncertain as to which exact features were used. This multifaceted notion of music similarity makes it difficult to evaluate similarity since there are so many different ways two music clips can be similar (c.f., [8]). In MIREX, we currently collect evaluators' opinions on how similar the query-candidate pairs are, but not on which aspects they thought were similar. One possible way to remedy this limitation would be to inform the evaluators which aspects they should focus on during the evaluation in order to obtain more objective results. Another measure would be to ask the evaluators to tell us which aspects made them think the results were similar or not. Although this proposed solution may slightly increase users' burden, we will be able to obtain more objective judgments as well as richer information on the relative importance of features for users.

4.2 Identification of Evaluator's Genre Preference and Familiarity

Collecting information about evaluator's preference and familiarity may enable us to gauge how much we can trust the response from each evaluator. As discussed in sections 3.2.2 and 3.2.6, participants did react differently to playlists of different genres based on their preference and familiarity. The background knowledge of the familiar and liked genre allowed the participants to evaluate the playlist based on a lot of contextual information (e.g., P1: "*Van Halen was kind of like **hair metal**"* P6: "*if you are going to input Nirvana, maybe you want some other **smaller**, **pacific Northwest, grungy**, **90s***"). On the other hand, they found it difficult to evaluate the lists if they did not know the genre very well. For instance, all 3 participants (P5, P7, P8) said that they listen to hip-hop but have limited knowledge of the genre which made it difficult to evaluate the playlists.

4.3 Providing Metadata with the Music Clips

In MIREX, human evaluators are not provided with metadata such as artist or title of the music clips they evaluate. Although this will help ensure that the similarity judgment is strictly based on the music itself not metadata, it does not reflect the real-life music experience of users. In any commercial system, the ultimate objective is to deliver music to users who will want to know what exactly they are purchasing. Even for non-commercial systems, metadata will be crucial for educating the users about music. Note that evaluation of music playlist is also affected by the availability of metadata.

P1: *(reacting to three songs from the same artist) I don't know. **If you didn't show me the metadata, I might not know, or I might not have that kind of reaction.***

Also 4 of our participants (P1, P6, P7, P8) discussed the connection of Nirvana to Foo Fighters when they evaluated the rock playlists. P7 said "*Foo Fighters is pretty obvious*" and P1 said "*it sounds different from the seed but it makes sense*," demonstrating the importance of artist information. P7 also said contextual information like "*the influences between bands*" was important in making a good playlist. Providing even the basic metadata such as artist, song title and genre with the music clips can help users better understand the context of music. This is also much closer to how users will respond to music playlists in real life.

4.4 Tasks Reflecting Real-life Use Scenarios

The current music similarity task relies on human evaluators for generating ground truth; however it is unclear how this information is going to be used in real life. Music similarity can be used for many user tasks other than just creating playlists; for instance, Lee [9] discusses the use of music similarity for known-item searches (e.g., trying to find a specific song by providing other song titles that sound very similar) on Google Answers. In this case, we suspect that candidates with higher similarity scores may be more useful for the user task.

We believe it is crucial that the MIR community as a whole think about how the current tasks can be evolved into tasks that are more user-centric, in other words, closer to the user tasks that actually happen in their everyday life. One possibility for evolving the current AMS task is to create different sub-tasks that use music similarity; for instance, playlist generation task, known-item search task, personal music collection management task, and so on.

5. CONCLUSION AND FUTURE WORK

The findings of our study suggest that similarity is only one of the many factors that affect how people feel about playlists. Although similarity does seem to affect the user rating of playlist, stronger similarity does not always make better playlists for users. Overall, participants had fairly clear ideas about what they expect from a good playlist and were able to articulate them. Their evaluation of playlists tended to be quite subjective as they were highly affected by personal preference and familiarity with the music on the list, although some common themes emerged, such as wanting a mix of familiar and new songs, more songs from the seed artist, etc. Many of the user behaviors observed during the interviews confirm and support various points that were raised in previous literature on music similarity, recommendation, playlists, and evaluation. This is promising as there do seem to be a set of features that we can implement in current systems to make them more user-centric.

We hope that findings from this study will provide useful information for redesigning the current AMS task and encourage the MIR community to think about how to evolve the current evaluation tasks. In our future studies, we plan to test more playlists generated by different algorithms submitted to MIREX based on different set of seed songs. Instead of researchers selecting random songs for users to test, we plan to have the users select the seed songs that they actually like and are more likely to use for eliciting playlists in real life.

6. REFERENCES

[1] P. Åman and L. A. Liikkanen: "A survey of music recommendation aids," *Proceedings of the Workshop on Music Recommendation and Discovery*, 2010.

[2] L. Barrington, R. Oda, and G. Lanckriet: "Smarter than genius? human evaluation of music recommender systems," *Proceedings of ISMIR*, pp. 357-362, 2009.

[3] S. J. Cunningham, D. Bainbridge, and A. Falconer: "More of an art than a science: supporting the creation of playlists and mixes," *Proceedings of the ISMIR*, pp. 240-245, 2006.

[4] B. Fields and P. Lamere: "Finding a path through the juke box," *tutorial presented at the ISMIR*, 2010.

[5] B. Fields, C. Rhodes, and M. d'Inverno: "Using song social tags and topic models to describe and compare playlists," *Proceedings of WOMRAD*, 2010.

[6] B. Glaser and A. Strauss: *The discovery of grounded theory: strategies for qualitative research*, Chicago, 1967.

[7] IMIRSEL: *"2010 AMS results,"* http://www. music-ir.org/mirex/wiki/2010:Audio_Music_Similarity_and_Retrieval_Results, 2010.

[8] M. C. Jones, J. S. Downie, A. F. Ehmann: "Human similarity judgments: implications for the design of formal evaluations, *Proceedings of the ISMIR*, pp. 539-542, 2007.

[9] J. H. Lee: "Analysis of user needs and information features in natural language queries seeking music information," *JASIS&T*, 61(5), pp. 1025-1045, 2010.

[10] B. Logan: "Content-based playlist generation: exploratory experiments. *Proceedings of the ISMIR*, pp. 295-296, 2002.

[11] S. Pauws and B. Eggen: "Pats: realization and user evaluation of an automatic playlist generator," *Proceedings of the ISMIR*, pp. 222-230, 2002.

[12] R. Ragno, C. J. C. Burges, and C. Herley: "Inferring similarity between music objects with application to playlist generation," *Proceedings of the 7th ACM SIGMM MIR*, pp. 73-80, 2005

[13] R. Sinha and K. Swearingen: "The role of transparency in recommender systems," *Proceedings of the ACM CHI*, pp. 830-831, 2002.

[14] M. Slaney: "Web-scale multimedia analysis: does content matter?" *IEEE Multimedia*, Vol. 18, No. 2. pp. 12-15, 2011.