TISMIR

Flexer, A., et al. (2021). On Evaluation of Inter- and Intra-Rater Agreement in Music Recommendation. *Transactions of the International Society for Music Information Retrieval,* 4(1), pp. 182–194. DOI: https://doi.org/10.5334/tismir.107

RESEARCH

On Evaluation of Inter- and Intra-Rater Agreement in Music Recommendation

Arthur Flexer, Taric Lallai and Katja Rašl

Our work is concerned with the subjective perception of music similarity in the context of music recommendation. We present two user studies to explore inter- and intra-rater agreement in quantification of general similarity between pieces of recommended music. Contrary to previous efforts, our test participants are of more uniform age and share a comparable musical background to lower variation within the participant group. The first study uses carefully curated song material from five distinct genres while the second uses songs from a single genre only, with almost all songs in both studies previously unknown to test participants. Repeating the listening tests with a two week lag shows that intra-rater agreement is higher than inter-rater agreement for both studies. Agreement for the single genre study is lower since genre of songs seems a major factor in judging similarity between songs. Mood of raters at test-time is found to have an influence on intra-rater agreement. We discuss the impacts of our results on evaluation of music recommenders and question the validity of experiments on general music similarity.

Keywords: music recommendation; music similarity; evaluation; listening test; user study

1. Introduction

The automatic recommendation of music or creation of playlists is one of the successful applications of Music Information Retrieval (MIR) and is now commonplace in music streaming services like Spotify, Deezer, Pandora and Tidal. Usually these services recommend music which is in some way similar to what users have been listening to previously. As a consequence, objective assessment of the quality of such services requires a quantification of similarity between recommended songs mirroring human perception of music similarity. A recent survey on challenges in music recommendation research also emphasized the point that recommender evaluation needs to go beyond mere system measures and include user experience (Schedl et al., 2018). However, previous research (Jones et al., 2007; Ni et al., 2013; Schedl et al., 2013; Flexer and Grill, 2016; Koops et al., 2019) made it clear that human perception and experience of music similarity is highly subjective with low inter-rater agreement. This is particularly true for perception of *general* music similarity, i.e. when listeners are asked simply to evaluate similarity between songs without any more specific explanations of what aspects of the music they should focus on. Since it is not meaningful to have computational models that go beyond the level of human agreement, such levels of inter-rater agreement present a natural upper bound for any algorithmic approach (Smith and Chew, 2013; Nieto et al., 2014; Serra et al., 2014; Flexer and Grill, 2016; Koops et al., 2019).

A possible solution is to not have one model of music similarity for all kinds of generic users but to personalize music recommendation. This however requires a high level of intra-rater agreement, i.e. that experience of music similarity is fairly stable over time within individual users, since otherwise such a constantly changing individual music perception would present yet another upper bound for algorithms. One important factor influencing individual music perception is mood of participants at test time (Schedl et al., 2013). Other proposed solutions to overcome lack of inter-rater agreement are better control of participant groups and song material, analysis of more specific aspects of music similarity, and holistic evaluation of complete MIR systems in specific use cases (Schedl et al., 2013; Flexer and Grill, 2016). It is also a drawback of previous studies (Jones et al., 2007; Flexer and Grill, 2016) of music similarity that they analysed similarity across genres, whereas most commercial services recommend music within genres.

In this article we still aim at quantification of general similarity among song pairs by human graders, but introduce a number of changes which should help us to explore and understand levels of inter- and also intra-rater agreement. Specifically we:

- use a more controlled group of human graders to introduce less variance than the largely unregulated

Johannes Kepler University Linz, AT
Corresponding author: Arthur Flexer (arthur.flexer@jku.at)

graders employed so far;
- carefully curate song material to determine the influence of genre on levels of both inter- and intra-rater agreement;
- measure graders' mood at test time to quantify the effect of their emotional state on intra-rater agreement, i.e. agreement when listening tests are repeated.

This article is an expanded version of a conference paper (Flexer and Lallai, 2019) with the main new contributions being the second larger user study including results on mood of listeners as well as an overall increased depth of coverage.

2. Related Work

It seems a basic insight that human perception of music is highly subjective with potentially low **inter-rater agreement**. For instance, if different human participants are asked to rate identical song pairs according to their perceived similarity, only a limited amount of agreement can be expected due to a number of subjective factors (Schedl et al., 2013; Flexer and Grill, 2016) like personal taste, musical expertise, familiarity with the music, listening history, current mood, etc. This also applies to annotation of music where different human participants will not always agree on genre labels or other semantic tags. For instance, the performance of humans classifying songs into nineteen genres ranges from a modest 26% to 71% accuracy, depending on the test participant (Seyerlehner et al., 2010). For transcription of chords annotators may disagree on about 10% of harmonic annotations (Ni et al., 2013). In case annotators are given full freedom to choose chords they tend to use different chord-label vocabularies, with overlap among all annotators being less than 20% (Koops et al., 2019). Related problematic results have been shown for audio-based grounding of everyday musical terms (Aucouturier, 2009).

On a more fundamental note, the argument has been brought forward (Wiggins, 2009) that music itself does not exist without the psycho-physiological effect of a stimulus on a human. As a consequence there is no immovable 'ground truth' in the context of music, which is rather subjective, highly context-dependent and not constant. A comparable conclusion was drawn in a study on automatically annotating acousmatic music (Klien et al., 2012).

Connected to these problems, a certain level of inter-rater agreement naturally presents an **upper bound** for any algorithmic approach trying to provide models which are valid for a multitude of generic users. Whenever these models are tested by new users, there will be a certain amount of disagreement rendering it impossible that these computational models surpass the level of human agreement. This has been documented (Jones et al., 2007; Schedl et al., 2013; Flexer and Grill, 2016) for the MIREX[1] tasks of 'Audio Music Similarity and Retrieval' (AMS) and 'Music Structural Segmentation' (MSS). MIREX ('Music Information Retrieval Evaluation eXchange') is an annual evaluation campaign for MIR algorithms (Downie, 2006). Because our experiments in Section 4 are closely connected to the MIREX task of 'Audio Music Similarity and Retrieval' (AMS), we now review previous results concerning rater agreement in the AMS task (Flexer and Grill, 2016). For the AMS task human graders evaluated pairs of query/candidate songs according to their general similarity. The query songs were randomly chosen from a test database and the candidate songs are recommendations automatically computed by participating algorithms. The human graders rated whether these query/candidate pairs "sound similar" using both a BROAD ('not similar', 'somewhat similar', 'very similar') and a FINE score (from 0 to 10 or from 0 to 100, depending on the year the AMS task was held, indicating degrees of similarity ranging from failure to perfection).

Interestingly, 2006[2] was the only year in which every query/candidate pair in the AMS task was evaluated by three different human graders. This makes 2006 the only year inter-rater agreement can be assessed. The average Pearson correlation between pairs of graders was found to be at the rather low level of 0.40. An upper bound based on ratings within the highest interval of scores has also been derived (Flexer and Grill, 2016): query/candidate pairs that were rated between 9 and 10 by one grader received an average rating of 6.54 from the respective other two graders. This constitutes an upper bound B^{AMS} as the maximum of average scores that can be achieved within the AMS evaluation setting, based on a considerable lack of agreement between human graders. What appears very similar to one of the graders will on average not receive equally high scores from other graders. Please note that such a lack of agreement could in principle also be due to different grading styles, e.g. graders either using just lower or higher parts of the available scale. This is however not the case for AMS, which is further discussed in Section 5.4. For AMS the upper bound was already reached in 2009 by a number of algorithms (Flexer and Grill, 2016). It is also important to notice for AMS 2006 that songs belong to nine genres. All evaluations were based on 60 randomly chosen query songs with a distribution heavily skewed towards rock music: 22 Rock songs, 6 Jazz, 6 Rap&HipHop, 5 Electronica&Dance, 5 R&B, 4 Reggae, 4 Country, 4 Latin, 4 Newage. Unfortunately the distribution of genres across the full data base of 5000 songs is not available, but there is some information concerning the "excessively skewed distribution of examples in the database (roughly 50% of examples are labelled as Rock/Pop, while a further 25% are Rap & Hip-Hop)".[3] An explanation for the rather low level of inter-rater agreement might be the lack of control over the participant groups. For all renditions of AMS over the years, test participants were drawn from the larger MIR community without any demographic limitations and without even recording basic sociographic data. The first question we try to answer in this paper therefore is: ***"Does a more controlled group of graders show increased levels of inter-rater agreement?"***.

Related results exist for the MIREX 'Music Structural Segmentation' (MSS) task, where a reported upper bound for MSS is already within reach for some genres of music (Flexer and Grill, 2016). Additional results for

music structure analysis (Smith and Chew, 2013; Nieto et al., 2014) and segment boundary recognition (Serra et al., 2014) exist. The level of inter-rater agreement has also been explored for melody extraction (Salamon et al., 2014; Bosch and Gómez, 2014; Balke et al., 2016), metrical structure (Quinton et al., 2015), rhythm and timbre similarity (Panteli et al., 2017), chord estimation (Koops et al., 2019; Selway et al., 2020), key estimation (Weiß et al., 2020), music emotion recognition (Gómez-Cañón et al., 2020), as well as perception of diversity of electronic music playlists (Porcaro et al., 2021). Deep learning has been used to account for different annotator styles in the task of chord labeling, essentially personalizing chord labels for individual annotators (Koops et al., 2020).

To the best of our knowledge, notions of **intra-rater agreement** in MIR have so far only been explored in the context of analyzing historical playlist data, focusing on either short-term (time of day) or long-term (in years) temporal dynamics of genre preferences (Aizenberg et al., 2012; Moore et al., 2013; Lex et al., 2020). Reports of J.S. Bach being sometimes inconsistent with himself when annotating his own music is a somewhat particular result concerning intra-rater agreement (Ju et al., 2020). In general IR it is however a well documented fact that items are judged differently over time, even by the same people (Schamber, 1994). The second question we want to answer in this paper therefore is: ***"What is the level of intra-rater agreement and is it higher than inter-rater agreement?"***.

Mood of participants at test time has been identified as a so-called 'user-context' influencing individual music perception (Schedl et al., 2013). Although most research on the interplay of music and mood is concerned with the arousal of emotions via music (Juslin and Sloboda, 2013), research on the influence of mood on music choice (Konečni, 2010) is an ongoing endeavour and success in developing emotion-aware music recommenders corroborates the importance of mood (Selvi and Sivasankar, 2019). Therefore the third question we want to answer in this paper is: ***"Is intra-rater agreement influenced by mood at test time?"***.

The concept of **general music similarity**, as it is being used in the MIREX AMS task, is in itself problematic. Criticism of this unclear abstract notion of general music similarity brings us to the concept of 'validity'. A valid experiment is an experiment that actually measures what the experimenter intended to measure (see e.g. Trochim, 2000; Urbano et al., 2013, for discussion in relation to MIR). Precisely this intention of the experimenter in the original MIREX AMS task is rather unclear, since it is rather dubious what general music similarity is supposed to mean in the first place. The argument that users apply very different, individual notions of similarity when assessing the output of music retrieval systems has been made before (Schedl et al., 2013). After all, music similarity is a multi-dimensional notion including timbre, melody, harmony, tempo, rhythm, lyrics, mood, etc., with many of these dimensions meaning different things to different people. The evaluation of abstract music similarity without reference to a specific usage scenario has been criticised as not being very meaningful (Hu and Liu, 2010; Serra et al., 2013). One way to make the intention of a music similarity experiment clearer is linking it to a user scenario, e.g. creating a playlist for a specific occasion. Identifying specific use cases has already been advocated as a method for better problem definition (Sturm, 2014) in MIR. Previous reviews (Lee and Cunningham, 2013) of user studies in MIR could serve as valuable input for designing these use cases. The focus of this article is rater agreement in the context of music recommendation and therefore recommendation is an obvious use case to be considered in our work. Since most commercial services recommend music within genres and not across genres, which is what studies so far have considered (Jones et al., 2007; Flexer and Grill, 2016), the fourth question we want to answer in this paper is: ***"Does genre influence inter- and intra-rater agreement?"***.

Such more **holistic user centred evaluation** has already been conducted within the 'MIREX Grand Challenge 2014: User Experience (GC14UX)'.[4] Here the task was to create web-based interfaces supporting users looking for background music for a short video. Systems were evaluated by human users with analysis showing that statistically significant differences between participating music similarity systems are very hard to obtain (Lee et al., 2015; Hu et al., 2017). Differences in user interfaces seemed quite important for evaluators and obscured differences in models of music similarity at the heart of the complete MIR systems. Another grand challenge user experience task was announced[5] for the following year, 2015, but regrettably no one participated.

The issue of holistic evaluation points to a fundamental problem in both MIR and general IR research. It is customary to conduct evaluation within the so-called **Cranfield paradigm**, where instead of conducting experiments with human users, only computer systems are being evaluated (see Cleverdon, 1991; Urbano et al., 2013, for a discussion related to MIR). The sole user input included in these evaluations is the ground truth data, which may range from low-level annotations (e.g. beat marks, tempo, frequency) to high-level concepts (e.g. genre, emotion or general tag information). This helps to save costs and allow for reproducibility, large scale comparisons, and tuning of algorithms. As a consequence of the Cranfield paradigm, a system response is characterized instead of a real user experience, assuming of course that they are correlated (see e.g. Urbano et al., 2013). For the task of finding similar music, one of the rare examples of comparing system responses to user-centered measures in MIR revealed only weak correlations (Hu and Kando, 2012). In general IR it was already documented fifty years ago that the implicit use orientation strongly influences manual rating of retrieved items (Cuadra and Katter, 1967) and hence a Cranfield-like disentanglement of users and IR systems is problematic.

3. Experimental Settings

Our two user studies are connected to the MIREX task of 'Audio Music Similarity and Retrieval' (AMS) as reviewed in Section 2. We still aim at quantification of general

similarity among song pairs by human graders, but use a more controlled group of human graders and carefully curated song material with query/candidate song pairs being based on Spotify recommendations instead of AMS recommendations, which also used a much smaller data base. We also measure graders' mood at test time to explore the effect of their emotional state on results. An overview of studies A and B is provided in **Table 1**.

3.1 Participants

In selecting test participants for our two studies, we were targeting a more controlled and uniform group of persons compared to the MIREX AMS task. For AMS, participants were recruited from the larger MIR community without any limitations and without recording any sociographic data. For both our studies test participants were required to have had some musical training in the past, giving all participants a comparable musical background. In addition we selected participants from an essentially young age group, making it more unlikely that they are overly familiar with the music material which has largely been published before 1990.

For **study A** all participants were born after 1984 and their age ranged from 26 to 34 years with an average of 28.2. The sample consisted of three females and three males. All participants were personal contacts of one of the authors. The study was conducted in 2019.

For **study B** all participants were born after 1985 and their age ranged from 21 to 35 years with an average of 25.6. The sample consisted of 20 females and 8 males. Please note that study A did not show any gender effects, hence the gender imbalance in study B is not expected to be problematic. All participants were personal contacts or colleagues of two of the authors. The study was conducted in 2021. There is no overlap between the groups of participants in study A and study B.

3.2 Song material

The original MIREX AMS task 2006 used songs from nine genres, with distribution highly skewed towards Rock/Pop and HipHop music which together comprised 75% of songs (see Section 2). In contrast our studies use either five equally distributed genres or just one.

For **study A**, songs belonged to five genres: (i) American **Soul** from the 1960s and 1970s with only male singers singing; (ii) **Bebop**, the main jazz style of the 1940s and 1950s, with excerpts containing trumpet, saxophone and piano parts; (iii) **High Energy** (Hi-NRG) dance music from the 1980s, typically with continuous eighth note bass lines, aggressive synthesizer sounds and staccato rhythms; (iv) **Power Pop**, a Rock style from the 1970s and 1980s, with chosen songs being guitar-heavy and with male singers; (v) **Rocksteady**, which is a precursor of Reggae with a somewhat soulful basis. The full list of 5 × 18 songs can be found in Section A in the appendix.

The five genres were chosen to have small stylistic overlap. All songs except two originate between the 1940s and early 1990s, making it more unlikely that participants are overly familiar with the music since all of them were born after 1984. Two songs were published in 2010, with one being a new release from a Hi-NRG artist from the 1980s and the other being a retro Rocksteady song. For each genre, we chose 18 songs. To further ensure unfamiliarity of song material to participants, we proceeded as follows. The songs were mainly chosen with the help of the recommendations of similar songs and artists on the music platform Spotify. We always started with one stereotypical artist of a genre and then searched for other similar songs using the similar artist function of Spotify, with the goal of finding similar music from rather unknown artists. Our search did not rely on personalized Spotify playlists and genres were validated via respective Wikipedia artist pages as well as by listening to all songs and corroborating the genre. The criterion for each song's degree of popularity was to have under 50,000 accesses on Spotify. Post-experiment questioning confirmed that on average only 2.17 songs out of 90 were familiar to the participants. No artists appears more than once on the song list. Within genres, we tried to find a homogeneous set of songs in order to evoke high similarity ratings which are crucial for determination of upper bounds in rater agreement.

For **study B** we used 90 songs all belonging to the genre **Power Pop**, including those already used in study A. Starting from each of the 18 **Power Pop** songs from study A, we chose additional songs from the respective Spotify recommendation lists. All songs are from the 1970s and 1980s with the exception of eight songs which are from the 1990s. Genres were corroborated the same way as for study A. All songs have under 50,000 accesses on Spotify and are in general not too well-known. Only one artist appears more than once in the song list. Post-experiment questioning confirmed that on average only 1.43 songs out of 90 were familiar to the participants. The full list of 90 songs can be found in Section B in the appendix.

For presentation in the questionnaire for both studies A and B, 15 seconds of a representative part of every song (usually the refrain) were chosen and normalized to 89db to control for volume effects using the "ReplayGain" plug-in of Audacity.

3.3 Mood scale

Only for **study B** we measured participants' mood at test time via the 'Brief Mood Introspection Scale' (BMIS) (Mayer and Gaschke, 1988). The BMIS scale consists of 16 mood-adjectives to which a person responds: lively, happy,

Table 1: Overview of song material and demographics in studies A and B.

	genres	songs	participants	female	male	average age	age range
Study A	5	90	6	3	3	28.2	26 to 34
Study B	1	90	28	20	8	25.6	21 to 35

sad, tired, caring, content, gloomy, jittery, drowsy, grouchy, peppy, nervous, calm, loving, fed up, active. In order to indicate how well each adjective describes their present mood participants choose one of the following answers: definitely do not feel, do not feel, slightly feel, definitely feel. The 16 answers can be summarized to yield measures of overall pleasant-unpleasant mood and arousal-calm mood.

3.4 Questionnaire

The studies were conducted online at www.soscisurvey.com, which is a free-access platform to compile questionnaires also allowing to include audio files. The first page of the questionnaires contained an introduction that explains the purpose of the study, the expected temporal effort as well as the note that the collection of data is held completely anonymous. Informed consent to participate in the study was obtained from participants in accordance with university and international regulations. Subsequently, the procedure of the study was explained to the participants. The participants were asked to "assess the similarity between the query song and each of the five candidate songs by adjusting the slider" and "to answer intuitively since there are no wrong answers".

For **study A**, before starting with the assessment, a test page was shown consisting of one randomly chosen query song and five randomly chosen candidate songs of all five genres. That was done to introduce all five genres and to give an idea of the variation of the song material used in the study. For the main part of the questionnaire the pairings of query and candidate songs were determined as follows. The complete song material consists of excerpts of 90 songs, each with a duration of 15 seconds, with 18 songs belonging to each of the 5 genres. We randomly drew 3 songs of each genre as query songs yielding a total of 15 query songs. For every query song we randomly chose five candidate songs with the constraint that at least one of them belongs to the same genre as the corresponding query song. This yields 15 groups of 6 songs each, with every song appearing exactly once in the whole questionnaire. The sequential order is held constant for all participants. Each group contains one query song paired with each of the five candidate songs of the group. In sum, comparisons of five pairs had to be made for every group yielding a total of $15 \times 5 = 75$ pairs. The participants were asked to indicate their rating of the similarity on a slider ranging from 0 to 100 %. The more similar a pair was assessed, the higher the percentage was, and the further to the right the slider had to be shifted.

For **study B**, the BMIS mood scale was administered at the beginning of the questionnaire. The song material again consists of excerpts of 90 songs with a duration of 15 seconds, but all belonging to the **Power Pop** genre. All 90 songs were randomly divided into query and candidate songs, resulting in 15 groups of 6 songs each, with the sequential order being held constant for all participants. Again every song appears exactly once in the whole questionnaire. In sum, comparisons of five pairs had to be made for every group yielding a total of $15 \times 5 = 75$ pairs. Again the participants were asked to indicate their rating of the similarity on a slider ranging from 0 to 100 %.

For both studies, at the end of the questionnaire, data regarding gender, age and musical education and experience was collected. On the last page of every questionnaire, the participants had the possibility to leave a comment. About two weeks after filling in the first questionnaire at time point **t1**, all participants filled in the same questionnaire with identical randomized items a second time (time point **t2**).

4. Results

In what follows we will detail results of the listening tests performed for studies A and B. Study A was planned as a pilot experiment to better understand the influence of genre on rater agreement. Study B is closer to a real-life music recommendation scenario with all songs belonging to a single genre.

4.1 Study A – five genres

First we analyse the degree of **inter-rater agreement** by computing the Pearson correlation ρ between graders for the 75 pairs of query/candidate songs. The 15 correlations between the six graders range from 0.59 to 0.86, with an average of 0.73 at t1 and 0.75 at t2 (see also **Table 2** for an overview of results). This is considerably higher than $\rho^{AMS} = 0.40$ which was reported for the MIREX AMS task 2006 (Flexer and Grill, 2016). The differences in correlation between ρ^{AMS} and correlations in our experiment are also statistically significant at both t1 ($t(16) = 8.33$, $p = 0.00$) and t2 ($t(16) = 8.85$, $p = 0.00$). Therefore the inter-rater agreement over the full range of scores in our experiment is increased compared to the MIREX AMS task.

In accordance with a previous study (Flexer and Grill, 2016), we next explore the level of agreement for specific intervals of scores. To illustrate this process, we

Table 2: Overview of results for time points t1, t2 and between t1 and t2 (t1 → t2), for both study A and B. Shown are average correlations ρ and upper bounds $B_{80} \pm$ standard deviations, also for MIREX AMS task (last two lines).

	study A – five genres			study B – one genre		
	t1	t2	t1 → t2	t1	t2	t1 → t2
ρ	0.73 ± .065	0.75 ± .065	0.80 ± .103	0.26 ± 0.146	0.22 ± 0.153	0.38 ± 0.175
B_{80}	67.7 ± 19.5	57.5 ± 25.6	82.1 ± 14.6	49.0 ± 22.8	45.0 ± 21.4	58.3 ± 21.6
ρ^{AMS}	0.40 ± .027					
B_{80}^{AMS}	61.65 ± 27.0					

plot the average score of a rater i for all query/candidate pairs, which he or she rated within a certain interval of scores v, versus the average scores achieved by the other five raters $j \neq i$ for the same query/candidate pairs. The average results across all $i = 1,...,6$ raters and for intervals v ranging from [0, 10], [10, 20]... to [90, 100] are plotted in **Figure 1** for t1 (top plot) and for t2 (bottom plot). We therefore explore how human graders rate pairs of songs which another human grader rated at a specific level of similarity. It is evident that there is a certain deviation from the theoretical perfect agreement indicated as a dashed line, especially for time t2. Pairs of query/candidate songs which are rated as being very similar (score between 90 and 100) by one grader are on average only rated at around 72.9 by the five other raters at t1 and at only 55.17 at t2.

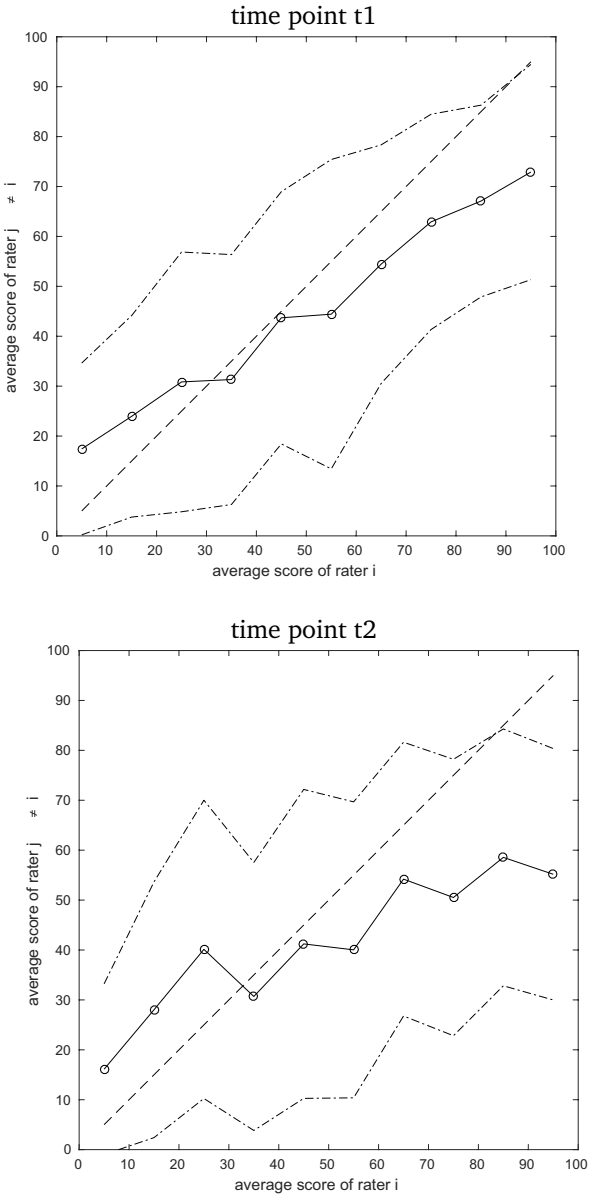

Figure 1: Average score inter-rater agreement for different intervals of scores (solid line) ± one standard deviation (dash-dot lines), at time point **t1** (top) and **t2** (bottom) for study A. The dashed line indicates theoretical perfect agreement.

On the other end of the scale, query/candidate pairs rated as being not similar at all (score between 0 and 10) receive average scores of 17.4 (t1) and 16.1 (t2) by the respective other raters.

In a previous study (Flexer and Grill, 2016) (see also Section 2) an upper bound B^{AMS} was reported based on scores within the highest interval. In the left plot of **Figure 2** we show a histogram of all scores for both time points t1 and t2 together. The scores cover the full range from 0 to 100, with an average of 34.43, indicating a slight preference for lower scores. Since for our experiment only 4 (out of 75 song pairs × 6 raters = 450) scores are higher than 90 for t1 and only 15 for t2, we compute a broader upper bound based on all scores between 80 and 100. There are 37 (8.22%) such scores for t1 and 47 (10.44%) for t2. This upper bound B_{80} is 67.7 for t1 and 57.5 for t2 (see **Table 2**). Multiplication of AMS 2006 scores by ten for better comparability and computation of a similar upper bound B_{80}^{AMS} yields 61.65. Since our upper bounds B_{80} are either above (t1) or below (t2) the upper bound B_{80}^{AMS}, we have to conclude that our experiment was not able to raise the upper bound in modeling general music similarity.

Next we looked at **intra-rater agreement** measured between time points t1 and t2. The six Pearson correlations ρ between t1 and t2 for the 75 pairs of query/candidate songs for the six graders range from 0.64 to 0.95, with an average of 0.80, which is somewhat higher than inter-rater correlation of 0.73 and 0.75 at time t1 and t2 (see **Table 2**). The differences between inter-agreement correlations and intra-agreement correlation are however not statistically significant, neither for t1 ($t(19) = -1.97$, $p = 0.06$) nor for t2 ($t(19) = -1.37$, $p = 0.19$).

Similarly to what we did for inter-rater agreement, we also computed an upper bound B^{80} based on ratings within the interval (80,100], where query/candidate pairs that were rated between 80 and 100 by a grader i at t1 received an average rating of 82.1 by the same grader i at t2. This is higher than the upper bound for inter-rater agreement at both t1 (67.7) and t2 (57.5). The differences between inter-agreement upper bounds and intra-agreement upper bound are statistically significant, both for t1 ($t(220) = -4.25$, $p = 0.00$) and t2 ($t(220) = -5.71$, $p = 0.00$). Therefore we conclude that the upper bound within participants measured with a two week time lag is higher than the upper bound based on inter-rater agreement.

Because a number of participants commented that the **genre of the songs was an important factor** when evaluating the similarity of songs, we analysed the results with respect to genre also. In **Figure 2** we show histogram plots of all scores within genres (middle plot) and scores between genres (right plot) for both time points t1 and t2 together. Although the scores in both histograms cover almost the full range from 0 to 100, scores for query/candidate pairs are on average higher when both songs belong to the same genre (within: 43.09) as compared to song pairs from different genres (between: 30.10). As a consequence the majority of scores in the between condition is in the leftmost third of the respective histogram. In addition, in **Table 3** we present

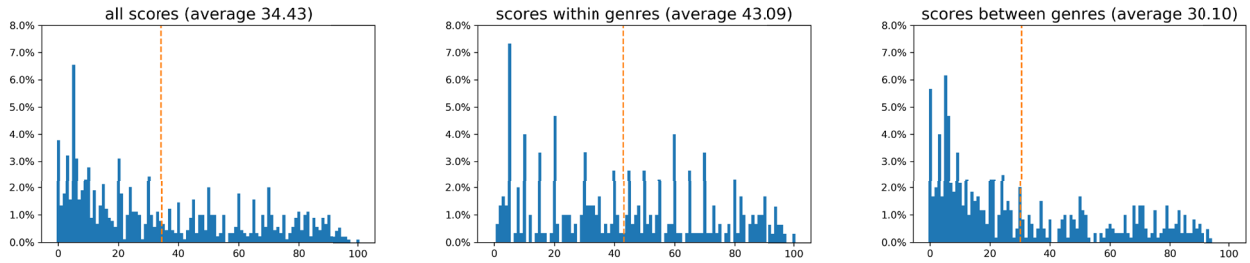

Figure 2: Histogram plots of all scores (left), scores within genres (middle) and scores between genres (right) for both time points t1 and t2 of study A. Average scores are plotted as vertical dashed lines and also given in the respective titles.

Table 3: Genre score matrix at time **t1** (top) and **t2** (bottom) for study **A**, showing average scores per genre combination. Left-most column: genre of query song; top line: genre of candidate song.

time point t1

Query \ Candidate	Soul	Bebop	High Energy	Power Pop	Rocksteady
Soul	**46.9**	16.2	–	38.3	25.1
Bebop	19.3	**73.4**	10.4	6.7	14.3
High Energy	30.4	8.1	**71.2**	32.0	15.5
Power Pop	17.4	–	20.9	**48.2**	11.0
Rocksteady	35.0	–	23.3	13.3	**66.1**

time point t2

Query \ Candidate	Soul	Bebop	High Energy	Power Pop	Rocksteady
Soul	**46.6**	13.5	–	36.4	19.6
Bebop	16.5	**64.7**	10.2	9.5	11.6
High Energy	33.3	5.8	**58.6**	29.8	15.8
Power Pop	18.6	–	16.8	**50.6**	11.0
Rocksteady	26.1	–	15.9	8.8	**62.5**

genre score matrices at times t1 and t2. Just to give one example, the first entry in the first line of the top sub table in **Table 3** shows that whenever both query and candidate song were from genre 'Soul', on average such pairs were judged at 46.9 by the graders. The average score for query songs from 'Soul' and candidate songs from 'Bebop' was 16.2, etc. An entry with a dash ('–') signifies that none such query/candidate pairs existed in our questionnaire. The genre score matrices are not symmetric, since there is a difference whether a song from a certain genre is used as a query or a candidate song.

At both times t1 and t2, average scores in the main diagonals are higher than all off-diagonal entries. This again shows that participants indeed rated similarities within genres higher than between genres, at least on average. For both times t1 and t2 average scores are highest within genre 'Bebop', followed by 'Rocksteady' and 'High Energy'. Genre 'Soul' has lowest within genre scores and considerable off-diagonal confusion with e.g. 'Power Pop'. Genres like 'Bebop' and 'High Energy', which are quite dissimilar music styles with different instrumentation and rhythm, show very little confusion.

To make clearer how often query/candidate pairs within one genre were rated higher than pairs with mixed genres, we computed average R-precision. In our scenario, for each of the 15 groups of songs (consisting of one query and five candidate songs), R is equal the number of candidate songs with genre identical to the respective query song. For our questionnaire R ranged from 1 to 4. When we order all five candidate songs from highest to lowest score, R-precision is then the fraction of candidate songs with matching genre among the first R candidate songs. Average R-precision across questionnaires of all all six participants is 0.86 for t1 and 0.88 for t2. These high values corroborate self-reports by participants that genre was an important aspect when rating similarity of songs.

4.2 Study B – one genre

First we again analyse the degree of **inter-rater agreement** by computing the Pearson correlation ρ between graders for the 75 pairs of query/candidate songs. For t1, the 378 correlations between the 28 graders range from –0.30 to 0.57 with an average of 0.26 (see also **Table 2** for an overview of results). For t2, the correlations range from –0.22 to 0.55 with an average of 0.22. This is lower than $\rho^{AMS} = 0.40$ which was reported for the MIREX AMS task 2006 (Flexer and Grill, 2016). The difference in correlation between ρ^{AMS} and correlations in our experiment is statistically significant at t2 ($t(379) = -2.11$, $p = 0.04$), but not at t1 ($t(379) = -1.69$, $p = 0.09$). Therefore the inter-rater agreement over the full range of scores in our experiment is somewhat decreased compared to the MIREX AMS task. These levels of inter-rater agreement of average 0.26 (t1) and 0.22 (t2) are also much lower than the inter-rater agreement in study A which is at 0.73 (t1) and 0.75 (t2). We pooled correlations for time points t1 and t2 for both studies A and B which yields a highly significant test result of $t(784) = 18.06$, $p = 0.00$. The fact that inter-rater agreement is below that of AMS 2006 and that of study A could be due to genre: songs from study B are all from one genre as opposed to AMS 2006 and study A with songs from multiple genres.

Next we again explore the level of agreement for only highly scored query/candidate pairs by computing upper bounds based on all scores between 80 and 100. There are 103 such scores for t1 and 62 for t2. This upper bound B_{80} is 49.0 for t1 and 45.0 for t2 (see **Table 2**). Since our upper bounds B_{80} for t1 and t2 are both below AMS 2006 upper bound $B_{80}^{AMS} = 61.65$, we have to conclude that study B was not able to raise the upper bound in modeling general music similarity. The upper bounds B_{80} are also both below those reported for study A, which are at 67.7 for t1 and 57.5 for t2. The overall reduced level of inter-rater agreement in study B is therefore also visible for scores between 80 and 100, which are the basis for the upper bounds.

Next we again looked at **intra-rater agreement** measured between time points t1 and t2. The 28 Pearson correlations ρ between t1 and t2 for the 75 pairs of query/candidate songs range from –0.22 to 0.62, with an average of 0.38, which is higher than inter-rater correlation of 0.26 and 0.22 at time t1 and t2 (see **Table 2**). The differences between inter-agreement correlations and intra-agreement correlation are statistically significant for both t1 ($t(404) = -4.24$, $p = 0.00$) and t2 ($t(404) = -5.50$, $p = 0.00$). Different from study A, where the difference between inter- and intra-rater agreement was not significant, participants of study B indeed showed a higher agreement with themselves than with other persons. The absolute level of agreement is however low with an average of only 0.38.

To better understand the intra-rater agreement we provide scatter plots (scores at t1 vs. t2) for the two raters who achieved maximal and minimal intra correlation in **Figure 3**. For both the maximal correlation of 0.62 (plot at top) and the minimal correlation of –0.22 (plot at bottom) it seems evident that the level of correlation is not due to one or a few outliers, but due to the accumulation of annotation differences across all query/candidate pairs. Scatter plots for other raters show similar behaviour.

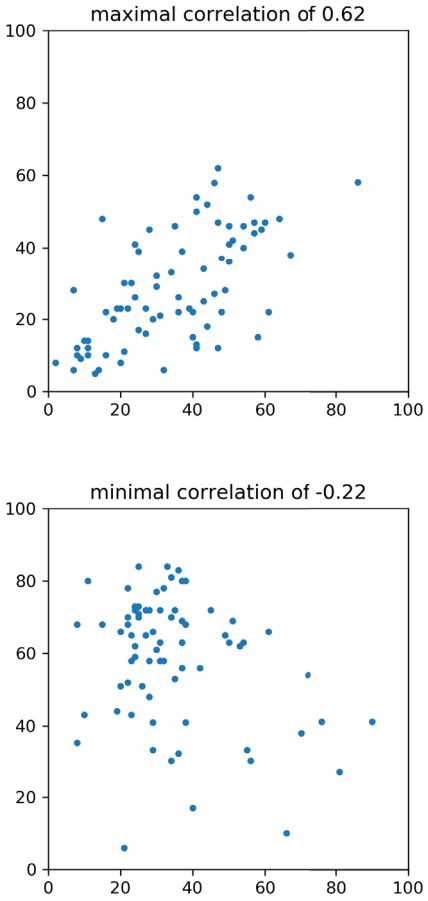

Figure 3: Scores at t1 (x-axis) vs. t2 (y-axis) for the two raters who achieved maximal (top) and minimal (bottom) intra-rater correlation in study B.

We also computed an upper bound B^{80} based on ratings within the interval (80,100], where query/candidate pairs that were rated between 80 and 100 by a grader i at t1 received an average rating of 58.3 by the same grader i at t2. This is higher than the upper bound for inter-rater agreement at both t1 (49.0) and t2 (45.0). The differences between inter-agreement upper bounds and intra-agreement upper bound are statistically significant, both for t1 ($t(2882) = -4.08$, $p = 0.00$) and t2 ($t(1775) = -6.13$, $p = 0.00$). Therefore we conclude that the upper bound within participants measured with a two week time lag is higher than the upper bound based on inter-rater agreement. This is in line with the results from study A, where the upper bounds within participants were also higher than those between participants. The absolute level of the upper bound within participants is however again rather low at only 58.3.

We also measured participants' mood at the beginning of the questionnaire by asking how well 16 mood-adjectives (BMIS (Mayer and Gaschke, 1988) describe their current mood. After all, the influence of current mood and emotional arousal on music appreciation has long been documented (Cantor and Zillmann, 1973) and might therefore also affect the degree of intra-rater agreement.

For analysis the four possible answers (definitely do not feel, do not feel, slightly feel, definitely feel) are converted to numbers from one to four. We computed Pearson correlations for these 16 mood values between time points t1 and t2 for each of the 28 participants. These correlations range from –0.22 to 0.93 with an average of 0.46 and a standard deviation of 0.31. It seems evident that some of the participants' self evaluation of mood was very different at times t1 and t2, while others showed very good agreement. When plotting these BMIS correlations versus the intra-rater agreement correlations described above, one can see a relationship (see **Figure 4**). Participants with higher BMIS correlations tend to also show higher intra-rater agreement. This opens up the possibility to focus further analysis only on participants whose BMIS scores did not change too much between t1 and t2. If all participants with a BMIS correlation smaller than 0.4 are excluded (horizontal dashed line in **Figure 4**), the remaining 17 (out of 28) participants' average intra-rater agreement correlation is 0.45 instead of 0.38 for all 28 participants. Comparable results can be achieved when basing such a focused analysis on an aggregation of all BMIS scores to one pleasant-unpleasant mood as suggested in the BMIS manual. This is less successful when using aggregated arousal-calm mood.

5. Discussion

Our goal for conducting this research and experiments was to explore the level of rater agreement when judging music similarity and to determine what factors influence the degree to which participants agree with each other and themselves. This is needed to quantify the success of computational models of music similarity which are used in automatic music recommendation services. Previous research has criticized the low level of inter-rater agreement and derived an upper bound for algorithms modeling music similarity, which was reached already in 2009 (Flexer and Grill, 2016). As a matter of fact, the upper bound was reached afterwards by other algorithms also but has never been outperformed. Consequently the respective MIREX task (AMS) has been dormant since 2015.

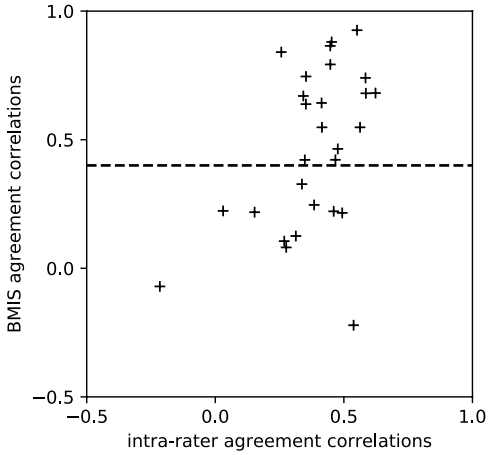

Figure 4: Intra-rater agreement correlations (x-axis) versus BMIS agreement correlations (y-axis). The horizontal dashed line illustrates the proposed exemption of participants with BMIS correlation < 0.4.

In trying to better understand and improve this situation, we have conducted two experiments with more controlled groups of participants and with more carefully curated song material. The participants were all of rather young age and all had a musical background. Most songs used in the questionnaires were published before the birth of the participants, making it unlikely that the music is known to the test participants and hence resulting in less personal connotations to the music. The songs in the MIREX AMS task were from nine different genres, but with only two genres making up 75% of all music. To explore the influence of genre on rater-agreement we first conducted study A with music from five very distinct genres. Study A was a pilot study with only six participants designed to test our hypothesis that genre helps raters to be consistent with other raters as well as with themselves. Study B had 28 participants and all songs belonged to a single genre. This situation corresponds much better to how real-life music recommenders work: they usually recommend music that is quite similar and stays within one genre. Actually curation of song material for study B is based on recommendations by Spotify.

5.1 Study A – five genres

For study A, the overall inter-rater agreement measured via correlation is indeed higher when compared to the AMS task. However the upper bound, which is only based on higher scores of music similarity, is not improved compared to AMS. One possibility to overcome the problem of upper bounds in measuring music similarity is to personalize models, i.e. to have separate models of music similarity for individual persons. Essentially this is what many commercial services do by offering individual recommendations tailored to their users. This of course brings us to the question of how stable assessment of music similarity is within persons over time, i.e. when the same persons have to judge music similarity repeatedly. For study A the result concerning this intra-rater agreement is divided. The overall agreement as measured via correlation could not be improved, at least not sufficiently to allow for statistical significance. On the other hand the upper bound could indeed be raised, opening up the possibility to better measure progress in computational models of personalized music similarity.

However, additional analysis of results from study A as well as comments by participants made it also clear that genres were so distinct that at least some participants used membership as a criterion when assessing similarity between songs.

5.2 Study B – one genre

Inter-rater agreement in study B, which used only music from a single genre, was lower than those observed in both study A and the AMS task. The same is true for upper bounds. The rather clear ranking of highest results for study A (five distinct genres), followed by intermediate results for AMS (two dominating genres out of nine

possible) and concluded by lowest results for study B (one genre only) suggests that genre indeed is a decisive factor influencing consistency of graders with each other.

Unfortunately, for study B both the overall agreement measured via correlation and the upper bound are much lower than for study A. This suggests that improvements of intra-rater agreement might again be due to the five very distinct genres in study A. When participants are not able to use genre cues when judging music similarity, it seems to be much harder for them to stay consistent with themselves over time.

Our results concerning mood of participants suggest that the current emotional state of users influences their perception of music similarity and hence this is something to take into account when evaluating music recommenders.

5.3 Control of participant groups and song material

Although we cannot quantify this effect, we are convinced that employment of our more controlled and uniform group of participants with a common musical expertise did lower variation of results. The same holds for the rather young age of participants and the choice of generally not well known song material. Post experiment questioning of participants corroborated that very few songs were known to them, hence less connotations to influence assessment of similarity existed. Such connotations could be specific memories or significance a person attributes to a certain song because of past experiences connected with this piece of music. Please note that all song material in studies A and B is nevertheless from genres which are familiar to most people in Western culture (Soul, Jazz, Disco, Rock, Reggae), but they have been chosen carefully to be older than participants and from rather specialized niches to prevent test participants from actually recognizing individual songs.

5.4 Limitations

In what follows we would like to discuss a number of limitations of our work which are due to certain choices in our experimental design.

The time lag of only two weeks between studies A and B was rather short and music recommenders are certainly being used for much longer time periods. We would expect that longer gaps might lower results on intra-rater agreement even further, possibly indicating long-range shifts in users' perception of music similarity (Lex et al., 2020; Moore et al., 2013).

In study B we are considering only one specific genre ('Power Pop'), which to a certain extent limits generalizability of our conclusions. Future work should repeat this part of our experiment with a genre that is maybe very distinct from 'Power Pop' to corroborate our main findings.

We would also like to consider the implications of using Pearson's correlation to measure inter- and intra rater agreement. Using Pearson's correlation implicitly normalizes for different rating styles; for example a Pearson correlation is perfect if two raters give identical answers, but also if all answers of one of the raters are always shifted by e.g. 10 units. For both user studies most raters used large parts of the scale which ranged from 0 to 100, which already is an indication that rating styles did not differ extensively (see also **Figure 2**). On average the utilized scoring ranges were: 88.5 (t1) and 79.3 (t2) for study A, 78.9 (t1) and 76.6 (t2) for study B. Replacing Pearson correlation with the mean absolute error, which is more sensitive to different rating styles, allows quite comparable overall conclusions with however smaller observed differences. Again intra-rater agreement is stronger than inter-rater agreement for both studies A and B. Again agreement is stronger for study A compared to study B. The exact numbers are given in Section C in the appendix.

5.5 Ways to move forward

In what follows we would like to give a number of clear suggestions to consider when evaluating notions of music similarity in the context of music recommendation. These suggestions are based on the main findings in our reported results and are intended to provide ways to move this field of research forward.

Pay attention to the demographics of your test participants. It is important to know what the target population of a music recommender evaluation experiment is. Is it the intention to draw conclusions about generic users or is there a specific target population, e.g. young people? Very often test participants from more homogeneous populations will show less variation and hence larger agreement concerning questions of music similarity.

Pay attention to the emotional state of your test participants. Human experience of music similarity is not only dependent on the demographics of test participants but also on their current emotional state. Therefore it is necessary to make this state part of the overall experimental design and record it or try keeping it stable.

Pay attention to the genre distribution in the song material. Music genres are characterised by certain highly salient properties (e.g. instrumentation, rhythm, tempo, theme of lyrics, etc.) which are likely to serve as reference points for evaluation of music similarity between different genres. In case recommendations within a single genre are being evaluated, test participants will have to focus on aspects of music beyond these salient features, probably resulting in overall lower agreement.

Focus on a specific aspect of music similarity. Music similarity is a multi-dimensional notion and without explicit instructions, test participants will focus their attention on different aspects of it, resulting in large variation and low agreement. Only clear instructions, e.g. asking whether music is similar according to the perceived tempo, can prevent such ambiguity.

Tie your research question to a use case. The reference to a specific use case, e.g. finding music for a special occasion, will clarify the intention of the overall experiment and will help to further focus test participants'

attention on specific aspects of music. It will also make clear what the real intention of the experiment is and hence settle its validity.

6. Conclusion

We have presented two studies exploring the level of agreement and disagreement between graders when measuring general music similarity. In our first study, using a more controlled group of participants and music material from five well defined genres resulted in increased levels of overall inter-rater agreement. We did not succeed in raising an upper bound for models of music similarity, which constitutes an obstructive glass ceiling for any machine learning approach. We did however succeed in raising this upper bound for intra-rater agreement, which corroborates the rationale of personalizing music services. Comparison to results from our second study, which used music from a single genre, made it clear that without perceptive cues from genre, test participants were much less consistent with others and themselves. As a consequence, inter- and intra-rater agreement as well as upper bounds were decisively diminished. Since real-world music recommendation systems very often stay within genre borders when recommending songs, results from our second study expose a serious problem for evaluation of such systems. Our results also cast doubt on the validity of experiments on general music similarity making it clear that focusing on a specific aspect of music similarity and definition of a specific use case might be necessary for conduction of truly valid experiments.

This article is also the first in-depth exploration of intra-rater agreement when annotating music in MIR. Future work should research whether other MIR tasks exhibit comparable problems of inter- and intra-rater consistency and what this uncertain ground-truth entails for evaluation of MIR systems in general.

Notes

[1] http://www.music-ir.org/mirex.

[2] https://www.music-ir.org/mirex/wiki/2006:Audio_Music_Similarity_and_Retrieval.

[3] See Note 2.

[4] https://www.music-ir.org/mirex/wiki/2014:GC14UX.

[5] https://www.music-ir.org/mirex/wiki/2015:GC15UX:JDISC.

Additional File

The additional file for this article can be found as follows:

- **Appendices.** Appendix A to C. DOI: https://doi.org/10.5334/tismir.107.s1

Acknowledgements

This research was funded in whole, or in part, by the Austrian Science Fund (FWF, P 31988). For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Competing Interests

Arthur Flexer is a member of the editorial board of the Transactions of the International Society for Music Information Retrieval. He was completely removed from all editorial processing. There are no other competing interests to declare.

References

Aizenberg, N., Koren, Y., and **Somekh, O.** (2012). Build your own music recommender by modeling Internet radio streams. In *Proceedings of the 21st International Conference on World Wide Web*, pages 1–10. DOI: https://doi.org/10.1145/2187836.2187838

Aucouturier, J.-J. (2009). Sounds like teen spirit: Computational insights into the grounding of everyday musical terms. *Language, Evolution and the Brain*, pages 35–64.

Balke, S., Driedger, J., Abeßer, J., Dittmar, C., and **Müller, M.** (2016). Towards evaluating multiple predominant melody annotations in jazz recordings. In *Proceedings of the 17th International Society for Music Information Retrieval Conference*, pages 246–252.

Bosch, J., and **Gómez, E.** (2014). Melody extraction in symphonic classical music: A comparative study of mutual agreement between humans and algorithms. In *Proceedings of the 9th Conference on Interdisciplinary Musicology*.

Cantor, J. R., and **Zillmann, D.** (1973). The effect of affective state and emotional arousal on music appreciation. *The Journal of General Psychology*, 89(1): 97–108. PMID: 4715319. DOI: https://doi.org/10.1080/00221309.1973.9710822

Cleverdon, C. W. (1991). The significance of the Cranfield tests on index languages. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12. DOI: https://doi.org/10.1145/122860.122861

Cuadra, C. A., and **Katter, R. V.** (1967). Opening the black box of 'relevance'. *Journal of Documentation*. DOI: https://doi.org/10.1108/eb026436

Downie, J. S. (2006). The Music Information Retrieval Evaluation eXchange (MIREX). *D-Lib Magazine*, 12(12). DOI: https://doi.org/10.1045/december2006-downie

Flexer, A., and **Grill, T.** (2016). The problem of limited inter-rater agreement in modelling music similarity. *Journal of New Music Research*, 45(3): 239–251. DOI: https://doi.org/10.1080/09298215.2016.1200631

Flexer, A., and **Lallai, T.** (2019). Can we increase interand intra-rater agreement in modeling general music similarity? In *Proceedings of the 20th International Society for Music Information Retrieval Conference*, pages 494–500.

Gómez-Cañón, J. S., Cano, E., Herrera Boyer, P., and Gómez Gutiérrez, E. (2020). Joyful for you and tender for us: The influence of individual characteristics and language on emotion labeling and classification. In *Proceedings of the 21st International Society for Music Information Retrieval Conference*.

Hu, X., and **Kando, N.** (2012). User-centered measures vs. system effectiveness in finding similar songs. In

Proceedings of the 13th International Society for Music Information Retrieval Conference, pages 331–336.

Hu, X., Lee, J. H., Bainbridge, D., Choi, K., Organisciak, P., and **Downie, J. S.** (2017). The MIREX Grand Challenge: A framework of holistic userexperience evaluation in music information retrieval. *Journal of the Association for Information Science and Technology*, 68(1): 97–112. DOI: https://doi.org/10.1002/asi.23618

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

Jones, M. C., Downie, J. S., and **Ehmann, A. F.** (2007). Human similarity judgments: Implications for the design of formal evaluations. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 539–542.

Ju, Y., Margot, S., McKay, C., Dahn, L., and **Fujinaga, I.** (2020). Automatic figured bass annotation using the new Bach Chorales Figured Bass Dataset. In *Proceedings of the 21st International Society for Music Information Retrieval Conference*.

Juslin, P. N., and **Sloboda, J. A.** (2013). Music and emotion. In Deutsch, D., editor, *The Psychology of Music*, pages 583–645. Academic Press, third edition. DOI: https://doi.org/10.1016/B978-0-12-381460-9.00015-8

Klien, V., Grill, T., and **Flexer, A.** (2012). On automated annotation of acousmatic music. *Journal of New Music Research*, 41(2): 153–173. DOI: https://doi.org/10.1080/09298215.2011.618226

Konečni, V. J. (2010). The influence of affect on music choice. In Juslin, P. N. and Sloboda, J. A., editors, *Handbook of Music and Emotion: Theory, Research, Applications*, pages 698–723. Oxford University Press.

Koops, H. V., de Haas, W. B., Bransen, J., and **Volk, A.** (2020). Automatic chord label personalization through deep learning of shared harmonic interval profiles. *Neural Computing and Applications*, 32(4): 929–939. DOI: https://doi.org/10.1007/s00521-018-3703-y

Koops, H. V., de Haas, W. B., Burgoyne, J. A., Bransen, J., Kent-Muller, A., and **Volk, A.** (2019). Annotator subjectivity in harmony annotations of popular music. *Journal of New Music Research*, 48(3): 232–252. DOI: https://doi.org/10.1080/09298215.2019.1613436

Lee, J. H., and **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. DOI: https://doi.org/10.1007/s10844-013-0259-2

Lee, J. H., Hu, X., Choi, K., and **Downie, J. S.** (2015). MIREX Grand Challenge 2014 user experience: Qualitative analysis of user feedback. In *Proceedings of the 16th International Society for Music Information Retrieval Conference*, pages 779–785.

Lex, E., Kowald, D., and **Schedl, M.** (2020). Modeling popularity and temporal drift of music genre preferences. *Transactions of the International Society for Music Information Retrieval*, 3(1): 17–30. DOI: https://doi.org/10.5334/tismir.39

Mayer, J. D., and **Gaschke, Y. N.** (1988). The experience and meta-experience of mood. *Journal of Personality and Social Psychology*, 55(1): 102. DOI: https://doi.org/10.1037/0022-3514.55.1.102

Moore, J. L., Chen, S., Turnbull, D., and **Joachims, T.** (2013). Taste over time: The temporal dynamics of user preferences. In *Proceedings of the 14th International Society for Music Information Retrieval Conference*, pages 401–406.

Ni, Y., McVicar, M., Santos-Rodriguez, R., and **De Bie, T.** (2013). Understanding effects of subjectivity in measuring chord estimation accuracy. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(12): 2607–2615. DOI: https://doi.org/10.1109/TASL.2013.2280218

Nieto, O., Farbood, M. M., Jehan, T., and **Bello, J. P.** (2014). Perceptual analysis of the f-measure for evaluating section boundaries in music. In *Proceedings of the 15th International Society for Music Information Retrieval Conference*, pages 265–270.

Panteli, M., Rocha, B., Bogaards, N., and **Honingh, A.** (2017). A model for rhythm and timbre similarity in electronic dance music. *Musicae Scientiae*, 21(3): 338–361. DOI: https://doi.org/10.1177/1029864916655596

Porcaro, L., Gómez, E., and **Castillo, C.** (2021). Perceptions of diversity in electronic music: The impact of listener, artist, and track characteristics. *arXiv preprint arXiv:2101.11916*.

Quinton, E., Harte, C., and **Sandler, M.** (2015). Extraction of metrical structure from music recordings. In *Proceedings of the 18th International Conference on Digital Audio Effects*.

Salamon, J., Gómez, E., Ellis, D. P., and **Richard, G.** (2014). Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine*, 31(2): 118–134. DOI: https://doi.org/10.1109/MSP.2013.2271648

Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology*, 29: 3–48.

Schedl, M., Flexer, A., and **Urbano, J.** (2013). The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41(3): 523–539. DOI: https://doi.org/10.1007/s10844-013-0247-6

Schedl, M., Zamani, H., Chen, C.-W., Deldjoo, Y., and **Elahi, M.** (2018). Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2): 95–116. DOI: https://doi.org/10.1007/s13735-018-0154-2

Selvi, C., and **Sivasankar, E.** (2019). An efficient context-aware music recommendation based on emotion and time context. In *Data Science and Big Data Analytics*, pages 215–228. Springer. DOI: https://doi.org/10.1007/978-981-10-7641-1_18

Selway, A., Koops, H. V., Volk, A., Bretherton, D., Gibbins, N., and **Polfreman, R.** (2020). Explaining harmonic inter-annotator disagreement using Hugo Riemann's theory of 'harmonic function'. *Journal of*

New Music Research, 49(2): 136–150. DOI: https://doi.org/10.1080/09298215.2020.1716811

Serra, J., Müller, M., Grosche, P., and **Arcos, J. L.** (2014). Unsupervised music structure annotation by time series structure features and segment similarity. *IEEE Transactions on Multimedia*, 16(5): 1229–1240. DOI: https://doi.org/10.1109/TMM.2014.2310701

Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gómez Gutiérrez, E., Gouyon, F., Herrera Boyer, P., Jordà Puig, S., Paytuvi, O., Peeters, G., Schlüter, J., Vinet, H., and **Widmer, G.** (2013). Roadmap for music information research. http://mires.eecs.qmul.ac.uk/files/MIRES_Roadmap_ver_1.0.0.pdf.

Seyerlehner, K., Widmer, G., and **Knees, P.** (2010). A comparison of human, automatic and collaborative music genre classification and user centric evaluation of genre classification systems. In *International Workshop on Adaptive Multimedia Retrieval*, pages 118–131. Springer. DOI: https://doi.org/10.1007/978-3-642-27169-4_9

Smith, J. B., and **Chew, E.** (2013). A meta-analysis of the MIREX structure segmentation task. In *Proceedings of the 14th International Society for Music Information Retrieval Conference*, pages 45–47.

Sturm, B. L. (2014). The state of the art ten years after a state of the art: Future research in music information retrieval. *Journal of New Music Research*, 43(2): 147–172. DOI: https://doi.org/10.1080/09298215.2014.894533

Trochim, W. (2000). *The Research Methods Knowledge Base*. Atomic Dog Publishing, Cincinnati, OH, 2nd edition.

Urbano, J., Schedl, M., and **Serra, X.** (2013). Evaluation in music information retrieval. *Journal of Intelligent Information Systems*, 41(3): 345–369. DOI: https://doi.org/10.1007/s10844-013-0249-4

Weiß, C., Schreiber, H., and **Müller, M.** (2020). Local key estimation in music recordings: A case study across songs, versions, and annotators. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2919–2932. DOI: https://doi.org/10.1109/TASLP.2020.3030485

Wiggins, G. A. (2009). Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music. In *Proceedings of the 11th IEEE International Symposium on Multimedia*, pages 477–482. IEEE. DOI: https://doi.org/10.1109/ISM.2009.36

How to cite this article: Flexer, A., Lallai, T., and Rašl, K. (2021). On Evaluation of Inter- and Intra-Rater Agreement in Music Recommendation. *Transactions of the International Society for Music Information Retrieval,* 4(1), pp. 182–194. DOI: https://doi.org/10.5334/tismir.107

Submitted: 26 March 2021 **Accepted:** 12 October 2021 **Published:** 24 November 2021

]u[*Transactions of the International Society for Music Information Retrieval* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS